Haberdashers' Aske's School

Occasional Papers Series in the Sciences

Occasional Paper Number Nine

Could Sickle Cell Anaemia save your life?

Exploring the intersection of heterozygote advantage with population genetics

Aske Project Winning Paper 2016

Sohum Patel

PAT500@habsboys.org.uk

February 2017

Haberdashers' Aske's Occasional Paper Number Nine

February 2017

Could Sickle Cell Anaemia save your life?

Exploring the intersection of heterozygote advantage with population genetics.

Sohum Patel

Abstract

This paper will consider the paradoxical relationship between sickle cell anaemia and malaria, two diseases ostensibly far removed from one another, but which in reality have greatly influenced one another through the centuries. Why is it that Nigerians don't get malaria quite as much as you might expect? The answer lies in the increased prevalence of sickle cell anaemia and sickle cell trait in the country, which affords individuals relative immunity to the infectious disease. In this paper, I analyse the frequency of the allele which makes all of this possible in a Nigerian population, creating a mathematical model based on evolutionary game theory which aims to predict the rate at which sickle cell anaemia will progress through society over the coming centuries.

Introduction

In its efforts to communicate the paradoxical nature of heterozygote advantage, this paper will refer to sickle cell anaemia and malaria, two ostensibly unrelated conditions; one a blood disorder, the other an infectious disease, but whose fates have been intertwined with one another through the centuries.

Mosquitoes are known as the world's deadliest animal (Gates 2016), and their role as vector for the *Plasmodium* parasite has caused untold misery around the world throughout the course of human history, killing hundreds of thousands of children every year (WHO, 2016a). But astonishingly, research (Brain, 1952) has shown that having the genes for sickle cell anaemia, or better still, sickle cell trait, will lower the number of deaths per hundred thousand caused by malaria.

Studying sickle cell anaemia is crucial to our efforts to find a cure for malaria, as by understanding the mechanisms by which the sickled haemoglobin prevent *P. falciparum* and other *Plasmodium* parasites from effectively acting upon an individual, we can attempt to simulate such mechanisms through pharmaceutical means. However, it is even more important that we understand the genetics of sickle cell anaemia and have a broader understanding of evolutionary game theory and population genetics, in order to monitor the relative frequencies of the *A* and *S* allele in society, thus allowing us to predict and prepare for potential outbreaks of malaria or indeed a rapidly increasing prevalence of sickle cell anaemia. This work can also be applied to other examples of heterozygote advantage, such as the relationship between cystic fibrosis and tuberculosis (Poolman and Galvani, 2007).

Heterozygote advantage is rooted in the theory of evolution by the mechanism of natural selection, a view propagated by the works of Darwin and Wallace. It is typically used to comment upon the higher relative evolutionary fitness of an organism that is heterozygous for a particular gene, compared to those who are homozygous for the same gene. For an individual's genes to be determined as heterozygous, one of the alleles must be dominant, whilst the other must be recessive. Homozygous genes contain alleles which are either both dominant or both recessive. As sickle cell anaemia is caused by a mutation at a single locus of the HBB gene, heterozygote advantage may be referred to in this paper as over dominance (Charlesworth and Willis, 2009).

When discussing evolutionary fitness, it will be predominantly used to refer to *reproductive success*, which is the typically assumed definition for the somewhat ambiguous term, with myriad variations and slight nuances existing in the field of biology. Reproductive success will in turn refer specifically to the 'number of offspring reaching sexual maturity' (Crognier, 2003). This definition is strengthened by the inclusion of 'reaching sexual maturity' as it is a common misconception that the mechanism of natural selection is driven by the number of children an individual produces. Rather, it is the number of *grandchildren* produced by an individual which is the true measure of evolutionary fitness, as throughout human history, on average, due to high rates of infant mortality, a very high proportion of children do not survive long enough to reproduce themselves. Consequently, whilst it may appear that an individual with a dozen offspring in the 15th century may have a greater chance of their particular advantageous alleles

surviving through the generations, an unrelated individual in separate circumstances but with only four offspring, is more likely to have their alleles survive through their lineage, as the same resources may be concentrated upon a smaller number of offspring, thus decreasing the childhood mortality rate and increasing the number of offspring who may reach the age of sexual maturity and so pass on these alleles down the generations.

Sickle Cell Anaemia

Sickle cell disease is a group of disorders related to the structure of the red blood cell in which abnormal haemoglobin is produced with a 'sickle' shape. Whilst there are many types of erythrocytes which may be present in a human, haemoglobin A (HbA) is typically understood to be 'normal' haemoglobin, containing two α-globin and two β-globin chains with a biconcave tertiary structure.

Of the three key types of point mutations (Lodish et al. 2000), sickle cell anaemia is caused by a missense mutation (Frenette and Atweh, 2007). These are substitution mutations, in which there is a change in the nitrogenous base sequence of the HBB gene such that a new amino acid is coded for in place of another in the same position. There is a single change in the nitrogenous base sequence of the HBB gene, a length of DNA coding for β-globin primary structure. Consequently, the GAG codon, which codes for the amino acid glutamic acid, is mutated into GTG, which codes for valine, which is the 6^{th} amino acid in the primary structure of HbS' β-globin chain. This results in HbS subunits being substituted for the β-globin chains and homozygous have two such sub-units, which aggregate together in low pO_2 conditions, forming long chains which disrupt the tertiary structure of the haemoglobin molecule (Nelson, 2015).

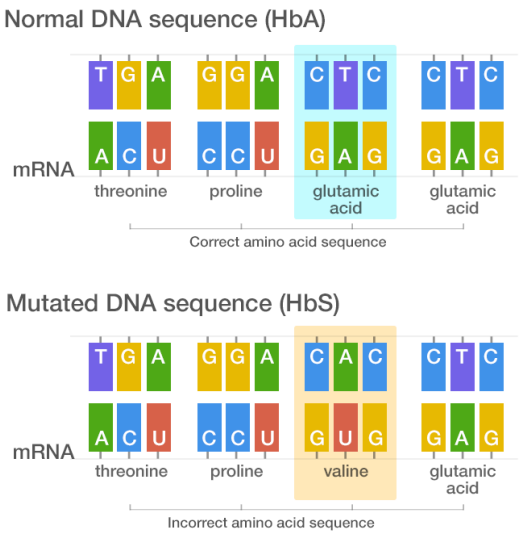


Figure 1: Diagram detailing the change in primary sequence (order of amino acids) caused by a mutation of the HBB gene; (Sickle Cell DNA, courtesy of PME, 2016)

Firstly, before exploring the potential implications of heterozygote advantage in natural selection, one must first ascertain whether sickle cell anaemia is actually inheritable, and not caused by environmental conditions.

We can do so by using the following formula in order to find the sibling recurrence risk ratio:

Figure 2: formula for sibling recurrence risk ratio, courtesy of (Pritchard, 2015)

K_s is the probability that the sibling (or another family member, provided that the results are standardised when used for comparison) of an individual who carries the condition also displays the phenotype for the same condition themselves, where K is the prevalence of the disease to the general population. If $\lambda s \sim 1$, then it may be said with a high level of confidence that the disease is not inheritable.

$$\lambda_s = \frac{K_s}{K}$$

In the absence of any raw statistical data existing for K_s, we will use Mendel's Law of Segregation in order to obtain an accurate estimate value.

K_s = P(sibling | individual)

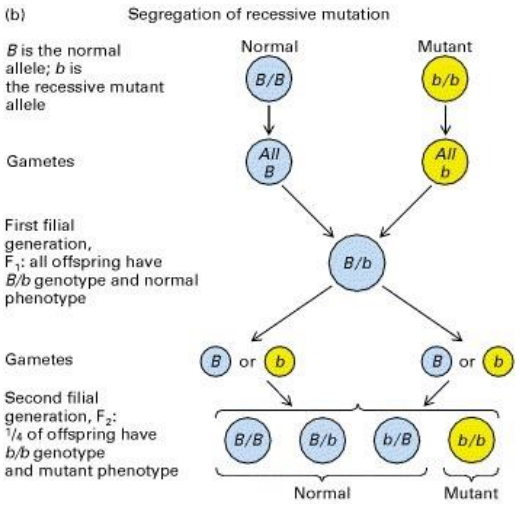


Figure 3: Segregation of recessive allele, courtesy of Lodish et al. (2000)

Sickle cell anaemia is autosomal recessive, which means that in order for one to suffer from the symptoms of the condition, they must possess both alleles of the HBB gene which codes for HbS. Individuals which are heterozygous for sickle cell anaemia are known as carriers and possess only one allele for the HbS coding gene and do not outwardly exhibit any of the symptoms.

The following notation will be employed throughout to denote the various genotypes:

AA: normal
AS: sickle cell trait (heterozygote)
SS: sickle cell anaemia

Therefore, for an individual to have the genotype SS, there are two potential cases:

I. Both parents have the SS genotype:

	S	**S**
S	SS	SS
S	SS	SS

Table 1: Punnett square created by this author detailing the genotypes produced by two parents of the SS genotype

This will result in there being a 100% chance that any offspring produced by the parents will display the sickle cell anaemia phenotype. Consequently, given that there are more than one offspring, the K_s value is 1.

II. One parent has the SS genotype, whilst the other has the AS genotype:

	S	**S**
A	AS	AS
S	SS	SS

Table 2: Punnett square created by this author detailing the genotypes produced by two parents of the AS and SS genotypes

There is a 50% chance that the sibling will have sickle cell trait and a 50% chance that they will have sickle cell anaemia, so the K_s value is 0.5.

III. Both parents have the AS genotype:

	A	**S**
A	AA	AS
S	AS	SS

Table 3: Punnett square created by this author detailing the genotypes produced by two parents of the AS genotype

There is a 25% chance the sibling will be normal, 50% chance that they will have sickle cell trait, and a 25% chance that they have sickle cell anaemia. Consequently the K_s value is 0.25.

Assuming that each of these three cases is equally likely to occur, the probability that an individual with sickle cell anaemia has a sibling with the same condition (given that they do in fact have siblings), is 0.583.

Whilst official figures vary, the global prevalence of the disease is ~ 0.231% (WHO, 2016b), with a value for K of 0.00231.

$$\lambda s = 0.583 / 0.00231 = 252.8$$

The extremely high value for λs leaves little doubt that sickle cell anaemia is an inheritable disease, as the probability that a sibling carries the disease given that an individual does is far greater (252.8 times greater, to be exact) than the prevalence of the disease in the general population.

That being said, due to the lack of data available, as a primary consequence of genome mapping techniques such as linkage analysis being used over the past few decades as opposed to non-genetic means of identifying inheritance, there are certain limitations in the use of the sibling recurrence risk ratio value:

- Circular logic - By using the Mendelian Law of Segregation to prove the inheritance of sickle cell anaemia, it could be argued that the K_s value is constructed on the assumption that the disease *is* inheritable, as opposed to attempting to prove that fact.

- Unequal proportions - By assigning all three cases by which a homozygous recessive SS trait is created with equal 1/3 proportions, the value of K_s is less accurate than if data on the relationships between parents with certain allele combinations had been accounted for. For example, there is intuitively a greater chance that two carriers of sickle cell trait, both with an AS genotype, will produce offspring than two people who both have the disease. This assumption is given that the inheritable nature of the disease is well known to the public, and they are more likely to be unaware that they are carriers given that there are no obvious phenotypical markers for heterozygotes. Therefore, a weighting, proportional to the frequency of long-term, sexual relationships between individuals of the different genotypes, should be given. However, without raw data, such musings are subjective and unsubstantiated, and so the decision was made to avoid speculation and to simply give equal weightings to all three cases.

Even in a simple model created by this author, attempting to find the sibling recurrence risk ratio given only a theoretical knowledge of the disease, but without empirical data, various underlying assumptions have decreased its accuracy. As the field of genetics intersects with geography and sociology, with all of the complex human factors associated with the social sciences, it is made increasingly apparent that the more complex models used later can only provide evidence of a trend, a vague direction, as opposed to a definitive answer.

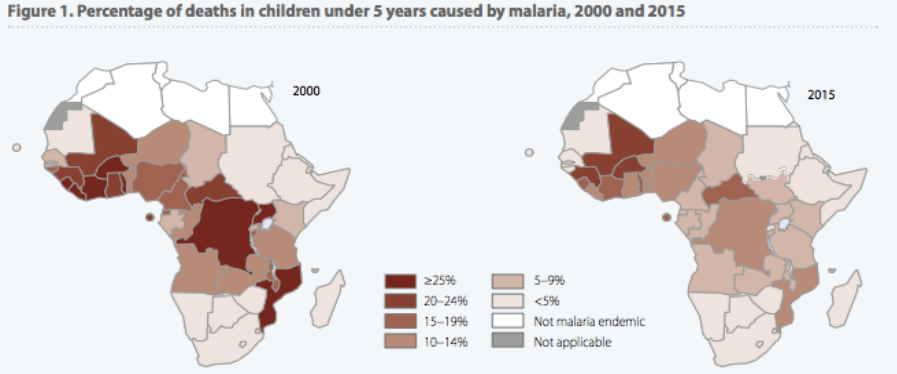
Malaria & 'The Haldane Hypothesis'

In a 1948 Stockholm conference, as recorded by (Bengtsson and Tunlid, 2010), JBS Haldane spoke about the reasoning behind his belief about the key role malaria plays as a selective pressure driving evolution, a topic about which he later wrote (Haldane, 1949):

I. *High prevalence in children:* For a disease to act as a selective pressure, it must in some way limit the reproductive success of the individual. As such, late-onset diseases such as Alzheimer's are far less effective selective pressures than those which affect individuals before or during their reproductive age, spanning approximately 'from ages 12 to 49' (Farlex, 2003). By the estimations of the WHO (WHO, 2016a), 69% of the approximately 438,000 deaths caused by malaria in 2015 were by children under five years old. Evolutionary fitness is determined by the number of gametes which can be produced by an individual. As malaria resulted in the deaths of 302,220 children in 2015, it proves an extremely strong selective pressure, as these individuals will not be able to produce any gametes whatsoever, having died as a result of the disease before reaching sexual maturity.

II. *High mortality rate:* In order for a disease to have an affect the evolutionary fitness in individuals, it must restrict the number of gametes produced and it

may do so by either rendering organisms impotent or killing them before they reach their reproductive age. Malaria is a leading cause of death in Africa, accounting for an estimated 22% (IVCC, 2016) of all deaths by children under the age of five. Consequently, fewer children without resistance-providing genes survive in high risk regions, and therefore they reproduce less and the proportion of individuals without the aforementioned genes will decrease over



the following generations.

Figure 4: courtesy of (WHO and Maternal and Child Health Epidemiology Estimation Group, 2015)

III. *Widespread impact:* Malaria, unlike other intermittent diseases such as the bubonic plague, has been a constant presence throughout the history of *Homosapiens* in Africa, Asia and South America (CDC 2016). It is caused by four separate pathogens: *P. malariae*, *P. vivax*, *P. ovale* and *P. falciparum* (malaria.com, 2011). The former three species are believed to have been present and evolved with humans, with the earliest samples of a malaria pathogen being found from the Paleogene period over 25 million years ago (Poinar, 2005). Therefore, the high presence of malarial parasites in society throughout the history of mankind, almost entirely independent of geographical location, having been present in almost every major landmass on Earth, barring Antarctica, has resulted in a strong selection pressure being placed by the disease. It allows for each generation to gradually increase in the proportion of those with resistance-conferring alleles. In stark contrast, a very specific disease with a high mortality rate and high prevalence in children, but which occurs once for a ten year period every millennia, would not impose a strong selection pressure. Though the immediate next generation may have a higher proportion of resistant alleles, with the absence of a selection pressure over the following centuries, this initially increased proportion will gradually decrease to its former state of equilibrium. The increased energy demand required for the resistant alleles to act will, in the absence of the disease's selective pressure, act in its own way as a selective pressure, and select *against* the disease-resistant genes. These organisms will have a greater energy demand than its competitors and so be more likely to die before reproducing.

Therefore, we can see that malaria serves as the quintessential example of a selection pressure due to the three key attributes outlined above. Whilst none of the qualities are in any way unique to malaria, it is the combination of all three which result in the disease providing such a strong pressure.

Heterozygote advantage

This refers to a case in which a heterozygote for a particular gene, such as Aa, has a greater evolutionary fitness and reproductive success than either homozygote, AA and aa.

It occurs due to the fact that:

i. AA: those with the normal phenotype, in malarial regions, are more susceptible to having the infectious disease, and so are more likely to die before reproducing.

ii. SS: those with sickle-cell anaemia might survive malaria, but are likely to die due to their condition before reproducing.

iii. AS: heterozygotes with sickle cell trait are both less susceptible to malaria, and do not have sickle cell anaemia, so will have a longer average life expectancy than either of the other two genotypes, and will therefore be more likely to reproduce to maximal capacity.

There are a number of reasons why the AS genotype is less susceptible to malaria, as detailed by (Luzzatto, 2012):

i. *Plasmodium falciparum* is extremely specific and cannot digest sickled haemoglobin. Therefore, it does not receive any nutrition and consequently will die.

ii. The presence of the *Plasmodium* parasite leads to sickling as shown by the research of Luzzatto, (1970) and Lang et al. (2009) as RBC stick to the capillary wall. This decreases the O_2 tension, resulting in the structure being more likely to be engulfed by neutrophils in the process of phagocytosis (Ayi et al. 2004).

iii. Studies have also shown that heterozygous individuals with sickle cell trait acquire immunity to malaria at a greater rate after initial infection, and have a far more efficient secondary immune response (Verra, 2007).

iv. A greater number of antibodies specific to the PfEMP-1 antigen are produced, as shown by (Marsh et al. 1989) and (Cabrera et al. 2005). PfEMP-1 is an antigen of *P. falciparum* (Pasternak and Dzikowski, 2009) and the increased specific antibody production results in a greater rate of the *Plasmodium* parasite being removed from the system.

v. Although heterozygotes can be infected by *Plasmodium* and be diagnosed with malaria, they have a decreased frequency of erythrocytes (Luzzatto 2012) which are affected by the parasite.

vi. Sickled haemoglobin causes the catalyst heme oxygenase-1 to be formed (Ferreira et al. 2009) which in turn breaks down heme to release carbon monoxide Kikuchi et al. (2005) which, after the initial infection of *P.*

falciparum prevents heme, which is not bound to a protein, from collecting together which would otherwise cause cells to undergo apoptosis.

Evolutionary game theory

'An ounce of algebra is worth a ton of verbal argument' - J.B.S. Haldane

Whilst the underlying theory of heterozygote advantage has been explained, it is difficult to fully understand and appreciate its significance without using logic to unravel its complexity. Furthermore, without explaining the mathematics driving evolution, arguments explaining the prevalence of overdominance would be purely conceptual and speculative, lacking scientific rigour.

John von Neumann (Easley and Kleinberg, 2009) created the concept of game theory in order to mathematically model and determine the optimum strategies which should be employed when two separate entities are competing with one another.

But classical game theory assumes that the competing players are entirely rational in their decision making process and possess free will. Therefore individual entities can be trusted to select the course of action, selfish though it may be, which serves themselves best. An example of such a strategy explored in classical game theory is whether or not one state should go to war with another. Both states have, to a certain practical extent, a choice in their own strategy, and can be trusted to make a logically rational decision. However, John Maynard Smith, (Maynard Smith and Price, 1973), noted that in nature individual organisms do not get a choice in their strategy; regardless of one's take on the concept of free will. Despite the potentially rational nature of a mouse, it cannot choose whether or not it is small or large, with the aforementioned being two possible strategies which evolutionary game theory has been created to model. Furthermore, in order for the outcomes of the game to affect the $n+1$ generation, these strategies must be inheritable in nature.

Natural selection, it can be argued, is driven by the competition for resources, as success in such endeavours is directly correlated to reproductive success and therefore, evolutionary fitness. However, as evolution is both process and consequence, but not an independent directly-acting body, it must test out all possible strategies before the optimum strategy is 'decided' upon, by its very nature. Therefore, we can study the alleles and genotypes of the HBB gene in order to determine the best possible combination of alleles which results in an evolutionarily stable population (Easley and Kleinberg, 2009).

To model heterozygote advantage, the relative payoffs for alleles arranging themselves in a particular manner will be tested. The idea of modelling relative fitness with specific reference to sickle cell anaemia is broached by Sober and Lewontint, (1982) and Maynard Smith, (1987) in the collected works of Dupré, (1987). However, their ideas and payoff values have been adapted to account for new research into the various mortality rates associated with sickle-cell anaemia and sickle cell trait in malarial regions, as well as specifically applying their broader interest in evolutionary fitness. With regards to sickle cell anaemia wth evolutionary

game theory, this approach is, to this author's belief, entirely novel in the study of sickle cell anaemia.

As has been earlier articulated, many studies such as Luzzatto, (2012), Verra, (2007) and Marsh et al. (1989) have shown there to be, in malarial regions, a clear advantage in being heterozygous for the HBB gene. Therefore, the AS genotype has been given the outcome value of 10, with the AA and SS outcomes being a certain positive value, x and y respectively, less than 10.

Traditional payoff matrices will have two values for each 'competition', each relating to the payoff for the two competing strategies. However, in this case, both strategies are being used to form a new, different strategy (as alleles form a specific genotype) and so each strategy will be given the same payoff.

	A	**S**
A	10-x	10
S	10	10-y

Table 4: Simplified general form of the payoff matrix and created by this author

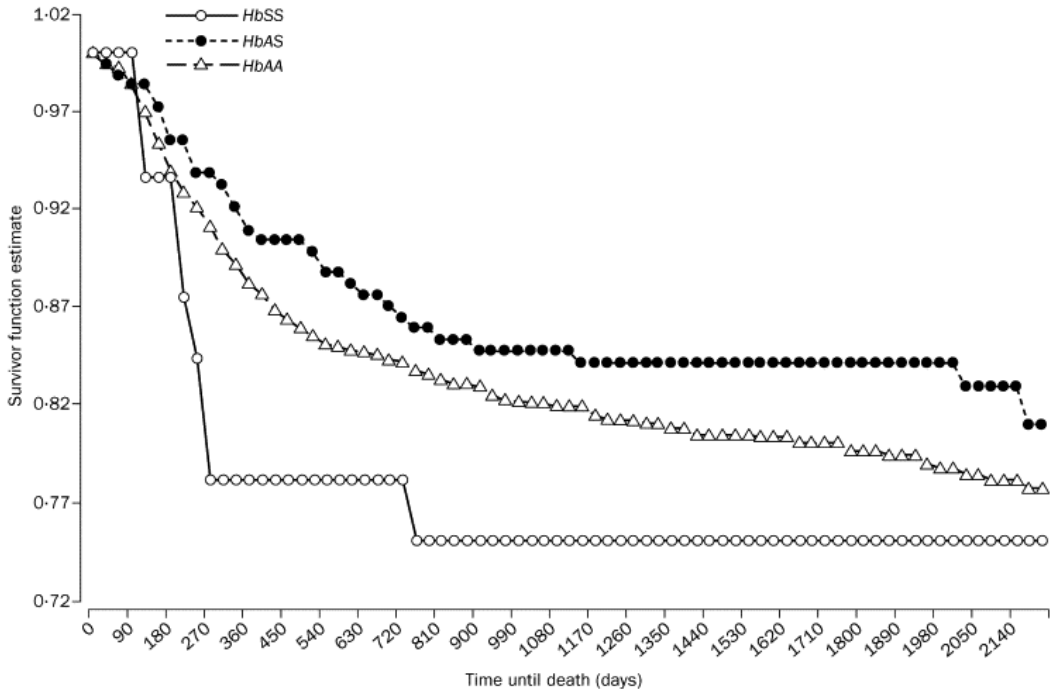


Figure 5: Kaplan-Meier survivor function from Aidoo et al. (2002)

Using Figure 5 which displays the infant mortality rates of babies in Kisumu, Kenya (Aidoo et al. 2002), it is clear that if AS is to be given an outcome of 10, with all

other genotype values relative to that value, AA should be given approximately 8 and SS have a payoff of 5. The SS payoff is lower than what one might expect in a malarial region, but is influenced by research (McAuley et al. 2010) which shows that amongst those homozygous recessive children with the *SS* genotype, their mortality rates upon being infected by malaria was over 1000% greater than that of those with the normal phenotype. The findings of Rucknagel and Neel (1961) were also accounted for, as they found that the ratio in fitness between those with the AA and AS genotype was 0.85:1. The value of 0.85, given the greater infant mortality rate, was decreased to 0.8 (and then scaled up to 8).

Therefore, $x=2$ and $y=5$.

It is important to note that these values are mere approximations based on data related to mortality, but in reality, many more variables must be accounted for including fecundity. However, for the purposes of this model such complexities may be ignored, as it is the *trend* in data, not a definitive quantifiable answer, which we hope will be determined.

	A	**S**
A	8 8	10 10
S	10 10	5 5

Table 5: specific payoff matrix written in conventional form, created by this author

Having created a payoff matrix, we can use it to test whether a particular strategy is evolutionarily stable.

Evolutionary stability is a term first used by Maynard Smith (Maynard Smith and Price, 1973), to describe a strategy which can successfully resist the invasion of a competing strategy and which 'tends to persist once it is prevalent in a population' (Easley and Kleinberg, 2009).

I. Is strategy *s* evolutionarily stable?

Strategy *s* is assigned probability *1-p* whilst strategy *a* is assigned probability *p*.

For *s* to be stable, if $p<1-p$, then $S_{fitness} > A_{fitness}$

The fitness of a strategy is equal to the expected payoff when it is utilised and so a probability distribution table is used to find the expected payoff.

	A	**S**
X	10	5
P(x = X)	p	1-p

Table 6: Expected payoff for strategy s, created by this author

$S_{fitness} = \Sigma x_i p_i$
$= 10p + 5(1-p)$
$= 5p + 5$

	A	**S**
Y	8	10
P(y = Y)	p	1-p

Table 7: Expected payoff for strategy a, created by this author

$A_{fitness} = \Sigma x_i p_i$
$= 8p + 10(1-p)$
$= 10 - 2p$

As stated earlier, for population *s* to be tested for stability, $p<1-p$. Given that $p + (1-p) = 1$, $0 \leq p < 0.5$.

Therefore, for the purpose of this calculation and future similar ones, let $p=0.2$.

$S_{fitness} = 6$
$A_{fitness} = 9.6$

We can then confirm that this is true of all probabilities *p*:

$S_{fitness} > A_{fitness}$
$5p + 5 > 10 - 2p$
$7p > 5$
$p > 0.714$

As $p < 0.5$, no values of *p* exist which can satisfy the above inequality.

Therefore, as $S_{fitness} < A_{fitness}$, strategy *s* is <u>not</u> evolutionarily stable.

II. Is strategy *a* evolutionarily stable?

Strategy *a* is assigned probability *1-p* whilst strategy *s* has probability *p*.

For *a* to be stable, if $p<1-p$, then $S_{fitness} < A_{fitness}$

	A	**S**
Y'	8	10
P(y = Y')	1-p	p

Table 8: Expected payoff for strategy a, created by this author, using general form of Easley and Kleinberg, (2009)

$A_{fitness} = 2p + 8 = 8.4 \ (w/\ p = 0.2)$

	A	**S**
X'	10	5
P(x = X')	1-p	p

Table 9: Expected payoff for strategy s, created by this author, using general form of Easley and Kleinberg, (2009)

$S_{fitness} = 10 - 5p = 9 \ (w/p = 0.2)$

With $p = 0.2$, $A_{fitness} = 8.4$ and $S_{fitness} = 9$ so therefore, $A_{fitness} < S_{fitness}$ so strategy *a* is unstable. However, this is not always the case.

For the population to be stable at strategy *a*:

$S_{fitness} < A_{fitness}$

$10 - 5p < 2p + 8$

$2 < 7p$
$p > 2/7$

For $p < 1-p$, and $p + (1-p) = 1$, $p < 0.5$.

Therefore, when $0.286 < p < 0.5$, strategy *a* is evolutionarily stable.

These results are extremely noteworthy, as though to some extent the values were arbitrary, they can still allude to many key insights into the current population dynamics of heterozygote advantage. They imply that for a population of *A* alleles

to be stable, between 28.6% and 50% of the alleles must be *S*. Therefore, it is clear that *S* alleles are actually a necessity in a stable population in malarial regions.

The implications of this are more evident once one reminds themselves of what it means to be an evolutionarily stable population, with *a* being a stable strategy. Generally, simple evolutionary game theoretic models would determine that for the stable strategy *a*, at any frequency of $a > 0.5$, the population would be stable. That is, if *a* is the majority strategy, it will be able to resist change and the *s* strategy will not be able to successively invade.

Nash equilibrium occurs when the optimum strategy as response to strategy *x* is strategy *x*.

Therefore, B. Thomas posited (Thomas, 1985) that for an evolutionarily stable strategy to be definitively classified as such, it must satisfy both Nash equilibrium, that Payoff(A,A) ≥ Payoff(S,A), and also the condition posited by Maynard Smith, (Maynard Smith and Price, 1973) with Payoff(A,S) > Payoff(S,S).

As is evident by looking at the payoff matrix in *Table 5*, only the second condition is satisfied:

Payoff(A,S) > Payoff(S,S)

Therefore, it is evident that this is a more complex system than those traditionally broached in evolutionary game theory, as it is not in Nash equilibrium.

This means that at allele frequencies, $0.5 \leq A \leq 0.714$, the population of *A* alleles 'playing' strategy *a* is stable and will resist the invasion of any *S* alleles. However, somewhat unusually in the context of evolutionary game theory, at higher frequencies $0.714 < A < 1$, the population then becomes unstable and becomes susceptible to an invasion of *S* alleles.

From this it appears evident that there is a minimum limit of *S* alleles required in a population in a malarial region, for strategy *A* to be stable, which is a frequency of 0.286.

When this is extended to real-life scenarios, it is clear that the frequency of *S* alleles in malarial regions will increase up to at least a minimum frequency of 0.286, and past that value. As the population has become stable, the frequency is unlikely to rise. This is because although the population is stable between $0.286 < S < 0.5$, once it reaches a specific stable state, by its very nature it will resist further invasion, and so though the *S* allele may increase in frequency to reach the minimum frequency required to population stability, once it reaches 0.286, no further invasion of *S* alleles will be successful. Furthermore, though it could be argued that a population with an *S* allele frequency greater than 0.5 could indeed *decrease* in frequency before stabilizing at the equilibrium value of 0.5, in practice this would not occur, as no sub-population of size great enough for natural selection to have any clearly observable impact on genotype frequencies possess the *S* allele in such abundance.

Therefore, given that in no sizeable sub-population in the world, $fq(S) \geq 0.286$, natural selection will drive the allele *S* to increase in frequency up until this value. Here it will remain constant until the parameters change of the game change, such as when anti-malarial medications become more effective and accessible, which would in turn decrease the selective advantage afforded by heterozygosity.

Furthermore, *f(S)* must be as small as is theoretically possible in a stable population (i.e. equal to 0.286) as even though *f(AS)* would decrease as a result, it would be many more times greater than *f(SS)*, which has a much lower payoff, as demonstrated by the Hardy-Weinberg equilibrium.

$f(AS) = nf(SS)$

$If\ f(S) = 0.3,\ f(AS) = 2pq = 0.42\ and\ f(SS) = p^2 = 0.09 \therefore n = 4.67$
$If\ f(S) = 0.4,\ f(AS) = 2pq = 0.48\ and\ f(SS) = p^2 = 0.16 \therefore n = 2.5$

Therefore, at higher frequencies of the *S* allele, the frequency of the detrimental sickle cell anaemia phenotype increases at a greater rate than that of the beneficial sickle cell trait phenotype, and thus decreases the average payoff, thus explaining the necessity for an upper limit of the allele's frequency.

Population genetics

'I can write of natural selection with authority because I am one of the three people who know most about its mathematical theory' - J.B.S Haldane, The Causes of Evolution (1932)

Population genetics is an overlapping field of evolutionary biology and statistics which attempts to use mathematical models in order to explore the relative frequencies of alleles in a given population. These models are created with base assumptions, such as rejecting the possibility of chance mutations or the effects of genetic drift. Many of these assumptions must immediately be rejected when applied in real-life scenarios. Though this fact may appear to entirely discredit models as being speculative musings diametrically opposed to the functions of mathematical rigour they profess to be, given the complexity of both biological organisms and biological symptoms, it must be accepted that such closed conditions may never be attainable. Once the models have been made, they can be compared against pre-existing empirical data, with a high correlation between both data sets implying that the model is accurate. Population genetics provides a macroscopic view of biology, applying intimate knowledge of genetics to the study of entire populations over the course of centuries. They provide a means by which Darwinian and Mendelian theories can be quantitatively tested, and as such are an invaluable tool to modern biologists.

Models created by population geneticists will be manipulated and used in this paper to quantify the heterozygote advantage of sickle-cell anaemia in malaria-prone regions as well as modelling historical data of the relative frequencies of the AA, AS and SS alleles in order to forecast changes in the frequency of the aforementioned alleles in the future, and in doing so, predict the rate at which natural selection will operate and whether the AS allele frequency will continue to

increase through the generations or oscillate about and tend towards an equilibrium frequency.

The Hardy-Weinberg equilibrium was independently determined by both G.H. Hardy and Wilhelm Weinberg in 1908 and it is both astonishingly simple and readily derivable. The formula serves as a fundamental cornerstone on the field of population genetics, commenting upon the relative frequency of alleles in a population and demonstrating that, in the absence of other mechanisms of evolution, such as natural selection and mutations, genetic variation is maintained in a population of infinite size (this stipulation is necessary in order to make negligible the effects of genetic drift, which would otherwise influence the results by way of random changes in the sample).

In a simple, two-allele, single locus system, as is present in the case of sickle cell anaemia, the relative frequencies of the various genotypes are easily determined:

$f(AA) + f(AS) + f(SS) = 1$

let $f(A) = p$ *and* $f(S) = q$, *such that* $p+q = 1$

$f(AA) = p^2$
$f(AS) = 2pq$ (Doubled as order does not matter. There is a pq probability of either an AS or a SA genotype)
$f(SS) = q^2$

$p^2 + 2pq + q^2 = 1$

Above is the basic form of the Hardy-Weinberg equilibrium.

The expected heterozygosity (H_E) of a locus, $2pq$, is therefore given by:

$H_E = 1-(p^2+q^2)$

Heterozygosity is highest in a population when $p = q$, which, at a two-allele locus, means that $p = q = 0.5$.

We will use the H_E value obtained through the use of the Hardy-Weinberg equilibrium to find the fixation index of various sub-populations, in order to quantify the differentiation between subpopulations in Sub-Saharan Africa and North America.

Firstly, population statistics about the prevalence of sickle cell anaemia and sickle cell trait must be obtained. To increase the reliability of the results, given that the precise frequency is constantly changing, a range of different sources were used to calculate this data.

In Nigeria, a 1988 study (Akinyanju, 1989) into newborns revealed that of the 5.4 million births in the year, 1.1 million infants has sickle cell trait (20.4%) and 90,000 were born with the condition itself (1.67%). A study into the genotype's prevalence in the entire population (Uzoegwu and Onwurah, 2003), found that 26.94% of the

population have sickle cell trait and 3.52% are carriers, whilst a study of Benin City, Nigeria (Nwogoh et al, 2012) discovered that 23.19% of the population have the AS genotype, and 2.23% have the SS genotype. The mean average of these values will be put in the table below, and the relative frequencies of the genotype will be scaled up to sum to a sub-population of 100 individuals.

Less study has been conducted into the prevalence of sickle cell *trait* in the USA, but statistics about those with the condition are abundant, and this data can be used with the Hardy-Weinberg principle to derive the frequencies of the other genotypes.

Whilst the specific number of cases of sickle cell anaemia in the USA is unknown, and therefore the frequency of the SS genotype cannot be ascertained through empirical evidence alone, a 2014 CDC report (CDC, 2014) on sickle cell trait in 2010 revealed that the incidence rate is '15.5 per 1,000 newborns overall'. Therefore, $f(AS) = 0.0155$.

$p^2 + 0.0155 + q^2 = 1$
$p^2 + q^2 - 0.9845 = 0$
$p^2 + (1-p)^2 - 0.9845 = 0$
$2p^2 - 2p + 0.0155 = 0$
$p = 0.9922$ $\Rightarrow f(AA) = p^2 = 0.9845$
$q = 0.007811$ $\Rightarrow f(SS) = q^2 = 0.00006102$

Sub-population	AA	AS	SS
Nigeria	74.01	23.51	2.48
USA	98.45	1.55	0.0061

Table 10: adjusted sample population size with observed genotype frequencies and created by this author

Firstly, we must ascertain whether the obvious difference in the genotype frequencies seen in Table 1 are merely the result of chance or not, and to do so, the Chi-Squared Test will be employed.

To do so, a null hypothesis must be established, which must consequently be refuted if any further assessment into the relative frequencies between the two sub-populations is to be accurately assessed.

The null hypothesis is that there is no statistically significant difference between the frequency of the three genotypes, AA, AS and SS, of the HBB gene in the sub-populations of Nigeria and the USA.

1) Allele frequency

Nigeria: $f(A) = (2 \times 74.01 + 23.51) / 200 = 0.858$
$f(S) = (2 \times 2.48 + 23.51) / 200 = 0.142$

USA: $f(A) = (2 \times 98.45 + 1.55) / 200 = 0.992$
$f(S) = (2 \times 0.0061 + 1.55) / 200 = 0.008$

2) Expected dominance

This can be determined through use of the Hardy-Weinberg equilibrium function, being equal to *2pq*.

Nigeria: $H_E = 2 \times 0.858 \times 0.142 = 0.243672$
USA: $H_E = 2 \times 0.992 \times 0.008 = 0.01597$

Table 11: adjusted population size with expected[1] genotype frequencies, created by this author

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Figure 6: Chi-squared test formula; OCR Biology Mathematical Skills Handbook Extract (OCR, 2016)

Sub-population	AA	AS	SS	Total
Nigeria	86.03	12.96	1.01	*100*
USA	86.03	12.96	1.01	*100*
Total	*172.05*	*25.92*	*2.03*	***200***

Sub-population	AA	AS	SS
Nigeria	1.6794	8.5882	2.1395
USA	1.7931	10.0454	0.9975

Table 12: $(f_o - f_e)^2 \div E$ values for both sub-populations, created by this author using collected data

[1]The expected frequencies are calculated by: column total x row total / overall total. As both sub-populations have the same row total, and the column totals are also in the same proportion to one another, the expected values for each genotype within each sub-population is the same.

The Chi-squared value is the sum of the data values in the table:

$x^2 = 25.2431$

In order to understand the meaning of the Chi-squared value, it must be viewed in context, by calculation the degrees of freedom of the original data set, which essentially equates to the number of data points used in order to derive a Chi-squared value.

df = (no. rows - 1)(no. columns - 1) = 1 x 2 = 2

TABLE 6-1 Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5

Table 13: Critical values of x^2, courtesy of Carr, (2015)

We test at the 5% level, with a probability value of 0.05 against the df value of 2.

Reading off the table, the critical value is 5.991. As x^2 > critical value, with 25.2431 > 5.991, it is with 95% certainty that the null hypothesis can be rejected. With a less than 5% probability that the difference between the two subpopulations' genotype frequencies is due to chance.

Though the 5% level is often used in biology as convention, Table 4 shows that the x^2 value is greater than the critical value of 10.597 at the 0.5% level, with probability 0.005. Therefore, the certainty value increases to 99.5% and the difference is not due to chance.

Using the Chi-squared test, we can reject the null hypothesis with a high level of confidence, and determine that it is very likely that there is a significant difference between the genotype frequencies of the USA and Nigeria.

This is likely due to the fact that the risk of malaria differs drastically amounts the two populations. Whilst the US Embassy in Nigeria reports ('Nigeria Malaria Fact Sheet', 2011) that Nigeria has approximately a 100,000,000 cases of malaria a year, the USA had 1925 cases of malaria in 2011 (CDC, 2016), quoted by the CDC as a '40 year high', which demonstrates that the prevalence of malaria in the USA is estimated to be 99.998% less than that in Nigeria.

Now, fixation indices for the genotype frequency values will be calculated in order to find the genetic differentiation between the two sub-populations, as well as the

degree to which the heterozygosity values correlate with the Hardy-Weinberg equilibrium model.

In lieu of statistical modelling software, the University of Wyoming's manual work-through method will be used with this author's own assimilated data to find the values of the F-statistics.

3) Observed dominance

Nigeria: H_O: 23.51 / 100 = 0.2351
USA: H_O: 1.55 / 100 = 0.0155

4) Inbreeding Coefficient (Mallet, 2016)

This is used in order to test whether inbreeding occurs within a population, by the difference between expected and observed dominance, divided by the expected dominance. This is because inbreeding reduces the heterozygosity of a population, as recessive deleterious alleles which are typically unique within a family's pedigree but which are not expressed in the phenotype can combine with one another, resulting in inbreeding depression (UC Berkeley, 2014) and the expression of the potentially dangerous phenotype.

Therefore, if $H_E > H_O$, then inbreeding occurs, whilst if $H_O > H_E$, then outbreeding occurs, with entirely random mating.

Nigeria: $F = (0.243672 - 0.2351) / 0.243672 = 0.03518$
USA: $F = (0.01597 - 0.0155) / 0.01597 = 0.02943$

In both sub-populations, as F>0, a degree of inbreeding occurs, which slightly limits the heterozygosity of the HBB gene. However, both values are extremely small, and it is clear from the results that neither inbreeding, nor certainly outbreeding, have any great effect on the prevalence of the AS genotype.

5) Allele Frequencies

These will be conducted across both sub-populations, with ā denoting the frequency of the A allele and x̄ denoting the frequency of the S allele.

ā = (2 x 74.01 + 2 x 98.45 + 23.51 + 1.55) / 400 = 0.9250
x̄ = (2 x 2.48 + 2 x 0.0061 + 23.51 + 1.55) / 400 = 0.0750

ā + x̄ = 1; 0.9250 + 0.0750 = 1 ✓

6) Heterozygote Index (with paraphrased definitions for each index courtesy of Shane, 2016)

A. H_O average of each individual within Nigeria and USA :

H_I = (23.51 + 1.55) / 200 = 0.1253

B. H_E average of each individual within Nigeria and USA, provided that mating is entirely random:

$H_S = (0.243672 \times 100 + 0.01597 \times 100) / 200 = 0.129821$

C. H_E across both sub-populations of Nigeria and USA:

$H_T = 2\bar{a}\bar{x} = 2 \times 0.9250 \times 0.0750 = 0.13875$

7) F-Statistics

D. Overall inbreeding coefficient

$F_{IS} = (0.129885 - 0.1253) / 0.129885 = 0.03530$

Used to measure the extent to which the genotypes of the entire population of the USA and Nigeria combined deviate from the Hardy-Weinberg model. The slight positive value calculated implies that a small degree of inbreeding occurs within the population, which decreases the heterozygosity of the entire population.

E. Overall fixation index

$F_{IT} = (0.13875 - 0.1253) / 0.13875 = 0.09694$

F. Fixation index

$F_{ST} = (0.13875 - 0.129821) / 0.13875 = 0.06435$

This is the key F-statistic, and measures the average decrease in heterozygosity due to the process of genetic drift, which occurs as individuals move between the USA and Nigeria.

This slight decrease in heterozygosity across the entire population can be explained by the Wahlund effect (USP, 2013) as the different allele frequencies between the Nigerian and USA sub-populations will result inevitably in lowered heterozygosity, even though, as has been proven by the Hardy-Weinberg model, both sub-populations are in a genetic equilibrium.

This is because *H* is equal to the mean of both sub-populations' AS frequency, given in the Hardy-Weinberg model as *2pq*, or rather, as is being used in this case, *2as*.

Therefore, in a population divided into two sub-populations: $H = a_1s_1 + a_2s_2 = a_1(1-a_1) + a_2(1-a_2)$

In a single-population system without any sub-populations: $H = 2as = 2a(1-a)$

$a_1(1-a_1) + a_2(1-a_2) < 2a(1-a)$ at all values of a except for when $a_1 = a_2$

This shows that due to the non-random mating naturally enforced by the geographic distance between Nigeria and the USA, heterozygosity will naturally decrease and be slightly lower than its expected value.

There is moderately high genetic differentiation (Shane, 2016) between the two sub-populations of individuals within Nigeria and USA due to the positive, and relatively large, fixation index value of 0.06389. The value means that of the genetic variation both within and between the two subpopulations, 6.389% is *between* Nigeria and USA whilst 93.611% of the variation is between individuals within the same population.

The relatively slight values of the F-statistics mean that the Hardy-Weinberg equilibrium, although not a perfect predictor of genotype frequency, given that it assumes entirely random mating whilst the F_{IS} coefficient values of 0.0353 proves that a slight degree of inbreeding occurs, is still a valid model, which can therefore continue to be used in relation to the HBB gene.

Given the fitness values of the various genotypes of the HBB gene explained in the discussion on payoff matrices in relation to evolutionary game theory, the Hardy-Weinberg equilibrium model can be manipulated. The general form used to find *q'* is illustrated by Okasha, (2012), but this function will be extended in order to model the Nigerian population's allele changes over many generations. This is possible as we have earlier ascertained that the data collected from both the Nigerian and USA sub-populations agrees to a great degree with the Hardy-Weinberg model.

	AA	**AS**	**SS**
Phenotype	Normal	Sickle cell trait	Sickle-cell anaemia
H-W expression	p^2	2pq	q^2
Fitness	$w_{11} = 8$	$w_{12} = 10$	$w_{22} = 5$

Table 14: Hardy-Weinberg model and evolutionary fitness, using fitness values created by this author

As reproductive fitness and the number of gametes produced are proportional to one another, the mean fitness, 7.67 multiplied by the population size, N, will calculate the total number of gametes produced.

Therefore, we can use the relative reproductive fitness values to calculate the number of *S* gametes in the second generation, and consequently find the frequency of the S allele.

$q' = (q^2 w_{22} + pq\, w_{12}) / E(w)$

$E(w) = w = p^2 w_{11} + 2pq\, w_{12} + q^2 w_{22} = 8.426852$

In the Nigerian sub-population: $q' = (0.142^2 \times 5 + 0.142 \times 0.858 \times 10) / 8.43$

$$q' = 0.1565448$$

q' represents the frequency of the *S* allele in the second generation.

$$\Delta q = q' - q$$
$$= 0.0145448$$

Δq is the increase in the frequency of the *S* allele between generations and as Δq > 0, it has been shown that the *S* allele will indeed increase in frequency. When Δq is equal to 0, it is implied that an equilibrium state has been reached.

As has been shown through the use of evolutionary game theory, in malarial regions such as Nigeria, the optimum frequency for the *S* allele, *q*, is 0.286.

Therefore, we can continue to manipulate the Hardy-Weinberg model, together with

Generation	q (S)	p (A)	Δq
1	0.142	0.858	
2	0.157	0.843	0.015
3	0.171	0.829	0.014
4	0.184	0.816	0.013
5	0.197	0.803	0.013
6	0.208	0.792	0.012
7	0.219	0.781	0.010
8	0.228	0.772	0.009
9	0.236	0.764	0.008
10	0.244	0.756	0.007
11	0.250	0.750	0.006
12	0.256	0.744	0.005
13	0.260	0.740	0.005
14	0.264	0.736	0.004
15	0.268	0.732	0.003
16	0.271	0.729	0.003
17	0.273	0.727	0.002
18	0.275	0.725	0.002
19	0.277	0.723	0.002
20	0.278	0.722	0.001
21	0.279	0.721	0.001
22	0.280	0.720	0.001
23	0.281	0.719	0.001
24	0.282	0.718	0.001

spreadsheet software, in order to predict the number of generations required to reach this value.

Table 15: values generated by this author using the recursive function outlined earlier and Numbers spreadsheet software[2]

This table of values shows that in the 24th generation, counting the current present day generation to be the 1st generation, the Nigerian population will be within 2% of reaching its optimum allele frequency of 0.286 in order to maximise the expected

[2] A variation on the original formula is to replace 'E(w)' with '$8p^2 + 20pq + 5q^2$' in order to prevent the mistake of assuming that E(w) is constant through the generations, whilst in fact, as the generations progress, though the relative fitness of each genotype remains the same, the changing frequencies in the alleles affects the mean fitness.

payoffs, as evidenced through the game theoretic model broached earlier. However, modelling the future generations on the spreadsheet software has shown that the frequency of the *S* allele in Nigeria will never be equal to 0.286, even though every generation has a positive Δq value, demonstrating that each generation has a greater *fq(S)* than that which came before it. The frequency value tends towards the equilibrium value of 0.286 (or rather, more specifically, the value of 2/7), but it will never reach said value as the Δq value decreases. Using the approximate value of 25 years per generation (ancestry.co.uk 2012), it will take 600 years for Nigeria to reach this allele frequency.

However, it must be noted that over time and multiple generations, the relative payoff values for the various genotypes are likely to change, particularly as anti-malarial drugs are developed, as they will decrease the selective advantage provided by being heterozygote, and so would lower the evolutionary fitness provided by the *AS* genotype. That being said, accounting for such changes over time in the future would be purely speculative, and other unforeseen events may occur which increases the selective advantage of heterozygosity. Consequently, it was decided not to form conjectures and to work with the data afforded to this author at this particular time.

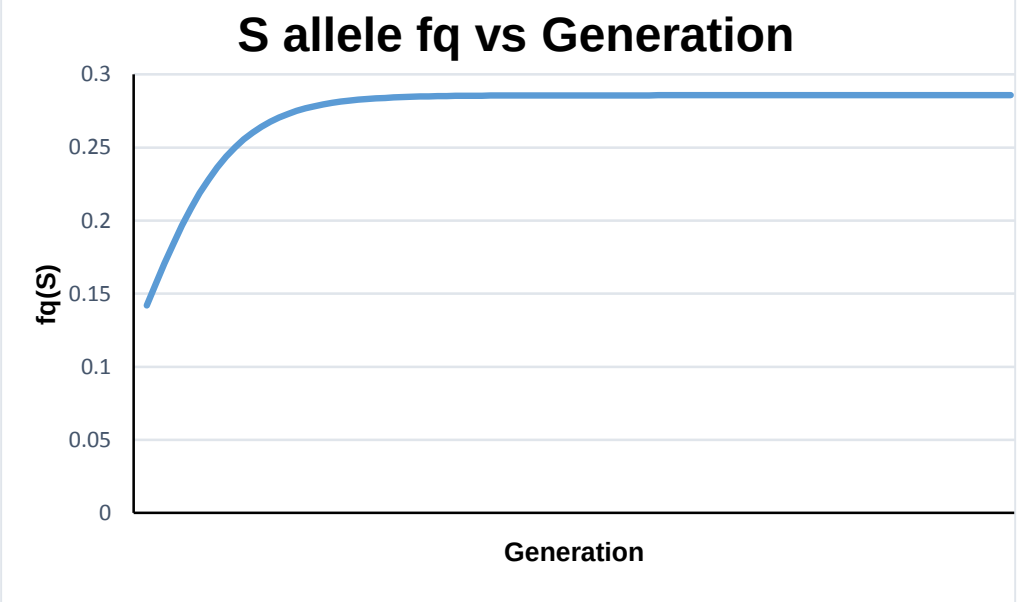


Figure 7: Created by this author using Microsoft Excel and Apple Numbers with the data collected in Table 15 to track the increasing frequency of the S allele in Nigeria over time as it tends to the value of 0.286.

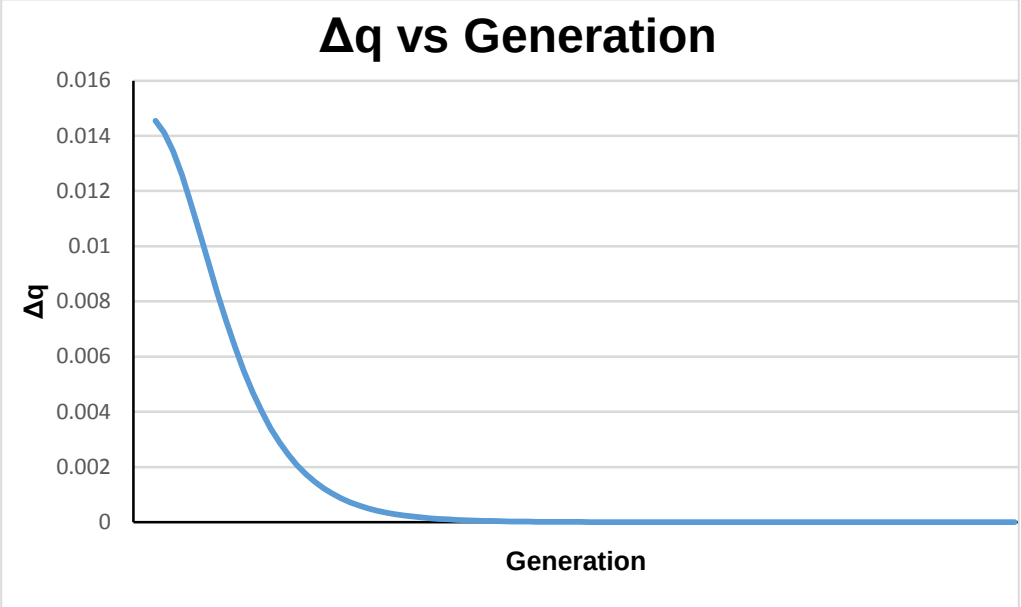


Figure 8: Created by this author using Microsoft Excel and Apple Numbers with the data collected in Table 15 to track the decreasing change in allele frequency of the S allele in Nigeria over time, as the graph tends to a Δq value of 0, which will result in an equilibrium state.

For a system in equilibrium, the following is true. With credit to Okasha, (2012) for the derivation:

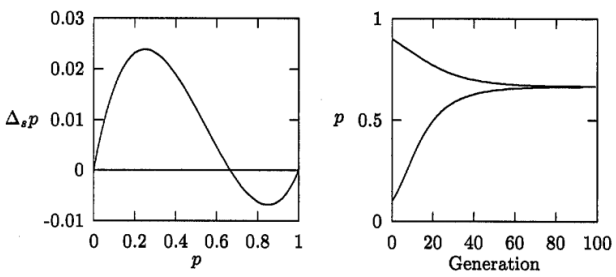


Figure 3.4: Properties of balancing selection with $h = -0.5$ and $s = 0.1$.

$p^* = (w_{12} - w_{22}) / (w_{12} - w_{22}) + (w_{12} - w_{11})$
$p^* = 5/(5+2) = 5/7 \therefore q^* = 2/7$

Figure 9: changing rates of allele frequency through the generations with arbitrary values for h and s, credit to Gillespie, (2004)

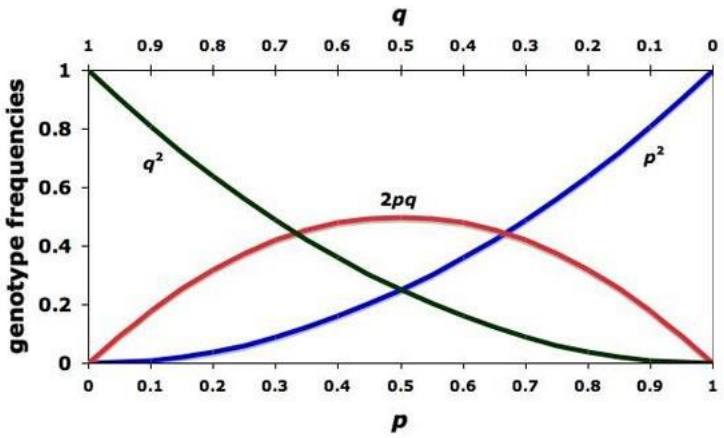
Another method to find the value of q* (otherwise denoted by 1- $\hat{p}$) is to apply the selection coefficients and heterozygous effect values into the formula, a method favoured by John Gillespie in his guide to population genetics (Gillespie, 2004):

$$\hat{p} = \frac{h-1}{2h-1}.$$

Figure 10: equilibrium point equation mentioned by Gillespie, (2004) with $\hat{p}$ referring to the equilibrium point of the S allele.

AA	AS	SS
1	1-hs	1-s
8	10	5
1	1.25	0.625

Table 16: finding the fitness of genotypes relative to the AA genotype.



Therefore, from this we can use simple algebra to ascertain that: $s = 0.375$ and $h = -2/3$.

$\hat{p} = (-5/3) / (-7/) = 5/7 \therefore q^* = 2/7$

All of the above supports the argument that 2/7 being the optimum allele frequency for *S*, supporting the game theoretic model's argument for *f(S)* to be of the same value.

By analysing the differences in the Δq values between generations, and more specifically the decreasing nature of the increase in allele frequency, as well as identifying the asymptotic nature of the optimum frequency value 2/7 for *S*, it is clear that the Nigerian population will evolve over time to select for the *S* allele, but

the rate at which this selection will occur will decrease over time, as though the *S* allele is beneficial. It is only so in relatively smaller proportions, as the population will then become stable and so will resist any further increase in the allele's prevalence.
Figure 11: allele frequencies vs. genotype frequencies, credit to Nature (Andrews, 2010)

This conveys both the benefits and drawbacks associated with heterozygote advantage: though it is true that to be a heterozygote in Nigeria will likely result in the greatest evolutionary payoff, to maximise one's chances of being a heterozygote, as shown in the graph above, the *S* allele must have a frequency of 0.5. However, at such a high frequency, the prevalence of sickle cell anaemia, shown to deliver the lowest payoff even in malarial regions despite the resistance it affords to the infectious disease, will inevitably increase to affect 25%[3] of the population, a simply unfeasible statistic, seeing as the condition is associated with the greatest mortality rates of all three genotypes. Therefore, even though the *f(S)* value of 0.286 may appear relatively low, given that heterozygosity will be greater at any values $0.286 < f(S) \leq 0.5$, values which would all, as has been earlier demonstrated, result in evolutionarily stable population able to resist any further invasion by alleles, the average payoff or mean fitness of the population would decrease, thus necessitating a compromise of 0.286 to be used for the equilibrium value.

It must be noted that the majority of the mathematical modelling stems from the evolutionary fitness values appointed by this author. This is a major cause of inaccuracy as despite the subsequent complexity of the model, the model was still based on great assumptions regarding the relative fitness values and reproductive successes of the three genotypes. Ultimately, despite the many different, slightly nuanced models used to find the equilibrium point of 0.286, they all rest on the aforementioned fitness values. That being said, much thought went into the assigning of these values, with recent mortality rates being analysed and offset by historical findings (Rucknagel and Neel, 1961). Whilst to a certain extent, it is impossible to identify the payoff values with 100% certainty due to the fact that fitness values are largely subjective and influenced by a myriad of other uncontrollable variables – (indeed this is a problem which has plagued the study of evolutionary game theory for decades) - the relative accuracy of the assigned values still result in a valid answer, and at the very least demonstrate the idea that an optimum frequency of the *S* allele in a malarial region lies at an intermediate value above 0 and below 0.5.

Conclusion

By first introducing the two diseases, and then delving in to the relative advantages the sickle cell anaemia and sickle cell trait genotypes confer upon an individual, we have been able to model the optimum prevalence of the *S* allele in malarial regions, and found it, using evolutionary game theory, to be 0.286 (or, more specifically,

[3] This figure is obtained through the use of the Hardy-Weinberg equilibrium, with q^2 being equal to the frequency of the SS genotype (sickle cell anaemia phenotype) in a population, with 0.5^2 resulting in a frequency of 0.25 of the genetic condition.

2/7). This value was then confirmed through analysis of the sub-populations of Nigeria and USA, which also provided insight into the effect of a selective pressure, namely the presence of malaria, into the relative frequencies of the *S* allele. Using the data on Nigeria, and the model created by this author to estimate the allele equilibrium value, it was estimated that it will take 24 generations to reach a frequency within 2% of the optimum frequency, with the equilibrium point being shown to be an unattainable ideal[4].

By applying the idea of an equilibrium point, discussed in its general form by the theoretical population geneticist John H. Gillespie (Gillespie, 2004), to the data collected by this author from various sources and assembled on the sub-population of Nigeria, and in turn applying myriad statistical tests and mathematical models, I have been able to forecast, admittedly subject to somewhat significant uncertainty, the changing allele frequencies of a country.

This data is naturally invaluable to population geneticists, but also has wider implications to pharmaceutical companies, who must determine long term strategies for drug production and indeed for public health administrators. They will be able to understand that as the *S* allele frequency tends to 0.286, though heterozygote numbers will increase, cases of sickle cell anaemia will increase at a greater rate, as is known through use of the Hardy-Weinberg model, and so they must prepare to invest resources and a greater share of the healthcare budget into improving treatments for the genetic condition.

So can sickle cell anaemia really save a life? Simply put, no - even in malarial regions, as the relative fitness of the disease is lower than either of the other two genotypes, despite the resistance to malaria which the normal phenotype would not have, given the oft ignored fact that sickle cell anaemia is life-threatening in its own right. However, in order to maximise the reproductive success of an entire population, we have shown through analysis of the relative fitness of various genotypes that the presence of sickle cell anaemia in a population is of vital importance. Indeed at the optimum *S* allele frequency, 8.2%[5] of the population will have the disease, a figure 35.5 times greater than its current global prevalence (WHO, 2016b). Therefore, whilst the disease should not be counted upon to save the life of any one particular individual, its increasing prevalence will be a necessary, if somewhat unpleasant, byproduct of the increase in frequency of the *S* allele, which will save many lives, in that its presence will increase not only the prevalence of sickle cell anaemia, but also the prevalence of heterozygosity, a genotype which does have a very high relative fitness and reproductive success associated with it.

[4] Please note that all payoff values and consequent deductions as well as values used for mathematical modelling and data manipulation thereof are this author's own work unless otherwise specified, and to his knowledge no such estimation of optimum frequency exists in the wide body of research perused.
[5] Estimate derived through use of the Hardy-Weinberg Equilibrium, with q^2 representing the frequency of the disease.

Bibliography

Adewowoyin, N.B., Iheanacho, O.E. and Bazuaye, G.N. (2012*), Prevalence of haemoglobin variants in Benin City, Nigeria*, Annals of Biomedical Sciences, 11(2), pp. 60–64.

Aidoo, M., Terlouw, D.J., Kolczak, M.S., McElroy, P.D., ter Kuile, F.O., Kariuki, S., Nahlen, B.L., Lal, A.A. and Udhayakumar, V. (2002), *Protective effects of the sickle cell gene against malaria morbidity and mortality*, The Lancet, 359(9314), pp. 1311–1312. doi: 10.1016/s0140-6736(02)08273-9.

Akinyanju, O. (1989), *A profile of sickle cell disease in Nigeria*, Annals of the New York Academy of Sciences., 565, pp. 126–36.

Ancestry (2012) Available at: http://www.ancestry.co.uk/cs/learning/HowLongIsGeneration (Accessed 26 June 2016)

Andrews, C.A. (2010), *The Hardy-Weinberg Principle*, Nature Education Knowledge, 3(10):65.

Ayi K, Turrini F, Piga A, Arese P (2004), *Enhanced phagocytosis of ring-parasitized mutant erythrocytes: a common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait*, Blood. 104: 3364-3371.

Bengtsson, B. and Tunlid, A. (2010), *The 1948 international congress of genetics in Sweden: People and politics*, Genetics., 185(3), pp. 709–15.

Brain, P. (1952), *Sickle-cell Anaemia in Africa*, Br Med J., 2(4789).

Cabrera G, Cot M, Migot-Nabias F, Kremsner PG, Deloron P, Luty AJF (2005), *The sickle cell trait is associated with enhanced immunoglobulin G antibody responses to Plasmodium falciparum variant surface antigens.*, J Infect Dis. 191: 1631-1638.

Carr, S. (2015) Chi Squared Critical Values. Available at: https://www.mun.ca/biology/scarr/MGA2-06-Table1.jpg (Accessed: 23 June 2016).

CDC (2014) Incidence of sickle cell trait — United States, 2010. Available at: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6349a3.htm. (Accessed: 20 June 2016)

CDC (2016a) Frequently asked questions (FAQs). Available at: https://www.cdc.gov/malaria/about/faqs.html (Accessed: 17 June 2016).

CDC (2016b) Malaria facts. Available at: https://www.cdc.gov/malaria/about/facts.html . (Accessed: 19 June 2016)

Charlesworth, D. and Willis, J.H. (2009), *The genetics of inbreeding depression*, Nature Reviews Genetics, 10(11), pp. 783–796. doi: 10.1038/nrg2664.

Crognier, E. (2003), *Reproductive success: Which meaning?*, American journal of human biology: the official journal of the Human Biology Council., 15(3), pp. 352–60.

Dupré, J. (1987), *Latest on the best: Essays on evolution and optimality*, Edited by John Dupré, Cambridge, MA: MIT Press.

Easley, D. and Kleinberg, J. (2009), *Networks, crowds, and markets*., Available at: https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book.pdf (Accessed: 20 June 2016).

Evolution of Genetic Diversity (no date) Available at: http://www.ucl.ac.uk/~ucbhdjm/courses/b242/MaintVar/MaintVar.html (Accessed: 23 June 2016).

Farlex (2003), *Reproductive age*, in Available at: http://medical-dictionary.thefreedictionary.com/reproductive+age (Accessed: 23 June 2016).

Ferreira, A., Marguti, I., Bechmann, I., Anatomy, I. of, Liebigstr, Leipzig, 04103, Anatomie, S., Neuroanatomy, C., Goethe-University, J.W., Frankfurt, 60 596 -, Main, G., Jeney, V., Chora, Â., Palha, N.R., Rebelo, S., Henri, A., d'Hematologie, I.U., 7, U.P., Diseases, I. of E., Therapies, I., Beuzard, Y. and Soares, M.P. (2011), *Sickle hemoglobin confers tolerance to plasmodium infection*, Cell, 145(3), pp. 398–409. doi: 10.1016/j.cell.2011.03.049.

Frenette, P.S. and Atweh, G.F. (2007), *Sickle cell disease: Old discoveries, new concepts, and future promise*, The Journal of Clinical Investigation, 117(4), pp. 850–858. doi: 10.1172/JCI30920.

Gates, B. (2016), *The deadliest animal in the world*, Available at: https://www.gatesnotes.com/Health/Most-Lethal-Animal-Mosquito-Week (Accessed: 29 May 2016).

Gillespie, J.H. (2004) Population genetics. The Johns Hopkins University Press.

Haldane, J. (1949), *Disease and Evolution*, Available at: http://nfs.unipv.it/NFS/MINF/DISPENSE/PATGEN/lectures/files/readings/haldane_1949.pdf (Accessed: 23 June 2016).

Inbreeding depression (2014) Available at: http://evolution.berkeley.edu/evolibrary/article/conservation_03 . (Accessed: 27 June 2016)

IVCC (2016) *WHO malaria fact sheet*. Available at: http://www.ivcc.com/who-malaria-fact-sheet (Accessed: 13 June 2016).

Kikuchi, G., Yoshida, T. and Noguchi, M. (2005), *Heme oxygenase and heme degradation*, Biochemical and biophysical research communications., 338(1), pp. 558–67.

Lang PA, Kasinathan RS, Brand VB, Duranton C, Lang C, Koka S, Shumilina E, Kempe DS, Tanneur V, Akel A, Lang KS, Foller M, Kun JFJ, Kremsner PG, Wesselborg S, Laufer S, Clemen CS, Herr C, Noegel AA, Wieder T, Gulbins E, Lang F, Huber SM (2009): *Accelerated clearance of Plasmodium-infected erythrocytes in sickle cell trait and annexin-A7 deficiency*. Cell Physiol Biochem. 24: 415-428.

Lodish, H., Berk, A., Zipursky, L.S., Matsudaira, P., Baltimore, D. and Darnell, J. (2000) *Mutations: Types and causes*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21578/ (Accessed: 23 June 2016).

Luzzatto, L. (2012), *Sickle cell Anaemia and malaria*, Mediterr J Hematol Infect Dis, 4(1).

Luzzatto L, Nwachuku-Jarrett ES, Reddy S (1970): *Increased sickling of parasitised erythrocytes as mechanism of resistance against malaria in the sickle-cell trait*. Lancet. 1: 319-321.

Mallet, J. (2016) *Inbreeding and Genetic Drift*. Available at: http://www.ucl.ac.uk/~ucbhdjm/courses/b242/InbrDrift/InbrDrift.html (Accessed: 24 June 2016)

McAuley, C.F., Webb, C., Makani, J., Macharia, A., Uyoga, S., Opi, D.H., Ndila, C., Ngatia, A., Scott, J.A.G., Marsh, K. and Williams, T.N. (2010), *High mortality from plasmodium falciparum malaria in children living with sickle cell anemia on the coast of Kenya*, Blood, 116(10), pp. 1663–1668. doi: 10.1182/blood-2010-01-265249.

Nelson, N. (2015), *Sickle cell anaemia and malarial resistance* Available at: https://www.physics.ohio-state.edu/~wilkins/writing/Samples/shortmed/nelson/sickle.html (Accessed: 25 June 2016).

Nigeria malaria fact sheet (2011) Available at: http://photos.state.gov/libraries/nigeria/231771/Public/December-MalariaFactSheet2.pdf (Accessed: 27 June 2016).

OCR (2016), *AS level and A level biology A and biology B (advancing biology) mathematical skills handbook*. Available at: http://www.ocr.org.uk/Images/294471-mathematical-skills-handbook-.pdf (Accessed: 22 June 2016).

Okasha, S. (2012), *Population genetics*. Available at: http://plato.stanford.edu/entries/population-genetics/ (Accessed: 24 June 2016)

Oxford (2006), *Human population genetics lecture 2 the Wright-Fisher model of random genetic drift concepts explained in this lecture identity by descent neutral Polymorphism A Gene Pool model for genetic drift why there is more turnover of genetic polymorphism in small*., Available at: http://www.stats.ox.ac.uk/~harding/files/MT_module/MTLect2_Thus2006.pdf (Accessed: 13 June 2016).

Pasternak, N.D. and Dzikowski, R. (2009), *PfEMP1: An antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite*, The International Journal of Biochemistry & Cell Biology, 41(7), pp. 1463–1466. doi: 10.1016/j.biocel.2008.12.012.

Poinar, G. (2005), *Plasmodium dominicana n. Sp. (Plasmodiidae: Haemospororida) from tertiary Dominican amber*, Systematic Parasitology, 61(1), pp. 47–52. doi: 10.1007/s11230-004-6354-6.

Poolman, E.M. and Galvani, A.P. (2007), *Evaluating candidate agents of selective pressure for cystic fibrosis*, Journal of The Royal Society Interface, 4(12), pp. 91–98. doi: 10.1098/rsif.2006.0154.

Pritchard, J. (2015), *Medical applications of population genetics*. Available at: http://www.stats.ox.ac.uk/~mcvean/slides8.pdf (Accessed: 18 June 2016).

Malaria.com. (2011) Videos. Available at: http://www.malaria.com/questions/how-long-has-malaria-existed (Accessed: 28 June 2016).

Marsh K, Otoo L, Hayes RJ, Carson DC, Greenwood BM (1989), *Antibodies to blood stage antigens of Plasmodium falciparum in rural Gambians and their relation to protection against infection*. Trans R Soc Trop Med Hyg. 83: 293-303.

Shane (2016) *Guide to f-statistics*. Available at: http://www.library.auckland.ac.nz/subject-guides/bio/pdfs/733Pop-g-stats2.pdf (Accessed: 21 June 2016).

Sickle Cell DNA (2016) Available at: http://www.passmyexams.co.uk/GCSE/biology/images/dna-sickle-cell.png (Accessed: 27 June 2016).

Smith, J.M. and Price, G.R. (1973), *The logic of animal conflict*, Nature, 246(5427), pp. 15–18. doi: 10.1038/246015a0.

Sober, E. and Lewontint, R.C. (1982), *Philosophy of science association*, Philosophy of Science Lewontin Source: Philosophy of Science Philosophy of Science Association, 49(2), pp. 157–180.

Thomas, B. (1985), *On evolutionarily stable sets*, Journal of Mathematical Biology, 22(1). doi: 10.1007/bf00276549.

University of São Paulo (2013), *The Wahlund effect and F statistics --the interaction of drift and Gene Flow*. Available at: http://www.ib.usp.br/~otto/downloads/pdf_templeton/wahlund.pdf (Accessed: 23 June 2016).

University of Wyoming (no date), *F-statistics worked example*. Available at: http://www.uwyo.edu/dbmcd/molmark/practica/fst.html . (Accessed: 24 June 2016)

Uzoegwu, P.N. and Onwurah, A.E. (2003*), Prevalence of haemoglobinopathy and malaria diseases in the population of old Aguata Division, Anambra State, Nigeria*, Biokemistri, 15(2), pp. 57–66.

Verra, F., Simpore, J., Warimwe, G.M., Tetteh, K.K., Howard, T., Osier, F.H.A., Bancone, G., Avellino, P., Blot, I., Fegan, G., Bull, P.C., Williams, T.N., Conway, D.J., Marsh, K. and Modiano, D. (2007), *Haemoglobin C and S role in acquired immunity against plasmodium falciparum malaria*, PLoS ONE, 2(10), p. e978. doi: 10.1371/journal.pone.0000978.

WHO (2016a) *Malaria in children under five*. Available at: http://www.who.int/malaria/areas/high_risk_groups/children/en/ (Accessed: 22 June 2016).

WHO (2016b) *Sickle-cell disease and other haemoglobin disorders*. Available at: http://www.who.int/mediacentre/factsheets/fs308/en/ (Accessed: 27 June 2016).