Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Online Biomedical Resources for Malaria-Related Red Cell Disorders

Frédéric B. Piel,[1]* Rosalind E. Howes,[1] Oscar A. Nyangiri,[2] Catherine L. Moyes,[1] Thomas N. Williams,[2,3] David J. Weatherall,[4] and Simon I. Hay[1]

[1]Spatial Ecology and Epidemiology Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford, United Kingdom; [2]Kenya Medical Research Institute/Wellcome Trust Programme, Centre for Geographic Medicine Research-Coast, Kilifi District Hospital, Kilifi, Kenya; [3]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom; [4]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom

Communicated by Alastair Brown

Received 19 November 2012; accepted revised manuscript 1 April 2013.
Published online 8 April 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22330

ABSTRACT: Warnings about the expected increase of the global public health burden of malaria-related red cell disorders are accruing. Past and present epidemiological data are necessary to track spatial and temporal changes in the frequencies of these genetic disorders. A number of open access biomedical databases including data on malaria-related red cell disorders have been launched over the last two decades. Here, we review the content of these databases, most of which focus on genetic diversity, and we describe a new epidemiological resource developed by the Malaria Atlas Project. To tackle upcoming public health challenges, the integration of epidemiological and genetic data is important. As many countries are considering implementing national screening programs, strategies to make such data more accessible are also needed.
Hum Mutat 34:937–944, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: G6PD; HbC; HbE; HbS; malaria; database; thalassemia

Introduction

Between the 19th century and the mid-1980s, epidemiology became a fast-evolving discipline aimed at the study of risk factors in human diseases and their spread. Clinicians and researchers rapidly became aware that, by gathering information on populations, diseases and the environment, they could improve knowledge of how to control, treat, eliminate, or even eradicate diseases and communicate this knowledge to the relevant authorities [Cerda and Valdivia, 2007; Whitelaw, 1923]. The recent surge of attention toward eliminating malaria illustrates the importance of epidemiology in reducing disease burden [Das and Horton, 2010]. Genetic epidemiology brings genetic factors into our understanding of disease risk and transmission. The spread of genetic disorders is driven by processes such as migration and selection over timescales of generations. These timescales make understanding these epidemiological features equally as important as those of rapidly transmitted infectious diseases because of their long-term burden, both at the individual and population level. Assembling epidemiological data about these disorders and implementing appropriate health policies is highly involved and may require population screening, genetic counseling, and antenatal diagnosis. A good illustration of the implementation of such interventions to prevent thalassemias is the example of Cyprus in the 1970s [Angastiniotis and Hadjiminas, 1981; Weatherall and Clegg, 2001].

Variation in the hemoglobin gene family is amongst the best-characterized genetic systems [Marks, 1989]. SCA (MIM #603903) was the first disease linked to the hemoglobin protein [Pauling et al., 1949] and genetically characterized [Ingram, 1956]. Many abnormal hemoglobin variants were discovered in the 1950s and 1960s because of the differences in their electrophoretic properties. Since the 1980s, DNA sequencing and genotyping studies have enabled confirmation of the large diversity of these variants, identification of many more new variants, and investigation of the underlying selection mechanisms driving this diversity.

Despite significant progress in our knowledge of malaria-related red cell disorders [Hedrick, 2011; Kwiatkowski, 2005], researchers, clinicians, and the public health community are currently confronted with important challenges. First, there is growing evidence that the global health burden of these inherited disorders is likely to increase [Weatherall, 2010]. Because of population growth and higher reproduction rates in areas of high frequencies for malaria-related red cell disorders, the incidence and prevalence of individuals affected globally are both expected to increase. This trend is reinforced in low- and middle-income countries by important reductions in child mortality resulting from significant improvements in hygiene, nutrition, and other public health interventions. The survival of children affected by genetic disorders who would have previously died undiagnosed would thus translate into a further increase in the incidence of these genetic disorders [Akinyanju, 2010; Balgir, 2000; Makani et al., 2007; Weatherall, 2011]. In high-income countries, growing fluxes of migrants from areas with high frequencies of these disorders results in increasing demands for screening programs, genetic counseling and healthcare services [Hassell, 2010; Streetly et al., 2009]. To quantify these spatial and temporal changes, it is important to gather detailed epidemiological information to assess the current and future burdens faced by countries.

*Correspondence to: Frédéric B. Piel, Spatial Ecology and Epidemiology Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom. E-mail: fred.piel@zoo.ox.ac.uk

Contract grant sponsor: Wellcome Trust (#085406, #091835, #095066, #091758); Fondation Wiener-Anspach and the European Research Council (grant DIVERSITY).

This is particularly needed in areas in which population surveys for these disorders are lacking, but also in areas where only out of date information is available. Second, these disorders present considerable geographical heterogeneity in terms of prevalence and genetic diversity, resulting in complex phenotype–genotype relationships, particularly for the thalassemias [Weatherall, 2001] and G6PD deficiency [Mason et al., 2007]. Individuals with similar genotypes, exposed to different environmental conditions, can present very different phenotypes. Furthermore, societal behaviors or customs (e.g., consanguinity) can also lead to higher frequencies of severe forms of a disease and the recent globalization of human migrations has led to the appearance of new compound statuses of different disorders, which would be highly unlikely to have occurred otherwise. To implement appropriate measures to reduce the clinical and economic burden associated with these disorders, it is crucial to improve our knowledge about these relationships. This relies on collecting reliable epidemiological and genetic data based on accurate diagnostics.

Until recently, most of the afore-mentioned disorders have not enjoyed a high priority on public health agendas, despite a relatively high impact on childhood mortality, particularly in Africa [Grosse et al., 2011]. Following decades of advocacy, the United Nations finally recognized hemoglobinopathies as a public health problem in 2006 [United Nations, 2008; World Health Organization, 2006]. These disorders have also been included in the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010, http://www.globalburden.org) as part of the Non-Communicable Diseases Cluster [Murray et al., 2012]. In parallel with this growing awareness, epidemiological evidence, supported by genetic studies, is increasingly needed by Ministries of Health and public health organizations to implement appropriate policies.

Here, we provide (1) a short description of seven malaria-related red cell disorders; (2) we review online global databases that include significant data on at least one of these disorders and we describe several challenges related to keeping these resources up to date and accessible; (3) we present a new open-access spatial database developed by the Malaria Atlas Project (MAP, http://www.map.ox.ac.uk) that currently holds data on four of these disorders and describes future work to expand this resource; and (4) we advocate the value of including spatial data in genetic databases to be used for epidemiological studies.

Malaria-Related Red Cell Disorders

Many hundreds of genetic conditions affecting the red blood cells (RBCs) have been identified worldwide [Weatherall et al., 2006]; however, we concentrate here on those sufficiently common to be of public health significance. We have a particular interest in genetic variants that interact with the parasitic disease malaria because the public health importance of these variants extends from their own clinical impact to the effect they have on the burden of malaria within populations in endemic countries [Flint et al., 1998; Hill, 1987; Hedrick, 2011; Williams, 2006). We have recently investigated the spatial support for such a relationship by looking at the distribution of hemoglobin S (HbS) and malaria endemicity [Piel et al., 2010]. The particular disorders we have focused on include the following hemoglobin mutations, enzymopathies, and RBC surface loci: the main structural hemoglobin variants: HbS (MIM #141900.0243) [Piel et al., 2010; 2013b; Serjeant and Serjeant, 2001], hemoglobin C (HbC; MIM #141900.0038) [Piel et al., 2013a], hemoglobin E (HbE; MIM #141900.0071) [Flint et al., 1998; Makani et al., 2007; Williams et al., 2005], and the thalassemias: respectively α- (MIM #604131) and β-thalassemias (MIM #613985) [Weatherall and Clegg, 2001]; glucose-6-phosphate-dehydrogenase (G6PD; MIM #305900) deficiency [Cappellini and Fiorelli, 2008; Howes et al., 2012; Luzzatto and Notaro, 2001; Ruwende et al., 1995]; Southeast Asian ovalocytosis (MIM #109270.0002) [Kidson et al., 1981], and Duffy negativity (MIM #613665.0002) [Carter, 2003; Howes et al., 2011; Livingstone, 1984; Menard et al., 2010; Miller et al., 1976]. For descriptive simplicity, we have grouped all of these conditions under the malaria-related red cell disorders terminology. Many detailed reviews have described each of these disorders comprehensively (e.g., Carter and Mendis, 2002; Hedrick, 2011; Kwiatkowski, 2005; Weatherall et al., 2006] and we provide here a very brief summary of their genetics and epidemiology.

(1) Sickle hemoglobin or HbS is a structural variant of normal adult hemoglobin (HbA) caused by an amino acid substitution at position 6 of the β-globin chain (HBB c.20A>T; p.Glu6-Val) and is inherited as a Mendelian trait. Carriers or heterozygotes (HbAS) are almost always asymptomatic. Homozygotes (HbSS) suffer from sickle cell anemia (SCA), which often leads to acute and chronic complications including vaso-occlusive crisis, acute chest crisis or hemolytic crisis [Serjeant and Serjeant, 2001]. Sickle hemoglobin was largely restricted to Africa, the Middle East and parts of India but nowadays it is also common in the Americas, the Caribbean and Europe following human diasporas [Piel et al., 2013b].

(2) HbC is another structural variant of HbA caused by an amino acid substitution (HBB c.19G>A; p.Glu6Lys) occurring at the same position. HbAC carriers are asymptomatic. HbCC causes clinically mild hemolytic anemia, because of the reduced solubility of the RBCs, which can lead to crystal formation. HbC is mainly of clinical significance when inherited in combination with HbS (sickle-HbC disease), causing chronic hemolytic anemia and intermittent sickle cell crises, slightly less severe or frequent than in HbSS [Powars et al., 2002], and when co-inherited with β-thalassemia (HbC-β thalassemia), causing moderate hemolytic anemia with splenomegaly [Weatherall and Clegg, 2001]. Previously, HbC was prevalent only in Western Africa but carriers can now be found much more widely [Piel et al., 2013a].

(3) HbE is a structural variant of normal hemoglobin (HBB c.79G>A; p.Glu26Lys) affecting the production rate of HbA. Heterozygotes with HbAE are asymptomatic, whereas homozygotes can present some mild clinical features similar to individuals with β-thalassemia trait. Globally, compound individuals with HbE and β-thalassemia represent the highest burden with a wide range of clinical severity. The most severely affected individuals are transfusion dependent [Weatherall and Clegg, 2001]. HbE reaches frequencies up to 60% in parts of Thailand, Laos, and Cambodia, and is highly prevalent in India, Sri Lanka and Malaysia [Vichinsky, 2007; Weatherall and Clegg, 2001].

(4) The thalassemias affect the rate of production of either the α- or β-globin chains that form the subunits of adult hemoglobin, leading to α- and β-thalassemia, respectively [Weatherall and Clegg, 2001]. Thalassemias are caused by a large variety of mutations and deletions, causing severity proportional to the inability to synthesize globin chains. Although α- and β-thalassemias are both genotypically classified into minor, intermediate and major forms, there is a continuum of phenotypes ranging from asymptomatic to lethal. They were originally found across the "thalassemia belt," which extends from the Mediterranean area through the Middle East and India, to Southeast Asia. They are

Table 1. Overview of the Characteristics of Existing Resources on Malaria-Related Red Cell Disorders (Termed IBDs) and of the New Resource Launched by the Malaria Atlas Project (MAP-IBD)

			Survey locations provided				IBDs included						
Resource	URL	Years covered	National	Subnational	Communlity	Geographic coordinates (lat/lon)	HbS	HbC	Duffy	G6PD	Thal	SAO	Prevalence data
Livingstone	/	1904–1985	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HGHG	/	1949–1994	Yes	n/a	n/a		Yes	Yes	Yes	Yes	No	Yes	No
FIDD	http://medic.cardiff.ac.uk/fidd/	n/a	Yes	Variable	No	No	Yes	No	No	Yes	Yes	No	Very limited
HbVar	http://globin.bx.psu.edu/hbvar			No	No		Yes	Yes	No	No	Yes	No	No
FINDbase	http://www.findbase.orR	n/a	Yes	Variable	No	Yes	Yes	No	No	Yes	Yes	No	No
ALFRED	http://alfred.med.vale.edu			n/a	Yes		Yes	Yes	No	Yes	Yes	No	Very limited
G6PD deficiency database	http://www.bioinf.org.uk/g6pd/	1986–2000	No	No	No	No	No	No	No	Yes	No	No	No
G6PD MutDB	http://bminfor.tongji.edu.cn/mutdb	1988–2009	No	No	No	No	No	No	No	Yes	No	No	No
MAP-IBD	http://www.map.ox.ac.uk	1950–2011	Yes	n/a	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes

now also commonly found in many other parts of the world [Weatherall and Clegg, 2001].

(5) The Duffy blood group is characterized by Duffy antigens, which are expressed on the surface of RBCs. Repeated exposure to different blood types may trigger immunogenic transfusion reactions in recipients, though this is rare. The Duffy gene has two main variants, which express the Fy^a and Fy^b antigens [Livingstone, 1984]. These antigens differ by a single amino acid (Gly42Asp), encoded by alleles $FY^{*}A$ and $FY^{*}B$. A null 'erythrocyte silent' (ES) phenotype, caused by a point substitution in the gene's promoter region prevents gene expression, encoding the $FY^{*}A^{ES}$ and $FY^{*}B^{ES}$ alleles, the former ($FY^{*}A^{ES}$) having only been occasionally reported [Kasehagen et al., 2007; Sellami et al., 2008]. The Duffy negative phenotype (most commonly encoded by $FY^{*}B^{ES}/^{*}B^{ES}$) was thought to be fully protective against *Plasmodium vivax* infection, as the parasite was found to depend on the Duffy antigen for RBC entry [Miller et al., 1976], but recent evidence of *P. vivax* infected-Duffy negative individuals has brought the universality of this relationship into question [Menard et al., 2010; Mercereau-Puijalon and Menard, 2010; Wurtz et al., 2011; Zimmerman et al., 2013]. The most prevalent Duffy allele globally is $FY^{*}A$, which reaches high frequencies (>90%) across East Asia, whereas in sub-Saharan Africa, the predominant allele is the silent $FY^{*}B^{ES}$ variant, commonly reaching frequencies approaching 100% and encoding the Duffy negative phenotype [Howes et al., 2011].

(6) Glucose-6-phosphate dehydrogenase (G6PD) deficiency is the most common human enzyme disorder [Cappellini and Fiorelli, 2008], found throughout malarious regions, with an estimated overall allele frequency of 8.0% (50% uncertainty interval: 7.4%–8.8%) across malaria endemic countries [Howes et al., 2012]. Mutations in the gene cause reduced enzyme activity, leaving RBCs vulnerable to oxidative stress. Although the condition is typically asymptomatic, severe acute hemolysis can be triggered by certain foods, infections, and drugs. One such drug is primaquine: the only drug currently licensed to clear the relapsing stages of *P. vivax* malaria from the liver [Howes et al., 2013]. G6PD deficiency is also a main cause of neonatal jaundice in some regions. A large number of mutations cause this condition, and these vary in their clinical characteristics from none to highly severe. The most clinically severe G6PD deficient variants are found across Asia. Although the variants common among sub-Saharan African populations are considered less severe, the high prevalence of the deficient phenotype in this region means that the public health risks associated with G6PD deficiency are also high in this region [Howes et al., 2012].

(7) Southeast Asian ovalocytosis (SAO) is an elliptocytosis, a genetic defect affecting the structural and functional properties of RBCs [Liu et al., 1990]. Heterozygotes are totally asymptomatic, whereas homozygotes are not viable [Delaunay, 2007]. SAO is mostly found in the malarious regions of Southeast Asia and the western Pacific [Rosanas-Urgell et al., 2012].

Existing Resources

Although a number of national and regional resources have also been launched (for example the Centre for Arab Genomic Studies Database (http://www.cags.org.ae/) or the Indian Genetic Disease Database (http://www.igdd.iicb.res.in/)), the present review focuses only on global databases. Table 1 provides an overview of key characteristics of each of the resources described here.

As early as the 1960s, Frank B. Livingstone started assembling a global database of the frequencies of hemoglobin variants, thalassemias, glucose-6-phosphate dehydrogenase deficiency, G6PD variants, and ovalocytosis in human populations. His dedication led to the publications of updated versions in the 1970s [Livingstone, 1973] and 1980s [Livingstone, 1985], which still represent a unique source of information on the prevalence of malaria-related red cell disorder variants among different population groups. Although Livingstone's last database has recently been reprinted [Livingstone and Marks, 2009], the absence of an electronic version of his tables and the crude spatial information associated with the location of each population survey hinder contemporary use of his databases. In the 1990s, Modell and Darlison updated his work on hemoglobin variants and gathered additional data on the thalassemias from research reviews, country visits, and the former WHO Working Group on Haemoglobin Disorders into an almanac [Modell and Darlison, 2008]. A significant amount of data have nevertheless been published since this study was conducted.

The History and Geography of Human Genes (HGHG) [Cavalli-Sforza et al., 1994], aimed at reconstructing human population history, using a novel approach combining population genetics and geography. Although now twenty years outdated, this book represents an invaluable resource containing 76,676 human gene frequency estimates. Investigations of hemoglobin disorders were limited because of their interactions with malaria. The HGHG included maps of the global distribution of HbS, HbC, Duffy negativity and G6PD deficiency, but the input data and the methodology used are not fully described, making it almost impossible to use them for quantitative analysis. An online version of the HGHG, named the Human Population Genetics Database (HPGD), was temporarily available on the Human Population Genetics Laboratory's Website (http://hpgl.stanford.edu/index.html) but, at the time of writing, was no longer operational.

The Frequency of Inherited Disorders Database (FIDD, http://medic.cardiff.ac.uk/fidd/), launched in 1998, aims to provide a systematic literature search summary on the prevalence and incidence of human Mendelian disorders [Al-Jader et al., 2001]. Survey information and phenotype prevalence are provided but this database only includes limited data on hematological disorders imported from the Online Mendelian Inheritance in Man compendium (14 records for sickle cell, 41 for alpha-thalassemia, 50 for beta-thalassemia, and 20 for G6PD deficiency). Locations are usually national or occasionally subnational but precise coordinate for the surveys reported are not given.

With the development of genomics in the 1990s, databases started to focus more on the genetic diversity of the new variants being regularly discovered than on their epidemiology and frequencies. HbVar (http://globin.bx.psu.edu/hbvar), launched in 2001, is a relational database on the genomic sequence changes leading to human hemoglobin variants and types of thalassemia [Hardison et al., 2002; Giardine et al., 2007; 2011; Patrinos et al., 2004]. It provides extensive information for each variant and mutation, including a description of the variant and associated pathology, hematology, electrophoretic mobility, methods of isolation, stability information, ethnic background, structure studies, functional studies, and references, but gives no details of gene frequencies or survey locations.

The Frequency of Inherited Disorders Database (FINDbase, http://www.findbase.org, launched 2005) is an online repository of information on the frequency of mutations causing inherited disorders [van Baal et al., 2007]. Data on thalassemias and G6PD deficiency from HbVar are included. Some mutation data are spatially referenced to the national level and occasionally to individual cities. It is far from comprehensive containing 768 records for beta-thalassemia, 18 for G6PD deficiency and none for other malaria-related red cell disorders.

The ALlele FREquency Database (ALFRED, http://alfred.med.yale.edu), launched in 1999, is a resource of gene frequency data on human populations supported by the U.S. National Science Foundation [Cheung et al., 2000]. The website allows users to visualize existing data and submit new data. The site includes extensive unpublished data, but only a tiny fraction (<1.5%) is on malaria-related red cell disorders (2 and 7 populations for Hb A/S/C and G6PD deficiency, respectively). A text description of the population surveyed is given but no coordinates are assigned to the location.

Finally, we identified a couple of resources specific to G6PD deficiency: (1) the G6PD deficiency database created by Dr Andrew Martin's group at University College London (http://www.bioinf.org.uk/g6pd/index.html) provides information on mutations leading to deficiency but no geographical information is provided; and (2) G6PD MutDB (http://bminfor.tongji.edu.cn/mutdb/) links mutations to deficiency phenotypes [Zhao et al., 2010] and while information on ethnic backgrounds is included, geographical location is not.

None of the above databases provide geographical coordinates for the communities surveyed. They do all provide citations that allow users to go back to the original source for further information and they allow users to extract key information from each record but they do not allow users to download integrated sets of community-level survey data from multiple records for use in epidemiological analyses.

Alongside these resources, others focus on increasing awareness of malaria-related red cell disorders. First, the Accessible Publishing of Genetic Information (APoGI, http://www.chime.ucl.ac.uk/APoGI, launched 2000) provides information and education materials on hemoglobin gene variants to help healthcare professionals on hemoglobin disorders to provide accurate counseling. The development of this resource was funded by the Wellcome Trust. No data on gene frequencies are available.

Second, the eInfrastructure for Thalassaemia Research Network (Ithanet, http://www.ithanet.eu, launched 2006) is an electronic infrastructure for a thalassemia research network developed within the European Union. Ithanet initially focused on the European community to facilitate contacts between researchers and data sharing. All hemoglobinopathies have recently been included in an extension of the project. It provides a community portal for experts, organizations and networks on thalassemias and other hemoglobinopathies. No data on gene frequencies are available.

Third, the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010, http://www.healthmetricsandevaluation.org/gbd) currently represents the most comprehensive effort to produce complete and comparable estimates of the burden of diseases, injuries, and risk factors for the years 1990, 2005, and 2010 for 21 regions globally. Sickle cell, G6PD deficiency, and the thalassemias have been included in the leading causes and risks, based on deaths, years of live lost, years lost to disability, and disability-adjusted life years for 1990 and 2010.

The MAP's Contribution: A New Database

Although each of the resources described above represents invaluable sources of information, none of them provides sufficiently disaggregated information on the incidence and prevalence of malaria-related red cell disorders that could potentially be used by health policy makers in developing or targeting policies. Over the last four years, the MAP, which aims to disseminate free, accurate and up-to-date information on malaria and associated topics, organized on a geographical basis, has assembled data on the distribution and prevalence of selected malaria-related red cell disorders to create an open-access biomedical resource for researchers, clinicians and members of the public health community. This database was conceived to complement the existing online resources described above. Its main features include: (1) data based on detailed contemporary searches conducted across various online bibliographic databases, including Pubmed, ISI Web of Science and Scopus, as well as cross-referencing with existing databases (e.g., Livingstone 1985, HGHG 1994) and unpublished sources of data accessed through personal communications. Details of the protocols used have been previously published [Howes et al., 2012; 2011; Piel et al., 2010; 2013a; 2013b]; (2) surveys that are representative of local communities (i.e., excluding patient surveys or surveys targeting specific ethnic group(s), which risk being biased samples); and (3) geographic coordinates of all surveys, mapped to the highest precision possible to encapsulate spatial heterogeneity in the distribution and prevalence of these disorders (Fig. 1). Survey data from all sources can be downloaded as a single output ready for use in epidemiological analyses.

As summarized in Table 2, this database includes 1,211, 445, 922 and 665 data points for sickle hemoglobin, HbC, G6PD deficiency, and the Duffy blood group, respectively. Similar work on the thalassemias and HbE is ongoing. For sickle hemoglobin [Piel et al., 2010; 2013b], the data provided include the sample size and the number of individuals with the HbAA, HbAS, and HbSS genotypes. A subset of the data, corresponding only to representative surveys

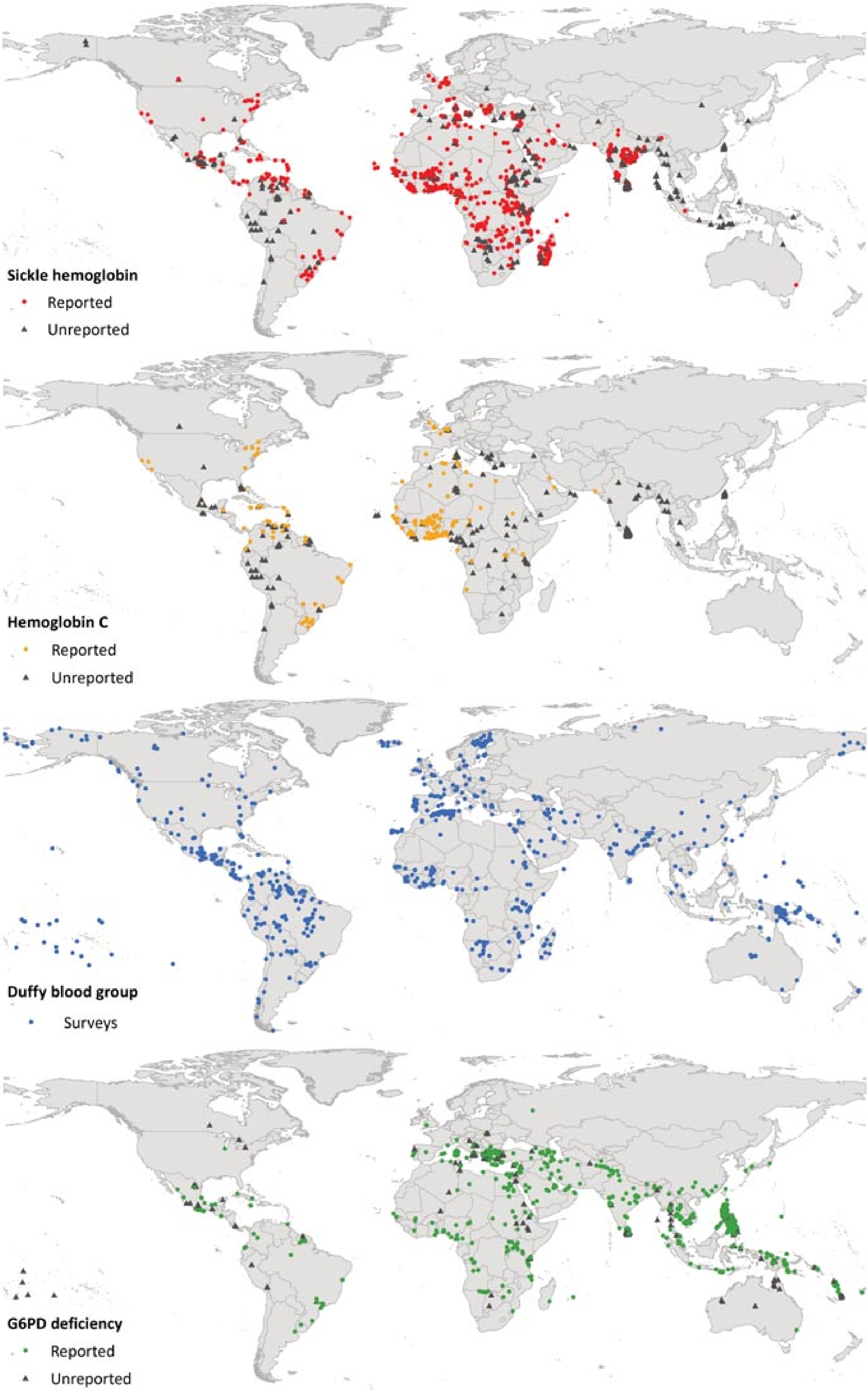

Figure 1. Global Maps of Surveys Included on Sickle Hemoglobin (HbS), Hemoglobin C (HbC), the Duffy Blood Group and G6PD Deficiency.

Table 2. Number of Data Sources, Spatially Unique Data Points and Individuals Tested for Each Malaria-Related Red Cell Disorder Included in the Database

Malaria-related red cell disorder	Data sources	Spatially unique data points released [*used in the model*]	Overall number of individuals tested
Sickle hemoglobin			
Autochthonous populations	278	730 [*773*]	347,434
All populations	435	1,129 [*1,211*]	9,032,377
Hemoglobin C	174	445 [*445*]	7,540,983
G6PD deficiency			
Males	254	910 [*1,720*]	316,448
Females	138	337 [*1,067*]	106,510
Total	258	922 [*1,734*]	422,958
Duffy blood group	301	665 [*821*]	114,563

Some of the surveys were included in the MAP modeling analyses, but authorization for public release was not granted. The figures in this table include all those freely available on the MAP Website.

of autochthonous communities, reflects the original distribution of sickle hemoglobin, which was used to validate the malaria hypothesis spatially [Piel et al., 2010]. Similarly for HbC [Piel et al., 2013a], data are summarized as HbAA, HbAC and HbCC genotypes. For the Duffy blood group, five different categories of data are presented, corresponding to the type of diagnostic used: serological (includes three data types depending upon the antigens tested for) and molecular (two data types depending upon which loci were examined) [Howes et al., 2011]. National-level estimates of Duffy negative populations are also available. For G6PD deficiency, data are presented by sex, with the number of individuals tested and deficient at each survey site. Only phenotypically diagnosed samples are included in the database. These datasets for each disorder provide the density of population surveys and the observed spatial heterogeneity of each disorder. Survey publication date is also listed. This could be used, for example, to identify obvious gaps in the data currently available or to measure deviation from Hardy–Weinberg equilibrium. Protocols detailing the data collection and inclusion criteria are available for each disorder on the associated map pages in the “Browse Resources” section.

The survey data described above form one part of a Web portal created to disseminate a range of data on malaria-related red cell disorders. As well as assembling these datasets, we have developed geostatistical models to make continuous predictions of the frequencies of each of these disorders in areas where no surveys were available [Diggle and Ribeiro Jr, 2007]. The models predicted continuous mapped surfaces of these disorders, as well as population estimates of affected individuals. All predictions are generated with associated uncertainty metrics [Patil et al., 2011]. Further information about these methodological aspects is available in the associated publications [Howes et al., 2011; 2012; Piel et al., 2010; 2013a; 2013b], linked from all of the MAP pages.

In addition to the raw survey data, we release model outputs in the form of map images (in PDF and PNG formats), GIS surfaces (in binary float and GeoTiff formats), and population estimates. These include continuous frequency maps of the distribution of sickle hemoglobin, HbC, Duffy negativity and G6PD deficiency, and their prediction uncertainty. For the first time, the Bayesian model-based geostatistical (MBG) approach used allowed us to account for the uncertainty associated with our predictions in a probabilistic way [Patil et al., 2010]. The maps presented show the mean or median predicted frequency and the interquartile range, which is a measure of uncertainty.

Such geographical maps can be used to support public health decisions through providing a spatially continuous estimate of the heterogeneity of the frequencies of these conditions. Furthermore, the maps permit estimates of individuals or newborns affected by these malaria-related red cell disorders to be made. Global, regional, national, and some subnational estimates in newborns are provided via the Web portal for sickle hemoglobin. Similar data are available for HbC in African countries. National population estimates are available for G6PD deficiency and Duffy negativity. The Duffy negativity map has been used to refine estimates of the population at risk of *P. vivax* malaria [Gething et al., 2012; 2010], whereas G6PD deficiency estimates provide information on areas in which primaquine therapy should be considered with caution [Howes et al., 2012].

Data searches can be performed for a given region (e.g., World Health Organization regions), country, topic and/or subtopic using the Resource Browser (http://www.map.ox.ac.uk/browse-resources/). Brief descriptions and links to complementary external resources, described above, are also available at http://www.map.ox.ac.uk/explore/inherited-blood-disorders/resources/. Individuals and organizations who have generously contributed unpublished data for use in the mapping models are listed at http://www.map.ox.ac.uk/acknowledgements/. Only data for which open-release permission was granted are included in the online database.

The Ways Forward

The creation of this new resource by the MAP is a first step toward assembling a contemporary database of epidemiological data on malaria-related red cell disorders, alongside data on malaria parasites and vectors. During the data collection process, it became obvious that only a fraction of survey data is easily accessible in the public domain; much being unpublished or published in journals or reports with limited visibility, even with modern search and access tools. This applies to data from universal screening programs in the United States of America [National Newborn Screening and Genetics Resource Center (NNSGRC), 2011], the United Kingdom [Streetly et al., 2009], and the French overseas territories [Bardakdjian-Michau et al., 2009]. Although data of high quality are collected, we have been unable to access reliable aggregated data for these countries. At a time when several developing countries, particularly in Africa, are moving toward the implementation of such programs [Ohene-Frempong et al., 2008], the development of guidelines and tools allowing optimal use of high volumes of data is going to be a major challenge.

Furthermore, existing online resources face various challenges [Galperin and Fernandez-Suarez, 2012]. Several of the resources reviewed here have stopped being maintained and updated, or are simply inaccessible, usually due to a lack of funding. We believe that, by improving collaborations between the different groups leading these resources and developing a better integration of the various databases, it will be possible in the long term to reduce the likelihood of such events, as well as to minimize the costs associated with their development and maintenance. In the short term, we hope to make the resource presented here sustainable, to provide regular updates and to add further data on the thalassemias and HbE [Colah et al., 2010; Olivieri et al., 2008], as we assemble data and develop tailored mapping models for these disorders.

Conclusions

We need epidemiological evidence to assess disease burden and to target interventions in an evidence-based manner, and this is as true for genetic diseases as it is for infectious diseases. It is crucial to gather data on the distribution and prevalence of these disorders and we present a new resource created by the MAP. A concerted effort from public health bodies, researchers and clinicians involved in malaria-related red cell disorders is necessary to scale up this work, and better integration of online resources would make it possible for public health workers and data modelers to find a comprehensive suite of information in one place. We strongly advocate the inclusion of geographical information in databases of genetic disorders to allow the assessment of the distribution of these disorders and highlight areas where their burden is highest.

Acknowledgements

We wish to thank the members of the technical advisory group associated with this project, including Dr J. Kevin Baird, Dr Suthat Fucharoen, Professor Dominic Kwiatkowski, Dr Julie Makani, Professor Kevin Marsh, Dr Kirk Rockett, Professor Graham Serjeant, Professor David Weatherall, Professor Thomas Williams, and Professor Bill Wood. This paper was published with the permission of the director of KEMRI.

Disclosure statement: The authors declare no conflict of interest.

References

Akinyanju O. 2010. *The National Burden of Sickle Cell Disorder and the Way Forward.* Sickle Cell Foundation Nigeria. Available at: http://www.sicklecellfoundation.com/Nat%20Burden%20SCD.pdf. Accessed 15 November 2012.

Al-Jader LN, Harper PS, Krawczak M, Palmer SR. 2001. The frequency of inherited disorders database: prevalence of Huntington disease. Community Genet 4:148–157.

Angastiniotis MA, Hadjiminas MG. 1981. Prevention of thalassaemia in Cyprus. Lancet 317:369–371.

Balgir RS. 2000. The burden of haemoglobinopathies in India and the challenges ahead. Curr Sci 79:1536–1547.

Bardakdjian-Michau J, Bahuau M, Hurtrel D, Godart C, Riou J, Mathis M, Goossens M. 2009. Neonatal screening for sickle cell disease in France. J Clin Pathol 62:31–33.

Cappellini MD, Fiorelli G. 2008. Glucose-6-phosphate dehydrogenase deficiency. Lancet 371:64–74.

Carter R. 2003. Speculations on the origins of *Plasmodium vivax* malaria. Trends Parasitol 19:214–219.

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

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The history and geography of human genes.* Princeton, NJ: Princeton University Press.

Cerda LJ, Valdivia CG. 2007. John Snow, the cholera epidemic and the foundation of modern epidemiology. Rev Chilena Infectol 24:331–334.

Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2000. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 28:361–363.

Colah R, Gorakshakar A, Nadkarni A. 2010. Global burden, distribution and prevention of beta-thalassemias and hemoglobin E disorders. Expert Rev Hematol 3:103–117.

Das P, Horton R. 2010. Malaria elimination: worthy, challenging, and just possible. Lancet 376:1515–1517.

Delaunay J. 2007. The molecular basis of hereditary red cell membrane disorders. Blood Rev 21:1–20.

Diggle PJ, Ribeiro Jr PJ. 2007. *Model-Based Geostatistics.* Springer, New York, USA.

Flint J, Harding RM, Boyce AJ, Clegg JB. 1998. The population genetics of the haemoglobinopathies. Bailliere Clin Haem 11:1–51.

Galperin MY, Fernandez-Suarez XM. 2012. The 2012 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 40:D1–D8.

Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, Guerra CA, Patil AP, Tatem AJ, Howes RE, Myers MF, George DB, Horby P, et al. 2012. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. PLoS Negl Trop Dis 6:e1814.

Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, Maglott D, Singleton BK, Anstee DJ, Basak AN, Clark B, Costa FC, Faustino P, et al. 2011. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat Genet 43:295–301.

Giardine B, van Baal S, Kaimakis P, Riemer C, Miller W, Samara M, Kollia P, Anagnou NP, Chui DH, Wajcman H, Hardison RC, Patrinos GP. 2007. HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. Hum Mutat 28:206.

Grosse SD, Odame I, Atrash HK, Amendah DD, Piel FB, Williams TN. 2011. Sickle cell disease in Africa: a neglected cause of early childhood mortality. Am J Prev Med 41:S398–S405.

Guerra CA, Howes RE, Patil AP, Gething PW, Van Boeckel TP, Temperley WH, Kabaria CW, Tatem AJ, Manh BH, Elyazar IRF, Baird JK, Snow RW, et al. 2010. The international limits and population at risk of *Plasmodium vivax* transmission in 2009. PLoS Negl Trop Dis 4:e774.

Hardison RC, Chui DH, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W, Wajcman H. 2002. HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum Mutat 19:225–233.

Hassell KL. 2010. Population estimates of sickle cell disease in the U.S. Am J Prev Med 38:S512–S521.

Hedrick PW. 2011. Population genetics of malaria resistance in humans. Heredity 107:283–304.

Hill AV. 1987. Haemoglobinopathies and malaria: new approaches to an old hypothesis. Parasitol Today 3:83–85.

Howes RE, Battle KE, Satyagraha AW, Baird JK, Hay SI. 2013. G6PD deficiency: global distribution, genetic variants and primaquine therapy. Adv Parasitol 81:133–201.

Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria CW, Gething PW, Zimmerman PA, Barnadas C, Beall CM, Gebremedhin A, Menard D, Williams TN, et al. 2011. The global distribution of the Duffy blood group. Nat Commun 2:266.

Howes RE, Piel FB, Patil AP, Nyangiri OA, Gething PW, Dewi M, Hogg MM, Battle KE, Padilla CD, Baird JK, Hay SI. 2012. G6PD deficiency prevalence and estimates of affected populations in malaria endemic countries: a geostatistical model-based map. PLoS Med 9:e1001339.

Ingram VM. 1956. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178:792–794.

Kasehagen LJ, Mueller I, Kiniboro B, Bockarie MJ, Reeder JC, Kazura JW, Kastens W, McNamara DT, King CH, Whalen CC, Zimmerman PA. 2007. Reduced *Plasmodium vivax* erythrocyte infection in PNG Duffy-negative heterozygotes. PLoS ONE 2:e336.

Kidson C, Lamont G, Saul A, Nurse GT. 1981. Ovalocytic erythrocytes from Melanesians are resistant to invasion by malaria parasites in culture. Proc Natl Acad Sci USA 78:5829–5832.

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

Liu S-C, Zhai S, Palek J, Golan DE, Amato D, Hassan K, Nurse GT, Babona D, Coetzer T, Jarolim P, Zaik M, Borwein S. 1990. Molecular defect of the band 3 protein in Southeast Asian Ovalocytosis. N Engl J Med 323:1530–1538.

Livingstone FB. 1973. *Data on the Abnormal Hemoglobins and Glucose-6-Phosphate Dehydrogenase Deficiency in Human Populations, 1967–1973.* Ann Arbor: Museum of Anthropology, University of Michigan.

Livingstone FB. 1984. The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum Biol 56:413–425.

Livingstone FB. 1985. *Frequencies of hemoglobin variants: thalassemia, the glucose-6-phosphate dehydrogenase deficiency, G6PD variants, and ovalocytosis in human populations.* New York/Oxford: OUP.

Livingstone FB, Marks J. 2009. *Abnormal Hemoglobins in Human Populations*: Aldine Transaction, New Jersey, USA.

Luzzatto L, Notaro R. 2001. Protecting against bad air. Science 293:442–443.

Makani J, Williams TN, Marsh K. 2007. Sickle cell disease in Africa: burden and research priorities. Ann Trop Med Parasitol 101:3–14.

Marks J. 1989. Molecular micro- and macroevolution in the primate alpha-globin gene family. Am J Hum Biol 1:555–566.

Mason PJ, Bautista JM, Gilsanz F. 2007. G6PD deficiency: the genotype–phenotype association. Blood Rev 21:267–283.

Menard D, Barnadas C, Bouchier C, Henry-Halldin C, Gray LR, Ratsimbasoa A, Thonier V, Carod JF, Domarle O, Colin Y, Bertrand O, Picot J, et al. 2010. Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. Proc Natl Acad Sci USA 107:5967–5971.

Mercereau-Puijalon O, Menard D. 2010. Plasmodium vivax and the Duffy antigen: a paradigm revisited. Transfus Clin Biol: J de la Societe Francaise de Transfusion Sanguine 17:176–183.

Miller LH, Mason SJ, Clyde DF, McGinniss MH. 1976. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. N Engl J Med 295:302–304.

Modell B, Darlison M. 2008. Global epidemiology of haemoglobin disorders and derived service indicators. Bull World Health Organ 86:480–487.

Murray CJ, Ezzati M, Flaxman AD, Lim S, Lozano R, Michaud C, Naghavi M, Salomon JA, Shibuya K, Vos T, Lopez AD. 2012. GBD 2010: a multi-investigator collaboration for global comparative descriptive epidemiology. Lancet 380:2055–2058.

National Newborn Screening and Genetics Resource Center (NNSGRC). 2011. *National Newborn Screening Status Report*. Austin, TX, USA.

Ohene-Frempong K, Oduro J, Tetteh H, Nkrumah F. 2008. Screening newborns for sickle cell disease in Ghana. Pediatrics 121:S120-S121.

Olivieri NF, Muraca GM, O'Donnell A, Premawardhena A, Fisher C, Weatherall DJ. 2008. Studies in haemoglobin E beta-thalassaemia. Br J Haematol 141:388–397.

Patil AP, Gething PW, Piel FB, Hay SI. 2011. Bayesian geostatistics in health cartography: the perspective of malaria. Trends Parasitol 27:246–253.

Patil AP, Huard D, Fonnesbeck CJ. 2010. PyMC: Bayesian stochastic modelling in Python. J Statist Softw 35:1–81.

Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, Anagnou NP, Wajcman H, Hardison RC. 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.

Pauling L, Itano HA, Singer SJ, Wells IC. 1949. Sickle cell anemia a molecular disease. Science 110:543–548.

Piel FB, Howes RE, Patil AP, Nyangiri OA, Gething PW, Bhatt S, Williams TN, Weatherall DJ, Hay SI. 2013a. The distribution of haemoglobin C and its prevalence in newborns in Africa. Sci Rep 3:1671.

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Dewi M, Temperley WH, Williams TN, Weatherall DJ, Hay SI. 2013b. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. Lancet 381:142–151.

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, Weatherall DJ, Hay SI. 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat Commun 1:104.

Powars DR, Hiti A, Ramicone E, Johnson C, Chan L. 2002. Outcome in hemoglobin SC disease: A four-decade observational study of clinical, hematologic, and genetic factors. Am J Hematol 70:206–215.

Rosanas-Urgell A, Lin E, Manning L, Rarau P, Laman M, Senn N, Grimberg BT, Tavul L, Stanisic DI, Robinson LJ, Aponte JJ, Dabod E, et al. 2012. Reduced risk of *Plasmodium vivax* malaria in Papua New Guinean children with Southeast Asian Ovalocytosis in two cohorts and a case-control study. PLoS Med 9: e1001305.

Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, Gupta S, Warn P, Allsopp CE, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, et al. 1995. Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376:246–249.

Sellami MH, Kaabi H, Midouni B, Dridi A, Mojaat N, Boukef MK, Hmida S. 2008. Duffy blood group system genotyping in an urban Tunisian population. An Hum Biol 35:406–415.

Serjeant GR, Serjeant BE. 2001. *Sickle cell disease*. Oxford, UK: Oxford University Press.

Streetly A, Latinovic R, Hall K, Henthorn J. 2009. Implementation of universal newborn bloodspot screening for sickle cell disease and other clinically significant haemoglobinopathies in England: screening results for 2005–7. J Clin Path 62:26–30.

United Nations. 2008. *Recognition of Sickle-Cell Anaemia as a Public Health Problem*. Resolution adopted by the General Assembly.

van Baal S, Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, Macek M, Jr., Scriver CR, Patrinos GP. 2007. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 35:D690–D695.

Vichinsky E. 2007. Hemoglobin E syndromes. ASH Education Program Book 2007:79–83.

Weatherall D, Akinyanju O, Fucharoen S, Olivieri N, Musgrove P. 2006. *Inherited disorders of hemoglobin. Disease Control Priorities in Developing Countries* (2nd Edition). New York: Oxford University Press. p 663–680.

Weatherall DJ. 2001. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2:245–55.

Weatherall DJ. 2010. The inherited diseases of hemoglobin are an emerging global health burden. Blood 115:4331–4336.

Weatherall DJ. 2011. The challenge of haemoglobinopathies in resource-poor countries. Br J Haematol 154:736–744.

Weatherall DJ, Clegg JB. 2001. *The Thalassaemia Syndromes*. Oxford/Malden, MA: Blackwell Science.

Whitelaw TH. 1923. Modern epidemiology. Can Med Assoc J 13:428–431.

Williams TN. 2006. Human red blood cell polymorphisms and malaria. Curr Opin Microbiol 9:388–394.

Williams TN, Mwangi TW, Roberts DJ, Alexander ND, Weatherall DJ, Wambua S, Kortok M, Snow RW, Marsh K. 2005. An immune basis for malaria protection by the sickle cell trait. PLoS Med 2:e128.

World Health Organization. 2006. *Fifty-ninth World Health Assembly, Resolutions and Decisions*, Annexes (WHA59/2006/REC/1). Geneva.

Wurtz N, Mint Lekweiry K, Bogreau H, Pradines B, Rogier C, Ould Mohamed Salem Boukhary A, Hafid JE, Ould Ahmedou Salem MS, Trape JF, Basco LK, Briolant S. 2011. Vivax malaria in Mauritania includes infection of a Duffy-negative individual. Malar J 10:336.

Zhao X, Li Z, Zhang X. 2010. G6PD-MutDB: a mutation and phenotype database of glucose-6-phosphate (G6PD) deficiency. J Bioinformatics Comput Biol 8 Suppl 1:101–109.

Zimmerman PA, Ferreira MU, Howes RE, Mercereau-Puijalon O. 2013. Red blood cell polymorphism and susceptibility to *Plasmodium vivax*. Adv Parasitol 81:27–76.