EVOLUTIONARY GENETICS OF MALARIA: GENETIC SUSCEPTIBILITY AND NATURAL SELECTION

Martin Sikora

DOCTORAL THESIS UPF / 2010

THESIS DIRECTORS

Dr. Jaume Bertranpetit
Department of Experimental and Health Sciences

Dr. Ferran Casals
Department of Experimental and Health Sciences

UNIVERSITAT POMPEU FABRA

To my family

”All you really need to know for the moment is that the universe is a lot more complicated than you might think, even if you start from a position of thinking it’s pretty damn complicated in the first place.”

Douglas Adams, *Mostly Harmless*

Acknowledgements

It is the evening of the 28th of March 2010, and here I am sitting in front of the screen finishing the last, and arguably most important part of this thesis: expressing my gratitude to all the people that have helped make these last four and a half years the fantastic and rewarding experience that it was. When we arrived in October 2005 for my interview, I had never been to Barcelona before, and knew almost no Spanish and certainly no Catalan at all. Today as I am writing this, the derbi of Real Madrid against Atlético is running on the livestream on the second screen, and like thousands of culés around the city I am hoping that Real will leave some points tonight (Atlético leads 1-0!). I guess it just goes to show how much I feel at home here now.

My first thank you then is to all the old and new friends that helped us make this city our new home: Kristin and Alfonso, from day one, when we stayed at your tiny flat in the Raval and you helped us calling for flats with our non-existing Spanish, to all the fiestas and trips together in the following years. The group of people that became our first circle of friends here: Luisa & Hagen, Jessica & Henrik, Kriszti & Pere, Tamara & Rasa, Helena, the early nacho evenings at Fianna's, Calçotadas, Football ... All the other Austrian expats, Klaus (blancas at Bar Ramon!), Nadine, Günther ... Being abroad for a long time made me realise how important it is not to forget where you come from. All the other friends who stayed in touch despite the distance: Spezi, Alex & their always growing family; summer festivals with Hübsi, Bianca, Muck; the BOKU people, Tapio, Letti, Daniel; Also many friends visiting over the years: Romana & Thomas, Susi & Flo, Henrikka, Veronica coming all the way from Australia.

The next big *gràcies!* goes to the people at Uni, the constantly expanding BioEvo family, too many to thank each of you here. Most important of all, members of the *despacho perdido*, for both a great time and collaborations during my time here. The original members: Andrés brotha, it's great that I will again share an office with you on the other side of the world soon;

Anna, without your help much of this thesis would not have been possible; and Johannes, always with a good spirit and enthusiastic, and a true master of cubicle basketball. The new generation of BioEvos taking over the despacho: Giovanni & Ludovica, soon to be followed by Brandon and Pierre. Not to forget the remaining members of Jaume's group: Marta, and especially Hafid, for the countless times giving me valuable advice. To all the other BioEvos past & present that made it such a joy to work in our group, from Tuesday morning football games to 'Grazy Fridays' occupying half of the Bitacora and allowing me to refine my knowledge of Spanish pop history in some shady bars in the Raval at 3 am: Urko, Ixa, Karla, Mònica, Rui, Olga, Belén, Michelle, Chiara, Ángel, Txema, Elena, Laura, Oscar, Judit ... In particular, special thanks to Bruno, David & Ville, lots of great memories to be taken with me. Me gusta!

This list would of course not be complete without thanking the two most important people at the lab for me, my thesis directors. Ferran, moltísimes gràcies per a tot! Without your constant help and encouragement I would have never been able to learn so much during this time in the lab, and I am deeply grateful for that. This thesis is as much yours as it is mine. Outside of work, I am looking forward to meeting each other again at the football field for some more legendary running duels. And of course Jaume, for being a thesis director like every director should be: always enthusiastic, encouraging, and supportive. Thank you for always finding time to help even with your sometimes crazy schedules, and for showing me what it will take to become a great scientist. And above all, thank you for giving me the opportunity to do my PhD in your group when I came for that interview with you four and a half years ago. It has turned out to be the best possible PhD I could have ever imagined.

Finally, I wish to thank the people closest to me, my family. Without your love and support it would have never been possible to accomplish this work. Even if it was sometimes hard being away from you for so long, I feel like the distance has only further strengthened our relationship.

Mama, Papa, danke für alles, ohne Eure Unterstützung wäre das alles nie möglich gewesen. Ich hoffe wir werden für unsere Kinder einmal solche Eltern werden wie Ihr es für uns seid.

Above all, I wish to thank Renia, the most important person in my life. I know how difficult it was for you in the beginning here, and there are no words to thank you enough for being there with me throughout these years. I know that for all the uncertainties and changes that this new adventure of ours will bring, I will never have to worry about anything, as long as I will be with you.

Martin Sikora, March 2010

During this work I was supported by a PhD fellowship from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982).

Abstract

One of the strongest selective forces affecting human populations in recent history is the malaria parasite *Plasmodium falciparum*, which is the cause of a variety of well-established examples of pathogen-induced adaptation in humans. A special form of malaria is pregnancy-associated malaria, which is characterised by the accumulation of infected erythrocytes in the placenta, and causes up to 200,000 maternal and infant deaths every year. The aim of this work is to characterise how this particular form of malaria has shaped human genetic variation. To that end we use methods of both evolutionary genetics and molecular epidemiology, reporting the first large-scale investigation of the genetic basis of placental infection. Our results provide new insights into genes modulating the risk of infection, as well as natural selection acting on cellular pathways involved in the pathogenesis of the disease. Finally, we also provide new data on the genetic structure of affected populations in Sub-Saharan Africa.

Resum

Una de les forces selectives més fortes que han afectat a les poblacions humanes en la història més recent és el paràsit de la malària, *Plasmodium falciparum*, que és la causa de varis exemples de adaptació induïda per patògens en els éssers humans. Una forma especial de malària és l'associada a l'embaràs, que es caracteritza per l'acumulació d'eritròcits infectats en la placenta, i que pot arribar a causar fins a 200.000 morts maternoinfantils cada any. L'objectiu d'aquest treball és descriure com aquesta forma peculiar de malària ha afectat la variació genètica humana. Amb aquesta finalitat, hem utilitzat mètodes tant de la genètica evolutiva com de l'epidemiologia molecular, resultant en de la primera investigació a gran escala de la base genètica de la malària placentària. Els resultats ofereixen una nova visió sobre els gens que modulen el risc d'infecció, així com de la selecció natural actuant sobre les vies cellulars implicades en la patogènesi de la malaltia. Finalment, també aportem noves dades sobre l'estructura genètica de les poblacions sub-saharianes analitzades.

Preface

It is surely not an exaggeration to say that we are right now in the midst of a new revolution in the life sciences. Ever more sophisticated high throughput technologies are producing data at an astounding rate, so much so that not the production, but the analysis of the data has become the bottleneck in research. This past four and a half years of working on my PhD project I have witnessed this revolution first hand, from starting off with analysing a handful of SNPs to routinely handling millions of markers in multiple populations in the end. At the end of 2007, almost halfway through the PhD, Science Magazine named human genetic variation the breakthrough of the year, and it certainly felt exhilarating to be working in one of the fields at the cutting edge of science.

This thesis is about human genetic variation, but it is also about one of the worst diseases humankind has experienced, malaria. While almost every other week sees the publication of a study on *diseases of civilisation* such as diabetes, studies on the genetics of malaria are even today, after completion of this four years of research into its genetic basis, surprisingly scarce. The aim of this thesis was therefore to address this lack, and to provide a contribution to our understanding of the genetic basis of malaria.

In the cause of this work we have both produced our own data, as well as increasingly taken advantage of the vast amounts of public datasets available, in order to provide a picture that is as complete as possible. This document is a compilation of the results we have obtained in this endeavour, together with their discussion in light of some recent results, available thanks to the aforementioned revolution in the study of genetic variation.

Barcelona, March 2010

Index

1

Introduction

1.1 Human genetic diversity

In the year 2001, more than ten years after the human genome project was officially launched, two landmark papers published the first draft sequence of the human genome (Lander et al., 2001; Venter et al., 2001). Although far from being complete at the time of publication, the sequence of the roughly three billion basepairs that make up the human genome gave the first blueprint of our hereditary material, the DNA. Importantly, this blueprint is represented as a haploid consensus sequence, derived from the assembly of the sequences of a number of different individuals. One can therefore say that it represents to some extent the genetic information that is shared among all humans. However, if we want to understand what makes one human individual differ from another one, the consensus sequence alone is of little help. What is needed is a description of the variation in the sequence among individuals. It is this variation, defined as genetic diversity, that, in interaction with other, non-genetic factors, determines the phenotypic differences among humans, from morphological characteristics such as height to differences in susceptibility to disease. Linking phenotypic differences to genotype differences has therefore become a central question for biology in the genomic era.

1.1.1 Classes of genetic variation

Genetic variation in humans can be broadly classified in two categories: Variations in single base positions, also called single nucleotide polymorphisms (SNPs); and variation in larger portions of the sequence, generally referred to as structural variation (Frazer et al., 2009). An overview of the different classes and its members is shown in Figure 1.1.

Single nucleotide polymorphisms

Single nucleotide polymorphisms are both the smallest and most prevalent types of genetic variation. As noted above and seen in Figure 1.1, a SNP corresponds to individuals of a population showing different alleles at a single nucleotide position. At the time of writing, there were roughly

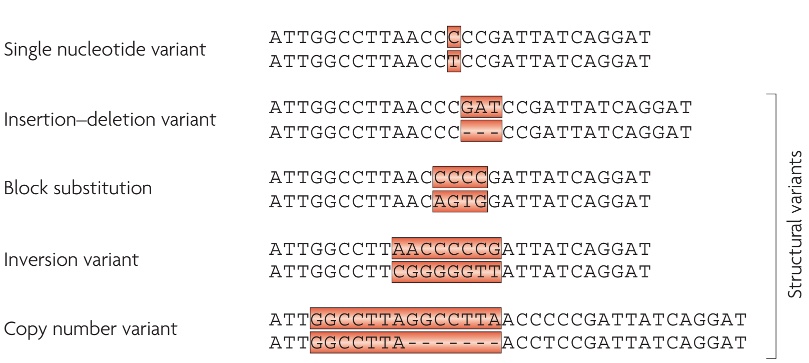

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Figure 1.1: Classes of genetic variation.The red boxes indicate parts of the sequence that differ between the chromosomes. (Frazer et al., 2009)

14 million validated human SNPs described in dbSNP (build 131, March 2010), and this number is set to increase further as continued sequencing of human populations will discover ever rarer and population specific variants. Most of those SNPs are biallelic markers, and the prevalence of the alleles within a population can be measured both as the frequency of the more frequent (major) or the less frequent (minor) allele. The latter one is more commonly used, and is referred to as *minor allele frequency* (MAF). If the ancestral state of a SNP is known, meaning that one can determine which of the two alleles arose through a new mutation, a distinction can be made between ancestral and derived alleles. In humans, it is common practice to assign the allele found in the chimpanzee as the derived allele, and its frequency is referred to as *derived allele frequency* (DAF). The fact that SNPs are both biallelic and very abundant in the genome has allowed to develop technologies to assess them with very high throughput, something not possible for other types of markers such as indels. They are therefore still the markers of choice in the majority of studies of human genetic diversity.

Depending on its position within the genome, a specific SNP can occur in different classes of sequence. Most of them fall into intergenic regions between known genes, and they are generally assumed to be non-functional.

However, some of those intergenic regions may harbour regulatory elements for nearby genes, and SNPs within those elements are usually referred to as regulatory SNPs. Among SNPs that fall within gene regions, the distinction is between SNPs within the non-coding regions of the gene and those in the actual coding regions (coding SNPs). Coding SNPs can be further distinguished based on their effect on the amino acid sequence of the protein product, where synonymous SNPs do not change the amino acid (due to the degeneracy of the genetic code), whereas non-synonymous SNPs lead to a change in the amino acid (missense change) or a premature stop codon and truncated protein (nonsense change).

In a diploid genome like the human one, the combination of the two allelic states of the two chromosomes is called a genotype. If we look at a particular stretch of a chromosome that harbours N SNPs, and each of those has two possible allelic states, we can in theory observe 2^N different allelic combinations for that stretch. This combination of allelic states among multiple SNPs is generally referred to as a haplotype.

Structural variation

In its broadest interpretation, the term structural variation encompasses a vast range of types of variations, ranging in size from small insertions/deletions (indels) of a few basepairs all the way to cytogenetically detectable large scale rearrangements that affect parts of or even entire chromosomes. Given the heterogeneity of this class of variation, a standardised way of defining some of the subclasses is still lacking, and their distinctions may be somewhat arbitrary (Scherer et al., 2007). Much recent effort has been aimed at the detection and characterisation of newly discovered types of structural variation such as copy number variations or segmental duplications (Conrad et al., 2009; Feuk et al., 2006). It has become evident that these variants play important roles in organismal function at all levels (Hurles et al., 2008), as well as in the evolutionary history of our species (Marques-Bonet et al., 2009). However, both the technology for assaying these variants, as well as the statistical methodology for analysing them is still lacking the maturity of those for assaying and analysing SNP genotypes. The advent of cheaply available whole-genome sequencing of individuals will certainly solve some of those problems, and they will without doubt continue to play an increasingly important part in the study of human genetic diversity.

1.1.2 Evolutionary processes shaping genetic diversity

Evolution can only occur if there is heritable genetic variation among individuals. On a microevolutionary scale, i.e. over generational time among individuals of a single species, the main processes that affect genetic diversity are mutation and recombination, genetic drift, natural selection, and gene flow.

Mutation and recombination

Mutation and recombination are the two process that are responsible for generating genetic diversity. Mutation in particular is the only process that can produce new alleles at a genetic locus. As briefly mentioned in the introduction to this section, evolution can only occur if genetic variation is heritable, from which follows that only mutations in the germ-line of an organism can contribute to evolutionary change. A variety of types of mutational changes exist, each of them exhibiting differing rates of mutation. Of particular interest for this work are single base substitutions, where one base of the DNA molecule is replaced by another one. If a particular substitution increases in frequency and segregates in the general population, it is called a SNP, as described earlier in this section.

In the germ-line of sexually reproducing organisms, pairs of parental chromosomes undergo meiotic recombination, a process whereby they reciprocally exchange parts among them. Recombination therefore also generates genetic diversity, albeit not by producing new alleles but by reshuffling the combinations of existing alleles that constitute a haplotype. At the level of human populations, it has been observed that the rate of

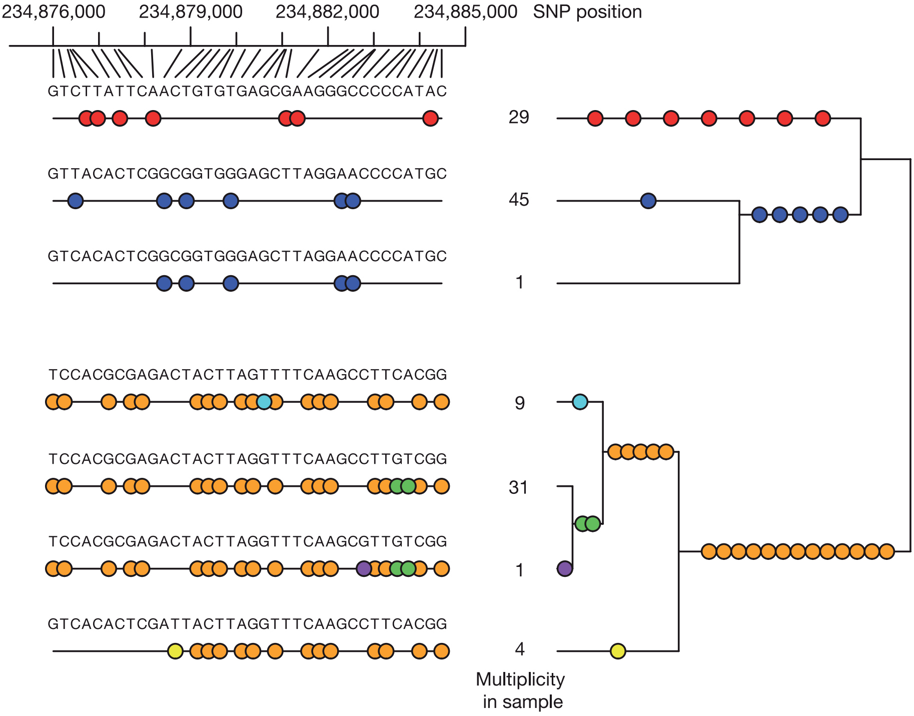

234,876,000
234,879,000
234,882,000
234,885,000
SNP position
GTCTTATTCAACTGTGTGAGCGAAGGGCCCCATAC
29
GTTACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
45
GTCACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
1
TCCACGCGAGACTACTTAGTTTTCAAGCCTTCACGG
9
TCCACGCGAGACTACTTAGGTTTCAAGCCTTGTCGG
31
TCCACGCGAGACTACTTAGGTTTCAAGCGTTGTCGG
1
GTCACACTCGATTACTTAGGTTTCAAGCCTTCACGG
4
Multiplicity
in sample

Figure 1.2: Relationship among haplotypes in a region without observed recombination events. On the left are the observed haplotypes in the region, whereas the right shows the genealogical relationship among them. SNPs with the same colour are in high LD with each other (The International HapMap Consortium, 2005).

recombination is not uniformly distributed among the genome, but rather shows a pattern of stretches of low rates separated by recombination hotspots exhibiting very high rates (Myers et al., 2005, 2008). In regions of low recombination, neighbouring alleles tend to be inherited as linked units, and the resulting non-random association of those alleles is referred to as linkage disequilibrium (LD) (see Figure 1.2). As a consequence of this, the human genome exhibits a mosaic like structure of haplotypic variation, with blocks of high LD, so-called haplotype blocks, separated by hotspots of recombination (Pääbo, 2003). The characterisation of this haplotype map of the human genome was the main objective of the still ongoing

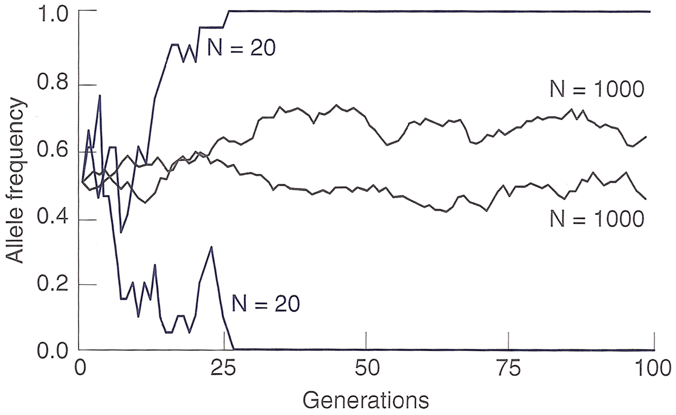


Figure 1.3: Genetic drift for different population sizes. Alleles starting off with an allele frequency of 0.5 become either lost or fixed much earlier in the smaller population (Jobling et al., 2004).

International HapMap Project (The International HapMap Consortium, 2003), which will be described in more detail later on.

Genetic drift

In the previous section we have seen how mutation gives rise to new alleles that can be passed on to the offspring. However, not every individual in a population will produce offspring, and each generation will constitute a random sample of the previous generation. As a consequence, the fate of a newly arisen allele will be partly governed by this stochastic process of random sampling, generally referred to as genetic drift. The action of genetic drift over time causes the slow eradication of genetic diversity, due to segregating alleles in a population being either lost or fixed. The magnitude of this effect will depend on the size of the population, with smaller populations being much more susceptible to this random fluctuations in allele frequencies (Figure 1.3). Demographic events such as founder effects or population bottlenecks can drastically reduce the size of a population, which in turn leads to increased genetic drift and a reduction in genetic diversity.

Natural selection

Natural selection, famously coined by Charles Darwin in *The Origin of the Species* (Darwin, 1859), is defined as the differential survival of individuals in a population with heritable trait variation in successive generations. Genotypic variation in a population results in individuals that vary in their ability to survive and reproduce (referred to as fitness) in a particular environment. Any genotype that increases the chances of survival of an individual, i.e increases its fitness, will increase in frequency over successive generations due to producing on average more offspring than individuals of a different genotype. Natural selection can therefore be said to manipulate genetic diversity, as it is the evolutionary process that causes the adaptation of organisms to a particular environment. Depending on how variants affect the fitness of the organism, we can distinguish different models of selection.

Negative selection If a variant reduces the fitness of an individual, it is said to be under negative selection. Sometimes a distinction is made between negative selection, keeping weakly deleterious variants at low frequency, and purifying selection, resulting in the complete elimination of the variant from the population.

Positive selection Variants that increase the fitness of an individual undergo positive selection, resulting in an increase in frequency in the population. Positive and negative selection are instances of *directional selection*, as they tend to remove variation in the population due to either fixation or elimination of the affected variant.

Balancing selection Balancing selection is the collective term for models of selection that results in balanced polymorphism, i.e. where it is advantageous to have multiple alleles segregating in a population. If we consider the genotypes of a biallelic SNP, *overdominant selection* occurs if both homozygote genotypes have reduced fitness compared to the heterozygote. Alternatively, a genotype whose fitness depends on its frequency in the population is said to be under *frequency-dependent selection*. If a genotype has higher fitness at low frequency than at high frequency, it will reach an equilibrium frequency and produce a balanced polymorphism. Another way for this to happen is when different selective forces favour different alleles at the same genetic locus.

Gene flow

Gene flow refers to the process of a transfer of alleles from one subpopulation to another, as a consequence of migration of individuals between them. If two populations that were previously isolated from each other start exchanging migrants, any existing genetic differentiation between them will tend to be erased due to the homogenisation of the collective gene pool. In practice, gene flow always acts in interplay with the other forces mentioned above, which may both increase or counteract its effect on the distribution of genetic variation.

1.1.3 Patterns of human genetic diversity

The patterning of genetic diversity among present day human populations is the result of the past action and complex interplay of the evolutionary forces described above. Studying present-day variation therefore enables us to search for the traces left by these forces in the human genome, from historical demographic events such as migrations to the effects of natural selection. The advent of cheap genotyping over the last decade has finally enabled the research community to characterise this patterning on an ever increasing scale, both in terms of number of individuals studied as well as the coverage of the genome.

Public repositories of genetic variation data

The first effort to obtain a catalog of human genetic diversity was launched in 1991, not long after the start of the Human Genome Project, under the name Human Genome Diversity Project (HGDP) (Cavalli-Sforza et al., 1991). Although initial plans of a more ambitious sampling scheme had to be abandoned, the project succeeded in establishing a repository of cell lines from 1,064 individuals of 52 worldwide populations (Cann et al., 2002). These samples have been used in a variety of studies of human genetic diversity since then, culminating in 2008 with the first analysis of genome-wide SNP genotype data in two separate publications (Jakobsson et al., 2008; Li et al., 2008).

The launch of the International HapMap Project in 2003 was the start of another, still ongoing effort to characterise human genetic variation (The International HapMap Consortium, 2003). Its focus was on characterising the haplotype structure of the human genome, also called haplotype map, and therefore differed in its approach to the HGDP. Initially, only four populations were chosen to be genotyped, albeit at a very high SNP density. The initial study populations were 30 parent-offspring trios from the Yoruba people of Ibadan, Nigeria; 30 trios of US citizens from Utah with northern and western European ancestry; 45 unrelated individuals from Tokyo, Japan; and 45 unrelated individuals from Beijing, China. During the initial two phases of the project, genotype data for more than three million SNPs was obtained and analysed (The International HapMap Consortium, 2005, 2007). For the still ongoing phase three of the project another seven populations were added, making it a total of 1,115 individuals from 11 populations.

The latest large-scale project, the so-called 1000 Genomes Project, was started in 2008. The ambitious goal of this project is to find most genetic variation with a frequency of at least 1% in the studied populations. At the time of writing the first data from three pilot projects have already been released. Once completed it will certainly be an extremely valuable resource for all research in human genetics for years to come.

Global patterns of genetic diversity

The availability of the above mentioned sample collections together with genotypic data covering multiple loci spread over the whole genome have resulted in a greater understanding of the global patterning of human genetic variation. The main result from numerous studies has been that African populations exhibit higher levels of both nucleotide and haplotype diversity compared to non-African populations, consistent with an African origin and subsequent out-of-Africa migration of modern humans. Non-African populations show a subset of the genetic variation present in Africa, which also show higher numbers of private alleles and haplotypes (Tishkoff and Verrelli, 2003; Hinds et al., 2005; The International HapMap Consortium, 2005; Conrad et al., 2006; Campbell and Tishkoff, 2008; Jakobsson et al., 2008; Li et al., 2008). Patterns of linkage disequilibrium show similar results, with levels of LD generally increasing with distance from East Africa (Jakobsson et al., 2008).

Many recent studies have taken advantage of high density genome-wide genotype data in order to get a clear and detailed picture of the amount of genetic structure in human populations, on both global and more regional scales. On a global scale, the results obtained show that individuals generally cluster together within populations according to geographical regions (Rosenberg et al., 2002; Jakobsson et al., 2008; Li et al., 2008). On a regional scale such as within Europe (Lao et al., 2008; Novembre et al., 2008), the Indian sub-continent (Reich et al., 2009) or Southeast and East Asia (The HUGO Pan-Asian SNP Consortium, 2009) results indicate a more gradual and clinal change in allele frequencies. Given these results, there has been much debate over whether the clusters observed on a global scale are real or just artefacts of uneven sampling. As often is the case, the answer seems to be a combination of the two possibilities, where patterns of human genetic variation could be mostly gradual, with a few sharp discontinuities due to geographical barriers to gene flow (Handley et al., 2007)

In the context of this work, the genetic structure present within Africa is of particular interest. Despite the fact that Africa harbours the majority of human genetic diversity, it was not until last year that an extensive survey of African diversity was published. Tishkoff et al. (2009) studied microsatellite variation in 121 African populations covering most of Sub-Saharan Africa, albeit lacking samples from the Southern part of the continent (with the exception of South Africa). Their study identified extensive genetic structure within African populations, which cluster into groups that correlate with ethnicity and cultural and linguistic properties. Another recent study analysed genome-wide SNP genotype data from 12 West African populations and found a similar result of genetic structure primarily reflecting linguistic groups (Bryc et al., 2010). Even though these studies have yielded a tremendous amount of new insights into the genetic diversity of Africa, more studies are clearly needed to adequately describe the large amount of genetic variation found within the continent.

1.2 Studying genetic diversity

An exhaustive description of the vast variety of studies of human genetic diversity and the methods used in them would be an impossible task. This section will therefore focus on three types of analyses that are relevant in the context of this work: The analysis of genetic structure of populations; linking genotype to phenotype in genetic association studies; and the detection of signatures of natural selection in the human genome.

1.2.1 Population structure

Two types of approaches are in widespread use for the inference of population structure from genetic data: Dimension reduction techniques such as principal component analysis (PCA) and multi-dimensional scaling (MDS), as well as model-based clustering algorithms using explicit population genetics models.

Dimension reduction techniques

Dimension reduction techniques like PCA have a long history in the analysis of genetic data, first introduced by Cavalli-Sforza and colleagues more than 30 years ago (Menozzi et al., 1978). The key feature of these methods is that they allow inference of any underlying structure in the data, without the need to assume a specific population genetic model. The raw genotype data is usually converted into a covariance matrix or a distance matrix among individuals, using for example the pairwise identity by state (IBS) of genotypes between individuals as a distance metric (Patterson et al., 2006). This matrix can then readily be subjected to the analysis method of choice like PCA or MDS. Recent years have seen a sort of a renaissance in the application of PCA to human genetic variation data, due to a variety of desirable features: The results are easily visualised by plotting the samples along the principal components, the data can be allowed to speak for itself without the need for any model assumptions, and it can be readily applied to large datasets of genome-wide genotype data (see Figure 1.4). Additionally, thanks to a number of recent studies it is

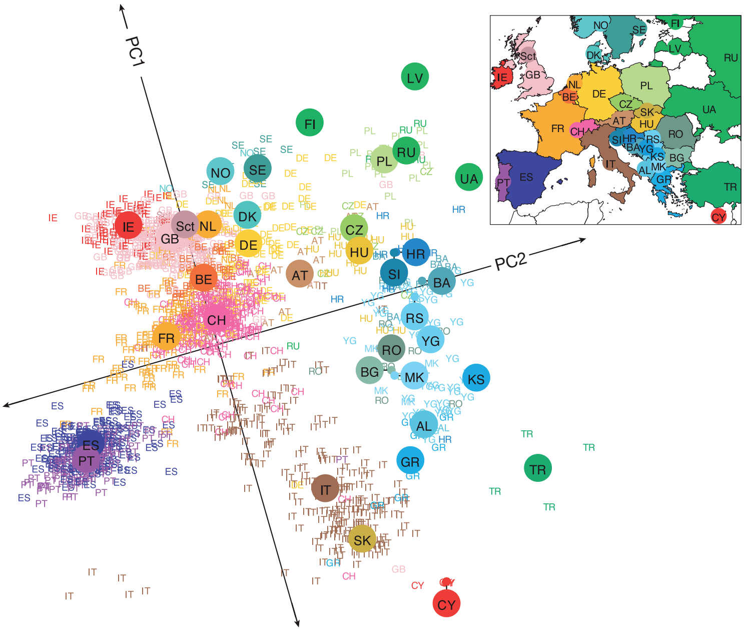


Figure 1.4: Results of a principal component analysis of European populations (Novembre et al., 2008).

becoming increasingly clear how the results obtained relate to the underlying evolutionary parameters of the populations in question (Novembre and Stephens, 2008; McVean, 2009).

Clustering algorithms

Clustering algorithms assume a model of K ancestral clusters or populations in the sample, each of them characterised by a set of allele frequencies at the analysed loci. Individuals are then probabilistically assigned to the K clusters in the way that best fits the observed data. As a result, proportions of cluster membership for each individual are obtained, which allow the identification of discrete subpopulations in the individuals analysed. A general issue with these methods is that the number of clusters is not known a priori. In a typical analysis, the value of K is therefore varied over a range of sensible values, and the appropriate value is chosen based on the improvement of the fit of the data. The most widely used cluster-

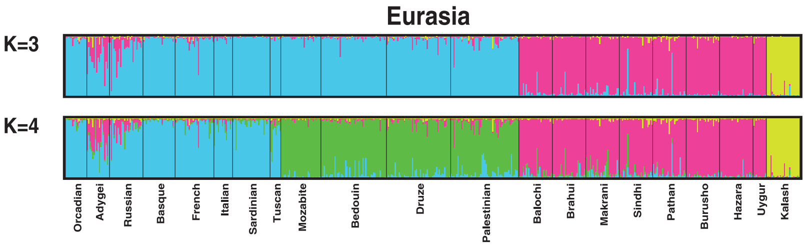


Figure 1.5: Example result from a *structure* analysis. Shown here are the cluster memberships for all Eurasian individuals from the HGDP, for the number of clusters K = 3 and K = 4, obtained from 377 autosomal microsatellite loci (Rosenberg et al., 2002).

ing software is *structure* (Pritchard et al., 2000; Falush et al., 2003), which uses a Bayesian Markov Chain Monte Carlo (MCMC) approach to infer the cluster memberships. A typical result from an analysis can be seen in Figure 1.5. Another, less computationally demanding method using maximum likelihood has been implemented in the package *frappe* (Tang et al., 2005).

1.2.2 Genetic association studies

The ultimate goal of genetic association studies is to link genotype to phenotype, i.e. to find the causal genetic variation responsible for a particular phenotypic outcome like disease status. In its simplest form, we can consider a situation as illustrated in Figure 1.6, where a single marker locus (e.g. a SNP) is causal for the phenotype under investigation. In the case

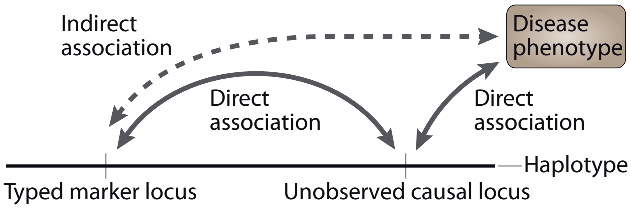


Figure 1.6: Basic principle of genetic association studies (Balding, 2006)

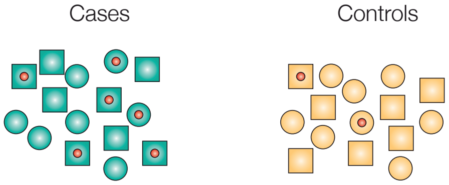


Figure 1.7: Association of a locus with a phenotype. Red circles indicate the causal allele of a locus (modified from Hirschhorn and Daly (2005))

of a phenotype with binary outcomes like disease, the allele frequency of the locus in question is then simply compared between cases (individuals with disease) and controls (disease-free individuals). If the locus has a sufficiently large effect on disease outcome (a high penetrance) there will be a significant difference in frequencies between the two groups (Figure 1.7). If the causal locus itself is not genotyped, associated regions can still be detected through indirect association of a locus that is in LD with the unobserved causal locus (Figure 1.6). Importantly, the power to detect an indirect association will, among other factors like the number of individuals analysed and the frequency of the causal allele, depend on the strength of LD between the two loci (Zondervan and Cardon, 2004). The concept of detection of disease causing alleles using indirect association through LD was the main motivation for the launch of the International HapMap project, described in Section 1.1.

Types of association studies

Different strategies exist for selection of the genetic markers to investigate. The simplest form is the case mentioned above, when a single candidate polymorphism is tested for association. If a particular gene has previous evidence of involvement, a *candidate gene* approach can be employed, where a number of SNPs distributed over the gene region are genotyped. This approach can be extended to collections of genes or markers that have related characteristics, for example to include all members of a particular pathway, or all SNPs that cause amino-acid changes. Finally, in a *genome-wide association study* (GWAS) a large number of SNPs distributed all over the genome are genotyped, eliminating the need of prior evidence for a particular gene or pathway (Hirschhorn and Daly, 2005). As the cost for SNP genotyping has been constantly dropping over the recent years, GWAS studies of thousands of individuals genotyped at hundreds of thousands of SNPs have become reality. The reference study for GWAS was published in 2007 by researchers from the Wellcome Trust Case Control Consortium, where 2,000 cases each of seven common diseases were assayed at 500,000 SNPs (The Wellcome Trust Case Control Consortium, 2007). Today, at the time of writing there are over 500 publications listed in the GWAS catalog of the National Human Genome Research Institute (NHGRI) (www.genome.gov/gwastudies). Despite some undeniable success, results in general have not been as good as initially expected, as the discovered variants generally show small effects and only explain a small fraction of the genetic variance of the phenotypes of interest (Manolio et al., 2009). The question of how this so-called 'missing heritability' can be explained is one of the key topics in the genetics research community today.

Analysis of genetic association studies

Before assessing the actual association of the studied markers with the trait of interest, a number of preliminary analyses are typically employed to avoid false positives due to spurious associations or biases in the data. One cause of such is low quality genotyping data, for example in the form of missing data, or miscalled genotypes. These problems are usually dealt with by setting stringent thresholds for genotype calling and removing any samples or SNPs not passing these thresholds prior to analysis, although more sophisticated methods also exist (Scheet and Stephens, 2008). Another potential source of spurious associations is undetected population structure. If the study population is made up of subpopulations that are genetically different, and also have different prevalence for the trait in question, spurious associations can arise at loci that show different frequencies in the two subpopulations (Figure 1.8) (Marchini et al., 2004).

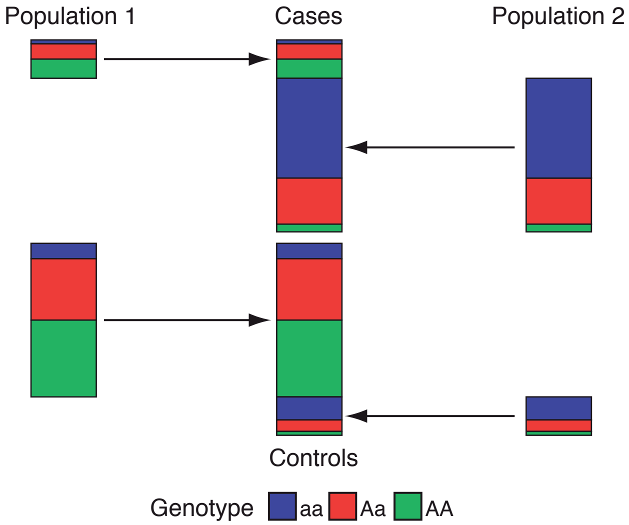


Figure 1.8: Effect of population structure in genetic association studies. In this example, a spurious association results from different genotype frequencies together with different proportions of cases and controls in the two populations (Marchini et al., 2004).

Population structure can be detected using approaches like PCA, as described above, and a number of methods exist to deal with any structure detected (Balding, 2006).

After preliminary analyses are completed one can do the actual testing for association. In a *single-marker analysis* each SNP is tested in turn for association with the trait. A number of different statistical tests, differing in their assumptions and power, can be employed, from a simple χ^2-test of allele frequencies to more sophisticated penetrance models using regression methods (Balding, 2006). *Multi-marker analysis* or haplotype analysis takes advantage of the additional information provided by considering the joint distribution of neighbouring SNPs. Analysing haplotypes has a variety of advantages, from biological interpretability to increased power to detect associations of rare alleles (Schaid, 2004). However, classical statistical tests using haplotypes instead of SNPs have their own problems, like defining the actual haplotypes and dealing with rare ones. More sophisticated multi-marker methods have been developed that allow to circumvent many of these problems (Browning, 2006; Tachmazidou et al., 2007).

Analysing a large number of SNPs inevitably means performing a large number of statistical tests. Considering a single marker analysis of an average sized GWAS with 500,000 SNPs, at a nominal significance threshold of 0.01 we would expect 5,000 of them to be associated due to chance alone. It is therefore necessary to correct the nominal p-values for each SNP for the number of tests performed. Only SNPs that remain significant after this correction can be considered truly associated with the trait. The simplest and most conservative method for correction is the Bonferroni procedure. In this method the corrected p - value is simply the nominal p-value multiplied by the number of tests if the resulting value is smaller than 1, and 1 if otherwise. However, this method is overly conservative for situations where the tests are not independent from each other, which is the case for SNPs that are in LD with each other. An alternative approach is to estimate the null-distribution of no association with a trait empirically by permutation. In this procedure, the labels of cases and controls are randomly permuted and the test statistic is calculated again for this permuted set. A corrected p-value can then be obtained by comparing the originally obtained statistic to a large enough number of permuted ones. Importantly, this procedure only destroys associations between phenotype and genotype, but keeps the correlation structure among markers intact (Hirschhorn and Daly, 2005). A disadvantage of this method is that it becomes computationally very demanding if applied to a large numbers of SNPs and samples.

1.2.3 Signatures of natural selection

As introduced in Section 1.1.2, natural selection is one of the main forces shaping genetic diversity. Even though it is clear that selection plays an important role in humans, ever since the introduction of the *neutral theory*

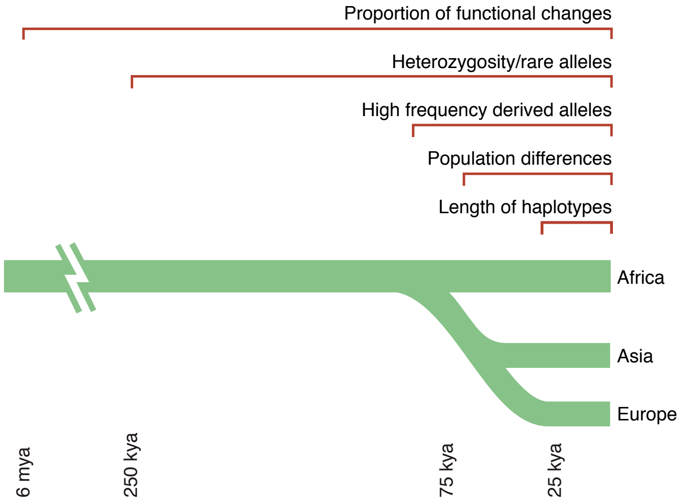

Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Figure 1.9: Molecular signatures of natural selection and the time scale to detect them (Sabeti et al., 2006).

of molecular evolution by Moto Kimura (1968) it was thought that it was mostly negative selection against deleterious alleles, and polymorphisms segregating in the population were in general neutral alleles governed by random genetic drift. The availability of DNA sequence data has allowed to reconsider this notion in recent years, and studies trying to infer the importance of all types of natural selection have become increasingly common. Positive selection in particular has seen much interest over the past years (Sabeti et al., 2006; Nielsen et al., 2007; Akey, 2009; Pickrell et al., 2009). The rationale for this is simple: Positive selection is the process that underlies human adaptation, so finding the targets of selection in the genome can reveal genes that are of functional importance for a particular trait (Bamshad and Wooding, 2003). Comparing differences of humans to closely related species like chimpanzees allows the inference of ancient selective events that were important in the early evolution of our species. In the context of this work however, our interest lies in more recent events, i.e. after the emergence of anatomically modern humans. The focus of this section will therefore be on the molecular signatures of these events, and the methods for detecting them (Figure 1.9).

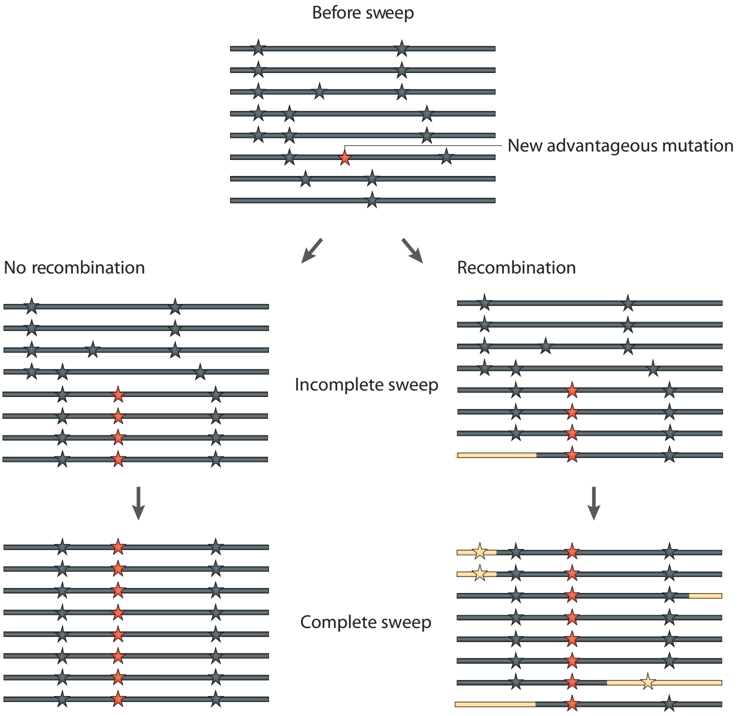


Figure 1.10: Example of a selective sweep involving a new advantageous mutation. Neutral haplotypes recombining onto the selected one are depicted in yellow (Nielsen et al., 2007).

Selective sweeps

An important concept in the methodology to detect positive selection from genetic data is that of a *selective sweep*. In a selective sweep, a newly arisen or previously rare advantageous mutation increases in frequency in the population. As a consequence, any neutral variation that is linked to the advantageous mutation will also be increasing in frequency, a process also known as *genetic hitchhiking* (Maynard Smith and Haigh, 1974). Its consequence are distinct signatures in the pattern of genetic variation in the selected region: A reduction in genetic diversity, a temporal increase in LD, as well as a skew in the distribution of allele frequencies (Nielsen, 2005). The magnitude of these signatures depends both on the strength of selection as well as the recombination rate in the selected region. In the absence of recombination, any genetic variation linked to the advantageous mutation will be eliminated when the sweep is complete. With recombination, neutral alleles on other haplotypes can recombine onto the selected one, so some neutral variation will remain after the sweep (Figure 1.10). The standard model for a selective sweep acting on a newly arisen mutation is sometimes also referred to as a *hard sweep*. In a *soft sweep*, selection is acting on genetic variation that was previously segregating in the population (standing genetic variation), which typically has a much weaker effect on linked neutral alleles, making it considerably more difficult to detect (Hermisson and Pennings, 2005; Przeworski et al., 2005; Pritchard et al., 2010).

Tests for natural selection

As described in the previous section, a selective sweep leaves distinct molecular signatures in the region surrounding the advantageous mutation. These signatures form the basis for many different tests of natural selection, also referred to as *neutrality tests*. In a typical neutrality test, some sort of summary statistic of the patterning of genetic variation in the analysed region is calculated, which is used to determine whether the result is compatible with the null hypothesis of neutral evolution. If the result is significantly different from the neutral expectation, the alternative hypothesis of natural selection is accepted. The general difficulty with these tests is the determination of the neutral expectation, since the patterning of genetic variation is a product of a complex interplay of many different processes. In particular, historical demographic events such as population reductions (bottlenecks) or expansions can produce molecular signatures very similar to the ones produced by natural selection. One solution is to determine the neutral expectation by simulating a large number of datasets, incorporating various demographic models (e.g. Schaffner et al., 2005). The significance of the statistic is then determined by comparing its value to the distribution obtained from the simulations. Another widely used approach is based on the notion that any event related to population history will affect the whole genome in a similar fashion, whereas selection acts at a specific location. Under the assumption that most regions of the genome will be evolving neutrally, the genome-wide distribution of the statistic can then be used as an empirical background to find regions in the extreme tails of the distribution (Akey, 2009). Importantly, this so-called outlier approach is not a formal statistical test, but rather a way to find regions that exhibit unusual patterns of variation. Simulation studies have shown that this approach can indeed find regions under selection, but care has to be taken with the interpretation as the number of false positives can be high depending upon the specific parameters and extent of selection (Kelley et al., 2006; Teshima et al., 2006).

Allele frequency spectrum As mentioned in the introduction to this section, natural selection leaves a distinct signature in the distribution of allele frequencies (i.e. the allele frequency spectrum) in the affected region. In particular, positive selection results both in a reduction in diversity combined with an excess of rare alleles, as well as an enrichment in high frequency derived alleles (Nielsen, 2005). The signature of balancing selection on the other hand is an excess of intermediate frequency alleles. These signatures form the basis of a variety of classical neutrality tests, the most famous one being Tajima's D test (Tajima, 1989). Even though in principle it allows the inference of both positive and balancing selection, the latter one is in general more difficult to detect (Andrés et al., 2009).

Population differentiation Another approach to detect positive selection is to investigate allele frequency differences between populations. If a locus is the target of selection in one population, but evolves neutrally in another one, it is expected to show a higher level of genetic differentiation than the average differentiation between the two populations (Novembre and Di Rienzo, 2009). A widely used approach is to calculate Wright's fixation index F_{ST} as a measure of population differentiation (Holsinger and Weir, 2009), and to identify regions with large values, although recently more sophisticated methods have also been developed (Chen et al., 2010).

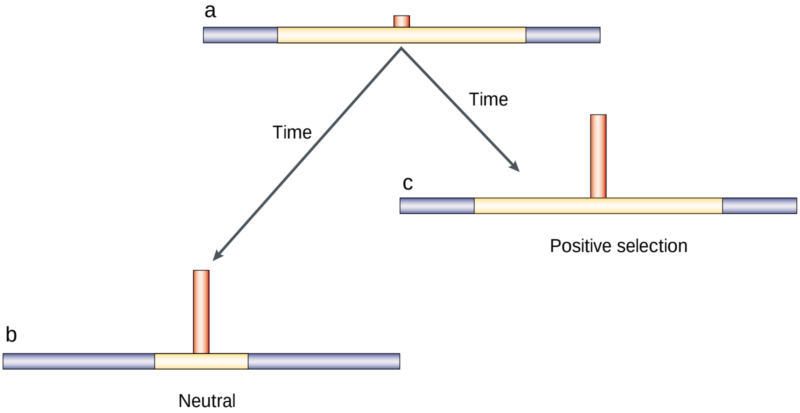


Figure 1.11: Linkage disequilibrium as a signature for recent positive selection (Bamshad and Wooding, 2003).

Long range linkage disequilibrium Powerful approaches for detecting positive selection are a class of methods that rely on linkage disequilibrium as a molecular signature. The basic concept of this type of tests is shown in Figure 1.11. A new allele is initially at low frequency in the population and in high LD with neighbouring markers (Figure 1.11a). If the allele is neutral, genetic drift can cause its increase in frequency over time, but recombination reduces the extent of LD surrounding it (Figure 1.11b). An advantageous allele on the other hand, can rise in frequency faster than recombination can break down LD with its surrounding markers, leading to a characteristic signature of high frequency alleles on a background haplotype with very long range LD (Figure 1.11c). A variety of methods have been developed to detect this signature, all of them based on a measure of long-range LD called *extended haplotype homozygosity* (EHH). The original test which first introduced this measure is the *long-range haplotype* (LRH) test (Sabeti et al., 2002a). This was subsequently much improved in the popular *integrated haplotype score* (iHS) test (Voight et al., 2006). Another class of tests is based on comparing EHH among different populations, as implemented in the *cross population EHH* (XP-EHH) test (Sabeti et al., 2007) and the Rsb test (Tang et al., 2007). These methods are particularly powerful for detecting very recent, incomplete or recently completed sweeps in the genome, back to around 30,000 years ago (Sabeti et al., 2006). They are therefore covering what is thought to be a key period for human evolution, when both population migrations and the onset of agriculture roughly 10,000 years ago exposed humans to a variety of new environments and selective agents. The example of the malaria parasite *Plasmodium falciparum* is one example of such a selective agent, and will be described in greater detail in the following section.

Developing methods for the detection of natural selection remains a very active field of research, with ever more powerful and robust methods being developed (Grossman et al., 2010). However, most of them are still based in the classical model of a hard sweep involving the spread of a new, advantageous mutation. A challenge for future studies will be to infer the targets of selection of other, likely equally important modes of adaptation, like soft sweeps on standing genetic variation or polygenic adaptation involving multiple interacting loci (Pritchard et al., 2010).

1.3 Malaria

The previous sections gave an introduction to human genetic diversity, and some of the motivations and methods for its study. The particular motivation of this present work has been to study human genetic diversity in the context of malaria. It is therefore necessary to introduce the disease in greater detail, which will be the goal of this following section.

1.3.1 Malaria parasites

Malaria is an infectious disease caused by protozoan parasites of the genus *Plasmodium*. It is a vector-borne disease, transmitted through bites of mosquitos of the genus *Anopheles*. *Plasmodium* parasites can infect a variety of host species, from reptiles, birds and rodents to the great apes. The four main species of *Plasmodium* that can cause human malaria are *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale*. Among them, *P. falciparum* is by far the most virulent, estimated to be responsible for more than 90% of cases worldwide (World Health Organization, 2009). Given this high virulence of *P. falciparum*, much research effort has been directed at the question of the origin of this species and its pathogenicity. Early phylogenetic studies indicated that human *Plasmodium* species were not a monophyletic group originating from a shared common ancestor, but rather that *P. falciparum* was most closely related to *P. reichenowi*, a parasite that infects chimpanzees (Escalante and Ayala, 1994; Escalante et al., 1995). These results suggested that human-infecting *P. falciparum* originated from a host-switch of a chimpanzee parasite, after which both parasite species diverged due to ongoing coevolution with their respective hosts. Recent studies support this hypothesis (Rich et al., 2009), although one study concluded that not chimpanzees but rather bonobos were the source host of the switch (Krief et al., 2010). A variety of studies investigating genetic variability of *P. falciparum* found low levels of neutral variation, indicating a recent population expansion of the parasite population estimated to have started around 10,000 years ago (Rich et al., 1998; Volkman et al., 2001; Joy et al., 2003). This would coincide with the time of the first expansion of malaria,

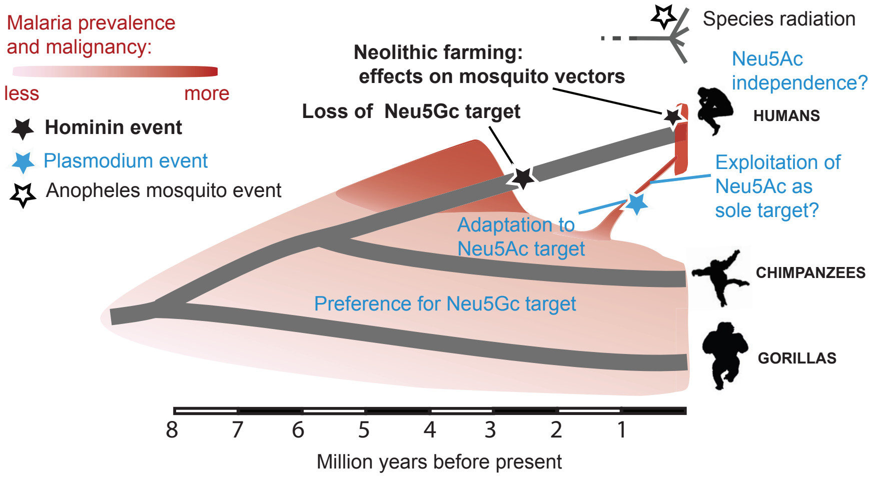

Malaria prevalence and malignancy:
less
more
Hominin event
Plasmodium event
Anopheles mosquito event
Species radiation
Neolithic farming: effects on mosquito vectors
Neu5Ac independence?
Loss of Neu5Gc target
HUMANS
Exploitation of Neu5Ac as sole target?
Adaptation to Neu5Ac target
CHIMPANZEES
Preference for Neu5Gc target
GORILLAS
8
7
6
5
4
3
2
1
Million years before present

Figure 1.12: A scenario for the origin of human malaria. Neu5Gc is the chimpanzee sialic acid, whereas Neu5Ac is the human variant (Varki and Gagneux, 2009).

which is thought to have occurred around the same time due to a variety of environmental changes that facilitated the spread of the disease: The onset of agriculture and subsequent increase in human population density, climate change towards warmer and more humid conditions as well as the rapid expansion of the population of the vector species *Anopheles* (Hartl, 2004). Even though this hypothesis is not undisputed (Mu et al., 2002), most of the data available to date seems to support it.

A possible scenario for the evolution of human malaria is shown in Figure 1.12. The authors of this hypothesis observed in a previous study that humans lack an enzyme in the biosynthesis pathway of a particular sialic acid, which are sugar molecules forming the terminal parts of the glycan chains of cell surface glycoproteins. This human-specific loss was estimated to have occurred around 2 million years ago. Strikingly, the authors also found that the chimpanzee parasite *P. reichenowi* preferentially bound to blood cells expressing the sialic acid that humans were missing, whereas *P. falciparum* preferentially bound the human variant (Martin et al., 2005). This strongly suggests that the enzyme function has been lost in humans as a consequence of coevolution with the parasite in order to escape infection. The changes in the environment described above could then have facilitated the spread of a newly arisen parasite mutant that exploited the human-specific glycan side chain, the present day highly pathogenic *P. falciparum* (Varki and Gagneux, 2009).

1.3.2 Epidemiology and burden of malaria

Depending on the observed annual parasite incidence, malaria-endemic regions can be classified into regions of unstable transmission, where transmission varies considerably from year to year, versus stable transmission, where generally intense transmission occurs year-round. Within regions of stable transmission, further subcategories of endemicity can be formed based on the prevalence of parasite infections. As can be seen in Figure 1.13, stable transmission occurs worldwide in tropical and subtropical regions, with highest levels of endemicity found in Sub-Saharan Africa. A recent effort to map malaria endemicity estimated a total of 1.3 billion people at risk of *P. falciparum* malaria, roughly half of those in Africa. Furthermore, even though a similar number of people are at risk in the Central and South East Asia regions, rates of endemicity are generally much higher in Africa, with only around 15% of people living in areas of low endemicity, compared to 80% in Central and South East Asia (Hay et al., 2009). This fact is also well reflected in the estimates of the global distribution of malaria cases and deaths. In 2008, there were an estimated 243 million cases of malaria worldwide, of which 208 million, or 85% were occurring in Africa. As mentioned briefly in the previous section, the lion share of those cases is due to *P. falciparum*, with 93% of cases worldwide and 98% in Africa (World Health Organization, 2009).

The distribution of deaths caused by malaria paints an even grimmer picture. The total number of deaths due to malaria worldwide was estimated to be up to one million, again with the majority occurring in Africa (89%) and due to *P. falciparum* infection. Even more disconcerting, the population group exhibiting by far the highest death toll is children under the

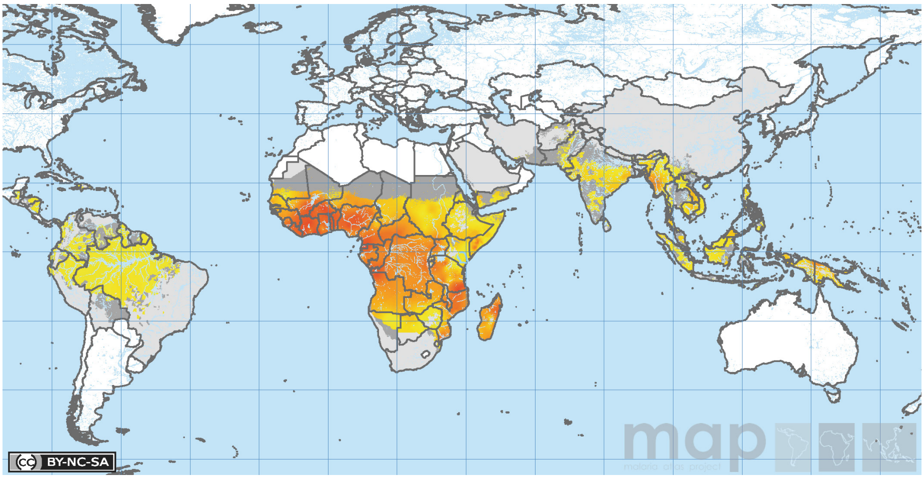


Figure 1.13: Worldwide malaria endemic regions. Orange colours show endemicity of malaria in regions of stable transmission, with darker colours indicating higher endemicity (Hay et al., 2009).

age of five, which account for 85% of deaths world-wide (88% in Africa). Given these figures, it is clear that malaria continues to cause a major burden for human health, particularly in Africa. Without neglecting the contribution of the other malaria parasites, it is evident that *P. falciparum* malaria is the main culprit of the global health toll, and therefore the main focus of this work.

1.3.3 Clinical manifestations

The consequences of infection with malaria parasites can be manifold, ranging from completely asymptotic infections or mild diseases to severe disease and death. Based on this wide range of outcomes, a distinction can be made between *uncomplicated* malaria and *severe* malaria. After being bitten by an infected mosquito, the incubation time until the first symptoms develop typically ranges from 6-14 days for *P. falciparum*, and slightly longer for other parasite species. A particular feature of *P. vivax* and *P. malariae* malaria is the possibility of a relapse of the disease even years after a previous infection, due to dormant parasite forms remaining in the liver of individuals (Trampuz et al., 2003). In the case of uncomplicated malaria, symptoms are similar to those of common viral infections, and include fever, headaches, chills, dizziness nausea and vomiting, among others. Severe malaria on the other hand is a complex disorder which can include a variety of potentially life-threatening complications. The typically observed major complications are severe anaemia due to destruction of red blood cells, and cerebral malaria, which can lead to loss or impairment of consciousness, convulsions and coma. However, many other complications also often occur in combination with any of the two mentioned above, including metabolic acidosis and acute respiratory stress, acidotic breathing, and renal failure (World Health Organization, 2000; Mackintosh et al., 2004; Rogerson et al., 2004). As seen in the previous section, the vast majority of cases and deaths due to severe malaria is caused by infection with *P. falciparum*, although recent studies show that severe disease due to *P. vivax* might be more common than previously thought (Genton et al., 2008; Tjitra et al., 2008).

As a consequence of constant exposure to infected mosquitos and multiple infections during adolescence, adults living in areas of stable transmission generally acquire semi-immunity to malaria. Although infections and subsequent parasitaemia continue to be observed, severe disease and death are essentially absent (Doolan et al., 2009). An exception to this are women during pregnancy, who show a marked increase in susceptibility to malaria even if they had previously acquired semi-immunity. This increased susceptibility occurs particularly in the first and second pregnancy, with a reduction in subsequent pregnancies due to increased acquired immunity (McGregor et al., 1983; Menendez, 1996). A key characteristic of malaria in pregnancy is infection of placental tissue, which will be described in more detail in the next section. Malaria in pregnancy can cause a wide variety of adverse outcomes for both the mothers and newborns, among them maternal anaemia, low birth weight and premature delivery, as well as increased susceptibility of infection for newborns (Figure 1.14) (Steketee et al., 1996; Menendez et al., 2000; Mutabingwa et al., 2005; Desai et al., 2007). Co-infection with HIV, which are common

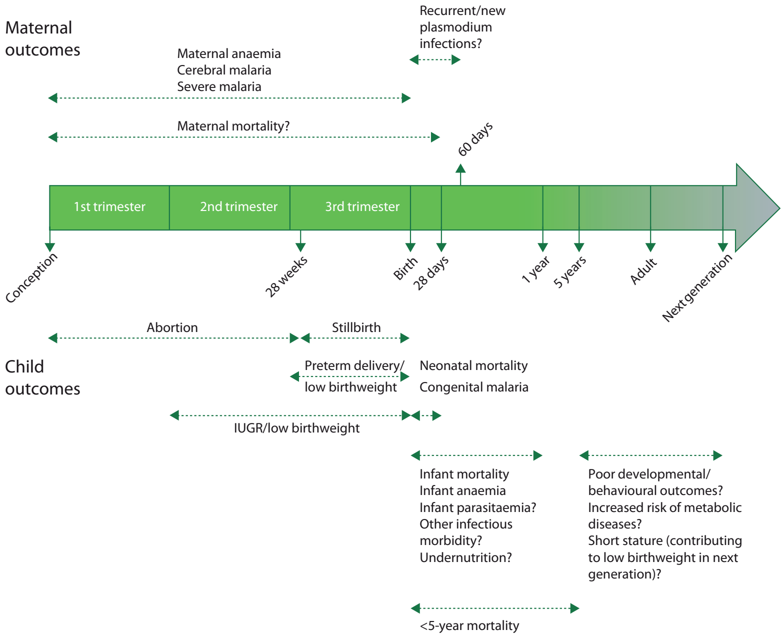


Figure 1.14: An overview of adverse outcomes for mothers and newborns due to malaria in pregnancy (Desai et al., 2007).

in Africa, can significantly worsen these outcomes. A substantial part of the yearly death toll associated with malaria is due to malaria in pregnancy, with up to 200,000 estimated infant deaths in Africa every year, as well as substantial maternal mortality (Steketee et al., 2001; Desai et al., 2007).

1.3.4 Life cycle and pathogenesis

The aim of this section is to give a more detailed description of the pathogenesis of malaria, with a particular focus on the life cycle of the malaria parasites and the key processes associated with infection. Figure 1.15 gives a schematic overview of the life cycle in both the human host as well as the mosquito vector.

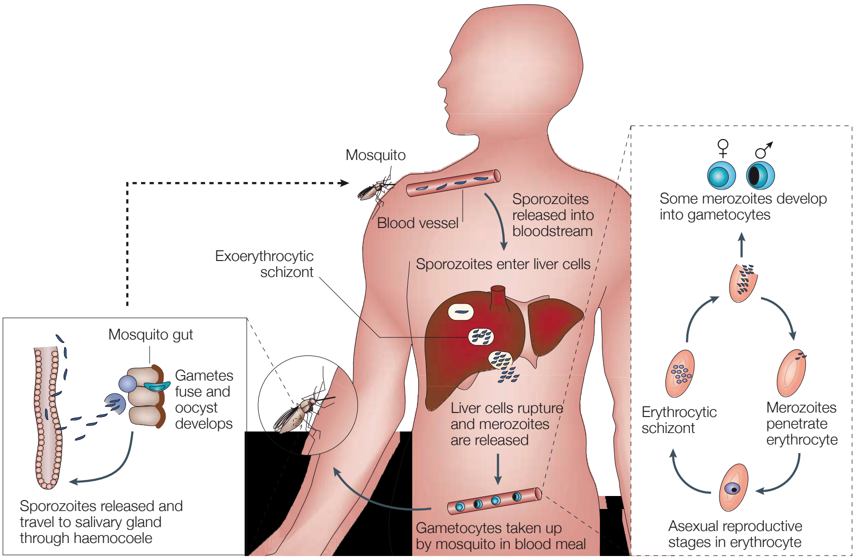

Mosquito
Blood vessel
Sporozoites released into bloodstream
Exoerythrocytic schizont
Sporozoites enter liver cells
Mosquito gut
Gametes fuse and oocyst develops
Liver cells rupture and merozoites are released
Sporozoites released and travel to salivary gland through haemocoele
Gametocytes taken up by mosquito in blood meal
♀ ♂
Some merozoites develop into gametocytes
Erythrocytic schizont
Merozoites penetrate erythrocyte
Asexual reproductive stages in erythrocyte

Figure 1.15: The life cycle of *Plasmodium falciparum* (Stevenson and Riley, 2004)

Liver stage

Infection begins with a bite from an infected female *Anopheles* mosquito, which results in the injection of a parasite form named *sporozoites* into the subcutaneous tissue and bloodstream of the affected individual. From there, sporozoites travel to the liver, where they rapidly invade liver cells (hepatocytes) through a complex process involving traversal of liver macrophages known as Kupffer cells (Vaughan et al., 2008). Once inside the hepatocytes, liver stage development is initiated, and parasites multiply by undergoing asexual reproduction. During this process, each sporozoite develops into thousands of merozoites, which then get released into the bloodstream to initiate the erythrocytic stage of infection. This release seems to involve interference of the parasite with host cell death mechanisms (apoptosis), one of the many strategies the parasite is thought to have evolved in order to evade the host defence (Sturm et al., 2006). Liver stage development typically takes around seven days, during which clinical symptoms of disease do not yet manifest themselves. Compared to other parasite stages, liver stage parasites are generally more difficult to study, due to both low numbers as well as being relatively inaccessible (Vaughan et al., 2008). Nonetheless, much recent effort has been aimed at them, as a consequence of being the target of the most promising malaria vaccine candidates (Alonso et al., 2005).

Erythrocyte invasion and parasite expansion

After their release into the bloodstream, merozoites initiate the erythrocytic stage of the disease by invasion of red blood cells. Erythrocyte invasion is a complex process involving molecular interactions between various parasite ligands and their respective erythrocyte cell surface receptors (Gaur et al., 2004). Host receptors used for invasion include glycophorins, as well as other, as of yet unknown molecules (Baum et al., 2005). The host receptors employed for invasion often possess glycan sidechains such as sialic acids that are important for ligand binding (see Section 1.3.1), but parasites can use a number of distinct invasion pathways that can be sialic acid independent (Stubbs et al., 2005; Baum et al., 2005). This is most likely a consequence of the ongoing evolutionary arms race between the parasite and the human host, and allows invasion also in the presence of other receptors and/or glycans. A famous example of this is the case of the Duffy antigen and erythrocyte invasion of *P. vivax*, which will be discussed in more detail in a later section. Once within the erythrocytes, merozoites develop and asexually reproduce during 48 hours. Each infected red blood cell (IRBC) produces around 20 daughter merozoites, which subsequently get released into the blood stream through rupture of the IRBCs, ready to newly infect erythrocytes themselves (Miller et al., 2002). It is at this stage that clinical manifestations of the disease appear, with bouts of high fever occurring in cycles coinciding with the massive synchronised release of merozoites into the bloodstream.

Cytoadherence of infected erythrocytes

A key process in the pathogenesis of malaria is the ability of parasites to modify the cell surface of IRBCs to facilitate adherence of the cells to

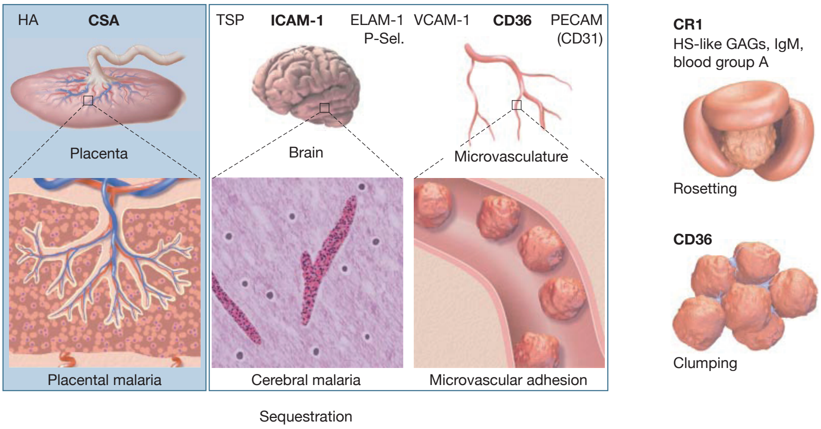


Figure 1.16: Cell adhesion processes in the pathogenesis of malaria. Host receptors molecules implicated with a particular process are indicated above the respective process (modified from Miller et al., 2002).

endothelial tissue throughout the body (Miller et al., 2002). Sequestration of infected blood cells in the vasculature prevents them from passing through the spleen and subsequent cleaning of the infected cells, an effective mechanism for the parasite to evade the host defence mechanisms. Cell adhesion processes of infected erythrocytes and their sequestration in other organs are also associated with complications observed in severe malaria, although the exact relationship with them is still not completely understood (Rogerson et al., 2004). For example, sequestration of IRBCs in cerebral blood vessels is a common occurrence in cerebral malaria, whereas placental malaria infection is characterised by sequestration and accumulation of IRBCs in the intervillous space of the placenta (Figure 1.16). Infected erythrocytes can also adhere to other non-infected erythrocytes (referred to as rosetting) or form clumps with each other, further contributing to the pathogenesis of *P. falciparum* (Miller et al., 2002).

Central to this adhesion processes is the *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) family, the members of which are expressed on the cell surface of IRBCs and mediate binding to the different host receptors. The members of this family of multi-domain proteins are encoded by a family of more than 50 different, highly variable genes, called *var* genes (Kraemer and Smith, 2006). The genomic organisation of *var* genes allows the generation of a repertoire of extreme antigenic diversity, through frequent recombination and gene conversion, thereby constantly generating new forms. At any particular time, only one var gene is expressed by an individual parasite, but expression is readily switched to another form during the course of an infection, which allows the parasite to evade clearance by the host immune system and the establishment of prolonged infection (Scherf et al., 2008). One exception to this is placental malaria, where it has been suggested that a single *var* gene, var2csa is expressed by most IRBCs that show cytoadherence with placental tissue. On the host side, the receptor molecule implicated to mediate the adherence of infected erythrocytes in the placenta is a glycosaminoglycan called chondroitin sulphate A (CSA) (Fried and Duffy, 1996; Fried et al., 2006). Although these *in vitro* studies clearly demonstrate the adhesion of IRBCs from placental tissue to CSA, knowledge on whether other receptors also play important roles *in vivo* is still limited.

Host inflammatory response

Another important aspect of the pathogenesis of infection is the inflammatory response of the host and its regulation. Both studies in humans, as well as model systems of malaria in mice have shed some light on the intricacies of this response. On one hand, a rapid and potent inflammatory response during early blood stage infection seems to be crucial for the control and eventual clearing of the infection. This response is mediated by pro-inflammatory cytokines such as tumor necrosis factor (TNF) and interferon-γ (IFN-γ). On the other hand, these pro-inflammatory cytokines have also been associated with severe disease states like cerebral malaria, whereas anti-inflammatory cytokines such as interleukin 10 (IL-10) and transforming growth factor-β (TGF-β) are protective. The general consensus from these results is that an efficient response to infection requires a delicate balance between a quick and intense early inflammatory response, and a subsequent efficient suppression of inflammation to avoid promotion of severe pathology of the disease (Artavanis-Tsakonas et al., 2003; Hunt and Grau, 2003; Stevenson and Riley, 2004; Riley et al., 2006).

Mosquito stage

During blood stage merozoite development, a small fraction of parasites develop into male and female gametocytes for subsequent sexual reproduction. Gametocytes then circulate in the peripheral bloodstream, from where they are taken up by mosquitos feeding on the infected host. Once within the mosquito, gametocytes relocate to the mosquito midgut, where the change in environment triggers their maturation into gametes, which subsequently fuse and undergo sexual reproduction. After several intermediate developmental stages, sporozoites appear as the last stage of development within the vector, and subsequently migrate to the salivary glands of the mosquito. The duration of this development varies depending on the species and temperature, taking around 14 days at 26 °C for *P. falciparum*. After finishing development, up to 1,000 sporozoites can be found in the salivary glands, ready to to reinitiate the life cycle after the mosquito feeds on a new host (Baton and Ranford-Cartwright, 2005).

1.3.5 Susceptibility to malaria

Susceptibility to malaria is, not surprisingly given the multitude of factors potentially involved in the disease, a complex issue. Being a system involving three different species - the malaria parasites, the human hosts, and the mosquito vectors - any factor affecting any one of them can potentially influence the final outcome in humans. But before trying to address the question of the relative importance of the various factors, one needs a clear definition of what kind of susceptibility is being investigated, i.e. what is the phenotype of interest. As seen in the previous sections, infection with parasites does not necessarily lead to a clinical manifestation of the disease, like it is most often the case for adults living in high transmission areas. A distinction can therefore be made between looking at an outcome based on the status of an infection, or one that is based on disease status. In the first case, the simplest case would be a binary cate-

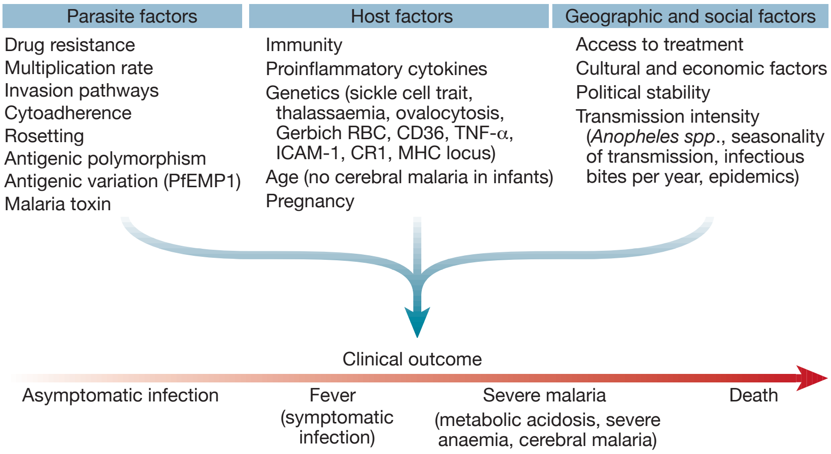

Parasite factors
Drug resistance
Multiplication rate
Invasion pathways
Cytoadherence
Rosetting
Antigenic polymorphism
Antigenic variation (PfEMP1)
Malaria toxin
Host factors
Immunity
Proinflammatory cytokines
Genetics (sickle cell trait, thalassaemia, ovalocytosis, Gerbich RBC, CD36, TNF-α, ICAM-1, CR1, MHC locus)
Age (no cerebral malaria in infants)
Pregnancy
Geographic and social factors
Access to treatment
Cultural and economic factors
Political stability
Transmission intensity (*Anopheles spp*., seasonality of transmission, infectious bites per year, epidemics)
Clinical outcome
Asymptomatic infection
Fever (symptomatic infection)
Severe malaria (metabolic acidosis, severe anaemia, cerebral malaria)
Death

Figure 1.17: The multitude of factors influencing the outcome of malaria infection (Miller et al., 2002).

gorisation of infection status, i.e. the presence or absence of parasites in the host. Additional information could be for example gained by quantitatively measuring levels of parasitaemia as a phenotype. On the other hand, researchers are most often interested in the clinical outcomes of disease, where again a multitude of different ways to define the phenotype of interest exist. Not surprisingly, a lot of effort has been aimed at discovering factors affecting the severity of the disease, given that severe malaria is the major cause of the death toll associated with malaria. From a clinical perspective, these phenotypes are arguably the most relevant to study, since the first goal of any intervention is to save the lives of patients. However, these high-level phenotypes are themselves complex entities, the final outcome of a complex interplay of cellular and organismal processes affected by the infection. It is therefore often difficult to properly define or measure what a particular phenotype like ’severe malaria’ actually entails. In the context of genetic susceptibility, studying more tractable phenotypes closer to basic organismal functions, so-called *endophenotypes*, could improve both the power to detect the genetic factors involved.

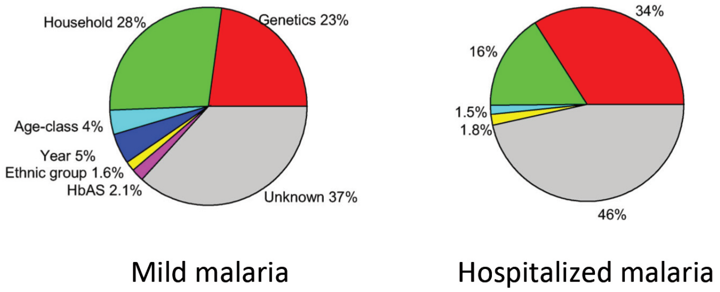


Figure 1.18: The relative contribution of factors in the incidence of mild and hospitalized malaria (modified from Mackinnon et al., 2005).

Figure 1.17 gives an overview of the many factors that can affect any of the many potential outcomes of interest. A first point of interest is the relative contribution of each of those factors on the outcome of interest, in particular genetic factors of the host, which are the focus of this work. Although the number of published studies is surprisingly scarce, an important work in this respect was published by Mackinnon et al. (2005). In order to investigate which factors explained the variation in malaria incidence, two separate pedigree-based heritability studies were conducted. In one study, the outcome of interest was incidence of mild, uncomplicated malaria, whereas the other one investigated hospitalised malaria. Figure 1.18 gives an overview of the results from both studies. The largest part of the variation in incidence corresponds to as-of-yet unknown factors. The other two factors with a large impact are the household of the study individuals as well as genetic factors, which account for roughly 50% of the variation when taken together. Genetic factors by themselves seem to be slightly more important in the more severe hospitalised malaria, where they account for a third of the variance, as opposed to 23% in mild malaria. Nevertheless, an important finding is the fact that the strongest protective genetic factor known, haemoglobin S, was not included in this and only accounted for 2.1% by itself. From these results it is therefore clear that susceptibility to malaria is a heritable trait, and that the most of its heritability is still unaccounted for. In the context of this work, this indicates that genetic susceptibility to malaria is a complex trait, with contribution of many different loci, the majority of which still unknown.

1.4 Malaria and the human genome

For any kind of disease to exert a strong selective pressure on human populations, it has to show two key characteristics: A large effect on mortality before reaching reproductive age; and exposure of this effect for a long period of time. As has been described in the previous sections, malaria clearly fulfils both of these requirements. It is therefore not surprising that malaria is today thought to be the strongest selective force in recent human history (Kwiatkowski, 2005). As a result of this strong selective pressure, it is therefore expected that any type of genetic variation that significantly increases the chances of survival of malaria would become more common in the population, i.e. be under the effects of natural selection. More than 60 years ago, J. B. S. Haldane first hypothesised that *thalassaemia*, a deleterious erythrocytic disorder that occurred at surprisingly high frequency, was in fact a protective factor against malaria and therefore under a selective advantage, inspired by the observation of the striking similarity of the geographical distributions of erythrocyte disorders and malaria transmission (Figure 1.19) (Haldane, 1949; Lederberg, 1999). The subsequent discovery of the protective effect of sickle cell anaemia in 1954 was the first evidence of natural selection acting in human populations (Allison, 1954). Since then, much has been learned about how the selective pressure of malaria affects human genetic variation (Kwiatkowski, 2005), and many loci influencing susceptibility have been described. Nevertheless, as mentioned in the previous section, the picture is far from complete. The aim of this section is then to give an overview over some of the loci that have been discovered, and the nature of selection acting upon them. The last part of this section will be devoted to introducing some very recent results of the first genome-wide studies of malaria.

1.4.1 Variants related to erythrocyte function

As mentioned in the introduction to this section, erythrocyte disorders like thalassaemia and sickle cell disease are the classical resistance factors for malaria. Given the importance of erythrocytes in the life cycle of the

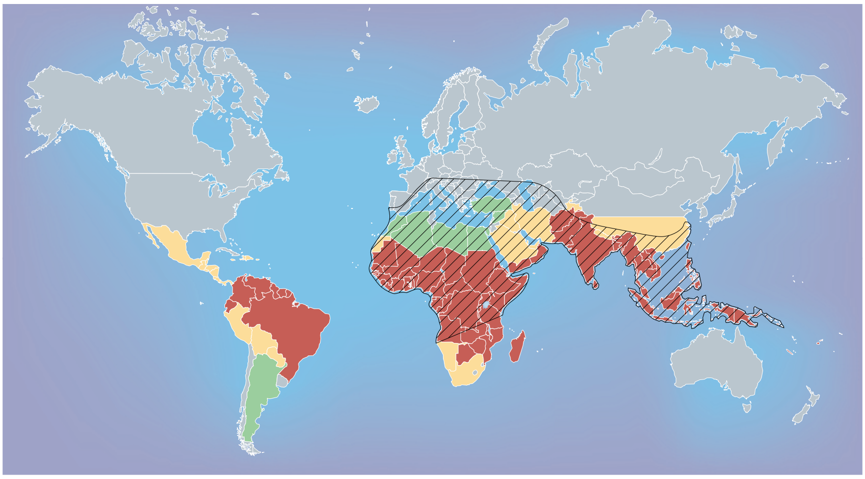


Figure 1.19: Worldwide malaria endemicity and distribution of erythrocyte disorders. The colors indicate rates of malaria transmission, with regions of high transmission in red (Cooke and Hill, 2001).

malaria parasites, it is not surprising that among the known resistance factors, many are associated with erythrocyte function.

Haemoglobin disorders

Adult haemoglobin (HbA) is a tetrameric protein composed of four protein chains, two chains of α globin (encoded by two identical HBA genes *HBA1* and *HBA2*), and two chains of β globin (encoded by *HBB*). Disorders of haemoglobin can be distinguished in two broad categories: structural variations of the HbA protein chains, as well as regulatory variation involving reduced levels of globin proteins.

Haemoglobin structural variants The three structural variants with protective effect that are found at appreciable frequencies are all due to non-synonymous SNPs in *HBB*: the sickle cell allele HbS (glutamic acid → valine at codon 6), HbC (glutamic acid → lysine at codon 6), and HbE (glutamic acid → lysine at codon 26). The sickle cell allele HbS results in deformed erythrocytes with a characteristic sickle-like shape (Figure 1.20), and individuals homozygous for the allele are affected by sickle cell

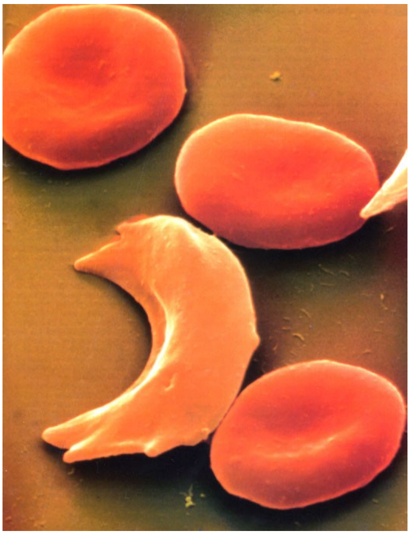


Figure 1.20: Microscopic image of a sickle cell erythrocyte

disease, a severe and very often fatal disorder. Heterozygote carriers on the other hand usually show only mild symptoms, but enjoy a roughly 10-fold reduction in risk of severe malaria (Allison, 1954; Aidoo et al., 2002; Ackerman et al., 2005; Williams et al., 2005a). HbS is found at frequencies up to 15 - 20% in some regions of Sub-Saharan Africa (Williams, 2006), despite the deleterious effect of the homozygote state, a clear indicator of its selective advantage. To date it is still the text-book example of balancing selection in human populations. Haemoglobin C is also found in Africa, although at lower frequencies than HbS (Kwiatkowski, 2005). As opposed to HbS, in the case of HbC homozygotes seem to exhibit the strongest protective effect, and only a mild clinical phenotype of anaemia (Modiano et al., 2001; Mockenhaupt et al., 2004a). The situation for HbE is similar to HbC, which is found at high frequencies in Southeast Asia, although epidemiological evidence for a protective effect is still somewhat limited (Hutagalung et al., 1999). However, population genetics analyses of both HbC and HbE alleles show signatures of recent positive selection in the affected populations, providing further support for a protective effect from malaria (Ohashi et al., 2004; Wood et al., 2005).

Reduced globin levels The main class of disorders caused by disruption of globin production are the thalassaemias. Thalassaemias are prevalent throughout Sub-Saharan Africa, the Mediterranean region and the Middle East, as well as Southeast Asia. Depending on which of the globins is affected, one can distinguish between α- and β-thalassaemias. A variety of subclasses of thalassaemias with varying degree in severity exist, depending on the nature and extent of the disruption or deletions of the globin genes (Weatherall, 2001). Population genetics analysis provided the first robust evidence for a protective effect of α-thalassaemia, as hypothesised by Haldane (Flint et al., 1986). More recently, this has also been confirmed by a number of epidemiological studies, which found a decreased risk of severe and fatal malaria in affected individuals (Mockenhaupt et al., 2004b; Williams et al., 2005c; Wambua et al., 2006). As an important reminder of the complexities involved in the susceptibility to malaria, another study reported compelling evidence of a negative interaction (epistasis) between α-thalassaemia and HbS in a Kenyan population (Williams et al., 2005b). Even though the two exhibit a protective effect when inherited by themselves, the protective effect was almost completely lost in individuals with both alleles together. These epistatic effects could explain why α-thalassaemia has not been fixed in African populations despite the strong selective advantage.

Erythrocyte surface variants

The blood stage of the life cycle of *Plasmodium* is initiated with the invasion of erythrocytes by the parasite, a process that involves extensive ligand-receptor interaction events between parasite and the erythrocyte surface. Not surprisingly, some of the most well-known examples of genetic resistance to malaria involve genes that produce erythrocyte cell surface molecules.

Duffy antigen The Duffy antigen is a chemokine receptor expressed on the cell surface of various cell types, encoded by the *DARC* gene (also known as *FY*). Allelic variants of the receptor expressed on erythrocytes

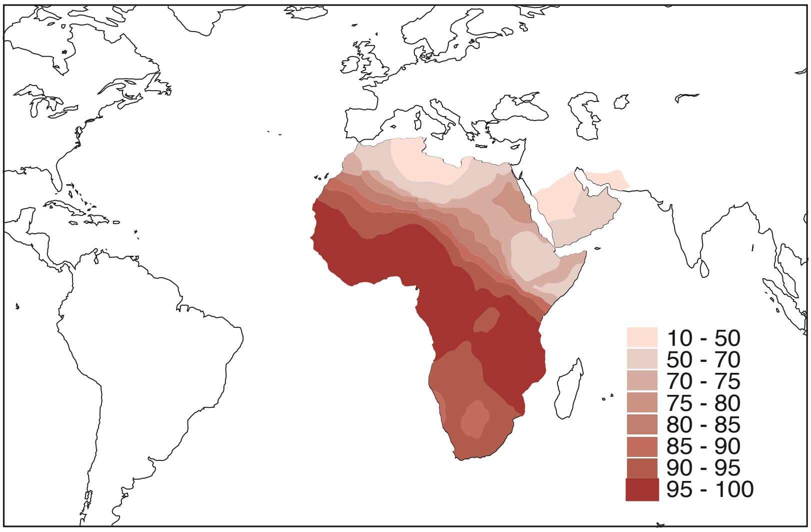


Figure 1.21: The global distribution and frequencies of the Duffy null allele (Sabeti et al., 2006).

form the basis of the Duffy blood group system, and three common forms have been described *(FY*A, FY*B, FY*0)*. The Duffy null allele *(FY*0)* causes repression of the expression of the receptor on the cell surface (Tournamille et al., 1995), and is at or close to fixation in most of Sub-Saharan Africa, but very rare in the rest of the world (Figure 1.21). The presence of the Duffy antigen is however required for erythrocyte invasion of *P. vivax*, which as a consequence is almost completely absent from Sub-Saharan Africa (Miller et al., 1976). Population genetic studies of the *DARC* locus have revealed strong signatures of positive selection at the locus, showing extreme population differentiation, much reduced levels of genetic variation and an excess of high frequency derived alleles in African populations, consistent with the hypothesis of a strong selective sweep causing the near-fixation of this allele (Hamblin and Di Rienzo, 2000; Hamblin et al., 2002).

ABO The ABO system is arguably the most important human blood group system. The three allelic forms A, B and O correspond to different glycan sidechains on erythrocyte membrane proteins. Although the global distribution of the different alleles is complex and could reflect selective pressures of a variety of pathogens, blood group O is consistently found at high frequency in Sub-Saharan Africa, whereas it is much less frequent in the rest of the world, which could indicate a protective effect against *P. falciparum* malaria (Cserti and Dzik, 2007). This hypothesised protective effect has indeed been confirmed recently, presumably due to a reduction of erythrocyte rosetting (Rowe et al., 2007; Fry et al., 2008b). In one of the studies, the authors also found extremely low levels of population differentiation in *ABO*, suggesting long-term balancing selection, a potential explanation for the absence of fixation of the O group in Sub-Saharan Africa (Fry et al., 2008b). Interestingly, another recent study found a highly significant association of a SNP at the *ABO* locus with levels of circulating soluble intercellular adhesion molecule-1 (sICAM-1), thereby providing a functional link between ABO polymorphisms and inflammatory adhesion processes involved in malaria pathogenesis (Paré et al., 2008).

Glycophorins Glycophorins are sialoglycoproteins expressed on the erythrocyte cell surface, which have been shown to be involved in *P. falciparum* erythrocyte invasion pathways (Miller et al., 2002; Kwiatkowski, 2005). Although there is to date no epidemiological evidence for a protective effect of glycophorin variants, a comparative study focusing on fast-evolving genes between humans and primates found three glycophorins (*GPA, GPB, GPE*) among the fastest evolving genes in humans. Intriguingly, the parasite ligand for the glycophorins (EBA-175) also shows high rates of substitutions, indicating a co-evolutionary arms race between humans and *P. falciparum* at this loci (Wang et al., 2003).

Erythrocyte enzymatic disorders

***G6PD* deficiency** Glucose-6-phosphate dehydrogenase (G6PD) is an enzyme involved in the pentose phosphate pathway, and plays an important role in the protection of the erythrocyte from oxidative stress. Patients with enzymatic deficiency caused by mutations in the gene suffer from *G6PD* deficiency, an X-linked recessive disorder that is one of the most common enzymatic disorders in humans. The most frequent deficient variant in Africa, *G6PD A*$^-$, is caused by two non-synonymous SNPs and results in G6PD with between 10% and 50% enzyme activity (Verra et al., 2009). The prevalence of the disorder in malaria endemic regions and its importance in erythrocyte metabolism had long suggested a protective effect. This has been confirmed by a number of studies, with a recent study showing that both hemizygous men and homozygous women are protected from severe malaria (Ruwende et al., 1995; Mockenhaupt et al., 2003; Guindo et al., 2007). However, another recent study analysing additional deficiency alleles in the Gambia detected the protective effect only when pooling the *G6PD A*$^-$ allele with the other deficiency alleles, showing the potential problems when faced with undetected allelic heterogeneity (Clark et al., 2009). In addition to the epidemiological results, evolutionary analyses of the *G6PD* region have shown strong signatures for recent positive selection for *G6PD A*$^-$ (Figure 1.22) (Tishkoff et al., 2001; Saunders et al., 2002; Sabeti et al., 2002a; Saunders et al., 2005). Additionally, a comparative study with nonhuman primates indicated that the signal of selection is specific to humans (Verrelli et al., 2006). Finally, one study investigated a different G6PD-deficiency allele common in Southeast Asia and also found strong signatures of recent positive selection. Surprisingly, the authors found that the allele reduced parasite density of *P. vivax*, and not *P. falciparum*, indicating that *P. vivax* has also been a strong selective force historically, despite its lower pathogenicity observed today (Louicharoen et al., 2009).

Pyruvate kinase deficiency Another common human enzymatic disorder is pyruvate kinase deficiency, an autosomal recessive disorder caused by loss-of-function mutations in the *PKLR* gene. An involvement of pyruvate kinase deficiency in resistance to malaria was first suggested after finding a protective effect in a mouse model of malaria (Min-Oo et al., 2003, 2004). This finding was subsequently also confirmed in humans, where it was found to protect erythrocytes against infection and replication of *P. falciparum* (Ayi et al., 2008).

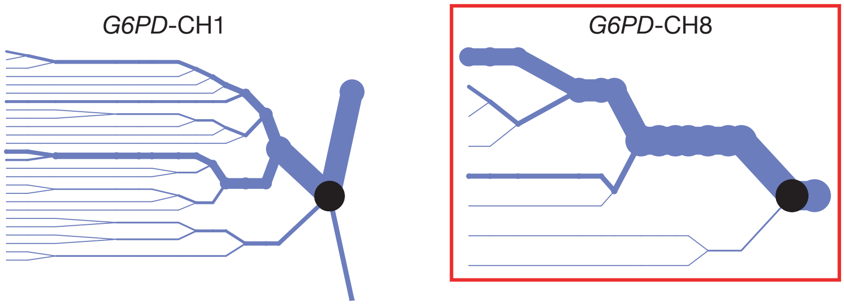


Figure 1.22: Signature for positive selection at *G6PD*. This so-called haplotype bifurcation diagrams shows the extent of LD breakdown at two haplotype variants of the *G6PD* region. The *G6PD A*$^{-}$ haplotype associated with protection is shown on the right. Compared with another haplotype from the same locus shown on the left, LD decays much slower for the *G6PD A*$^{-}$ haplotype, a clear indication of positive selection.

1.4.2 Cytoadherence, immunity and inflammation

As has been described in Section 1.3.4, cytoadherence, inflammation and immunity are all processes that are important in the pathogenesis of disease and its subsequent clinical manifestations. Not surprisingly, a number of genes encoding key members in this processes have been found to harbour variants that modulate the risk of disease, although results in general have been somewhat more inconclusive than those found in the erythrocytic variants.

Cytoadherence

ICAM1 Intercellular adhesion molecule-1 is encoded by the gene *ICAM1*, and is one of the receptors mediating the binding of infected erythrocytes to the endothelium of blood vessels (Berendt et al., 1989). In an early study, a non-synonymous SNP leading to a substitution in the N-terminal tail of the protein was found to increase susceptibility to cerebral malaria (Fernandez-Reyes et al., 1997), but subsequent studies found either a protective effect (Kun et al., 1999), or no effect at all (Bellamy et al., 1998). The most recent study also failed to find a significant association with severe malaria in three different African regions, and the SNP was not found to be an outlier in published genome-wide surveys of positive selection (Fry et al., 2008a).

CD36 Another endothelial receptor that binds infected erythrocytes is the CD36 antigen, encoded by *CD36* (Barnwell et al., 1989). A nonsense mutation in *CD36* that causes CD36 deficiency was found to be associated with susceptibility to cerebral malaria in homozygous individuals, but another study found a protective effect in heterozygous individuals (Aitman et al., 2000; Pain et al., 2001). A recent study found evidence for positive selection of the allele in Sub-Saharan Africa, but no association with severe malaria, suggesting that the signature of selection might be the consequence of a different selective agent (Fry et al., 2009).

Immunity and inflammation

Pro-inflammatory cytokines The pro-inflammatory cytokines TNF and IFN-γ play an important role in the inflammatory response to infection (see Section 1.3.4), and variants in both genes have been found to be associated with malaria. In *TNF*, a number of polymorphisms in the promoter region have been associated with cerebral malaria (McGuire et al., 1994, 1999) and levels of parasitaemia (Flori et al., 2005). In the case of IFN-γ, SNPs in the gene encoding it (*IFNG*) as well as the ligand-binding chain of the IFN-γ receptor (*IFNGR1*) have been found to be weakly associated with severe malaria (Koch et al., 2002, 2005). Another study found that reduced levels of IFN-γ were protective in cerebral malaria, but none of the SNPs analysed in the regions were associated with it (Cabantous et al., 2005).

CD40LG The protein CD40 ligand, encoded by the gene *CD40LG*, is expressed on the cell surface of T-cells and plays a role in B cell regulation. A SNP in the promoter region of the gene was found to be associated with severe malaria (Sabeti et al., 2002b). The haplotype carrying the allele conferring protection from severe malaria was also found to be under positive selection in African populations, in the study that originally introduced the long-range haplotype test for recent positive selection (Sabeti et al., 2002a).

Toll-like receptor signalling Toll-like receptors (TLRs) are key members of the human innate immune response, where they act by recognising a wide variety of different pathogens (Akira and Takeda, 2004). In recent years evidence for an involvement in malaria susceptibility has been building. In one study, two different mutations in the gene encoding Toll-like receptor 4 (*TLR4*) were found to increase susceptibility to severe malaria (Mockenhaupt et al., 2006), although individuals with one of the two mutations (Asp299Gly) seemed to exhibit lower mortality. Another study investigating the phenotypic consequence and global distribution of those mutations found an increased pro-inflammatory response caused by Asp299Gly (Ferwerda et al., 2007), which would be consistent with the increased susceptibility to severe malaria observed by Mockenhaupt et al. (2006). On the other hand, the high prevalence of the allele in Africa accompanied by an almost complete absence in the rest of the world suggest a selective advantage of the allele, presumably due to malaria. The authors also observed reduced mortality in individuals with the mutation, despite the increased susceptibility to malaria, which would be consistent with selection acting on this allele. Finally, a large case-control study found a non-synonymous SNP in the gene *TIRAP*, which encodes for an intracellular adaptor molecule for TLRs, associated with a variety of infectious diseases, among them severe malaria. The functional consequence of the variant seems to be an attenuation of TLR2 signalling (Khor et al., 2007).

1.4.3 Genome-wide approaches

With the publication of the first genome-wide association study in 2009, research into the genetic basis of malaria susceptibility has finally arrived in the genomic era. A large consortium of researchers conducted a large-scale case-control study in the Gambia in West Africa, backed by the Wellcome Trust (Jallow et al., 2009). The initial scan included 958 cases of severe malaria and 1,382 controls, genotyped at around 500,000 SNPs. A replication study to validate significant associations was sub-

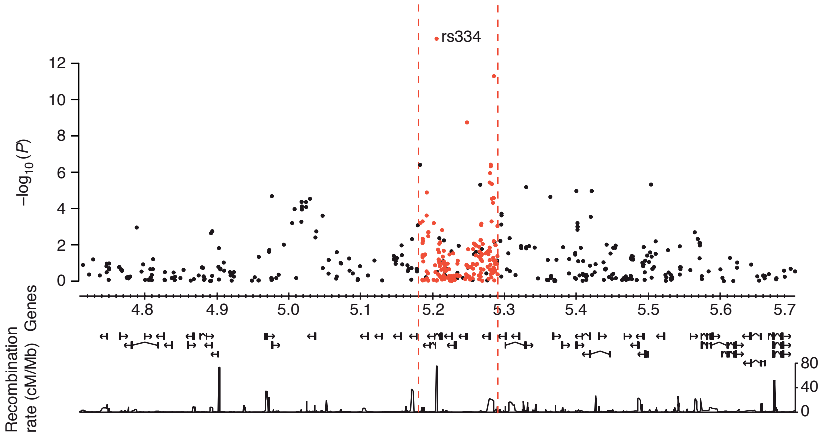


Figure 1.23: Results of the malaria GWAS in the *HBB* region. The black dots represent p-values of SNPs on the initially included SNPs, whereas red dots show results of imputation based on resequencing the region in the same population (Jallow et al., 2009).

sequently carried out in an independent sample of 1,087 cases and 2,376 controls. Despite the considerable expectations, results were somewhat disappointing, but highlighted a number of potential problems for carrying out large scale genetic studies in African populations. The strongest signal found in the study was found close to the *HBB* gene, but barely passed the suggested genome-wide significance p-value cutoff of 5×10^{-7} (The Wellcome Trust Case Control Consortium, 2007). When the causing variant of sickle cell anaemia was directly genotyped on the other hand, a much stronger signal was obtained ($p = 1.3 \times 10^{-28}$). The authors found that by resequencing the region in 62 individuals from the control group, and using this data to statistically infer genotypes at the discovered SNPs in the full analysis panel (an approach called *imputation*), a number of those imputed SNPs showed much stronger associations than any of the genotyped SNPs (Figure 1.23). The reason for the lack of strong associations in the original SNP array is the much reduced levels of LD in African populations, which require much higher density in genotyped SNPs for the indirect association approach of GWAS to have sufficient power. This explanation was also given for the observed lack of signals in other loci that had previously been associated with malaria (see previous sections). Only two other loci apart from the known HbS variant could successfully be replicated: a SNP close to the gene *SCO1*, involved in cytochrome oxidase function, as well as an intronic SNP in *DDC*, involved in dopamine synthesis. The replication study highlighted another potential pitfall in the analysis of African populations: due to to the large amount of genetic diversity within Africa, SNPs found by indirect association to a causal SNP will not be expected to replicate in a different population that shows a very different pattern of LD at the same locus.

In summary, although some progress has been made, there is still much to be learned in the genetics of resistance to malaria. A particular concern is the lack of diversity of phenotypes that have been analysed so far, with the majority of studies focussing on severe and/or cerebral malaria. The complete lack of studies investigating placental malaria for example was one of the motivations for this work. Nevertheless, the studies described in this section clearly demonstrate that malaria has had a strong impact on the human genome, and studies incorporating both evolutionary analysis and epidemiological data are without doubt the way forward in the genetics of malaria (Ayodo et al., 2007).

2

Objectives

The general goal of the work presented in this thesis is to gain a better comprehension of the way malaria has shaped the human genome. As outlined in the introduction, given the strong selective pressure that malaria parasites, in particular *P. falciparum*, have exerted on human populations in recent history, we expect the genome to harbour genetic variation related to the susceptibility to malaria, the majority of which still remains to be discovered. One aim of this work was therefore to identify novel genetic factors involved in malaria susceptibility, and characterise the nature and strength of selection acting on them. Furthermore, the majority of studies so far have focussed on severe malaria or related phenotypes, almost completely ignoring other, equally important phenotypes such as placental malaria. Another aim of this study was therefore to address this lack of knowledge and investigate placental malaria. Finally, a third aim for study was to provide a contribution to the understanding of the genetic structure of Sub-Saharan African populations.

In order to achieve these goals, two complementary approaches were applied: A classical genetic epidemiology approach, carried out as a case-control association study; and evolutionary analysis of genetic variation datasets, aimed at detecting evidence of natural selection due to malaria. In all of the analyses, we used a functional pathway approach for selection of the genes of interest. Below, we outline the specific aims for the studies carried out as part of this thesis work.

Natural selection in a candidate pathway for placental malaria infection

The aim of this part was to infer whether there was evidence for natural selection in two glycan receptor molecules that have previously been implicated in placental malaria infection. In particular, our aim was to provide a complete picture of the potential sources of variation in the receptors, by analysing all genes whose protein products are involved in their biosynthesis. The results of this part are presented in Section 3.1.

Genetic susceptibility to placental malaria infection

The aim of this study was to identify novel host genetic factors involved in placental malaria infection. To this end we conducted a case-control association analysis in population of pregnant women from Mozambique with and without placental malaria infection. The analysis was carried out in two phases, focussing on two different sets of candidate genes. In both of these phases we were also interested whether any positive results also showed evidence for natural selection.

Glycosylation In phase one, we chose to investigate SNPs in genes involved in glycosylation, augmented by a number of innate immunity genes. We decided to use a pathway-based approach, selecting genes whose products are the basis of the biosynthesis pathways of glycan structures. As above, the aim of this was to provide a complete picture of the potential sources of variation within the pathways. The results of this part are presented in Section 3.2.

Inflammation and Immunity In the second phase, our aim was to considerably increase the scope of the analysis while still employing a functionally informed approach in gene selection. We therefore chose to analyze SNPs in a set of around 1,000 genes involved in inflammation and immunity processes. Additionally, we were also interested in whether analysing the genes in the context of their network of interactions could give us any additional insights. The results of this part are presented in Section 3.3.

Genetic structure of Sub-Saharan African populations

The aim of the last part was to provide some new insights into the genetic structure of Sub-Saharan African populations, in particular the relationship of the population from Mozambique with other Sub-Saharan populations where genotype data was publicly available. In addition, we also wanted to investigate how the number of markers analysed influences the results. The results of this part are presented in Section 3.4.

3

Results

3.1 Evolutionary analysis of genes of two pathways involved in placental malaria infection

Martin Sikora, Anna Ferrer-Admetlla, Alfredo Mayor, Jaume Bertranpetit, Ferran Casals

Human Genetics 2008 May; 123(4): 343-57.

Sikora M, Ferrer-Admetlla A, Mayor A, Bertranpetit J, Casals F. Evolutionary analysis of genes of two pathways involved in placental malaria infection. Hum Genet. 2008; 123(4): 343-57.

3.2 A variant in the gene *FUT9* is associated with susceptibility to placental malaria infection

Martin Sikora, Anna Ferrer-Admetlla, Hafid Laayouni, Clara Menendez, Alfredo Mayor, Azucena Bardaji, Betuel Sigauque, Inacio Mandomando, Pedro L. Alonso, Jaume Bertranpetit, Ferran Casals

Human Molecular Genetics 2009 Aug 15; 18(16): 3136-44.

Sikora M, Ferrer-Admetlla A, Laayouni H, Menendez C, Mayor A, Bardaji A, et al. A variant in the gene FUT9 is associated with susceptibility to placental malaria infection. Hum Mol Genet. 2009; 18(16): 3136-44.

3.3 The role of immunity and inflammation genes and networks in placental malaria infection

Martin Sikora, Hafid Laayouni, Clara Menendez, Alfredo Mayor, Azucena Bardaji, Betuel Sigauque, Mihai G. Netea, Ferran Casals, Jaume Bertranpetit

Manuscript in preparation

The role of immunity and inflammation genes and networks in placental malaria infection

Martin Sikora[1], Hafid Laayouni[1,2], Clara Menendez[2,3,4], Alfredo Mayor[3,4], Azucena Bardaji[3,4], Betuel Sigauque[4] , Mihai G. Netea[5], Ferran Casals[1,6] and Jaume Bertranpetit[1,2,§]

[1] Institute of Evolutionary Biology (UPF-CSIC), CEXS – UPF – PRBB, Barcelona, Catalonia, Spain

[2] CIBER en Epidemiología y Salud Pública (CIBERESP, Spain)

[3] Barcelona Center for International Health Research (CRESIB), Hospital Clinic, Institut d'Investigacions Biomedicas August Pi i Sunyer (IDIBAPS), Universtitat de Barcelona, Spain

[4] The Manhiça Health Research Center (CISM), Manhiça, Mozambique

[5] Department of Medicine, Radboud University Nijmegen Medical Center, 6500 HB Nijmegen, The Netherlands

[6] Present address: Centre de Recherche, CHU Sainte-Justine, Université de Montréal, Montréal, Québec H3T 1C5, Canada

[§]Corresponding author: Jaume Bertranpetit

Phone: +34 933160845

Fax: +34 933160901

E-mail: jaume.bertranpetit@upf.edu

Abstract

A large proportion of the death toll associated with malaria is a consequence of malaria infection during pregnancy, causing up to infant 200,000 deaths annually. We previously published the first extensive genetic association study of placental malaria infection, and here we extend this analysis considerably, investigating genetic variation in over 9,000 SNPs in more than 1,000 genes involved in immunity and inflammation for their involvement in placental malaria infection. We applied a new approach incorporating results from both single gene analysis as well as gene-gene interactions on a protein-protein interaction network. Our results suggest a role for the IL-7/IL-7R signalling module in the susceptibility to infection. To our knowledge, this is the first large-scale genetic study on this particular form of malaria to date.

Introduction

Recent years have seen a substantial increase in efforts and funding directed at the control (1, 2) and eventual eradication of malaria (3). However, despite these efforts, it remains one of the deadliest diseases worldwide. The global death toll has been estimated at 700,000 to 1,000,000 in 2008 alone, from a total number of cases ranging from 208 million to 276 million (4). The majority of these deaths occur in Sub-Saharan Africa, as a consequence of infection with *Plasmodium falciparum*. A somewhat lesser known aspect of this statistics is the fact that a large fraction of this annual death toll is a consequence of malaria infection during pregnancy, with up to 200,000 deaths attributed to it (5, 6). Women in their first pregnancy are particularly at risk of infection (7, 8), and both mothers and their offspring face a number of potential life-threatening complications as a consequence of infection (9, 10). The key characteristic of the pathogenesis of this type of malaria is infection of the placenta, i.e. the sequestration and subsequent accumulation of infected erythrocytes in the intervillous space (11).

Genetic studies of malaria susceptibility have a long history, and a variety of host genetic factors have been implicated, most prominently the protective effect of the haemoglobin S (HbS) variant (12). Notably, even this strong effect has been estimated to account for only roughly 2% of the total variation, indicating that a large proportion of genetic resistance factors remain to be discovered (13). Efforts in discovering these factors have culminated in the recent publication of the first genome-wide association study (GWAS) of malaria, carried out in West Africa (14). This study highlighted both the potential and the possible pitfalls of large-scale genetic association studies using single nucleotide polymorphisms (SNP) genotyping in African populations. However, its focus was on severe malaria, a complex compound phenotype consisting of severe anaemia or cerebral malaria that occurs mostly in children under the age of five, which greatly differ in its physiopathology from malaria in pregnancy. Important differences between the pathophysiology of placental malaria and malaria occurring in non-pregnant patients have been extensively documented, both in terms of the erythrocyte surface antigens sequestered in the placenta that adhere specifically to chondroitin sulphate A, and the predominantly monocytic infiltrate present in placental malaria (for review see (11)). We previously published the first extensive genetic association study of placental malaria infection, analyzing genetic variation in 64 genes, and reporting a significant association in the gene *FUT9* (15). Here, we extend this analysis considerably, by investigating genetic variation in more than 1,000 genes involved in immunity and inflammation for their involvement in placental malaria infection. To our knowledge, this is the first large-scale genetic study on this particular form of malaria to date.

Materials and Methods

Study subjects

The subjects enrolled in this study have been described in detail elsewhere (15, 16). Briefly, 360 pregnant women from Manhiça District, southern Mozambique, were chosen among the subjects of a malaria control intervention trial to form nested case-control sample of 180 cases and 180 controls, with placental malaria infection as the phenotype of interest. Placental infection was defined as the presence of asexual *Plasmodium falciparum* parasites and/or malaria pigment in placental tissue samples. The study received ethical clearance by the National Mozambican Ethics Review Committee and the Hospital Clinic of Barcelona Ethics Review Committee.

Genotyping

Genotyping was performed using the Affymetrix GeneChip Human Immune and Inflammation 9K SNP Kit, which contains approximately 9,000 SNPs located in around 1,000 genes related to the human immune and inflammatory response. Sample preparation before genotyping consisted in DNA extraction from dried blood spots on filter paper, followed by whole genome amplification (GenomiPhi kit, GE Healthcare). Genotype calling was performed using the Affymetrix GeneChip Targeted Genotyping Analysis Software (Version 1.6). Samples were included for cluster genotyping according to the following criteria: QC call rate > 75%; QC half rate < 17%; Signal/background ratio > 20. Cluster genotyping was performed using the standard settings with the exception of parameter “MinCallConfidence”, which was reduced to 0.8 in order to increase the number of SNPs with raw genotype calls (before applying quality control data filtering, as described in the results).

Statistical analysis

Genotype data management and filtering, as well as single marker association tests and SNP-SNP interaction tests were carried out using PLINK (version 1.06) (17). Five tests of association were applied for each SNP: allelic, dominant, recessive, full genotypic, and the Cochran-Armitage trend test. Both pointwise and multiple testing corrected estimates of empirical p-values were obtained by permutation as implemented in PLINK (100,000 replicates). SNP-SNP interaction tests were performed for all pairwise combinations, using the logistic regression option in PLINK.

Genotype imputation and Bayesian association mapping for candidate regions were performed using BIMBAM (version 0.99) (18). Due to the lack of an appropriate Southeast African reference panel, we used the combined panel of all three HapMap populations as a reference for imputation to minimize error rates (19). All SNPs within 20 kb of the candidate regions were considered for imputation. Bayes factors were transformed to p-values using permutation (1,000,000 replicates) (20).

Multi-marker association analysis was performed using BEAGLE (version 3.0.2) (21, 22), using the allelic, dominant and recessive tests. In order to account for the expected greater haplotypic diversity and therefore increased number of distinct haplotype clusters in African samples, we reduced the 'scale' parameter to 2. Empirical p-values were obtained using permutation (10,000 replicates). Phase estimation prior to analysis was carried out using fastPHASE (version 1.4.0) (23) with default settings.

Protein interaction network visualization and analysis was performed using Cytoscape (version 2.6.3) (24), as well as R (version 2.9.0) and Bioconductor (version

2.4) (25). Protein interaction data was obtained using the Michigan Molecular Interactions (MiMI) plugin for Cytoscape (version 3.0.1) (26). Network motif search was carried out using the tYNA web tool (27). SNP-SNP interaction association statistics were mapped to a particular protein-protein interaction if both SNPs were found within 5 kb of the respective genes.

Gene-wise and interaction-wise p-values were obtained by combining the p-values of all SNPs within 5 kb of a gene, or of all SNP-SNP interactions mapped to a particular protein-protein interaction, respectively. P-values were combined using the Simes procedure:

$$p_{combined} = \min(\frac{p_i \cdot k}{i})$$

where p_i is the i^{th} ordered p-value of all k SNPs mapped to a particular gene or interaction. This was chosen instead of a more simple minimum p-value method, in order to avoid bias towards lower combined p-values in genes/interactions with larger number of SNPs.

Population structure was analyzed using EIGENSOFT (version 2.0) (28), with default settings and correcting for linkage disequilibrium (LD). Population differentiation (F_{ST}) and selection statistics (iHS) of the Illumina 650k Human Genome Diversity Panel (HGDP) data (29) as well as HapMap data (30) were calculated as previously described (15). Data visualization and additional statistical analysis was performed using R.

Results

Genotyping quality and population structure

Out of a total of 347 samples for which genotyping was attempted, 68 did not pass the quality control thresholds for cluster genotyping (see Methods) and were removed. The remaining 279 samples were assayed for a total of 9,178 SNPs. Out of those, we removed all SNPs with ambiguous mapping in the latest SNP annotation (56; Affymetrix annotation release 1.5), with a call rate below 0.8 (439), all that failed testing for Hardy-Weinberg equilibrium in controls (115; $p \leq 10^{-4}$), as well as all monomorphic SNPs (1,554). This resulted in a final analysis sample of 279 individuals (173 cases, 174 controls) and 7,442 SNPs, with an average call rate of 0.97. Testing for population stratification using EIGENSOFT revealed no significant differentiation between cases and controls (p = 0.16), therefore this final sample was used without correcting for stratification in all following analyses.

Association analysis

As a first step we performed single marker analysis for all SNPs, testing for five models of penetrance (see Methods). Figure 1 shows a quantile-quantile (QQ) plot of the results, confirming the absence of a bias in the distribution of p-values due to population structure, as expected from the results of the EIGENSOFT analysis. The genomic distribution of the results of the single marker association tests is shown in Figure 2. The strongest signal was seen in three neighbouring SNPs in the region of the gene *KLRK1* on chromosome 12 (see Table 1), with the best SNP showing $p = 5 \times 10^{-5}$ (rs12821887). However, none of the SNPs reached genome-wide significance after correcting for multiple testing. In the absence of a clear association signal, we chose to further investigate all regions showing multiple SNPs among the top signals.

The *KLRK1* region contains three of the top four among all tested SNPs within a region of 8 kb, and therefore merits a more thorough investigation. Another region of potential interest was found in the gene *IL7* on chromosome 8, where two neighbouring SNPs 33 kb apart were also found among the top genome wide 10 association signals.

In order to get a more fine scale resolution of the results in both of those regions, we used BIMBAM to impute genotypes at untyped SNPs and test them for association, using the combined HapMap 2 as a reference panel. We chose to use this reference panel following the suggestions of Guan and Stephens (19), who found that accuracy was improved for situations were the population analyzed was not well represented by any single reference panel. Nevertheless, results from this analysis have to be interpreted with care, as even the African population in the reference panel (Yorubans from Nigeria) shows considerable geographic as well as genetic distance from the study population from southern Mozambique (Sikora *et al.*, in preparation). Figure 3 shows the result of the analysis. As expected by the much higher density of the reference panel, both regions show imputed SNPs with stronger association signals than any of the genotyped SNPs. Around rs2583764, the top signal of the genotyped SNPs in the *IL7* region, a number of imputed SNPs with lower p-values are found. The top SNP in this region (rs2583763, $p = 9 \times 10^{-6}$), 392 bp upstream of rs2583764, is also the top hit among all imputed SNPs. In the *KLRK1* region, a number of imputed SNPs within the gene region also show stronger evidence (top SNP rs7962112, $p = 5 \times 10^{-5}$).

Finally, in order to investigate whether power to detect association in our sample could be improved by using combinations of multiple markers, we carried out haplotype association tests using BEAGLE. The strongest signal in this analysis was found in the gene *IL7*, where a cluster including rs2583764 was close to the significance level after a strict multiple test correction (minimum p = 3.2 x 10^{-6}; permutation p = 0.07). Dissecting this cluster association result further, we found that a haplotype of four SNPs (ATGA; rs1441438 - rs1036751 - rs6993386 - rs2583764) explained the observed signal. A standard association test for the ATGA haplotype against all other observed haplotypes at the locus indicated a susceptibility effect (OR_{ATGA} = 3.1; 95% CI = 1.9 - 5.0; p = 1.6 x 10^{-6}). Not surprisingly, this 22 kb haplotype also spans the SNP rs2583763 found in the imputation analysis.

Gene-gene interactions and network analysis

Given that our study focuses on genes related to specific organismal functions, namely inflammation and immunity, we next set out the analysis to include information on the interactions of those genes and their organization in cellular pathways in our analysis. We therefore performed SNP x SNP interaction association tests in our sample. In order to reduce the number of tests, we only analyzed those interacting SNP pairs that were found within 5 kb of genes that had evidence for interaction of their respective protein products. To that end, curated protein interactions were obtained from the Michigan Molecular Interactions database (see Methods). Querying the database with the genes included in the study resulted in a network consisting of a total of 892 genes (nodes), connected by 3789 interactions (edges), from now on referred to as "immunity network". We then tested all resulting SNP pairs for interaction effects using logistic regression. The strongest signal comes from a SNP pair located in the genes *IL7R* (rs1494558) and *JAK3* (rs6512227). However, this was again not significant after correcting for multiple testing (nominal

$p = 5.4 \times 10^{-5}$), but given the still considerable number of tests in this pair wise analysis, this is not unexpected.

Even so, two interesting properties of this SNP pair called our attention: firstly, one of the two SNPs, rs1494558, is a non-synonymous variant responsible for a non-polar to polar amino acid substitution (Ile -> Thr; Grantham distance D = 89 (32)); and second, *IL7R* is also interacting with *IL7*, one of the two regions found with the strongest signal in the single marker analysis. We therefore investigated the joint distribution of association results for both genes and interactions on the immunity network in more detail. As a first step, we assigned a single p-value to each gene and interaction, by combining all SNP p-values mapped to a particular gene or interaction using the Simes method (see Material and Methods). We chose this method over a more simple minimum p-value method in order to avoid systematic biases due to differing number of SNPs in the respective genes/interactions. Having obtained the gene- and interaction-wise p-values, we wanted to see how unusual it was to find a low interaction p-value at one edge distance to a low gene p-value in the immunity network. To this end, we employed a network motif search algorithm to first identify all distinct network motifs composed of three nodes connected by two edges (i.e. chain motifs of length three) in the network. After having obtained the list of motifs (116,728 total), we plotted the distribution of the minimum p-values of the genes versus the minimum p-values of the interaction, for each of the motifs. Results are shown in Figure 4. As can be seen in the distribution, the motif containing both *IL7* and the neighbouring interaction *IL7R – JAK3* is a clear outlier in the empirical distribution. Furthermore, the only other point that also behaves as an outlier for both genes and interactions is an overlapping motif to the former one, including the same *IL7R* - *JAK3* interaction together with *JAK2*, a gene downstream of *JAK3.* A more detailed look at the subnetwork containing all first neighbours of the *IL7R* - *JAK3* interaction is shown in Figure 5. As can be seen, the strongest signals in both genes and interactions are clustered in the module *IL7-IL7R-JAK3-JAK2-CNTFR*, which is part of the *IL7* signalling cascade. Taken together, these results could indicate a role for this signal-transduction module in the susceptibility to placental malaria infection.

Discussion

The present study is the first large-scale survey on human genetic variation in the immune system and inflammatory response and its relationship to malaria in pregnancy. It is clear that both of these cellular processes are important components in the response to malaria infection, making them strong candidates for harbouring loci that influence susceptibility. Nonetheless, our results show no single region standing out with a clear signal of association, if we consider the stringent statistical thresholds normally employed for GWAS. Explanations for the lack of a clear candidate can be manifold, from a genuine lack of susceptibility loci in the analyzed genes to a lack of power to detect loci with weak effects due to small sample size. It is however important to note that our results mirror the observations of Jallow *et al.* (14) in their malaria GWAS in West Africa, namely the difficulty of achieving the thresholds normally applied in studies with samples of European descent in Africa. In their study, even in the HBB region, with its known strong protective effect due to haemoglobin S, the strongest signal only barely passes the normally applied threshold of $p = 5 \times 10^{-7}$ ($p = 3.9 \times 10^{-7}$) (33). The reason is that due to generally lower levels of linkage disequilibrium (LD) in Africa and genotyping arrays being designed using tagSNPs derived from European populations, effective coverage of even common variants in African populations is low. The targeted genotyping array used in our study will evidently suffer from the same limitations. We therefore followed their example and considered loci with $p < 10^{-4}$ as interesting regions. The only region that achieves that threshold is *KLRK1*. In addition to that, *IL7* emerged as another potential candidate locus, based on the results of both imputation and haplotype association.

A more general concern with association studies is the focus on single genes, without taking into account the interactions among them and their organization into functional pathways. Recent effort has therefore been aimed at incorporating this knowledge into the analysis, both in the development of methods for detecting gene-gene interactions (reviewed in (31)) as well as in the analysis of genetic association studies in the context of biological networks (34-36). In this study, we took an empirical approach to integrate the results of both single gene and gene – gene interaction tests in the context of the known interactions of the immunity network. Results of this analysis give additional support to a role of IL-7 signalling in modulating susceptibility to placental infection. Integration of interaction data can therefore overcome some of the problems mentioned above, and can aid in prioritizing candidates in the absence of clear association signals. Nonetheless, there is still a general lack of powerful statistical tools to disentangle the effects of multiple interacting variants at different loci on a phenotype of interest, which, when becoming available, will certainly be of great impact in the mapping of genotypes to phenotypes.

Based on these results we therefore suggest a role of IL-7/IL-7R signalling in susceptibility of placental malaria infection. The module identified forms part of the JAK-STAT signalling pathway, which regulates cellular responses mediated by binding of cytokines like IL-7. Some of the responses mediated upon binding of cytokines include cell proliferation and differentiation, making it a key pathway in processes like haematopoiesis and immune development (37). IL-7 in particular is an important factor in B and T cell development. Looking in more detail at the interaction results, it was intriguing that the top result involved an interaction with a non-synonymous SNP in *IL7R*. The variant (rs1494558) causes a change from

Threonin to Isoleucin in the extracellular domain of the receptor. It is, together with other variants, implicated in autosomal recessive severe combined immunodeficiency, although impairment of IL-7 signalling was not observed (38). The observed effect is a dominant interaction with rs6512227 (Figure 6), a SNP located upstream of *JAK3,* indicating a potential regulatory effect. The gene codes for the tyrosine kinase JAK3, an intracellular adaptor protein that is involved in the transduction of signal induced by cytokine receptor binding.

The involvement of the IL-7/IL-17R pathway in placental malaria may be relatively unexpected, but not unlogical. Although the most studied activities of IL-7 are those related to B- and T-cell proliferation, IL-7 also exerts important proinflammatory effects. IL-7 has been shown to induce production of TNF by T- and B-cells (Roato et al, Ann N Y Acad Sci 2007; 1117: 377-84), an important proinflammatory cytokine with deleterious effects in placental malaria. Moreover, IL-7 has been described to drive inflammation in several prototypic inflammatory conditions such as reumatoid arthritis (39) or atherosclerosis (40). The inflammatory properties of IL-7/IL-7R pathway could influence susceptibility to placental malaria infection by acting at two separate levels: by modulating direct antimalarial immunity and resistance to infection, and by modulating the inflammatory reaction in the placenta during infection, with subsequent consequences for the outcome of the pregnancy.

In conclusion, our results point towards a possible role for IL-7 signalling through IL7R and the JAK/STAT intracellular adaptors in placental malaria infection. Our study is the first large-scale attempt to determine the genetic basis of placental infection in malaria, and suggests an important unexpected role of the IL-7/IL-7R pathway for the susceptibility of this important clinical condition.

Acknowledgements

We are grateful to the women who participated in this study without whom it would not have been possible. We thank the staff of the Manhiça District Hospital and the Manhiça Health Research Center, for their collaboration. We also thank Elisa Serra for her help and Sergi Sanz and John Aponte, for helping with the databases.

Funding

This work was supported by grant SAF-2007-63171 awarded by Ministerio de Educación y Ciencia (Spain) to JB and the CIBERESP (CIBER of Epidemiology and Public Health, Ministry of Health). Funds were also obtained from Direcció General de Recerca of Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608) and from Banco de Bilbao, Vizcaya, Argentaria Foundation (grant number BBVA 02-0). The CISM receives core support from the Spanish Agency for International Cooperation and Development (AECID).
All the genotyping was provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; www.cegen.org); Support on computations was provided by the National Institute for Bioinformatics (www.inab.org); both are platforms of Genoma España. MS is supported by a PhD fellowship from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982). AM was supported by the Spanish Ministry of Health (Program for the Promotion of Biomedical Research and Health Sciences, Instituto de Salud Carlos III, CP-04/00220).

References

1 Alonso, P.L., Sacarlal, J., Aponte, J.J., Leach, A., Macete, E., Milman, J., Mandomando, I., Spiessens, B., Guinovart, C., Espasa, M. *et al.* (2004) Efficacy of the RTS,S/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial. *Lancet*, **364**, 1411-1420.

2 Sacarlal, J., Aide, P., Aponte, J.J., Renom, M., Leach, A., Mandomando, I., Lievens, M., Bassat, Q., Lafuente, S., Macete, E. *et al.* (2009) Long-term safety and efficacy of the RTS,S/AS02A malaria vaccine in Mozambican children. *J. Infect. Dis.*, **200**, 329-336.

3 Roberts, L. and Enserink, M. (2007) Malaria. Did they really say ... eradication? *Science*, **318**, 1544-1545.

4 World Health Organization (2009) World Malaria Report 2009.

5 Steketee, R.W., Nahlen, B.L., Parise, M.E. and Menendez, C. (2001) The burden of malaria in pregnancy in malaria-endemic areas. *Am. J. Trop. Med. Hyg.*, **64**, 28-35.

6 Desai, M., ter Kuile, F.O., Nosten, F., McGready, R., Asamoa, K., Brabin, B. and Newman, R.D. (2007) Epidemiology and burden of malaria in pregnancy. *Lancet Infect. Dis.*, **7**, 93-104.

7 McGregor, I.A., Wilson, M.E. and Billewicz, W.Z. (1983) Malaria infection of the placenta in The Gambia, West Africa; its incidence and relationship to stillbirth, birthweight and placental weight. *Trans. R. Soc. Trop. Med. Hyg.*, **77**, 232-244.

8 Menendez, C. (1996) An immunological hypothesis to explain the enhanced susceptibility to malaria during pregnancy: Reply. *Parasitol Today*, **12**, 41-42.

9 Menendez, C., Ordi, J., Ismail, M.R., Ventura, P.J., Aponte, J.J., Kahigwa, E., Font, F. and Alonso, P.L. (2000) The impact of placental malaria on gestational age and birth weight. *J. Infect. Dis.*, **181**, 1740-1745.

10 Mutabingwa, T.K., Bolla, M.C., Li, J.L., Domingo, G.J., Li, X., Fried, M. and Duffy, P.E. (2005) Maternal malaria and gravidity interact to modify infant susceptibility to malaria. *PLoS Med.*, **2**, e407.

11 Rogerson, S.J., Hviid, L., Duffy, P.E., Leke, R.F. and Taylor, D.W. (2007) Malaria in pregnancy: pathogenesis and immunity. *Lancet Infect. Dis.*, **7**, 105-117.

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

13 Mackinnon, M.J., Mwangi, T.W., Snow, R.W., Marsh, K. and Williams, T.N. (2005) Heritability of malaria in Africa. *PLoS Med.*, **2**, e340.

14 Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M. *et al.* (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.*

15 Sikora, M., Ferrer-Admetlla, A., Laayouni, H., Menendez, C., Mayor, A., Bardaji, A., Sigauque, B., Mandomando, I., Alonso, P.L., Bertranpetit, J. *et al.* (2009) A variant in the gene FUT9 is associated with susceptibility to placental malaria infection. *Hum. Mol. Genet.*, **18**, 3136-3144.

16 Menendez, C., Bardaji, A., Sigauque, B., Romagosa, C., Sanz, S., Serra-Casas, E., Macete, E., Berenguera, A., David, C., Dobano, C. *et al.* (2008) A randomized placebo-controlled trial of intermittent preventive treatment in pregnant women in the context of insecticide treated nets delivered through the antenatal clinic. *PLoS ONE*, **3**, e1934.

17 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559-575.

18 Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, **3**, e114.

19 Guan, Y. and Stephens, M. (2008) Practical issues in imputation-based association mapping. *PLoS Genet.*, **4**, e1000279.

20 Stephens, M. and Balding, D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681-690.

21 Browning, B.L. and Browning, S.R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.*, **31**, 365-375.

22 Browning, S.R. (2006) Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.*, **78**, 903-913.

23 Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629-644.

24 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498-2504.

25 Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.

26 Gao, J., Ade, A.S., Tarcea, V.G., Weymouth, T.E., Mirel, B.R., Jagadish, H.V. and States, D.J. (2009) Integrating and annotating the interactome using the MiMI plugin for cytoscape. *Bioinformatics*, **25**, 137-138.

27 Yip, K.Y., Yu, H., Kim, P.M., Schultz, M. and Gerstein, M. (2006) The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, **22**, 2968-2970.

28 Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

29 Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100-1104.

30 The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-861.

31 Cordell, H.J. (2009) Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*

32 Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862-864.

33 The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661-678.

34 Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., Polman, C.H. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078-2090.

35 Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J.D., Jin, L. *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.*, **18**, 111-117.

36 Emily, M., Mailund, T., Hein, J., Schauser, L. and Schierup, M.H. (2009) Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.*, **17**, 1231-1240.

37 Rawlings, J.S., Rosler, K.M. and Harrison, D.A. (2004) The JAK/STAT signaling pathway. *J. Cell Sci.*, **117**, 1281-1283.

38 Puel, A., Ziegler, S.F., Buckley, R.H. and Leonard, W.J. (1998) Defective IL7R expression in T(-)B(+)NK(+) severe combined immunodeficiency. *Nat. Genet.*, **20**, 394-397.

39 van Amelsfort, J.M., van Roon, J.A., Noordegraaf, M., Jacobs, K.M., Bijlsma, J.W., Lafeber, F.P. and Taams, L.S. (2007) Proinflammatory mediator-induced reversal of CD4+,CD25+ regulatory T cell-mediated suppression in rheumatoid arthritis. *Arthritis Rheum.*, **56**, 732-742.

40 Damas, J.K., Waehre, T., Yndestad, A., Otterdal, K., Hognestad, A., Solum, N.O., Gullestad, L., Froland, S.S. and Aukrust, P. (2003) Interleukin-7-mediated inflammation in unstable angina: possible role of chemokines and platelets. *Circulation*, **107**, 2670-2676.

Figures

Figure 1. QQ plot of single SNP association statistics

The negative logarithm of the ordered empirical p-values obtained from the single marker association analysis is plotted against the negative logarithm of the ordered p-values from a uniform distribution, as expected under the null hypothesis. The grey shaded area indicates the 95% concentration band. No systematic inflation was observed in the test statistics, as noted by the majority of points falling on the diagonal (red line, y=x).

Figure 2. Genome-wide distribution of single SNP association statistics

Manhattan plot showing the distribution of the single SNP association statistics. Coloured circles indicate the negative logarithm of the empirical p-value for all QC+ SNPs, with colours differentiating the different chromosomes. The red dashed line indicates a significance level of 10^{-3}. The two candidate regions *IL7* and *KLRK1* are also indicated.

Figure 3. Imputation association results at *IL7* and *KLRK1*

Results of the association analysis using imputation in the two candidate regions *IL7* (A) and *KLRK1* (B). In both regions, the negative logarithm of the empirical p-values is plotted (see Methods for details), with filled circles indicating the SNPs genotyped in the respective regions, and empty circles indicating imputed SNPs using the full HapMap 2 panel as a reference. Chromosomal positions of the genes in the region are indicated below the respective panel.

Figure 4. Distribution of gene and interaction p-values for chain motifs

Plot showing the distribution of minimum gene p-value versus minimum interaction p-value for each chain motif of length three, on a negative logarithmic scale. For clarity, only interactions with $p <= 10^{-2}$ are shown. In order to deal with the considerable amount of overplotting due to the large number of data points with similar values, alpha transparency is used for colour, resulting in darker colours in regions with many overlapping points. Histograms show the marginal distributions of the minimum gene / interactions p-values, respectively. The structure of the two outlier motifs containing the interaction *IL7R – JAK3* are also depicted. The gene / interaction corresponding to the values on the plot are indicated in red.

Figure 5. Sub-network of first neighbours of *IL7R – JAK3* interaction.

A sub-network of the full immune network, containing all nodes separated by one edge from the *IL7R – JAK3* interaction (52), and all their interactions (425). Nodes in red indicate genes with $p < 10^{-2}$. Edges in red indicate interactions with $p < 10^{-3}$, with thicker lines corresponding to lower p-values. The *IL7* module is shown on the right part of the plot.

Figure 6. Interaction effect of rs1494558 - rs6512227 interaction

Log odds of disease for all allelic combinations of the two SNPs, estimated by logistic regression.

Tables

Table 1. Top 10 single SNP association signals

SNP	Chr[a]	Position	Alleles[b]	Best test[c]	Affected	Unaffected	df[d]	p_{test}[e]	$p_{nominal;perm}$[f]	$p_{corrected;perm}$[g]	Gene	Distance[h]
rs12821887	12	10,445,946	C/T	Dominant	47/92	82/57	1	2.56E-05	5.00E-05	0.21	*KLRK1*	0
rs728010	12	10,444,534	A/G	Dominant	47/91	80/58	1	6.74E-05	7.00E-05	0.47	*KLRK1*	0
rs2850760	18	59,117,803	C/T	Allelic	35/243	11/269	1	2.00E-04	2.50E-04	0.85	*BCL2*	0
rs7972757	12	10,437,407	G/A	Allelic	36/220	75/191	1	7.98E-05	2.70E-04	0.53	*KLRK1*	0
rs2583764	8	79,826,355	A/G	Allelic	128/144	85/189	1	1.22E-04	2.80E-04	0.69	*IL7*	0
rs1571344	1	205,737,551	G/A	Recessive	16/121	41/94	1	1.53E-04	4.30E-04	0.76	*CR1*	0
rs3917422*	1	167,965,384	G/T	Trend	15/261	2/278	1	1.02E-03	5.10E-04	1.00	*SELE*	0
rs2583762	8	79,860,038	A/T	Trend	128/144	90/190	1	3.03E-04	6.70E-04	0.96	*IL7*	0
rs3824433	9	5,103,577	C/T	Dominant	59/71	84/39	1	2.39E-04	7.00E-04	0.89	*JAK2*	0
rs7486905	12	105,684,299	G/A	Recessive	12/104	35/86	1	3.36E-04	7.00E-04	0.96	*RFX4*	3588

[a] Chromosome
[b] minor/major allele (positive strand)
[c] test model with lowest asymptotic p-value
[d] degrees of freedom
[e] asymptotic p-value from best test model
[f] empirical p-value for respective SNP, overall for all tested models; 100,000 permutations
[g] empirical p-value for respective SNP, corrected for all 7442 tested SNPs
[h] distance from gene (bp); 0 indicates within gene region
[*] rs3917422 is a non-synonymous coding SNP (Pro / Gln)

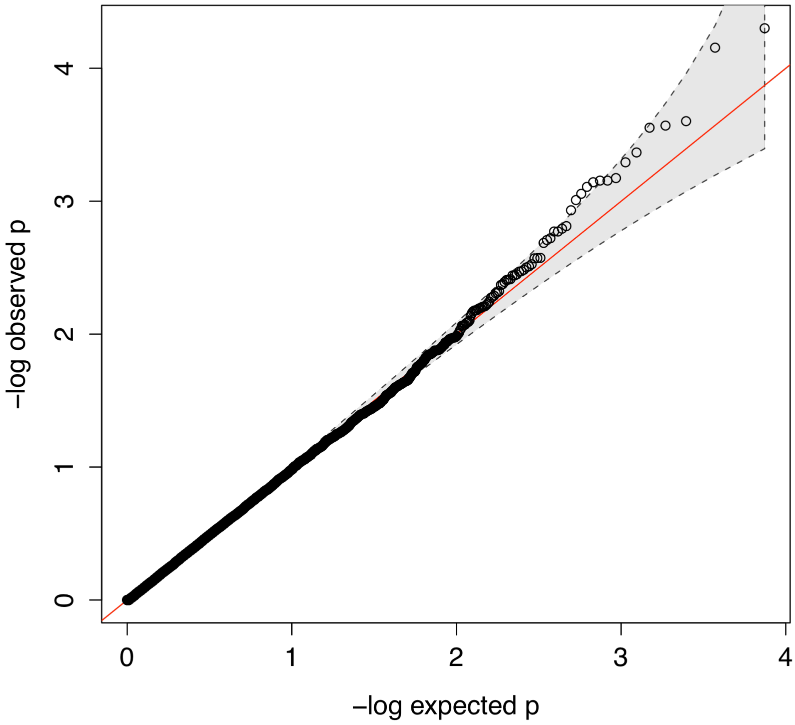


Figure 1

Figure 2

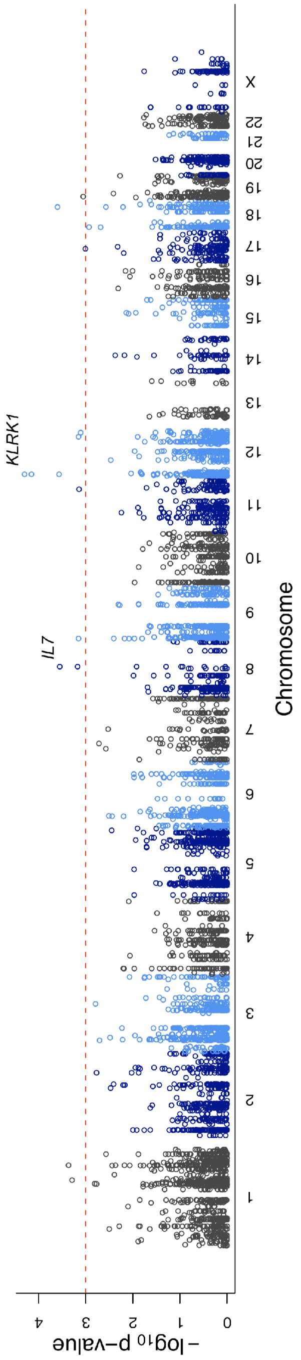
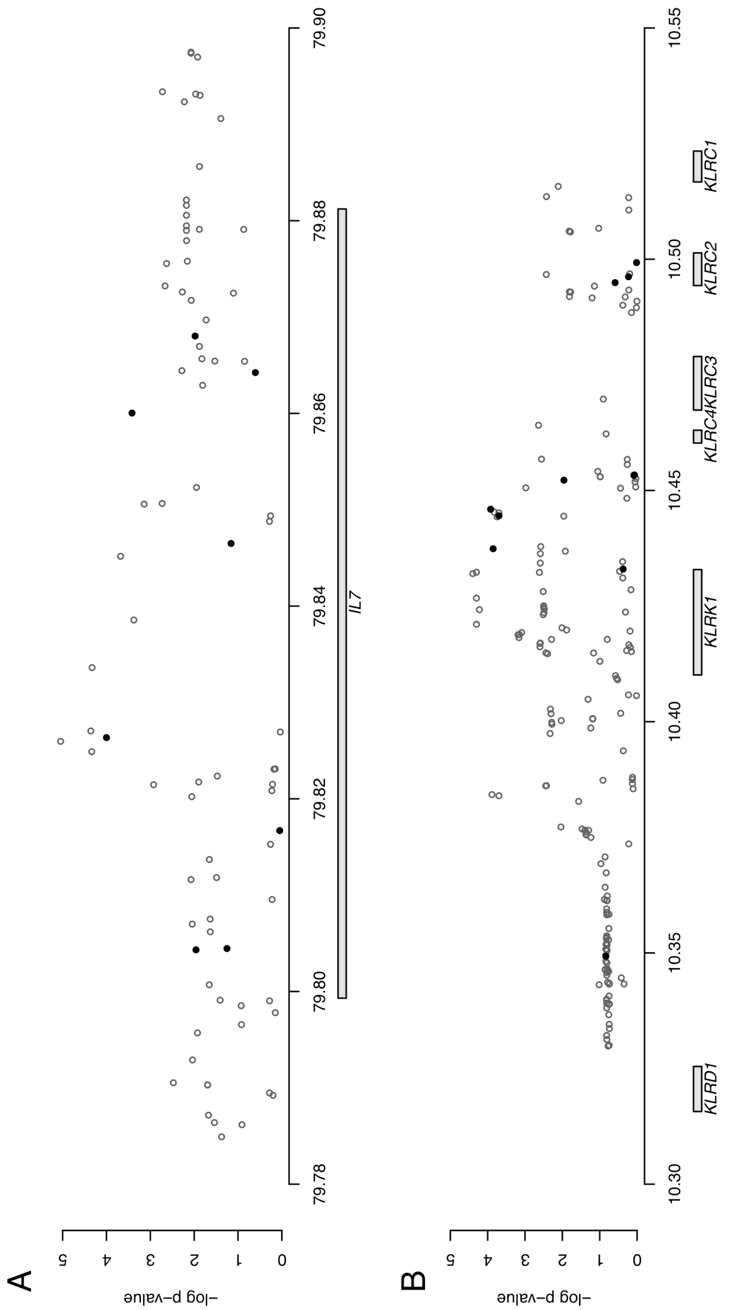

KLRK1
IL7
−log10 p−value
0
1
2
3
4
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X

Figure 3


A
−log p−value
0
1
2
3
4
5
79.78
79.80
79.82
79.84
79.86
79.88
79.90
IL7
B
−log p−value
0
1
2
3
4
5
10.30
10.35
10.40
10.45
10.50
10.55
KLRD1
KLRK1
KLRC4
KLRC3
KLRC2
KLRC1

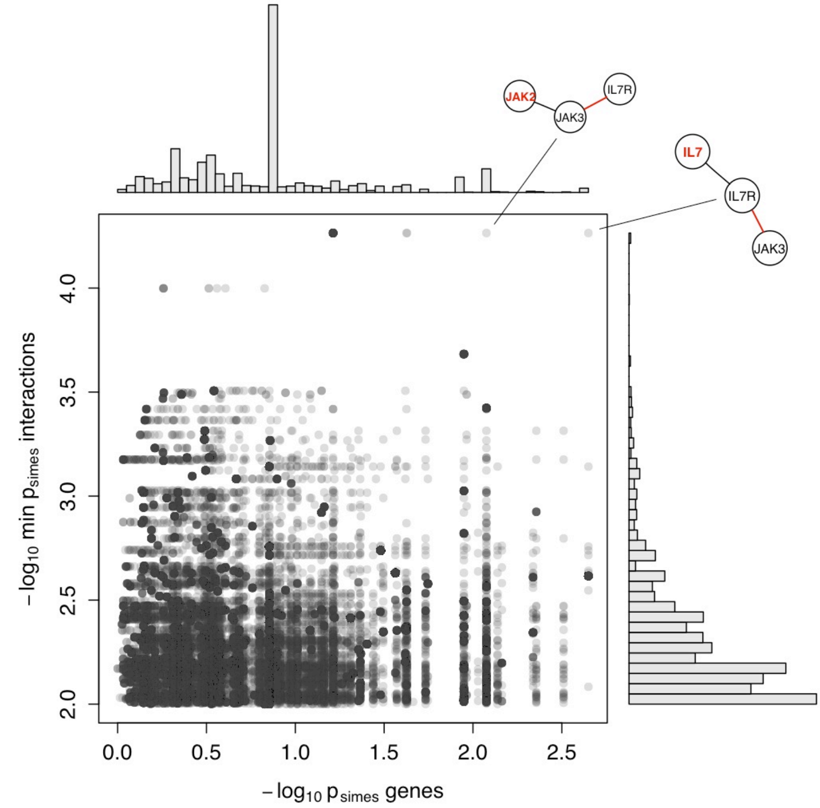

JAK2
JAK3
IL7R
IL7
IL7R
JAK3
4.0
3.5
3.0
2.5
2.0
$-\log_{10}$ min p_{simes} interactions
0.0
0.5
1.0
1.5
2.0
2.5
$-\log_{10} p_{simes}$ genes

Figure 4

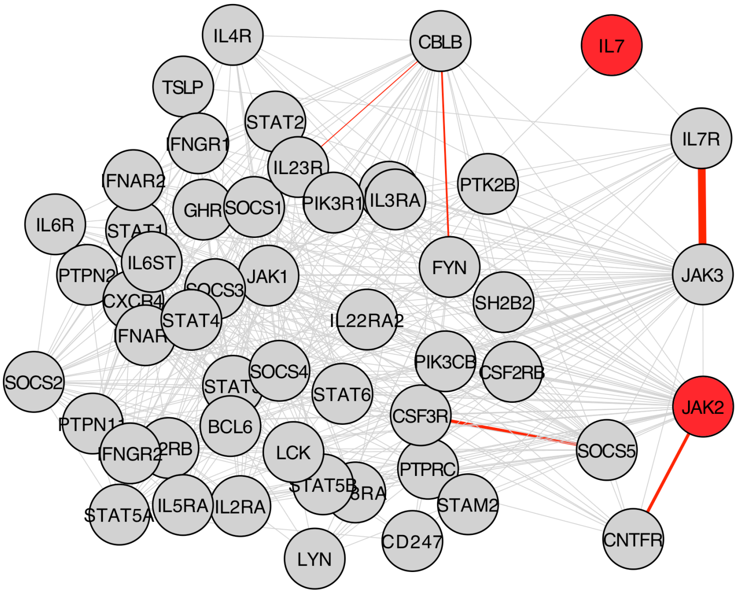

IL4R
CBLB
IL7
TSLP
STAT2
IFNGR1
IL7R
IL23R
IFNAR2
PTK2B
GHR
SOCS1
PIK3R1
IL3RA
IL6R
STAT1
IL6ST
FYN
PTPN2
JAK1
JAK3
SOCS3
CXCR4
SH2B2
STAT4
IFNAR
IL22RA2
PIK3CB
CSF2RB
SOCS2
SOCS4
STAT
STAT6
JAK2
PTPN11
BCL6
CSF3R
2RB
SOCS5
IFNGR2
LCK
PTPRC
STAT5B
3RA
STAM2
STAT5A
IL5RA
IL2RA
CNTFR
CD247
LYN

Figure 5

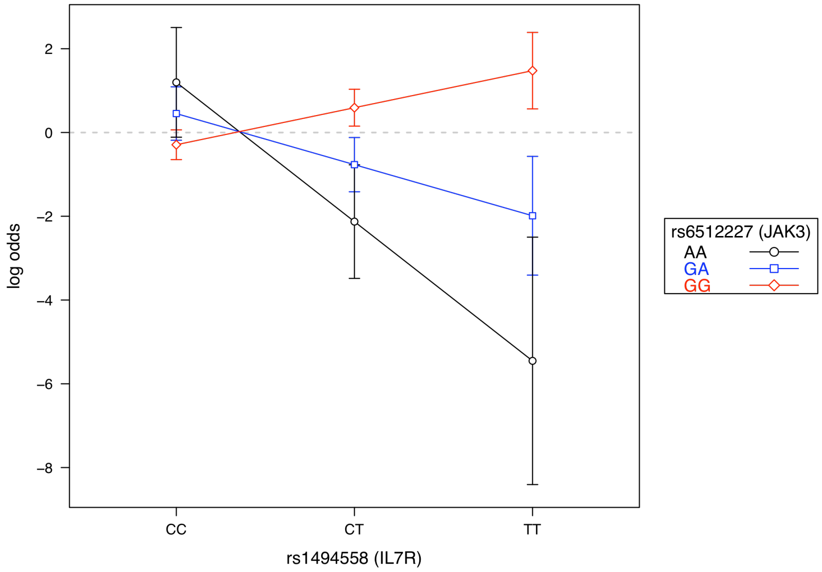

2
0
−2
−4
−6
−8
log odds
CC
CT
TT
rs1494558 (IL7R)
rs6512227 (JAK3)
AA
GA
GG

Figure 6

3.4 A genomic analysis identifies a novel component in the genetic structure of Sub-Saharan African populations

Martin Sikora, Hafid Laayouni, Francesc Calafell, David Comas and Jaume Bertranpetit

Manuscript submitted to European Journal of Human Genetics

A genomic analysis identifies a novel component in the genetic structure of Sub-Saharan African populations

Martin Sikora, Hafid Laayouni, Francesc Calafell, David Comas and Jaume Bertranpetit

Institut de Biologia Evolutiva (UPF-CSIC), CEXS-UPF-PRBB, Doctor Aiguader 88,
08003 Barcelona, Catalonia, Spain

Author for correspondence: jaume.bertranpetit@upf.edu

ABSTRACT

Studies of large sets of SNP are proven a powerful tool for population genetics. We show that SNP number above 1,000 give robust results in a word-wide perspective. A SNP analysis in 12 widespread Sub Saharan populations shows new and interesting features in its genetic complexity. There is a strong differentiation of Nilo-Saharans, much beyond what would be expected by geography. Also, a highly differentiated population is found in the southeast (Mozambique) where we detect a pre-Bantu substrate that was assimilated by Bantu speakers and completely erased, without any living population nowadays with this substrate. Populations of the West Africa present an unexpected similarity among them, result of a population expansion. Hunter-gatherer populations (Khoisan and Pygmies) show a clear distinctiveness with very intrinsic Pygmy (and not only Khoisan) genetic features.

INTRODUCTION

Human population genetics has recently completed a circle. It started with the so-called *classical* polymorphisms (i.e., blood groups and other protein polymorphisms) that were analyzed on the basis of their allele frequencies with potent statistical instruments such as principal component analysis (PCA). This era culminated with the landmark publication of the magnificent book by Cavalli-Sforza *et al.*[1]; however, a few shortcomings of classical polymorphisms can be pointed out: they are relatively few loci; their connection to actual DNA variation was mostly unknown; and they could have been subjected to confounding natural selection. PCR and automated sequencing heralded the uniparental marker era: mtDNA and the non-recombining region of the Y chromosome (NRY) could be routinely analyzed and a firm phylogeography could be established for both genomic regions, allowing the dissection of population structure with unprecedented precision and reliability. But yet, they behave as just two loci, in which natural selection cannot be ruled out and with the peculiarities associated to their sex-specific transmission. Another technological development spearheaded a new breakthrough in human population genetics: Single nucleotide polymorphism (SNP) array genotyping platforms have made it affordable to genotype hundreds of thousands of markers. The results are again treated in terms of allele frequencies and subjected to PCA and to newer techniques such as Bayesian classification algorithms. Now, the whole genome is covered, and the action of selection is masked by a vast majority of putatively neutral markers.

The genetics of African populations, of paramount interest given the recent African origin of humankind, has been through the full cycle of studies. Cavalli-Sforza *et al.*[1] identified a north-south gradient in the continent that could be attributed to the

Bantu expansion, while other principal components had a less clear interpretation. African mtDNA phylogeography was firmly established by Salas et al.[2], who described the structure of maternal lineages in the continent and identified some haplogroups involved in human expansions such as the Bantu expansion. Recently, deep analyses of maternal lineages in African hunter-gatherers (Khoisan speakers and Pygmies) have revealed a clearly structured phylogeny for the mtDNA[3-5]. A number of papers have approached both the general and the more local aspects of non-recombining part of the Y-chromosome in Africa[6; 7], although its phylogeographic structure has not been as refined as its maternal counterpart. Recently, Tishkoff *et al.*[8] in a landmark paper analyzed 1327 nuclear microsatellite markers in 121 African populations, and identified a number of layers in the African population structure that could be related to history, language, and geography. The continent, south of the Sahara, seems to be dominated by a component mostly correlated with Niger-Congo speaking populations, while other components are found in the Sahel, among Nilo-Saharan speakers, and in Afro-Asiatic speakers in the north and northeast. Among the hunter-gatherer populations, the Khoisan-speaking Hadza of Tanzania were clearly distinct, while Pygmies could not be discriminated from the Southern Africa Khoisan. At higher discrimination level, western Pygmies became distinct, but the eastern Pygmies remained similar to the Khoisan. In a more recent publication, Bryc *et al.* analyzed SNP data obtained from West Africans (and African Americans), which revealed a structure reflecting primarily language and secondarily geographical distances[9]. Unfortunately, this work is restricted mainly to populations in Central West Africa, around the Gulf of Guinea.

DATA

Results obtained in human population genetics are contingent on sampling. Frequently, apparent barriers or discontinuities are just a reflection of heterogeneous spatial sampling. Both in Tishkoff *et al*. and Bryc *et al*., as in previous works, the area between Central and South Africa remains under sampled. In particular, southeast Africa is a key geographical zone to understand the Bantu expansion routes. In a case-control study for placental malaria, we obtained, with appropriate informed consent, 180 cases and 180 controls from Mozambique, in southeast Africa. These samples were genotyped with the Affymetrix GeneChip® Human Immune and Inflammation 9K SNP Kit, resulting in a total of 279 samples with reliable data after stringent quality control. Other African samples with genowide surveys available are Biaka Pygmies, Mbuti Pygmies, Mandenka, Yoruba, San, and Bantu-speakers from the HGDP panel[10], and the Maasai, Luuya, Yoruba, and African-Americans from Hapmap Phase 3 (http://hapmap.ncbi.nlm.nih.gov/). The intersection of all arrays provides a common set of 2,841 SNPs with genotype data for all populations (see Supplementary Information for details).

RESULTS

First, we wanted to test if the number of SNPs available (2,841 SNPs) provides enough genetic resolution to detect any structure in African populations and provide a reference for the number of SNPs needed in population studies. To that effect, we combined the global HGDP and Hapmap Phase 3 genotype data (~460,000 SNPs) and subjected them to PCA (see Supplementary Information for details). Results are similar to those obtained with the HGDP samples, with the first and second PC (Figure 1a) separating East Asia (upper left corner of the plot) from Europe (bottom centre) and sub-Saharan Africa (upper right)[10; 11]. The same structure is recovered when random subsamples of 100,000 (Figure 1b), 10,000 (Fig. 1c), and 1,000 (Figure 1d) SNPs are considered, although inter-individual variation increases. A random set of 2,841 SNPs from this pooled HGDP-HapMap dataset (Figure 1e) performs similarly to the set of 2,841 SNPs related to immunity and inflammation (Figure 1f), despite of the slightly reduced inter-population differentiation of the latter, which is expected as they are gene-based SNPs[12]. We can conclude that the common set of 2,841 SNPs genotyped is an appropriate tool to study population structure in African populations; in general, worldwide patterns are evident and robust when using a minimum of 1,000 SNPs.

Next, we applied PCA[13] and STRUCTURE[14; 15] to 775 individuals in 11 populations of sub-Saharan African descent. The first PC (Figure 2a) and STRUCTURE with K=2 (Figure 3) separate the Nilo-Saharan-speaking Maasai from all other populations, with neighboring Luuya and African Americans in an intermediate position. Both the second PC and K=3 separate the hunter-gatherer samples, presumably ancestral Pygmy and San populations from the rest. The third PC allows us to discriminate between Western/Central (Mandenka, Yoruba), Eastern (Maasai, Luuya), and Southeastern populations (Mozambique), irrespectively of language family. This is the PC that is mostly correlated with geography (Figure 4), and the fact that it is the third rather than the first component, as would be expected if isolation by distance was the predominant force shaping genetic diversity[16], implies that directional population movements (such as the Bantu expansion) and barriers to gene flow (such as that between food producers and hunter-gatherers) are more relevant than geographic distance to understand the genetic landscape of sub-Saharan Africa. The distinction between West and Southeast Africa is also shown with at K=4; at K=5, the Niger-

Congo speaking Luuya are separated from the rest. The new component that appears at K=6 is restricted to African Americans and Biaka Pygmies, and is the last component that can be attributed to specific populations.

DISCUSSION

The preceding results are in agreement with what was found previously by Tishkoff *et al.* using microsatellites, and goes beyond with new findings and refinement of previous genetic studies:

i) The main distinction is among Niger-Congo groups and the rest, including Nilo-Saharan speakers and hunter-gatherers (with the Khoisan having preserved their ancestral language but not Pygmies). Among Niger-Congo, geography is the main factor explaining the genetic differences, with a remarkable similarity among western populations (Yorubas and Mandenka), which could reflect a burst in the expansion to the west, related to iron technology and Niger-Congo languages.

ii) The South-eastern Bantu from Mozambique are remarkably differentiated from the western Niger-Congo speaking populations, such as the Mandenka and the Yoruba, and also differentiated from geographically closer Eastern Bantu samples, such as Luuya. These results suggest that the Bantu expansion of languages, which started ~5,000 years ago at the present day border region of Nigeria and Cameroon, and was probably related to the spread of agriculture and the emergence of iron technology[17-19], was not a demographic homogeneous migration with population replacement in the southernmost part of the continent, but acquired more divergence, likely due to the integration of pre-Bantu people.

The complexity of the expansion of Bantu languages to the south (with an eastern and a western route[20]), might have produced differential degrees of assimilation of previous populations of hunter-gatherers. This assimilation has been detected through uniparental markers due to the genetic comparison of nowadays hunter-gatherers (Pygmies and Khoisan) with Bantu-speaker agriculturalists[2; 21-24]. Nonetheless, the singularity of the southeastern population of Mozambique (poorly related to present Khoisan) could be attributed to a complete assimilation of ancient genetically differentiated populations (presently unknown) by Bantu-speakers in southeastern Africa, without leaving any pre-Bantu population in the area to compare with.

iii) The difference between hunter-gatherers and the rest of South Saharan populations is important but it is not the main trait in the African genetics. To note is the strong similarity among the three studied populations, with no specific Pygmy component, but an important Bantu introgression (as seen in K=3) in Biaka Pygmies. Pygmies should be included along with Khoisan in the search for deep-rooted African and Human lineages. Moreover, the specific component that identifies the three hunter-gatherer populations is found at a small amount in all other African populations, as possible results of introgression with previous settlers of most African territory.

The genetic analysis of a large number of SNPs is thus providing a robust tool to refine our understanding of past populations history.

Acknowledgements

This research was supported by Dirección General de Investigación, Ministerio de Ciencia y Tecnología, Spain (grants SAF2007-63171, BFU2007-63657) and Direcció General de Recerca, Generalitat de Catalunya (2005SGR00608 and 2009 SGR 1101). SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; http://www.cegen.org); Bioinformatic services were kindly provided by the Genomic Diversity node, Spanish Bioinformatic Institute (www.inab.org). MS is supported by a PhD fellowship from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2005-3982).

REFERENCES

1. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, N.J.
2. Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111
3. Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F (2007) Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol 43:635-644
4. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105:1596-1601
5. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82:1130-1140
6. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70:1197-1214
7. Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D (2009) Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Mol Biol Evol 26:1581-1589
8. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic

structure and history of Africans and African Americans. Science 324:1035-1044

9. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 107:786-791
10. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104
11. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998-1003
12. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40:340-345
13. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190
14. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959
15. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587
16. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40:646-649
17. Vansina J (1995) New Linguistic Evidence and 'the Bantu Expansion'. The Journal of African History 36:173-195
18. Phillipson DW (1993) African archaeology. Cambridge University Press, Cambridge
19. Newman JL (1995) The peopling of Africa : a geographic interpretation. Yale University Press, New Haven ; London

20. Oslisly R (1995) The middle Ogooué valley, Gabon: cultural changes and palaeoclimatic implications of the last fourmillenia. Azania 39–40:324–331

21. Pereira L, Gusmao L, Alves C, Amorim A, Prata MJ (2002) Bantu and European Y-lineages in Sub-Saharan Africa. Ann Hum Genet 66:369-378

22. Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Ann Hum Genet 65:439-458

23. Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo A, Comas D (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet 115:439-447

24. Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A (2005) The genetic legacy of western Bantu migrations. Hum Genet 117:366-375

FIGURE LEGENDS

Figure 1. PCA of merged HGDP and HapMap 3 samples.

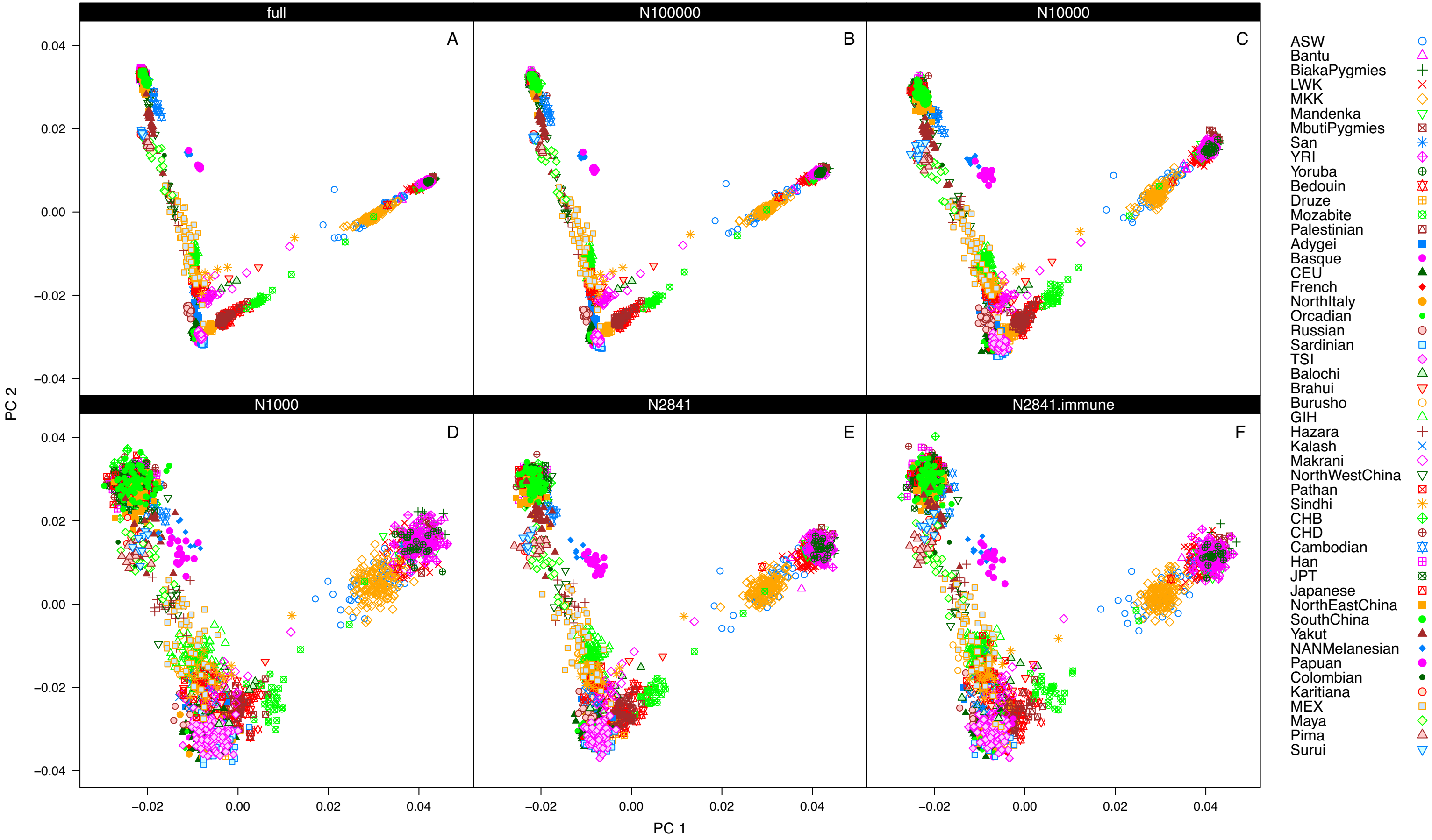
Panels show the results of the PCA for the full merged set of SNPs (a; 460,147 SNPs), for random subsets of 100,000 (b), 10,000 (c), 1,000 (d), and 2,841 SNPs (e), as well as for the 2,841 SNPs in the merged analysis set including the samples from Mozambique. As can be seen, the general pattern of differentiation is reproduced even using only a random subset of 1,000 SNPs. HapMap populations are indicated by their abbreviated names: ASW - African ancestry in Southwest USA; CEU - Utah residents with Northern and Western European ancestry from the CEPH collection; CHB - Han Chinese in Beijing, China; CHD - Chinese in Metropolitan Denver, Colorado; GIH - Gujarati Indians in Houston, Texas; JPT - Japanese in Tokyo, Japan; LWK - Luhya in Webuye, Kenya; MEX - Mexican ancestry in Los Angeles, California; MKK - Maasai in Kinyawa, Kenya; TSI Toscani in Italia; YRI - Yoruba in Ibadan, Nigeria.

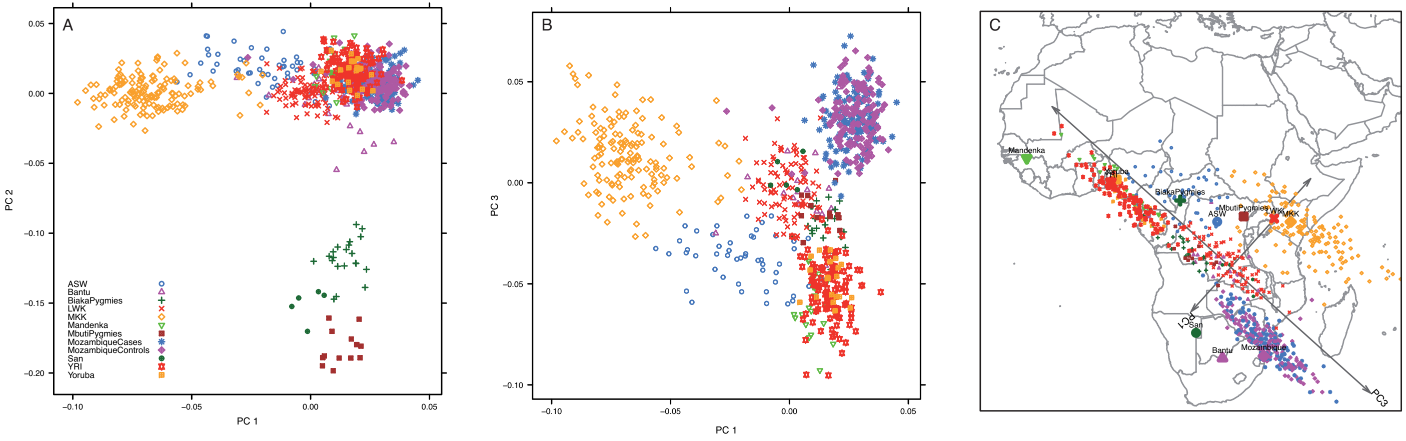
Figure 2. PCA of sub-Saharan African populations

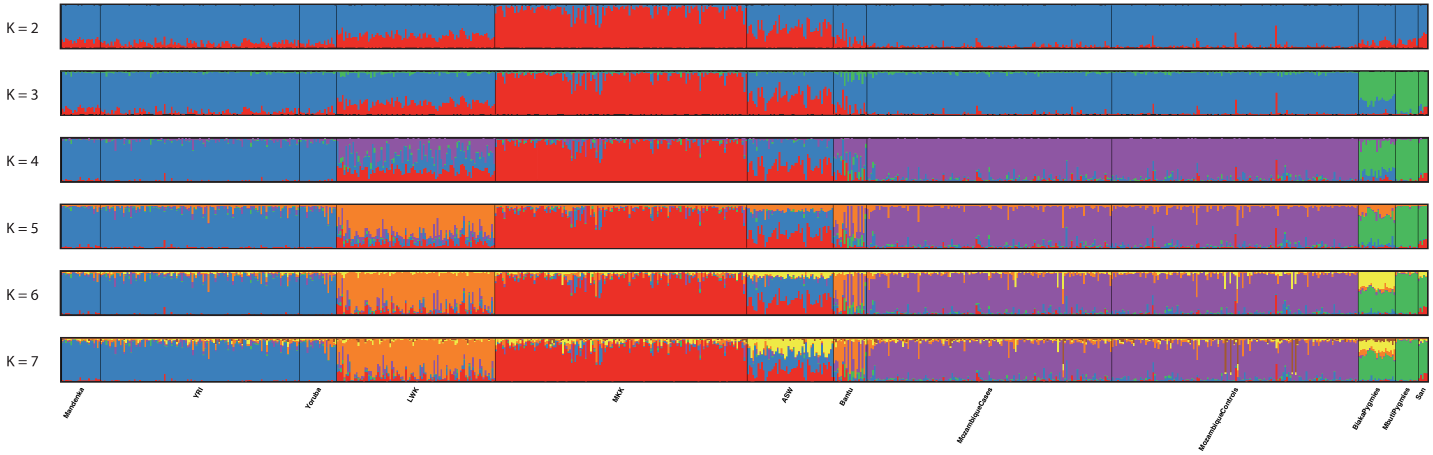
Panels show plots of the first three principal components obtained from the 11 sub-Saharan African populations. 2A, first and second. 2B, first and third. 2C.- Biplot of PC1 and PC3 superimposed onto a map of Africa. Geographical locations of the population are indicated by their names and their respective enlarged plot symbols.

Figure 3. STRUCTURE results for sub-Saharan African populations

Depicted are the results of 5 runs each for the number of clusters ranging from K = 2 to K = 7, combined using CLUMPP.


full
N100000
N10000
N1000
N2841
N2841.immune
A
B
C
D
E
F
PC 2
PC 1
0.04
0.02
0.00
−0.02
−0.04
−0.02
0.00
0.02
0.04
ASW
Bantu
BiakaPygmies
LWK
MKK
Mandenka
MbutiPygmies
San
YRI
Yoruba
Bedouin
Druze
Mozabite
Palestinian
Adygei
Basque
CEU
French
NorthItaly
Orcadian
Russian
Sardinian
TSI
Balochi
Brahui
Burusho
GIH
Hazara
Kalash
Makrani
NorthWestChina
Pathan
Sindhi
CHB
CHD
Cambodian
Han
JPT
Japanese
NorthEastChina
SouthChina
Yakut
NANMelanesian
Papuan
Colombian
Karitiana
MEX
Maya
Pima
Surui


A
PC 2
PC 1
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
ASW ○
Bantu △
BiakaPygmies +
LWK ×
MKK ◇
Mandenka ▽
MbutiPygmies ■
MozambiqueCases ✱
MozambiqueControls ◆
San ●
YRI ✡
Yoruba ■
B
PC 3
PC 1
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
C
Mandenka
Yoruba
YRI
BiakaPygmies
MbutiPygmies
ASW
LWK
MKK
San
Bantu
Mozambique
PC1
PC3


K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
Mandenka
YRI
Yoruba
LWK
MKK
ASW
Bantu
MozambiqueCases
MozambiqueControls
BiakaPygmies
MbutiPygmies
San

4

Discussion

The aim of the work as outlined in the objectives was to characterise how the selective pressure of malaria has shaped human genetic variation. In particular, the aim was to discover novel loci involved in the susceptibility to placental malaria infection. We approached this aim by two different, but complementary approaches, employing tools of both evolutionary genetics and genetic epidemiology analyses, the results of which take up the main part of this thesis. Another aim of this work was a more general description of the genetic structure of our study population in relationship to other African populations, which has implications for both the population history of those populations, as well as for any future genetic studies on them. This chapter will provide a discussion of our main findings, in the context of those initial aims.

Natural selection in a candidate pathway for placental malaria infection

The premise for the analysis presented in Section 3.1 is a simple one: In a population exposed to a particular selective agent exerting strong selective pressure, any genetic variant that significantly increases the chances of survival for an individual is expected to show evidence for positive selection. One should therefore be able to find those variants just by localising the targets of natural selection in the exposed populations (Bamshad and Wooding, 2003). The case of malaria should be particularly amenable to this approach, given the strong selective pressure it has exerted over the past 10,000 years of human population history. In fact, as mentioned in the introduction to this work, one of the most powerful tests for recent positive selection, the so-called long range haplotype test, used two malaria susceptibility loci as a proof-of-principle for the power to detect selection (Sabeti et al., 2002a).

In the case of placental malaria, the particular form of malaria that is our study subject, this premise should equally hold, as it is responsible for a substantial part of the yearly burden of malaria (Desai et al., 2007). The particularities of placental malaria also provided an obvious candidate for testing this hypothesis, as the key feature of its pathogenesis is the adhesion of infected erythrocytes in the intervillous space of the placenta. Previous studies suggested that this adhesion was mediated by two glycosaminoglycan receptors, hyaluronic acid (HA) and in particular chondroitin sulphate-A (CSA) (Rogerson et al., 2007). In this first study we therefore analysed all genes encoding for the proteins that are involved in the biosynthesis of this two receptors, taking advantage of the data produced by the International HapMap project. Importantly, the population of African origin in this dataset, the Yorubans (YRI) from Nigeria, originate from a region of stable high malaria transmission, so any signature of selection related to malaria should be detected in this population. In order to detect those signatures, three different methods based on allele frequencies, population differentiation and long range haplotypes were employed.

Our results indicated two regions exhibiting some evidence of selection in the African population, both of them based mainly on the results from the long-range haplotype (LRH) results. In the case of *UST*, we found two core haplotypes around 20 kb upstream of the gene that were significant in the LRH analysis. Furthermore, both of these cores also contained SNPs with large F_{ST} values, with the YRI being differentiated from the other populations, which would be consistent with a selective sweep in Africa. At the time of completion of the published paper, we hypothesised that the identified region would be involved in transcriptional regulation of *UST*, which was also supported by a number of peaks in regulatory potential scores calculated from an alignment of seven mammalian species (obtained from the UCSC genome browser). The availability of a number of large scale surveys of regulatory variation, referred to as expression quantitative trait loci (eQTL), using the same HapMap samples allows us to revisit this hypothesis (Stranger et al., 2007; Veyrieras et al., 2008). Nevertheless, neither of these studies reported any eQTL in the whole region

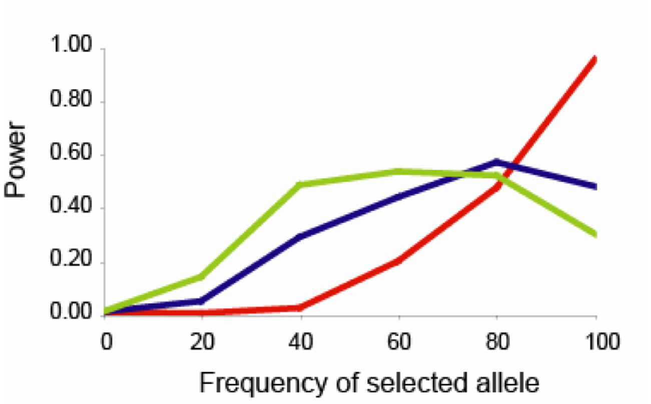


Figure 4.1: The power to detect a selective sweep for alleles of different frequencies and different tests of positive selection (iHS - blue; LRH - green; XP-EHH - red) (Sabeti et al., 2007)

between *UST* and its neighbouring gene upstream, *SASH1* (obtained from the Pritchard lab eQTL browser, http://eqtl.uchicago.edu). This apparent negative result would however not rule out a tissue specific effect (Dimas et al., 2009), which would be expected in the placenta, a tissue that has not been used in this surveys.

The other gene identified was *ChGn*, which was interesting due to the fact that a non-synonymous SNP only polymorphic in the YRI showed the highest score of all non-synonymous SNPs analysed in this population. An issue with this result is the low frequency of this SNP of 0.08 in the YRI. The test that yielded the high score for this SNP is very similar to the iHS score (Voight et al., 2006), which is expected to have very low power to detect alleles at this frequency (Figure 4.1). Additionally, in a recently developed composite method for detecting positive selection it was noted that the iHs statistic, which is calculated as a ratio, is very sensitive to fluctuations in the length of the ancestral haplotype, particularly at low frequency sweeps (Grossman et al., 2010). It is reasonable to assume that the score used in our analysis suffers from the same issue, since it is similarly based on a ratio of haplotype lengths.

The most striking result in our analysis was found in the gene *XYLT1*, which showed strong evidence for selection in all of the applied methods. However, this result was obtained in the combined sample of Japanese and Chinese individuals, and not the Africans. Even if we take into account that historically malaria was much more widespread, it is unlikely that it occurred with high transmission in these two regions in East Asia. Without discarding the possibility of a false-positive result, it is therefore likely that a different selective agent is responsible for the observed signatures. This result then highlights one of the problems of this approach: Even if an interesting result is obtained in a particular population or geographic region, there are usually a number of different selective pressures possible to explain the result. The premise of the employed approach, the identification of functional variation based on signatures of selection, is therefore still a valid one, as it is likely that this region harbours genetic variation that has contributed to regional adaptations of these two populations. However, in order to detect adaptations to a *particular* selective agent, follow-up studies will have to provide an in-depth analysis of both the patterns of genetic variation at the locus and its functional impact.

The question is then what can be concluded from the results in the African population. Given that both results lack the strength observed for *XYLT1*, our conclusion would be that it is unlikely that they are targets of strong recent selective events. This however should not rule out the possibility of selection in other populations from malaria-endemic regions, as malaria is likely to have provoked regionally different adaptations due to its strong pressure (Kwiatkowski, 2005; Jallow et al., 2009). Interestingly, the region surrounding *ChGn* was identified as among the top 10 signals for the XP-EHH statistic in a recently published world-wide survey of positive selection using the HGDP populations (Pickrell et al., 2009). This signal is found in the populations from Oceania, which is a malaria endemic region and would therefore merit a more in-depth investigation. The non-synonymous SNP found in this study is however only polymorphic in a few other African populations apart from the Yorubans, but not in Oceania. This would suggest an independent selective event in this geographic region.

Finally, it is also interesting to ask how our main candidate fares in comparison with the results of the newly available world-wide data. Confirming the result found in this study, *XYLT1* was also found among the top 1% iHs signals in the East Asian populations of the HGDP.

Genetic susceptibility to placental malaria infection

In the works presented in Sections 3.2 and 3.3, our aim was to identify novel host genetic factors involved in placental malaria infection by means of a genetic epidemiology approach. In particular, we carried out a case-control association study with patients and control individuals from Mozambique, a region of stable high malaria transmission. This study has been carried out in two phases, and below we will discuss the findings of both of them separately. Before that, a brief discussion on the study individuals will also be provided.

Study individuals

The individuals of this study were enrolled in the study at the Centro de Investigação em Saúde da Manhiça (CISM), in southern Mozambique. A total of 360 pregnant women were chosen from the participants of a malaria intervention trial, in order to form a case-control sample based on the infection status of the placenta assessed after delivery. One drawback of this sample is that due to the relatively small sample size, the power to detect variants with a small effect will be limited. On the other hand, the samples have a variety of desired characteristics. Because the CISM is the main health centre for the entire population of the district, we expect the samples to provide a good representation of the genetic diversity in the region. This should also avoid any biases which might occur in a situation of differential selection of cases and controls for example. The analysis of genetic structure we carried out as part of the preliminary analysis confirms this, as we could not detect any significant population stratification between cases and controls even with the large number of SNPs analysed in the second phase. It remains to be seen whether a genome-wide panel of SNPs will provide enough resolution to detect fine-scale population structure that we missed in our analysis, however, our results from Section 3.4, which will be discussed later on, indicate that this is likely not going to be the case.

Another important property of the samples is that they are epidemiologically well characterised, which is crucial for ruling out spurious results due to confounding of another factor influencing susceptibility to placental infection. An example for such a factor is the influence on the number of previous pregnancies, which has been shown to be an important factor in susceptibility, as outlined in the introduction.

Glycosylation

The first of the two studies was carried out analysing data for 719 SNPs in a set of 62 genes involved in glycosylation and immunity. As a first observation, results from the preliminary analysis show the importance of stringent quality control for both samples and SNPs, as the inclusion of lower quality data clearly resulted in substantially inflated test statistics. On the other hand, this naturally implies throwing away some of the data, which is of course never desirable. The somewhat crude approach of removing samples and SNPs based on a simple measure of genotyping success like the call rate clearly has much room for improvement, and recently some more sophisticated methods have been introduced (Scheet and Stephens, 2008). In combination with the ever increasing accuracy of genotyping assays this should enable researchers to make optimal use of the produced data.

The main result from this study was a significant association of a SNP in the gene *FUT9*. Importantly, this SNP (rs3811070) was located in a block of high LD with two neighbouring SNPs, which were together with rs3811070 the strongest signals found in our analysis. This rules out the possibility of a spurious association due to genotyping error, as is expected after the stringent quality control measures applied. Given the low p-value even in the small sample size we expect a large effect size for the SNP, which was confirmed with an odds ratio (OR) of 2.31. However, it should be noted that this effect is most likely an overestimation of the real effect, due to a bias known as the so-called *winner's curse*, which results from selecting the most significant statistic and subsequent estimation of the effect in the same sample (Göring et al., 2001).

As mentioned above, it is important to control for potential confounding effects, so we performed logistic regression to assess their influence. It was interesting to see that we could confirm that women with fewer pregnancies were more susceptible to infection in our sample, with parity by itself being significant in our model. Nevertheless, the effect of rs3811070 remained unchanged even after controlling for parity, which confirmed that its effect was an independent one. This result was also not influenced by different ways to categorise parity.

The analysis of haplotype association also yielded a number of interesting results. The four SNP haplotype block containing rs3811070 found to be associated with placental infection showed a pattern of two common haplotypes with mismatching alleles at every SNP, a pair of so-called *yin-yang haplotypes* (Zhang et al., 2003). We then looked at the same region in data from the HapMap populations, in order to take advantage of the much higher SNP density in this dataset. The yin-yang haplotype pair was also observed in those populations, which suggests that this pair already existed before the out-of-Africa migration of modern humans. However, the higher SNP density revealed a more complicated pattern than anticipated, with the non-risk haplotypes showing evidence of being recombinants made up of the two different yin-yang pairs (see Figure 3 in Section 3.2). The risk-type haplotypes were also much more common in the CEU population, where one of the two recombinant non-risk types was not observed. A possible explanation for this pattern could be that both recombinant types afford a protection from placental infection, and are therefore both under selection to increase in frequency in the Africans. This would be consistent with the reduced haplotype diversity observed in the YRI, which is not expected in an African population. Additionally, it would also provide an explanation for the lack of a clear signal of positive selection in the region, as it would involve selection on two different, previously segregating haplotypes, which is likely to be very difficult to detect. How could the protective effect for two different haplotypes be explained? One can for example imagine that, since both recombinants share the same breakpoint in their yin-yang patterns, there might be some regulatory site spanning that breakpoint, which, as a consequence, is not bound by its regulatory factor anymore. In such a situation the actual pattern of variation on both sides of the breakpoint would not be important, as long as the binding site itself is affected.

It should be clear from the above that, irrespectively of the plausibility of this hypothesis, this region would be a very interesting candidate for fine-mapping the observed association signal. From a functional perspective, the result also provides a number of interesting follow-up study one could envision. An obvious experiment would be for example to test whether there are differences in expression of the protein product of *FUT9*, FucT-IX, in placental tissue of samples with the different haplotypes. Using a next generation sequencing approach could even allow to simultaneously asses parasite DNA quantitatively, which would provide a much more detailed view of the state of infection of the placental tissue.

In summary, our results implicate *FUT9* as a novel host genetic factor in placental malaria infection. With respect to the expectation of signatures of positive selection the results were less clear, although the patterns of haplotype diversity described above suggest a complex pattern of evolution at the locus.

Inflammation and immunity

For phase two of the genetic association study of placental infection, we chose a commercially available SNP genotyping chip containing more than 9,000 SNPs in around 1,000 genes with involvement in the immune and inflammatory responses in humans. The rationale for this was that those processes are clearly playing an important role in the pathogenesis of both malaria in general as well as placental malaria, as outlined in the introduction to this work.

The results for both single SNP and multi SNP analyses showed the strongest signals in two different regions, *KLRK1* and *IL7*. However, both of them did not remain significant after correcting for multiple testing, even though *IL7* showed a marginally significant p-value in the multi-marker analysis. Although this result is somewhat discouraging, it is worth noting that even the recently published large scale genome-wide association study of severe malaria suffered from this lack of strong signals (Jallow et al., 2009). Even the large number of individuals can not sufficiently improve the power that is lost by the generally low levels of LD in African populations (see Section 1.4.3). This effect will also affect the targeted genotyping array used for this study. It also highlights one of the practical issues one faces when conducting a genetic study on African populations. Until the day when whole-genome sequencing of a large number of individuals becomes reality, it will be desirable to use much higher density arrays for those studies.

One potential way to circumvent the above mentioned problem is to use imputation based on a much denser set of genotypes or even resequencing data for the region of interest. Ideally, a researcher would sequence the region of interest in a small subset of the same study individuals and use this data as a template for imputation. However, as in the case of our study, this data is often not available, meaning that imputation can only be performed using a reference panel from a different population, or multiple reference populations. The results of our imputation analysis indicate a number of SNPs in both candidate regions identified above with stronger predicted signals than any of the genotyped SNPs. Nevertheless, these results have to be interpreted with caution, as it has been shown that imputation accuracy using a mismatched panel is particularly low in African populations, as expected due of the large genetic diversity (Huang et al., 2009).

Another aim of this study was to take advantage of the known relationships among the genes investigated, i.e. their interactions in the context of a functional network. It is generally accepted that gene-gene interactions or epistasis play an important role in susceptibility to disease, although its relative importance compared to other factors such as multiple rare alleles with large effects is still debated (Moore, 2003; Manolio et al., 2009). Not surprisingly, much effort is aimed at the development of statistical methods to detect such interactions (Cordell, 2009), but there is still a lack of powerful tools to assess whether a set of genes organised in a functional network is associated with a phenotype. In this study we approached this problem by simultaneously mapping evidence for association of single genes, as well as epistatic interactions among SNPs in genes whose products are interacting, onto a network of curated protein interactions. Interestingly, we found that the strongest results for both genes and interactions occur within the same neighbourhood of the network, which corresponds to the IL-7/IL-7R signalling module. Although it has to be kept in mind that statistical interaction between loci does not necessarily imply a biological interaction, this result clearly shows the potential of such an approach to identify modules of interacting genes that are involved in a phenotype of interest.

In the context of this thesis work, or, more general, in the study of genetic resistance to malaria, it would be particularly interesting to incorporate evidence for natural selection in a similar manner. In a situation as described above, when a module of interacting loci contributes to susceptibility, it is likely that any single gene will not show strong signatures of selection, as a number of different allelic combinations at the loci that form the module might produce a similar phenotype. One would therefore need a method to detect a compound signature for those allelic combinations that reduce susceptibility, for example by assessing whether particular combinations occur more often than expected by random in an exposed population as opposed to a non-exposed population. Event though this is today probably only feasible for a small number of loci, novel innovative statistical methods would certainly bring this goal a great step closer.

As a summary for this second phase, based on the overall evidence from both single gene as well as network analysis, IL-7/IL-7R signalling emerges as an interesting candidate for an involvement in placental malaria infection. One of the roles for this pathway in the organism is the induction of the pro-inflammatory cytokine TNF, which plays an important role in the early defence against *Plasmodium* infection and control of parasitaemia. A possible effect on placental infection can be explained by a general reduction in parasitaemia and faster clearance of infection, thereby reducing the chances of infection of the placenta. As a first follow-up study for a functional relevance of the result, it would be interesting to see whether the identified SNP-SNP interaction involving the non-synonymous SNP in *IL7R* has an effect in modulating the inflammation response described above.

Genetic structure of Sub-Saharan African populations

Our aim in the last work presented in Section 3.4 was to to get a better understanding of the genetic structure of Sub-Saharan African populations, in particular with respect to our study population from Mozambique. It is quite striking that the continent which harbours more genetic variation than the rest of the world combined has been so under-represented in studies of human genetic diversity to date. It was only as recently as last year that the first large-scale survey of African genetic diversity was published, followed this year by another study focussed on West African populations (Tishkoff et al., 2009; Bryc et al., 2010). Event though these studies give the first detailed picture of the genetic structure present in the continent, some regions remained undersampled, in particular in the southern part of the continent. As our study population originates from such an undersampled region we were interested in how it is related to other Sub-Saharan African populations. Adding to our own results genotype data from both HapMap and HGDP allowed us to investigate the genetic structure of a total of 12 populations within Sub-Saharan Africa.

In order to perform this analysis it was necessary to build a consensus set of SNPs that had genotype data in all 12 populations. The limiting factor was the data of the population from Mozambique, as we only typed around 9,000 SNPs in the course of the study presented in Section 3.3, providing a final dataset with 2,841 SNPs for all populations. In order to ensure that this number was sufficient, we evaluated the performance of different sized SNP sets to detect the known pattern of genetic structure in the shared HGDP and HapMap samples. It was surprising to find that even a seemingly low number of 1,000 randomly chosen SNPs was sufficient to recover the general structure obtained from analysing a genome-wide panel of hundreds of thousand of SNPs. The main consequence of reducing the number of markers is an increasing inter-individual variance, but the pattern for populations as a whole stays remarkably similar. This suggests that if the goal of an analysis is to get a general picture of the genetic structure among populations, genotyping a few thousand randomly distributed markers over the genome will be sufficient. Due to the increased inter-individual variance it will however be increasingly difficult to unambiguously assign individuals with unknown ancestry to a particular population, especially if those populations are closely related to each other. In the context of our work we can therefore be confident that our result will give us a true representation of the underlying genetic structure, even if it might lack a fine-scale resolution.

A number of interesting results were obtained from the analysis, both confirming previous ones and adding new insights to the understanding

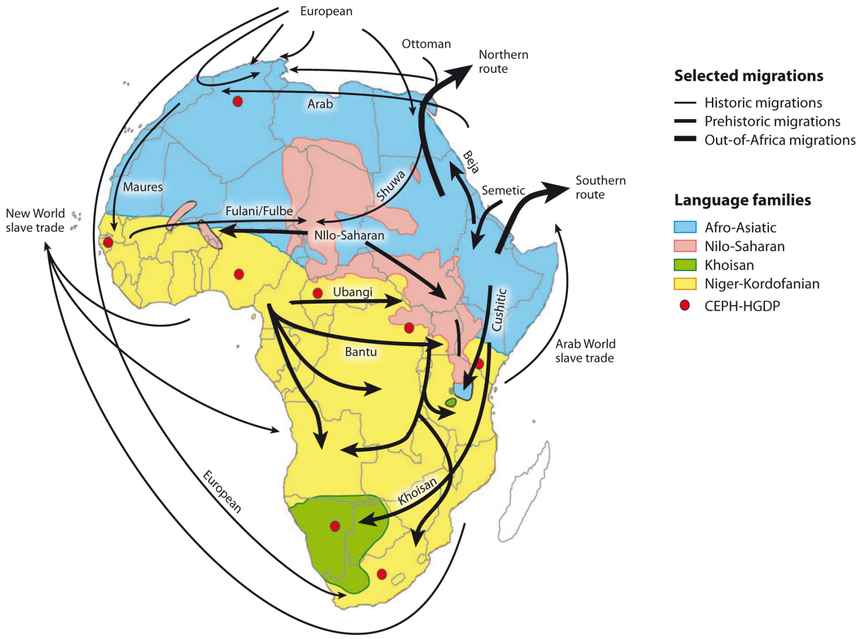


Figure 4.2: A map of the distribution of African language families and migration events (Campbell and Tishkoff, 2008).

of African genetic structure. As a first observation, the main distinctions among the populations investigated reflect primarily linguistic and cultural differences, as opposed to geographic differentiation (see Figure 4.2). The first genetic component separates members of the two large language families in Sub-Saharan Africa, the Nilo-Saharan and Niger-Kongo families, whereas the second component differentiates hunter-gatherer populations such as the Pygmies from the remaining, predominantly pastoralist populations. This result is in agreement with the findings of Tishkoff et al. (2009), again confirming that the set of SNPs is sufficient for uncovering the genetic structure of the study populations. However, the most interesting result was the position of our study population from Mozambique. This population is part of the Bantu family, and originates from a region that was one of the main routes for the Bantu expansion of languages. One could therefore hypothesise that they would be genetically similar to other Bantu populations. Nevertheless, our results show a remarkable degree of differentiation even from geographically comparably close Bantu populations such as the Luuya in Kenya. This raises the possibility that they represent a pre-Bantu population that was already present in the region at the time of the Bantu expansion, but subsequently got completely assimilated within the Bantu. As mentioned above, this region remains poorly sampled by genetic studies, making a more dense sampling of populations along the Eastern expansion route paramount for a further investigation of this hypothesis. A detailed picture of the genetic diversity in the region could also answer whether the observed differentiation could be explained by an isolation-by-distance model, as an alternative hypothesis.

Concluding remarks

Our aim as outlined in the objectives of this thesis was to aid our understanding of how malaria has shaped the human genome. The question is therefore now whether this aim has been achieved, in particular with respect to the two key features of our approach, the combination of evolutionary and genetic epidemiology analysis, as well as the focus on placental infection as a phenotype.

As to the first key feature, our results highlight both the potential but also the difficulties of a purely evolutionary genetics approach to malaria susceptibility. It is clear from Section 3.1 that interesting regions can be identified, but the difficulty lies in establishing a causal relationship with a particular selective agent. Another difficulty arises as a side-effect of the large selective pressure of malaria that justifies this kind of approach in the first place, as this pressure is likely to induce different molecular adaptations to malaria in different geographical regions. For example, the sickle cell variant HbS is thought to have arisen in at least four different populations independently (Kwiatkowski, 2005), which indicates that results from a particular geographic region might not be transferable to other regions. The way forward is then to perform an integrated approach within one single population, consisting in evolutionary analysis of genetic variation, epidemiological studies on well-phenotyped samples as well as functional assays to determine the functional impact of any interesting locus found.

With respect to the second key feature, both studies presented in Sections 3.2 and 3.3 are the first to provide a large-scale investigation of the genetic basis of placental malaria infection. We could identify one strong candidate locus for involvement in susceptibility to placental infection. Our results also indicate that studying more clearly defined phenotypes such as infection facilitates the detection of associations in comparison to complex compound phenotypes such as severe malaria. From an evolutionary perspective on the other hand, the phenotype that matters is the survival of an individual, which mostly depends on the severity of the disease. We would therefore expect natural selection to act mostly on loci that affect severity. However, as severity of disease is also related to the ability to control parasitaemia, studying variation in levels of parasitaemia in infected individuals might provide a phenotype that is more easily tractable than severe malaria itself, but nonetheless has a strong influence on it. Finally, animal models of malaria have shown that the host can evolve both strategies of resistance (i.e. the ability to limit parasitaemia) as well as tolerance (i.e. the ability to control severity for a particular level of parasitaemia) as a consequence of infection, which are moreover negatively interacting with each other (Råberg et al., 2007). It will be interesting to test this result in human malaria, for example to see whether one of the two general mechanisms is more common in humans, or if different populations have evolved different strategies as an answer to the selective pressure of malaria.

In summary, one can say that we are entering an exciting age for studying the host genetics of malaria, as cheap high-throughput sequencing combined with sophisticated statistical analysis tools will enable us to tackle these questions with unprecedented power and detail.

Bibliography

Ackerman H, Usen S, Jallow M, Sisay-Joof F, Pinder M, Kwiatkowski DP (2005). A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. Annals of Human Genetics 69(Pt 5):559–65.

Aidoo M, Terlouw DJ, Kolczak MS, McElroy PD, ter Kuile FO, Kariuki S, Nahlen BL, et al (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet 359(9314):1311–2.

Aitman TJ, Cooper LD, Norsworthy PJ, Wahid FN, Gray JK, Curtis BR, McKeigue PM, et al (2000). Malaria susceptibility and CD36 mutation. Nature 405(6790):1015–6.

Akey JM (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–22.

Akira S, Takeda K (2004). Toll-like receptor signalling. Nat Rev Immunol 4(7):499–511.

Allison A (1954). Protection afforded by sickle-cell trait against subtertian malarial infection. British Medical Journal 1(4857):290.

Alonso PL, Sacarlal J, Aponte JJ, Leach A, Macete E, Aide P, Sigauque B, et al (2005). Duration of protection with RTS,S/AS02A malaria vaccine in prevention of Plasmodium falciparum disease in Mozambican children: single-blind extended follow-up of a randomised controlled trial. Lancet 366(9502):2012–8.

Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, et al (2009). Targets of balancing selection in the human genome. Mol Biol Evol 26(12):2755–64.

Artavanis-Tsakonas K, Tongren JE, Riley EM (2003). The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. Clinical & Experimental Immunology 133(2):145–152.

Ayi K, Min-Oo G, Serghides L, Crockett M, Kirby-Allen M, Quirt I, Gros P, et al (2008). Pyruvate kinase deficiency and malaria. N Engl J Med 358(17):1805–10.

Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, Patterson N, et al (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. American Journal of Human Genetics 81(2):234–42.

Balding DJ (2006). A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7(10):781–91.

Bamshad M, Wooding SP (2003). Signatures of natural selection in the human genome. Nature Reviews Genetics 4(2):99–111.

Barnwell JW, Asch AS, Nachman RL, Yamaya M, Aikawa M, Ingravallo P (1989). A human 88-kD membrane glycoprotein (CD36) functions in vitro as a receptor for a cytoadherence ligand on Plasmodium falciparum-infected erythrocytes. J Clin Invest 84(3):765–72.

Baton LA, Ranford-Cartwright LC (2005). Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito. Trends Parasitol 21(12):573–80.

Baum J, Maier AG, Good RT, Simpson KM, Cowman AF (2005). Invasion by P. falciparum merozoites suggests a hierarchy of molecular interactions. PLoS Pathog 1(4):e37.

Bellamy R, Kwiatkowski D, Hill AV (1998). Absence of an association between intercellular adhesion molecule 1, complement receptor 1 and interleukin 1 receptor antagonist gene polymorphisms and severe malaria in a West African population. Trans R Soc Trop Med Hyg 92(3):312–6.

Berendt AR, Simmons DL, Tansey J, Newbold CI, Marsh K (1989). Intercellular adhesion molecule-1 is an endothelial cell adhesion receptor for Plasmodium falciparum. Nature 341(6237):57–9.

Browning SR (2006). Multilocus association mapping using variable-length Markov chains. American Journal of Human Genetics 78(6):903–13.

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, et al (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107(2):786–91.

Cabantous S, Poudiougou B, Traore A, Keita M, Cisse MB, Doumbo O, Dessein AJ, et al (2005). Evidence that interferon-gamma plays a protective role during cerebral malaria. J Infect Dis 192(5):854–60.

Campbell MC, Tishkoff SA (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–33.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al (2002). A human genome diversity cell line panel. Science 296(5566):261–2.

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King MC (1991). Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. Genomics 11(2):490–1.

Chen H, Patterson N, Reich D (2010). Population differentiation as a test for selective sweeps. Genome Res 20(3):393–402.

Clark TG, Fry AE, Auburn S, Campino S, Diakite M, Green A, Richardson A, et al (2009). Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. Eur J Hum Genet 17(8):1080–5.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics 38(11):1251–1260.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, et al (2009). Origins and functional impact of copy number variation in the human genome. Nature.

Cooke GS, Hill AV (2001). Genetics of susceptibility to human infectious disease. Nat Rev Genet 2(12):967–77.

Cordell HJ (2009). Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10(6):392–404.

Cserti CM, Dzik WH (2007). The ABO blood group system and Plasmodium falciparum malaria. Blood 110(7):2250–8.

Darwin C (1859). On the origin of species by means of natural selection. London: J. Murray.

Desai M, ter Kuile FO, Nosten F, McGready R, Asamoa K, Brabin B, Newman RD (2007). Epidemiology and burden of malaria in pregnancy. Lancet Infectious Diseases 7(2):93–104.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, et al (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. Science 325(5945):1246–1250.

Doolan DL, Dobaño C, Baird JK (2009). Acquired immunity to malaria. Clin Microbiol Rev 22(1):13–36, Table of Contents.

Escalante AA, Ayala FJ (1994). Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proc Natl Acad Sci U S A 91(24):11373–7.

Escalante AA, Barrio E, Ayala FJ (1995). Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol Biol Evol 12(4):616–26.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–87.

Fernandez-Reyes D, Craig AG, Kyes SA, Peshu N, Snow RW, Berendt AR, Marsh K, et al (1997). A high frequency African coding polymorphism in the N-terminal domain of ICAM-1 predisposing to cerebral malaria in Kenya. Hum Mol Genet 6(8):1357–60.

Ferwerda B, McCall MBB, Alonso S, Giamarellos-Bourboulis EJ, Mouktaroudi M, Izagirre N, Syafruddin D, et al (2007). TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans. Proc Natl Acad Sci U S A 104(42):16645–50.

Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. Nat Rev Genet 7(2):85–97.

Flint J, Hill AV, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, et al (1986). High frequencies of alpha-thalassaemia are the result of natural selection by malaria. Nature 321(6072):744–50.

Flori L, Delahaye NF, Iraqi FA, Hernandez-Valladares M, Fumoux F, Rihet P (2005). TNF as a malaria candidate gene: polymorphism-screening and family-based association analysis of mild malaria attack and parasitemia in Burkina Faso. Genes Immun 6(6):472–80.

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009). Human genetic variation and its contribution to complex traits. Nature Reviews Genetics 10(4):241–251.

Fried M, Domingo GJ, Gowda CD, Mutabingwa TK, Duffy PE (2006). Plasmodium falciparum: chondroitin sulfate A is the major receptor for adhesion of parasitized erythrocytes in the placenta. Experimental Parasitology 113(1):36–42.

Fried M, Duffy PE (1996). Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. Science 272(5267):1502–4.

Fry AE, Auburn S, Diakite M, Green A, Richardson A, Wilson J, Jallow M, et al (2008a). Variation in the ICAM1 gene is not associated with severe malaria phenotypes. Genes Immun 9(5):462–9.

Fry AE, Ghansa A, Small KS, Palma A, Auburn S, Diakite M, Green A, et al (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Hum Mol Genet 18(14):2683–92.

Fry AE, Griffiths MJ, Auburn S, Diakite M, Forton JT, Green A, Richardson A, et al (2008b). Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria. Hum Mol Genet 17(4):567–76.

Gaur D, Mayer DCG, Miller LH (2004). Parasite ligand-host receptor interactions during invasion of erythrocytes by Plasmodium merozoites. Int J Parasitol 34(13-14):1413–29.

Genton B, D'Acremont V, Rare L, Baea K, Reeder JC, Alpers MP, Müller I (2008). Plasmodium vivax and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea. PLoS Med 5(6):e127.

Göring HH, Terwilliger JD, Blangero J (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69(6):1357–69.

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, et al (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967):883–6.

Guindo A, Fairhurst RM, Doumbo OK, Wellems TE, Diallo DA (2007). X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. PLoS Med 4(3):e66.

Haldane J (1949). Disease and evolution. La ricerca scientifica 19(Suppl 1):3–10.

Hamblin MT, Di Rienzo A (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5):1669–79.

Hamblin MT, Thompson EE, Di Rienzo A (2002). Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70(2):369–83.

Handley LJL, Manica A, Goudet J, Balloux F (2007). Going the distance: human population genetics in a clinal world. Trends Genet 23(9):432–9.

Hartl DL (2004). The origin of malaria: mixed messages from genetic diversity. Nature Reviews Microbiology 2(1):15–22.

Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, Noor AM, Kabaria CW, et al (2009). A World Malaria Map: Plasmodium falciparum Endemicity in 2007. PLoS Medicine 6(3):e1000048.

Hermisson J, Pennings PS (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–52.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, et al (2005). Whole-genome patterns of common DNA variation in three human populations. Science 307(5712):1072–9.

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6(2):95–108.

Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics 10(9):639–650.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P (2009). Genotype-imputation accuracy across worldwide human populations. American Journal of Human Genetics 84(2):235–50.

Hunt NH, Grau GE (2003). Cytokines: accelerators and brakes in the pathogenesis of cerebral malaria. Trends Immunol 24(9):491–9.

Hurles ME, Dermitzakis ET, Tyler-Smith C (2008). The functional impact of structural variation in humans. Trends Genet 24(5):238–45.

Hutagalung R, Wilairatana P, Looareesuwan S, Brittenham GM, Aikawa M, Gordeuk VR (1999). Influence of hemoglobin E trait on the severity of Falciparum malaria. J Infect Dis 179(1):283–6.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, et al (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181):998–1003.

Jallow M, Teo Y, Small K, Rockett K, Deloukas P, Clark T, Kivinen K, et al (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. Nature Genetics 41(6):657–665.

Jobling MA, Hurles M, Tyler-Smith C (2004). Human evolutionary genetics: origins, peoples and disease. New York: Garland Science.

Joy DA, Feng X, Mu J, Furuya T, Chotivanich K, Krettli AU, Ho M, et al (2003). Early origin and recent expansion of Plasmodium falciparum. Science 300(5617):318–21.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16(8):980–9.

Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, Frodsham AJ, et al (2007). A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. Nat Genet 39(4):523–8.

Kimura M (1968). Evolutionary rate at the molecular level. Nature 217(5129):624–6.

Koch O, Awomoyi A, Usen S, Jallow M, Richardson A, Hull J, Pinder M, et al (2002). IFNGR1 gene promoter polymorphisms and susceptibility to cerebral malaria. J Infect Dis 185(11):1684–7.

Koch O, Rockett K, Jallow M, Pinder M, Sisay-Joof F, Kwiatkowski D (2005). Investigation of malaria susceptibility determinants in the IFNG/IL26/IL22 genomic region. Genes Immun 6(4):312–8.

Kraemer SM, Smith JD (2006). A family affair: var genes, PfEMP1 binding, and malaria disease. Curr Opin Microbiol 9(4):374–80.

Krief S, Escalante AA, Pacheco MA, Mugisha L, André C, Halbwax M, Fischer A, et al (2010). On the diversity of malaria parasites in African apes and the origin of Plasmodium falciparum from Bonobos. PLoS Pathog 6(2):e1000765.

Kun JF, Klabunde J, Lell B, Luckner D, Alpers M, May J, Meyer C, et al (1999). Association of the ICAM-1Kilifi mutation with protection against severe malaria in Lambaréné, Gabon. Am J Trop Med Hyg 61(5):776–9.

Kwiatkowski DP (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. American Journal of Human Genetics 77(2):171–92.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al (2001). Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, et al (2008). Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241–8.

Lederberg J (1999). J. B. S. Haldane (1949) on infectious disease and evolution. Genetics 153(1):1–3.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, et al (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–4.

Louicharoen C, Patin E, Paul R, Nuchprayoon I, Witoonpanich B, Peerapittayamongkol C, Casademont I, et al (2009). Positively Selected G6PD-Mahidol Mutation Reduces Plasmodium vivax Density in Southeast Asians. Science 326(5959):1546–1549.

Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN (2005). Heritability of malaria in Africa. PLoS Medicine 2(12):e340.

Mackintosh CL, Beeson JG, Marsh K (2004). Clinical features and pathogenesis of severe malaria. Trends Parasitol 20(12):597–603.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, et al (2009). Finding the missing heritability of complex diseases. Nature 461(7265):747–53.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). The effects of human population structure on large genetic association studies. Nature Genetics 36(5):512–7.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, et al (2009). A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457(7231):877–81.

Martin MJ, Rayner JC, Gagneux P, Barnwell JW, Varki A (2005). Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of N-glycolylneuraminic acid. Proc Natl Acad Sci U S A 102(36):12819–24.

Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favourable gene. Genet Res 23(1):23–35.

McGregor IA, Wilson ME, Billewicz WZ (1983). Malaria infection of the placenta in The Gambia, West Africa; its incidence and relationship to stillbirth, birthweight and placental weight. Transactions of the Royal Society of Tropical Medicine and Hygiene 77(2):232–44.

McGuire W, Hill AV, Allsopp CE, Greenwood BM, Kwiatkowski D (1994). Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria. Nature 371(6497):508–10.

McGuire W, Knight JC, Hill AV, Allsopp CE, Greenwood BM, Kwiatkowski D (1999). Severe malarial anemia and cerebral malaria are associated with different tumor necrosis factor promoter alleles. J Infect Dis 179(1):287–90.

McVean G (2009). A Genealogical Interpretation of Principal Components Analysis. PLoS Genetics 5(10):e1000686.

Menendez C (1996). An immunological hypothesis to explain the enhanced susceptibility to malaria during pregnancy: Reply. Parasitology Today 12(1):41–42.

Menendez C, Ordi J, Ismail MR, Ventura PJ, Aponte JJ, Kahigwa E, Font F, et al (2000). The impact of placental malaria on gestational age and birth weight. Journal of Infectious Diseases 181(5):1740–5.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. Science 201(4358):786–92.

Miller L, Mason S, Clyde D, McGinniss M (1976). The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. New England Journal of Medicine 295(6):302.

Miller LH, Baruch DI, Marsh K, Doumbo OK (2002). The pathogenic basis of malaria. Nature 415(6872):673–9.

Min-Oo G, Fortin A, Tam MF, Gros P, Stevenson MM (2004). Phenotypic expression of pyruvate kinase deficiency and protection against malaria in a mouse model. Genes Immun 5(3):168–75.

Min-Oo G, Fortin A, Tam MF, Nantel A, Stevenson MM, Gros P (2003). Pyruvate kinase deficiency in mice protects against malaria. Nat Genet 35(4):357–62.

Mockenhaupt FP, Cramer JP, Hamann L, Stegemann MS, Eckert J, Oh NR, Otchwemah RN, et al (2006). Toll-like receptor (TLR) polymorphisms in African children: Common TLR-4 variants predispose to severe malaria. Proc Natl Acad Sci U S A 103(1):177–82.

Mockenhaupt FP, Ehrhardt S, Cramer JP, Otchwemah RN, Anemana SD, Goltz K, Mylius F, et al (2004a). Hemoglobin C and resistance to severe malaria in Ghanaian children. J Infect Dis 190(5):1006–9.

Mockenhaupt FP, Ehrhardt S, Gellert S, Otchwemah RN, Dietz E, Anemana SD, Bienzle U (2004b). Alpha(+)-thalassemia protects African children from severe malaria. Blood 104(7):2003–6.

Mockenhaupt FP, Mandelkow J, Till H, Ehrhardt S, Eggelte TA, Bienzle U (2003). Reduced prevalence of Plasmodium falciparum infection and of concomitant anaemia in pregnant women with heterozygous G6PD deficiency. Trop Med Int Health 8(2):118–24.

Modiano D, Luoni G, Sirima BS, Simporé J, Verra F, Konaté A, Rastrelli E, et al (2001). Haemoglobin C protects against clinical Plasmodium falciparum malaria. Nature 414(6861):305–8.

Moore JH (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56(1-3):73–82.

Mu J, Duan J, Makova KD, Joy DA, Huynh CQ, Branch OH, Li WH, et al (2002). Chromosome-wide SNPs reveal an ancient origin for Plasmodium falciparum. Nature 418(6895):323–6.

Mutabingwa TK, Bolla MC, Li JL, Domingo GJ, Li X, Fried M, Duffy PE (2005). Maternal malaria and gravidity interact to modify infant susceptibility to malaria. PLoS Medicine 2(12):e407.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–4.

Myers S, Freeman C, Auton A, Donnelly P, McVean G (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet 40(9):1124–9.

Nielsen R (2005). Molecular signatures of natural selection. Annual Review of Genetics 39:197–218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007). Recent and ongoing selection in the human genome. Nature Reviews Genetics 8(11):857–68.

Novembre J, Di Rienzo A (2009). Spatial patterns of variation due to natural selection in humans. Nature Reviews Genetics 10(11):745–755.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, et al (2008). Genes mirror geography within Europe. Nature 456(7218):98–101.

Novembre J, Stephens M (2008). Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40(5):646–9.

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, et al (2004). Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am J Hum Genet 74(6):1198–208.

Pääbo S (2003). The mosaic that is our genome. Nature 421(6921):409–12.

Pain A, Urban BC, Kai O, Casals-Pascual C, Shafi J, Marsh K, Roberts DJ (2001). A non-sense mutation in Cd36 gene is associated with protection from severe malaria. Lancet 357(9267):1502–3.

Paré G, Chasman DI, Kellogg M, Zee RYL, Rifai N, Badola S, Miletich JP, et al (2008). Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. PLoS Genet 4(7):e1000118.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. PLoS Genetics 2(12):e190.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, et al (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19(5):826–37.

Pritchard JK, Pickrell JK, Coop G (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20(4):R208–15.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. Genetics 155(2):945–59.

Przeworski M, Coop G, Wall JD (2005). The signature of positive selection on standing genetic variation. Evolution 59(11):2312–23.

Råberg L, Sim D, Read AF (2007). Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. Science 318(5851):812–4.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009). Reconstructing Indian population history. Nature 461(7263):489–494.

Rich SM, Leendertz FH, Xu G, LeBreton M, Djoko CF, Aminake MN, Takang EE, et al (2009). The origin of malignant malaria. Proceedings of the National Academy of Sciences of the United States of America 106(35):14902–14907.

Rich SM, Licht MC, Hudson RR, Ayala FJ (1998). Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum. Proc Natl Acad Sci U S A 95(8):4425–30.

Riley EM, Wahl S, Perkins DJ, Schofield L (2006). Regulating immunity to malaria. Parasite Immunology 28(1-2):35–49.

Rogerson SJ, Grau GE, Hunt NH (2004). The microcirculation in severe malaria. Microcirculation 11(7):559–76.

Rogerson SJ, Hviid L, Duffy PE, Leke RF, Taylor DW (2007). Malaria in pregnancy: pathogenesis and immunity. Lancet Infectious Diseases 7(2):105–17.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002). Genetic structure of human populations. Science 298(5602):2381–5.

Rowe JA, Handel IG, Thera MA, Deans AM, Lyke KE, Koné A, Diallo DA, et al (2007). Blood group O protects against severe Plasmodium falciparum malaria through the mechanism of reduced rosetting. Proc Natl Acad Sci U S A 104(44):17471–6.

Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, Gupta S, Warn P, et al (1995). Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376(6537):246–9.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, et al (2002a). Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–7.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, et al (2006). Positive natural selection in the human lineage. Science 312(5780):1614–20.

Sabeti PC, Usen S, Farhadian S, Jallow M, Doherty T, Newport M, Pinder M, et al (2002b). CD40L association with protection from severe malaria. Genes and Immunity 3(5):286–91.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, et al (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918.

Saunders MA, Hammer MF, Nachman MW (2002). Nucleotide variability at G6pd and the signature of malarial selection in humans. Genetics 162(4):1849–61.

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (2005). The extent of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171(3):1219–29.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Research 15(11):1576–83.

Schaid DJ (2004). Evaluating associations of haplotypes with traits. Genetic Epidemiology 27(4):348–64.

Scheet P, Stephens M (2008). Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS Genetics 4(8):e1000147.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, et al (2007). Challenges and standards in integrating surveys of structural variation. Nat Genet 39(7 Suppl):S7–15.

Scherf A, Lopez-Rubio JJ, Riviere L (2008). Antigenic variation in Plasmodium falciparum. Annu Rev Microbiol 62:445–70.

Steketee RW, Nahlen BL, Parise ME, Menendez C (2001). The burden of malaria in pregnancy in malaria-endemic areas. American Journal of Tropical Medicine and Hygiene 64(1-2 Suppl):28–35.

Steketee RW, Wirima JJ, Slutsker L, Heymann DL, Breman JG (1996). The problem of malaria and malaria control in pregnancy in sub-Saharan Africa. American Journal of Tropical Medicine and Hygiene 55(1 Suppl):2–7.

Stevenson MM, Riley EM (2004). Innate immunity to malaria. Nat Rev Immunol 4(3):169–80.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, et al (2007). Population genomics of human gene expression. Nat Genet 39(10):1217–24.

Stubbs J, Simpson KM, Triglia T, Plouffe D, Tonkin CJ, Duraisingh MT, Maier AG, et al (2005). Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes. Science 309(5739):1384–7.

Sturm A, Amino R, van de Sand C, Regen T, Retzlaff S, Rennenberg A, Krueger A, et al (2006). Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. Science 313(5791):1287–90.

Tachmazidou I, Verzilli CJ, De Iorio M (2007). Genetic association mapping via evolution-based clustering of haplotypes. PLoS Genet 3(7):e111.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–95.

Tang H, Peng J, Wang P, Risch NJ (2005). Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28(4):289–301.

Tang K, Thornton KR, Stoneking M (2007). A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biology 5(7):e171.

Teshima KM, Coop G, Przeworski M (2006). How reliable are empirical genomic scans for selective sweeps? Genome Res 16(6):702–12.

The HUGO Pan-Asian SNP Consortium (2009). Mapping Human Genetic Diversity in Asia. Science 326(5959):1541–1545.

The International HapMap Consortium (2003). The International HapMap Project. Nature 426(6968):789–96.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437(7063):1299–320.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–78.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, et al (2009). The genetic structure and history of Africans and African Americans. Science 324(5930):1035–44.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, et al (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455–62.

Tishkoff SA, Verrelli BC (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340.

Tjitra E, Anstey NM, Sugiarto P, Warikar N, Kenangalem E, Karyana M, Lampah DA, et al (2008). Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia. PLoS Med 5(6):e128.

Tournamille C, Colin Y, Cartron JP, Le Van Kim C (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 10(2):224–8.

Trampuz A, Jereb M, Muzlovic I, Prabhu RM (2003). Clinical review: Severe malaria. Crit Care 7(4):315–23.

Varki A, Gagneux P (2009). Human-specific evolution of sialic acid targets: Explaining the malignant malaria mystery? Proceedings of the National Academy of Sciences of the United States of America 106(35):14739–14740.

Vaughan AM, Aly ASI, Kappe SHI (2008). Malaria parasite pre-erythrocytic stage infection: gliding and hiding. Cell Host Microbe 4(3):209–18.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al (2001). The sequence of the human genome. Science 291(5507):1304–51.

Verra F, Mangano VD, Modiano D (2009). Genetics of susceptibility to Plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. Parasite Immunol 31(5):234–53.

Verrelli BC, Tishkoff SA, Stone AC, Touchman JW (2006). Contrasting histories of G6PD molecular evolution and malarial resistance in humans and chimpanzees. Mol Biol Evol 23(8):1592–601.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4(10):e1000214.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006). A map of recent positive selection in the human genome. PLoS Biology 4(3):e72.

Volkman SK, Barry AE, Lyons EJ, Nielsen KM, Thomas SM, Choi M, Thakore SS, et al (2001). Recent origin of Plasmodium falciparum from a single progenitor. Science 293(5529):482–4.

Wambua S, Mwangi TW, Kortok M, Uyoga SM, Macharia AW, Mwacharo JK, Weatherall DJ, et al (2006). The effect of alpha+-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya. PLoS Med 3(5):e158.

Wang HY, Tang H, Shen CKJ, Wu CI (2003). Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Mol Biol Evol 20(11):1795–804.

Weatherall DJ (2001). Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2(4):245–55.

Williams TN (2006). Human red blood cell polymorphisms and malaria. Curr Opin Microbiol 9(4):388–94.

Williams TN, Mwangi TW, Wambua S, Alexander ND, Kortok M, Snow RW, Marsh K (2005a). Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. J Infect Dis 192(1):178–86.

Williams TN, Mwangi TW, Wambua S, Peto TEA, Weatherall DJ, Gupta S, Recker M, et al (2005b). Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. Nat Genet 37(11):1253–7.

Williams TN, Wambua S, Uyoga S, Macharia A, Mwacharo JK, Newton CRJC, Maitland K (2005c). Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. Blood 106(1):368–71.

Wood ET, Stover DA, Slatkin M, Nachman MW, Hammer MF (2005). The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. Am J Hum Genet 77(4):637–42.

World Health Organization (2000). Severe falciparum malaria. Trans R Soc Trop Med Hyg 94 Suppl 1:S1–90.

World Health Organization (2009). World Malaria Report 2009. Geneva, Switzerland: World Health Organization.

Zhang J, Rowe WL, Clark AG, Buetow KH (2003). Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet 73(5):1073–81.

Zondervan KT, Cardon LR (2004). The complex interplay among factors that influence allelic association. Nat Rev Genet 5(2):89–100.

Appendix A

Supplementary information for results

A.1 Supplementary information for Section 3.1

Supplementary methods

Validation of nHLS method

To assess whether our modified LRH method can indeed detect selection, we analyzed the *LCT* region in the CEU and the *HBB* region in the YRI sample as positive control, both being well-established examples of genes under positive selection in humans. *LCT* codes for lactase, the enzyme responsible for the ability to digest the milk sugar lactose. Most adults lose this ability due to decreasing enzyme activity after childhood, but carriers of a *LCT* variant causing lactase persistence retain it, particularly in cattle-domesticating populations. This variant has been shown to be under selection in Europeans [1, 2], most likely due to the advantage gained from the ability to feed on dairy products. Within the analyzed 2 Mb region centered on *LCT*, the SNP that showed the strongest association with lactase persistence, rs4988235 [3], has a nHLS of -4.48. Although not the most extreme score in the region, it is ranked at position 15 of all scores in the region. Additionally, the whole region displays very high proportions of extreme scores, a clear indicator of the strong selective pressure in the region (see Figure S6). The second region that was chosen as positive control is *HBB*, where a non-synonymous SNP (rs334) is responsible for the hemoglobin S variant of hemoglobin B, which causes sickle cell disease. This variant is regarded as the textbook example of a locus under selection in areas of high malaria transmission [4, 5], due to the roughly 10-fold reduction of risk of severe malaria for heterozygote carriers [6]. The SNP rs334 showed an nHLS of 3.49, which is the maximum score we observed for any non-synonymous SNP in the YRI sample. It is worth noting that in this case the surrounding region did not show particularly high proportions of high score SNPs (data not shown), indicating that it is important to also incorporate additional information like functional status of the SNP into the analysis.

References

1. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN: **Genetic signatures of strong recent positive selection at the lactase gene**. *Am J Hum Genet* 2004, **74**(6):1111-1120.
2. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome**. *PLoS biology* 2006, **4**(3):e72.

3. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I: **Identification of a variant associated with adult-type hypolactasia**. *Nat Genet* 2002, **30**(2):233-237.

4. Kwiatkowski DP: **How malaria has affected the human genome and what human genetics can teach us about malaria**. *Am J Hum Genet* 2005, **77**(2):171-192.

5. Tishkoff SA, Williams SM: **Genetic analysis of African populations: human evolution and complex disease**. *Nature reviews* 2002, **3**(8):611-621.

6. Ackerman H, Usen S, Jallow M, Sisay-Joof F, Pinder M, Kwiatkowski DP: **A comparison of case-control and family-based association methods: the example of sickle-cell and malaria**. *Annals of human genetics* 2005, **69**(Pt 5):559-565.

Supplementary figures

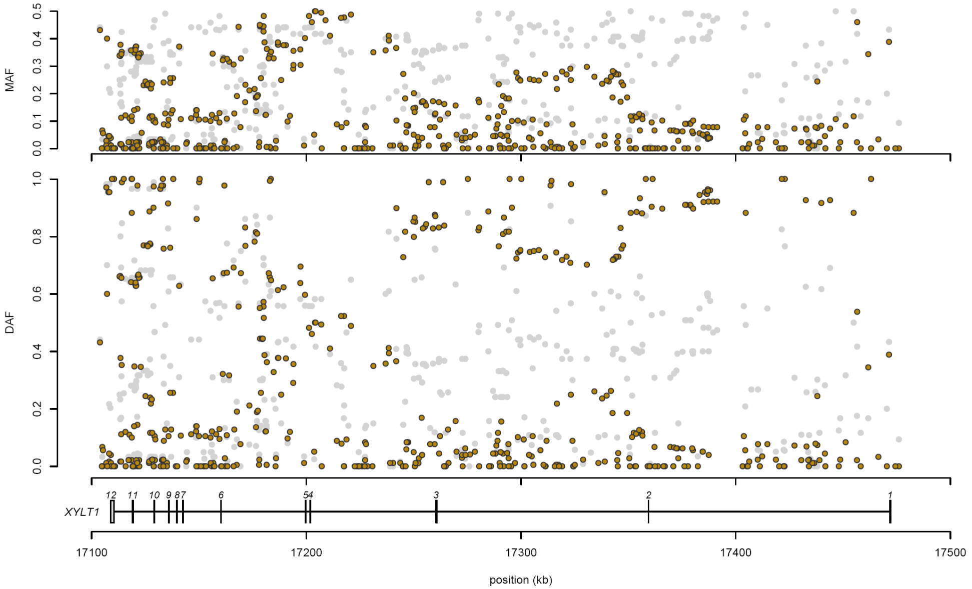


Figure S1 – Distribution of allele frequencies in *XYLT1*
Minor allele frequency (MAF, top) and derived allele frequency (DAF, bottom) for each SNP are plotted over the region of *XYLT1*. Colored points indicate SNPs in the JC (Japanese and Chinese) sample, while grey points indicate CEU (Central Europeans) SNPs. We chose to use CEU as a comparison due to their similar overall distribution of allele frequencies (see Figure 1). To note is the apparent lack of intermediate allele frequencies from exons 1 to 3 in JC, which is not observed in the CEU.

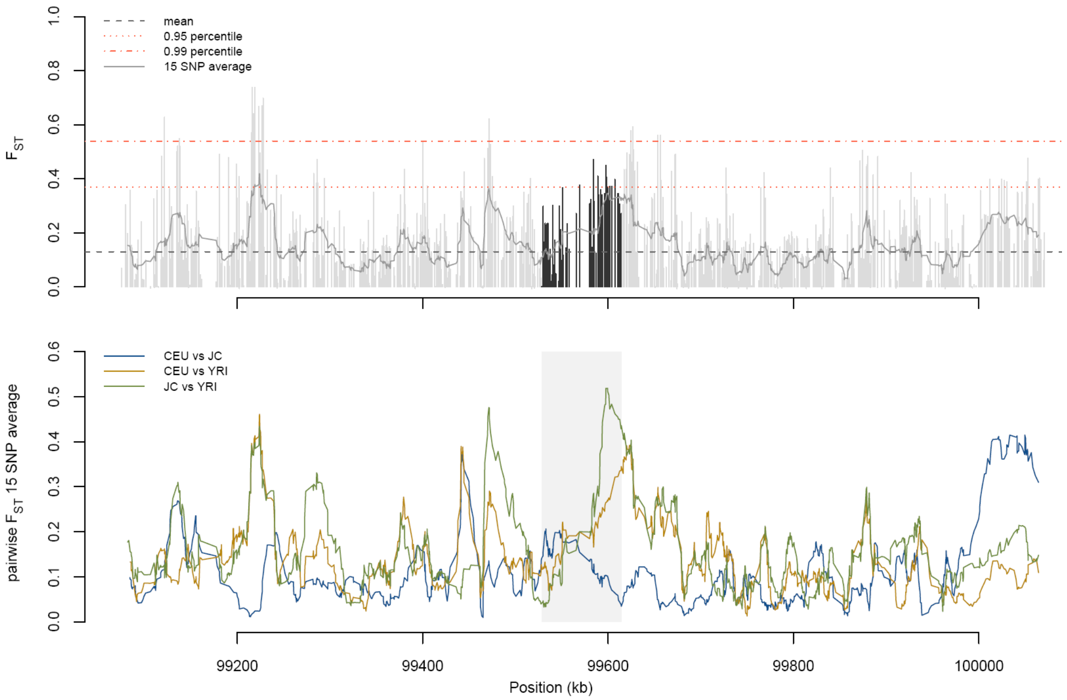


Figure S2 – Distribution of F_{ST} along CHSY1 region

(top) global F_{ST} for all SNPs (vertical bars) with lines indicating mean F_{ST}, 0.95 and 0.99 percentile as well as average F_{ST} for 15 SNP sliding windows; *(bottom)* 15 SNP sliding window average F_{ST} for the three pairwise populations comparisons. Shown are results over the full 1 Mb regions, with the region of the gene ±5 kb indicated through darker bars (top) or grey background (bottom).

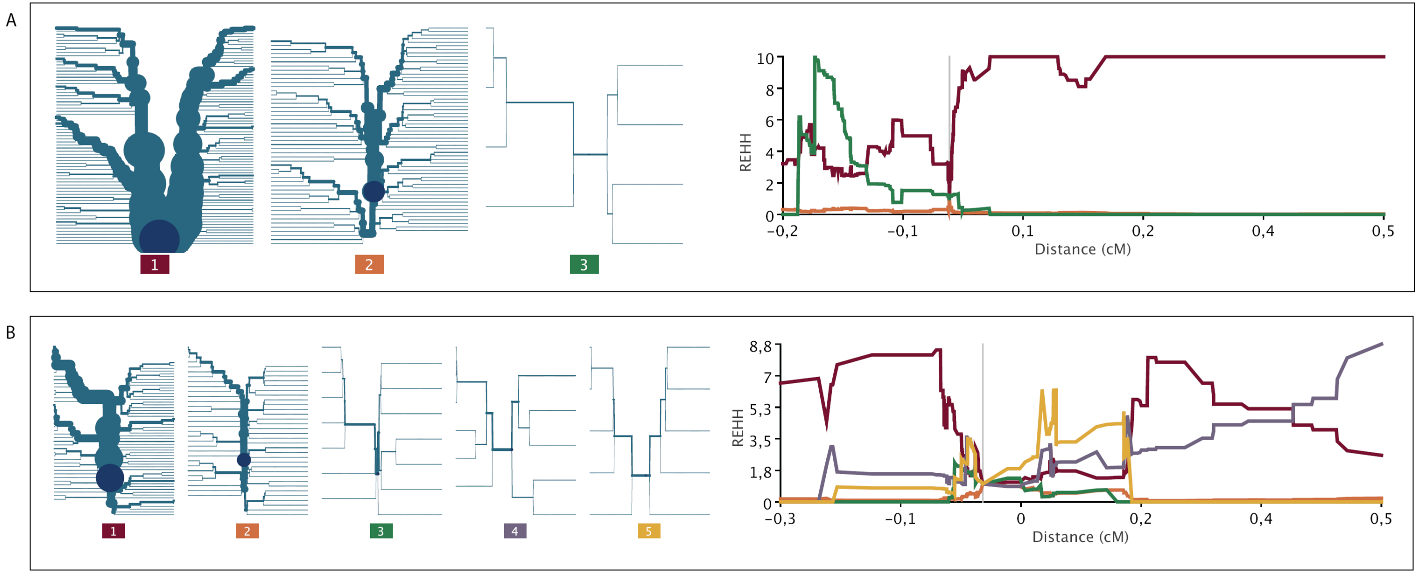


Figure S3 – Haplotype bifurcation and REHH decay plots for two core regions
Shown are haplotype bifurcation plots (left) as well as REHH decay (right) over genetic distance for all different core haplotypes observed at the position of two significant cores, XYLT1-1 in JC (A, core number 1) and UST-1 in YRI (B, core number 1). Note that XYLT1-1 only shows a strong signal in the + direction (to the right)

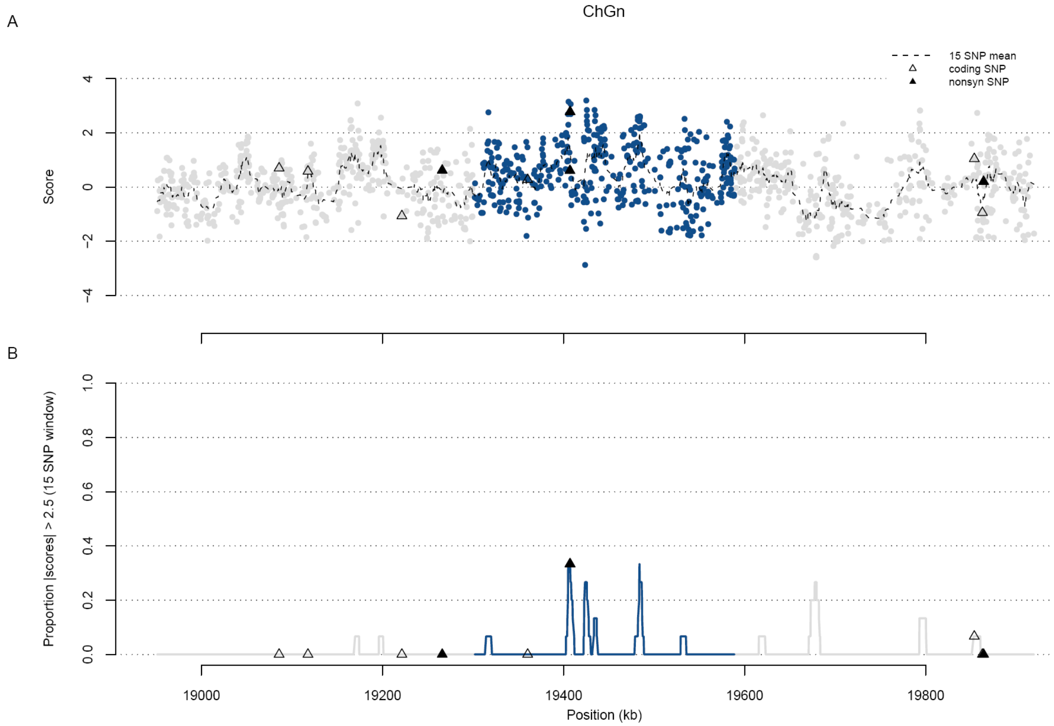


Figure S4 – nHL score profile over the *ChGn* region in YRI

nHL scores (A) as well as proportion of high scores (B) in the 1 Mb region centered on *ChGn*. SNPs within ±5 kb of the gene are indicated in blue. To note is the high scoring non-synonymous SNP (rs17128518) within a narrow cluster of high scores.

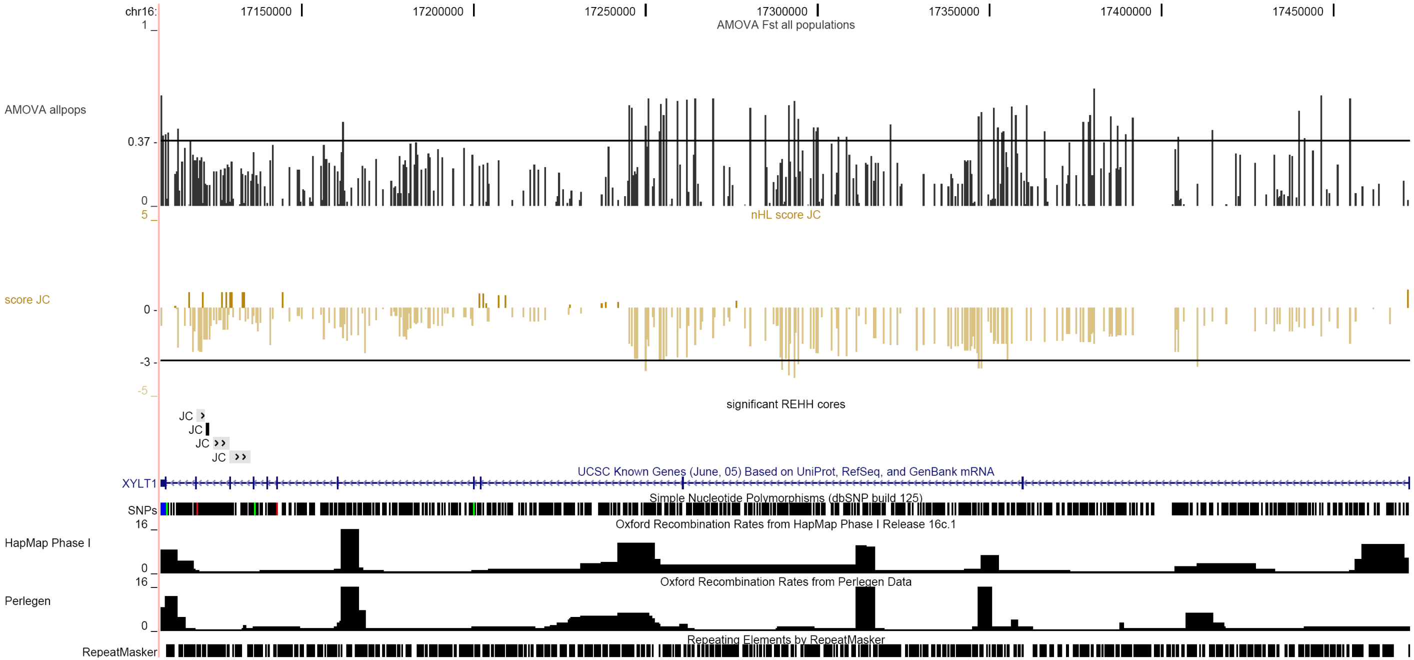


Figure S5 – Plot of the *XYLT1* gene from UCSC genome browser

A plot of the *XYLT1* gene obtained from the UCSC genome browser, showing gene structure, predicted recombination rates, as well as custom tracks with F_{ST} (AMOVA F_{ST} all populations), nHL score (nHL score JC) as well as position of significant core haplotypes in REHH analysis (significant REHH cores).

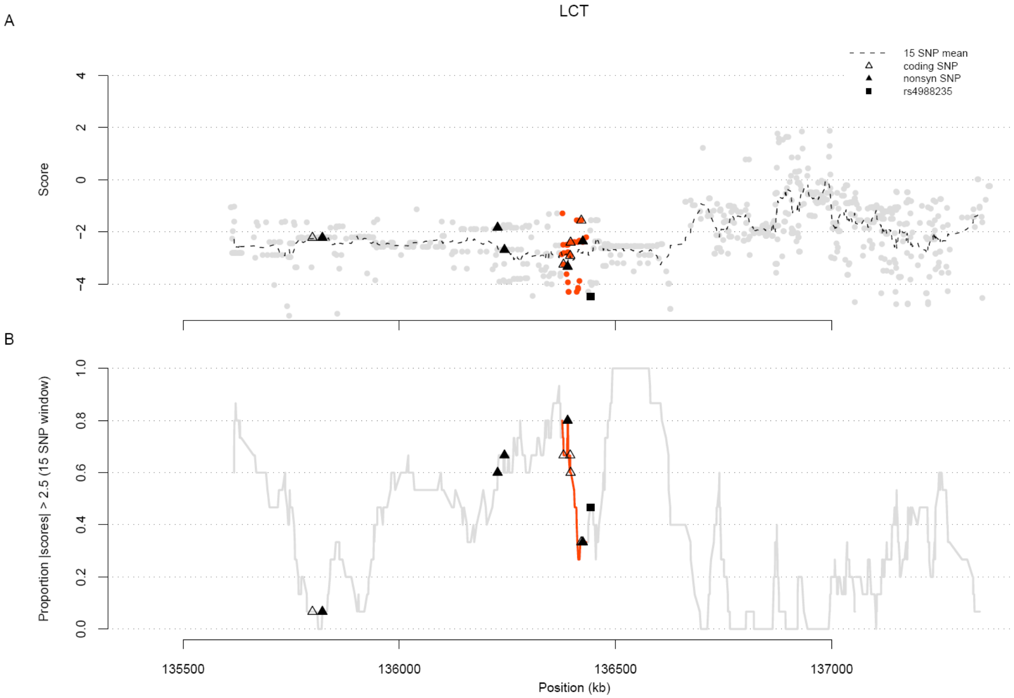


Figure S6 – nHL score profile over the LCT region in CEU
(A) Grey dots indicate scores for each SNP along a 2 Mb region centered on *LCT*, while coloured dots indicate SNPs within ±5 kb of the gene. The black square represents rs4988235, which show the strongest association with lactase persistence. (B) Lines show the proportion of SNPs with an absolute score > 2.5. To note is the high proportion of high score SNPs over most of the region.

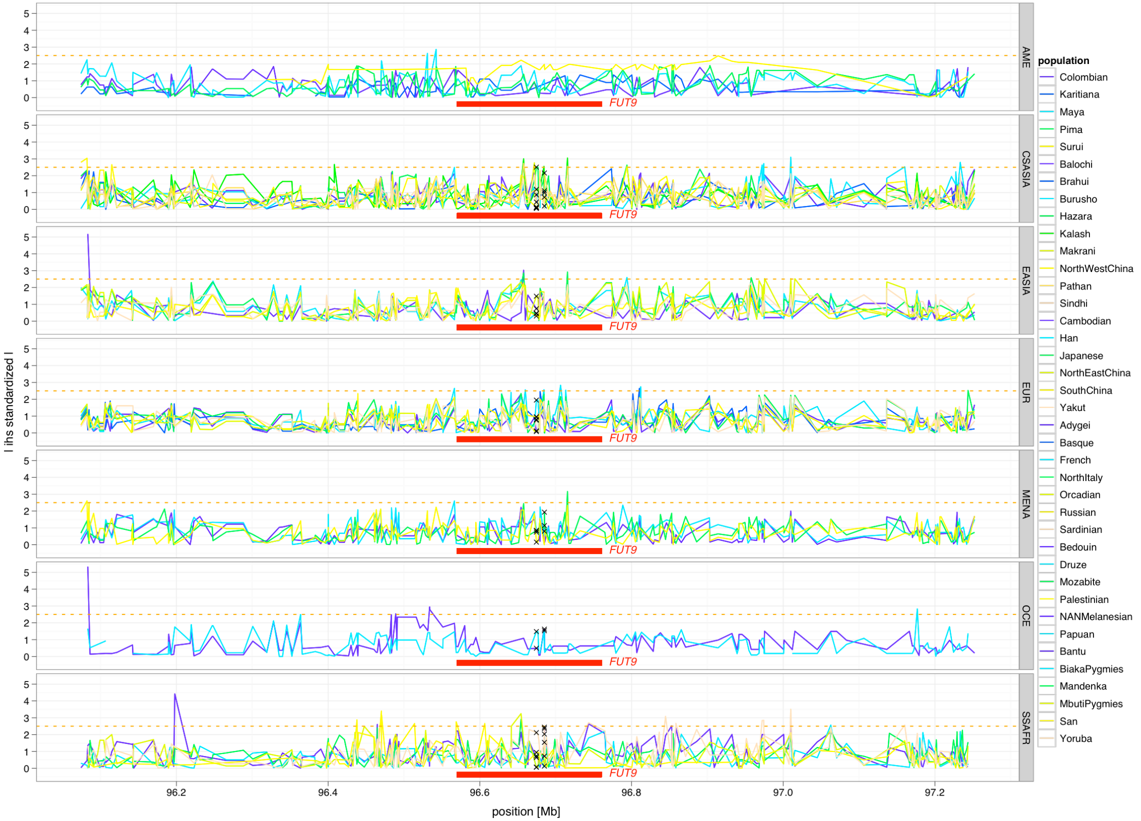
Supplementary tables

Table S1. List of candidate genes and their function

Locus	kg ID[a]	Name	Pathway	Molecular function[b]
B3GALT6	NM_080605	UDP-Gal:betaGal beta 1,3-galactosyltransferase polypeptide 6	Chondroitin sulphate biosynthesis	Glycosyltransferase
B3GAT1	NM_054025	Beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)	Chondroitin sulphate biosynthesis	Glycosyltransferase
B3GAT2	NM_080742	Beta-1,3-glucuronyltransferase 2 (glucuronosyltransferase S)	Chondroitin sulphate biosynthesis	Glycosyltransferase
B3GAT3	NM_012200	Beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)	Chondroitin sulphate biosynthesis	Glycosyltransferase
B4GALT7	NM_007255	Xylosylprotein beta 1,4-galactosyltransferase, polypeptide 7 (galactosyltransferase I)	Chondroitin sulphate biosynthesis	Glycosyltransferase
ChGn	NM_018371	Chondroitin beta1,4 N-acetylgalactosaminyltransferase	Chondroitin sulphate biosynthesis	Glycosyltransferase
CHPF	NM_024536	Chondroitin polymerizing factor	Chondroitin sulphate biosynthesis	Glycosyltransferase
CHST11	AB042326	Carbohydrate (chondroitin 4) sulfotransferase 11	Chondroitin sulphate biosynthesis	Other transferase
CHST12	NM_018641	Carbohydrate (chondroitin 4) sulfotransferase 12	Chondroitin sulphate biosynthesis	Other transferase
CHST13	NM_152889	Carbohydrate (chondroitin 4) sulfotransferase 13	Chondroitin sulphate biosynthesis	Other transferase
CHST3	NM_004273	Carbohydrate (chondroitin 6) sulfotransferase 3	Chondroitin sulphate biosynthesis	Other transferase
CHST7	NM_019886	Carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7	Chondroitin sulphate biosynthesis	Other transferase
CHSY1	NM_014918	Carbohydrate (chondroitin) synthase 1	Chondroitin sulphate biosynthesis	Glycosyltransferase
CSGlcA-T	NM_019015	Chondroitin sulfate glucuronyltransferase	Chondroitin sulphate biosynthesis	Glycosyltransferase
CSS3	AJ578034	Chondroitin sulfate synthase 3	Chondroitin sulphate biosynthesis	Glycosyltransferase
D4ST1	NM_130468	Dermatan 4 sulfotransferase 1	Chondroitin sulphate biosynthesis	Other transferase
GALNAC4S-6ST	NM_015892	B cell RAG associated protein	Chondroitin sulphate biosynthesis	Other transferase
GALNACT-2	NM_018590	Chondroitin sulfate GalNAcT-2	Chondroitin sulphate biosynthesis	Glycosyltransferase
UST	NM_005715	Uronyl-2-sulfotransferase	Chondroitin sulphate biosynthesis	Synthase; Glycosyltransferase
XYLT1	NM_022166	Xylosyltransferase I	Chondroitin sulphate biosynthesis	Synthase; Glycosyltransferase
XYLT2	NM_022167	Xylosyltransferase II	Chondroitin sulphate biosynthesis	Synthase; Glycosyltransferase
HAS1	NM_001523	Hyaluronan synthase 1	Hyaluronic acid biosynthesis	Other transferase
HAS2	NM_005328	Hyaluronan synthase 2	Hyaluronic acid biosynthesis	Glycosyltransferase
HAS3	NM_005329	Hyaluronan synthase 3 (Isoform a)	Hyaluronic acid biosynthesis	Glycosyltransferase

[a] UCSC Genome Browser known Gene ID
[b] from PANTHER Database

A.2 Supplementary information for Section 3.2



Supplementary Figure 1. iHs statistic for the *FUT9* region**.**

Shown are the absolute values for iHs over the region containing *FUT9* calculated from the 650k Illumina array data on the HGDP. Populations are ordered according to continental groups as indicated in the panels (AME Americas; CSASIA Central-South Asia; EASIA East Asia; EUR Europe; MENA Middle East & Northern Africa; OCE Oceania; SSAFR Sub-Saharan Africa). Black crosses indicate the values of the two SNPs included in the HGDP dataset (rs4452646, rs4555922). Red dashed line indicates iHs > 2.5.

Appendix B

Contributions to other articles

B.1 Human adaptation of antibacterial innate immunity genes

Ferran Casals, **Martin Sikora**, Hafid Laayouni, Ludovica Montanucci, Philip Awadalla, Mihai G. Netea, Jaume Bertranpetit

Manuscript in preparation

Human adaptation of antibacterial innate immunity genes

Ferran Casals[1,2] ,Martin Sikora[1], Hafid Laayouni[1,3], Ludovica Montanucci[1], Philip Awadalla[2], Mihai G. Netea[4], and Jaume Bertranpetit[1,3,§]

[1] Institute of Evolutionary Biology (UPF-CSIC), CEXS – UPF – PRBB, Barcelona, Catalonia, Spain

[2] Centre de Recherche, CHU Sainte-Justine, Université de Montréal, Montréal, Québec H3T 1C5, Canada

[3] CIBER en Epidemiología y Salud Pública (CIBERESP, Spain)

[4] Department of Medicine, Radboud University Nijmegen Medical Center, 6500 HB Nijmegen, The Netherlands

AUTHOR FOR CORRESPONDENCE:

Ferran Casals:
ferran.casals.lopez@umontreal.ca

Introduction

Infectious diseases are among the most important selective agents for any vertebrate species. In humans, they have represented a great challenge for the adaptation to new environments and social habits, with increasing population densities and cattle domestication favouring their emergence and spread in the last tens of thousands of years. The human immune system would then have played a main role in the adaption of the different populations to the emerging conditions. According to that expectation, genomic scans for signatures of adaptive selection have revealed that immune function is one of the classes enriched with genes under balancing or positive selection (BUSTAMANTE *et al.* 2005; NIELSEN *et al.* 2005; SABETI *et al.* 2006), the two evolutionary forces underlying adaptation.

Vertebrate's immune function can be divided in the adaptive, exclusive of this phylogenetic group, and the innate immune system. Innate immunity constitutes the first barrier of defence, and acts in a semi-specific way by recognizing pathogen-associated molecular patterns (PAMPs), which are essential and conserved components of the pathogens. Several works have reported evidences of positive or balancing selection acting on this genes (BARREIRO *et al.* 2005; CAGLIANI *et al.* 2008; FERRER-ADMETLLA *et al.* 2008; FUMAGALLI *et al.* 2009a; FUMAGALLI *et al.* 2009b), although some of this genes are not unequivocally assigned to the innate system.

In this manuscript, we address the analysis of the footprint of the adaptive selection in the innate immune mechanisms involved in antibacterial host defense. We have chosen to assess innate immunity to bacteria, as they are probably the most important human pathogens, with a major impact on morbidity and mortality. The same mechanisms are mostly also common to most human parasites, and in part to fungi and viruses. However, two major classes of gene families involved in the host defense to fungi (C-type lectins) and viruses (RiG-I helicases) have not been assessed in the present study. Beyond its molecular signatures in particular genes, we focus on the adaptation of the functional network, seeking whether human adaptation at the level of innate immunity system has occurred preferentially on some functional class or according to the molecular physiology, with more plasticity to respond to the pathogenic pressures. Previous work showed the existence of different selective pressures in the innate immune system of *Drosophila*, where adaptation seems to have mainly occurred at intracellular signalling molecules and purifying selection has acted on the receptors and secreted antimicrobial peptides (LAZZARO 2008). Similarly, strong purifying selection has also been reported to act in the *toll like* receptors genes (TLR) in humans (BARREIRO *et al.* 2009; MUKHERJEE *et al.* 2009), even if strong purifying selection is the rule in protein coding genes.

Recent times have seen an increased move towards characterizing the evolutionary dynamics of proteins in the context of their functional networks. The availability of genome wide data on biological networks, for example in the form of experimentally derived data sets on protein-protein interactions, or curated databases of functional pathways, have facilitated studies relating evolutionary parameters to network topology (ALVAREZ-PONCE *et al.* 2009; KIM *et al.* 2007). We therefore have also characterized this set of genes in a network context, where it is of interest to understand the evolutionary and adaptive process in a complexity framework in a setting towards evolutionary system biology.

The final goal of this paper is to unravel selective forces related to immunological responses to bacterial infections, having acted in humans; the approach is based in detecting adaptations through natural selection uncovered in resequencing information in two populations (Europeans and Africans) and the results are interpreted in a full functional context of gene products interacting in well identified networks.

Materials and Methods

Selection of genes and DNA sequences

As the final goal is to study as many genes as possible involved in innate immunity response, an initial list of genes were obtained from several databases, including KEEG, Reactome and Pathways-on-line. Not all genes could be included in the analysis, even if their place in the pathway was preserved.

Resequencing data for 132 were retrieved from different sources. 62 genes were from the Innate Immunity Program in Genomics Applications (IIPGA, http://innateimmunity.net) (LAZARUS *et al.* 2002), only considering those genes with available information on the sequencing primers used. In addition, data for seven genes was obtained from Environmental Genome Project database (NIEHS SNPs, http://egp.gs.washington.edu), and for 54 from the SeattleSNPs database (http://pga.gs.washington.edu). For the analysis, we also included the previously published results (FERRER-ADMETLLA *et al.* 2008) for eight additional genes from Innate Immunity Program in Genomics Applications and one from the SeattleSNPs database. All genes from the Innate Immunity Program in Genomics Applications, and 44 from SeattleSNPs were resequenced in the same 23 European-American and 24 African-American samples included in the Coriell CEPH / African American panel. The remaining 11 genes from Seattle SNPs were resequenced in 23 European (HapMap CEU) and 24 African individuals (HapMap YRI) individuals. For the seven genes from the Environmental Genome Project database we retrieved resequencing data in 22 European (HapMap CEU) and 15 Coriell African American individuals. Chimpanzee sequences were obtained from the GenBank (http://www.ncbi.nlm.nih.gov/) and Ensembl databases (http://www.ensembl.org/index.html). Sequence alignments were performed with ClustalW (THOMPSON *et al.* 1994).

Molecular Data Analysis

The following diversity statistics and neutrality tests were calculated for every gene: heterozygosity or π (TAJIMA 1983), Tajima's D (TAJIMA 1989), Fu and Li's F*, D*, F and D (FU and LI 1993), and Fay Wu's H (FAY and WU 2000) using DnaSP v5 (LIBRADO and ROZAS 2009). Indels and triallelic positions were not included in the analyses. The significance of these tests was calculated by means of coalescent simulations using the COSI software (SCHAFFNER *et al.* 2005), using the model which takes into consideration the demographic history of humans. 1,000 replicates were performed by using the local recombination rate estimate of each region obtained from HapMap (www.hapmap.org). The orthologous chimpanzee sequence was obtained (*panTro2*).

Network analyses

We used the MiMI plugin (GAO *et al.* 2009) for Cytoscape (SHANNON *et al.* 2003) to retrieve all known interactions among the genes in our dataset. Network statistics for this network were calculated using the NetworkAnalyzer plugin (ASSENOV *et al.* 2008) in Cytoscape.

Results

Gene Selection

We have analyzed publicly available resequencing data for 132 genes involved in innate immunity in African and European ancestry individuals (see Materials and Methods). 129 genes are autosomal and only three genes (IRAK1, TLR7 and TLR8) are located in the X chromosomes. All these genes have an unequivocally role in innate immunity against bacteria, and were classified according to their main function in this system into five categories: receptors, adaptors, modulators, cytokines and effectors molecules (Figure 1). We have also defined some subclasses in the case of the receptors (TLR2, TLR4, TLR5, NOD1/2, TLR3 and TLR7-9 modules), which can be differentiated according to their location at the plasmatic (the first four) or at the nuclear membrane (the remaining two), also differentiated for the ligand that they recognize. Cytokines have been classified according to their involvement in different immune processes (acute phase, cellular immunity, anti-inflammatory cytokines, neutrophil function, chemokines, extracellular bacterial parasites, and complement modules; see Figure1 and TableS1for a complete list). These in-depth structure allow a much comprehensive analysis of innate immunity function and will help the network approach.

Nucleotide diversity is reduced in adaptor proteins

Nucleotide diversity levels can provide a measure of the degree of conservation of the different genes and their comparison across functional classes can shed light on different intensities of purifying selection. We estimated the nucleotide diversity measured as the average number of differences between pairs of sequences (π) for the 132 genes in each population, and compared those levels of diversity among the five functional categories.

As expected and in agreement with previous reports, nucleotide diversity levels are higher in Africans than in Europeans (paired t-test $P < 0.001$), as seen by the higher number of points below the diagonal in Figure 2. When the different functional classes are compared, mainly the adaptor but also the complement proteins show a lower level of diversity compared to the rest of genes (Table 1). This difference is significant in the case of the adaptor proteins in Africans (unpaired t-test $P = 0.03$).

Nucleotide diversity within a network approach

Alternatively, changes in nucleotide diversity levels can be analyzed from a system perspective, considering the relationship of each gene product with the rest of the gene products, which finally defines a functional network. We carried out the network analysis at three levels: first, considering all interaction described in any of the protein-protein interactions databases with a total of 101 genes (out of the 132 total; for the remaining genes no interactions are annotated) and 279 interactions in total (Figure 3); second, focussing on the protein-protein interaction network underlying our functional network, and third considering the signalling network of Fig 1.

Considering the whole set of interaction, the degree distribution of our innate immunity protein network follows a power-law distribution, as previously observed in many other examples of biological networks ($r^2 = 0.80$ Figure S1) (BARABASI and OLTVAI 2004; YUAN *et al.* 2008). Here we show that also the considered immunity network, which is a sub-network of the whole protein-protein interaction web, is characterized by the scale-free architecture peculiar of most biological networks. This architecture is known to arise from an evolution process that follows the two mechanisms of growth (new nodes are added over time) and preferential attachment (new added nodes have a greater probability of being attached to already highly connected nodes) (Barabasi and Albert, 1999).

A frequent finding emerging from previous studies on the influence of network topology on evolutionary rate is the observation that highly connected proteins (“hubs“) are more constrained in their evolution (BARABASI and OLTVAI 2004). In order to test whether this is also true for the innate immunity protein network, we investigated the correlations between nucleotide diversity and two measures of network centrality: the degree centrality (defined as the number of links incident upon a node) as well as the betweenness centrality of the respective proteins or nodes (nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not). As a general observation, there is a negative correlation for both indices in both populations, and the correlations tend to be stronger in Africa than in Europe. The highest value for Africa could simply be a consequence of the much greater variability in African populations, leading to increased power of detection in this population. However, none of the correlations were significant, probably due to a lack of power given the small network size. Still, both African and European populations show a negative correlation coefficient of nucleotide diversity with degree (Figure S2), which would be consistent with the expectation of higher conservation of highly connected proteins.

In order to determine whether connected proteins were more likely to evolve under similar pressure, we performed a permutation procedure (similar that described in ALVAREZ-PONCE *et al.* 2009). Namely, for each connected protein pair, we calculated the absolute difference in their nucleotide diversities (Δ_{π}), and compared the mean of this difference to the mean of each of 10,000 permuted networks. The permuted networks were generated by randomly rewiring the nodes of the original network with the same number of edges. Both, African and European populations did not show significant difference in Δ_{π} as compared to the permuted networks (average Δ_{π}: AFR 4.4 x 10^{-4}, EUR 3.6 x 10^{-4}; p-values: AFR 0.79, EUR 0.08).

Hierarchical structure of innate immunity signalling

The third network considers the signalling network of innate immunity, as it is seen from a functional perspective. Looking at it, one can observe a ”bow-tie” shaped organization in terms of the number of components and the flow of information: a relatively large number of receptor molecules recognizing different classes of pathogens all signal to a limited number of intracellular adaptor proteins and transcription factors. These adaptors in turn can interact with proteins that act as modulators, and subsequently signal to a diverse array of downstream molecules, including cytokines and other effector molecules (Figure 1). This bow-tie structure characterized by a large number of “inputs”, a relatively small number of central control nodes which elaborates the information, and a large number of “outputs” seems to be a topological organization widely adopted by metabolic networks (Hong-Wu and Zeng 2003, Csete and Doyle 2004) and also by the immunity system signalling network (Oda and Kitano 2006). Given this structure, one can hypothesize that the amount of gene variation should follow a similar pattern, with the adaptors common to all different pathogen responses being more constrained than both up- and downstream proteins, as previously reported (Csete and Doyle 2004). We used a permutation procedure to test this hypothesis. Each protein was classified according to its level within the hierarchy of signalling, and the mean nucleotide diversity calculated for each level in the original data, as well as in 10,000 replicates with class labels randomly permuted. As expected, adaptor proteins are the only class showing significantly lower mean nucleotide diversity, both in

Africans and Europeans (Figure S3; $p_{AFR} = 0.010$, $p_{EUR} = 0.052$), indicating that they are indeed more constrained.

Some functional classes are preferential targets for adaptive selection

Next, we evaluated if some functional categories are also preferential targets for positive or balancing selection. We performed different neutrality tests to detect genes with an excess of rare or intermediate variants, which is suggestive of positive or balancing selection. The significance of these tests has been evaluated by comparing the results for each gene to the rest of the genes included in the study, and for the genes showing more evidences of positive or balancing selection in one of the two populations we performed coalescent simulation including the demographic history of Africans and Europeans; these methods should exclude the possibility of these signals being produced by demographic events such as expansion or population contractions. Moreover, our aim is to identify different selective pressures acting on the different functional classes, which minimizes these possible random effects of demography. Therefore, the analysis at the functional class level provides also a correction for these possible random effects.

We have considered a gene to show evidence of positive or balancing if it is included in the 95th upper or lower percentiles for two or more neutrality test in the same population. Table 2 shows the list of these genes, and significance has been validated by coalescent simulations in all cases. From the total of 132 genes, 30 show signs of adaptive selection, 14 of positive selection and 16 of balancing selection. The proportion of genes with balancing selection is very high. Europeans and Africans show a similar amount of genes under adaptive selection (16 in Africans and 17 in Europeans), an interesting trait showing the difference with the amount of neutral variation and indicative of the strong selective pressures in non-African populations.

In addition to signatures of adaptive selection on specific genes, we wanted to analyze the effect of functional class. Figure 5 shows the results of two of the neutrality tests, one that only uses the intraspecific information (Tajima's D), and one using the chimpanzee sequence as out-group (Fu and Li's F) for the two populations. We tested in each functional class if there is an excess of genes under positive, balancing or both kinds of selection, by comparing the number of genes with and without significant results in this class against the rest of the genes. Genes with statistical evidence for positive or balancing selection are not randomly distributed among the different functional classes, and tend to accumulate in a reduced number of classes. Specifically, there is an unexpected high number of acute phase proteins showing signatures of balancing selection, both in Africans and Europeans (Fisher's exact test p = 0.03 and 0.001, respectively). Remarkably, all genes in that category with signatures of balancing selection in Europeans are included in the IL1 family (Figure 1). Moreover MEFV and IL1F7, two other acute phase proteins, also show an excess of intermediate variants in one of the neutrality tests in Europeans, and signatures of balancing selection have been reported in the two genes in two independent studies (FUMAGALLI *et al.* 2009a; FUMAGALLI *et al.* 2009b).

On the other hand, the receptors are also overrepresented among the genes showing signatures of positive or balancing selection, suggesting adaptive selection events (Fisher's exact test p = 0.02). In this case, this tendency is only observed in Europeans, and mainly for the extracellular receptors. It is particularly interesting in the case of the TLR4 module, where four of the seven genes show evidences of selection (Fisher's exact test P = 0.006).

Discussion

The evolutionary pace of the innate immune system in humans, in its complex history of expansion all over the planet, with large differences in climatic, ecological and infectious conditions, has not been just a demographic event but natural selection has been shaping it in different ways according to the environment. No doubt infectious diseases have been a major selective factor and adaptation of the innate immune system has taken place in an intense manner. Out of the 132 genes analyzed with the most detailed genetic data (resequencing) has shown signs of adaptive selection in 30, selected using a rather restrictive criterion of being statistical significant (at $p<0.05$) for at least two of the selection tests performed. Even if data is not comparable for whole genome scans looking at the footprint of selection, the five studies published so far find from 460 to 1030 genome regions candidates for positive selection. Thus the proportion found here is very high. It is even higher if we consider the cases where balancing selection have been studied (Andrés et al., 2009)

In addition to some well known cases as the human major histocompatibility complex (SOLBERG *et al.* 2008), recently several works have described the action of balancing selection on genes related to the host pathogen interaction, including innate immunity genes (CAGLIANI *et al.* 2008; FERRER-ADMETLLA *et al.* 2008) and blood group antigen genes (CALAFELL *et al.* 2008; FUMAGALLI *et al.* 2008). Interactions with pathogens on one side and the development of pathogenic processes such as autoimmunity on the other side could be the balanced forces acting on these genes, with too strong a response causing the various adverse effects associated with infectious and autoimmune diseases, while a weak inflammatory response attenuates the subsequent immune response (FERRER-ADMETLLA *et al.* 2008).

Several important biological consequences may be inferred from the results presented in this study. Firstly, one important question that may be asked is which major class of bacteria could have exerted the evolutionary pressures observed here. As the innate immune system is relatively non-specific, any genetic information obtained in the present study cannot refer to specific species or even families, but rather to broad groups of pathogens such as Gram-positive or Gram-negative bacteria. Both Gram-positive bacteria (e.g. *Staphylococci* and *Sterptococci*, mycobacteria) and Gram-negative bacteria (*Salmonella* and other enterobacteriaceae, *Yersinia pestis*) contain pathogens that had a major impact on human populations during history. An answer to the question of whether both these two groups of bacteria exerted evolutionary pressures cannot be obtained from the selection profile of effector genes such as defensins or cytokines, as these proteins are generally necessary for host defence to both Gram-positive and Gram-negative bacteria. In contrast, a certain level of specificity can be seen by the recognition of either Gram-negative bacteria mainly by the TLR2/TLR1/TLR6/TLR10 cluster, while Gram-negative bacteria recognition is heavily reliant on the TLR4 pathway (Akira et al., Cell 2006; 124:783-801). In our study there is a clear selection of genes from both the TLR2 cluster (TLR2/TLR6/TLR10) and the TLR4 pathway (MD2, RP105, CD14), suggesting that both Gram-negative and Gram-positive bacteria have exerted pressure on innate immunity genes. This conclusion is supported by Barreiro and colleagues who also reported recent positive selection in the region encompassing TLR1/TLR6/TLR10 (with clearer effects on TLR1 in their hands), as well as TLR4 pathway (Barreiro et al., 2009). In addition, TLR9, a pathogen recognition receptor recognizing unmethylated DNA from both Gram-negative and Gram-positive bacteria (Krieg AM. 2002) is also under selection pressure in Europeans. Thus, these complementary data reinforce the hypothesis that both Gram-negative and Gram-positive bacteria have exerted important selective effects on the genes of innate immunity.

A second important observation refers to the genes that are under balancing selection, presumably due to a beneficial effect in infections that is conterbalanced by deleterious effects in autoinflammatory and autoimmune conditions. Interestingly, there is clear overrepresentation of three gene families: the IL-1 family (most abundantly and strongly under balanced selection, with 7 genes represented in both European and African populations), the TNF family (both TNFα and TNFβ/LTα being represented) and complement (three genes). While their role in host defence is well documented, these genes encode the main proinflammatory mediators of the two most important classes of sterile inflammation disorders: the autoinflammatory diseases (e.g. Crohn's disease, gout, autoinflammatory syndromes) in which inflammation is mediated by the IL-1 family, and the autoimmune diseases (e.g. rheumatoid arthritis, systemic lupus erythematous, type 1 diabetes) in which the TNF family and complement play a central role (Dinarello CA, 2009). The fact that the genes found to be under balancing selection are exactly those encoding the major mediators of human autoinflammatory and autoimmune pathologies is a strong argument for the biological relevance of our findings.

Thus, these results suggest that the innate immune system have had enough plasticity to play an important role in the adaptation of modern humans to new environments, but only specific parts of the innate response have been able to develop adaptive responses. In contrast, some specifics parts of the innate immune response seem to be more constricted, and it can be predicted both by their function and their position in the network.

On the other hand, we have only described an excess of receptors with signatures of selection in Europeans. This lack of correlation is due to the signatures of positive selection of two extracellular receptor proteins exclusive of Europeans. In this case, these signatures could be related to more recent responses, that for example could have been related to the introduction of agriculture and zoonotic diseases (e.g. *Brucella*, *Coxiella)* or the major epidemics of the medieval periods (e.g. *Yersinia pestis* – plague). Two recent works reported strong purifying selection acting on the *toll-lik*e receptors, more intense in the case of the intracellular receptors (BARREIRO *et al.* 2009; MUKHERJEE *et al.* 2009). Selection would be conserving the amino acid sequence in these genes, which points to adaptive selection possibly acting on regulatory regions...

Finally we analyzed the evolution of the immunity genes in relation to the networks of interactions they are involved, since the immunity response arise as a complex behavior which simultaneously involves of all these genes and their interactions.

We first investigated the protein-protein interaction network which links the physically interacting protein products of the considered immunity genes. We investigated whether highly connected proteins (hubs) had undergone a more constrained evolution in respect to peripherical ones, through the analysis of the correlation of the nucleotide diversity with two measures of topological centrality: the degree centrality which is a local measure since only takes into account the number of connected neighbors of a gene, and the betweenness centrality which, by considering all the shortest paths between each pair of nodes in the whole network, provides a global measure of centrality. Although these correlations are negative, suggesting a stronger structural and functional constraint on genes that occupy a central position in the network, none of them appears to be significant. Similarly, no significant evidence was found for interacting proteins to show signatures of similar levels of selective pressures. Conversely to what has been obtained for the protein protein interaction network, significant results are found when analyzing the signaling pathway that has a clear bow tie architecture. It seems that this topological organization is a predominant feature of the functional network that we are analyzing and it is able to explain much better the evolution landscape of its genes.

Indeed, in a analysis considering the signaling network depicted in Fig 1, we observed that this network is organized into a bow tie structure, consisting in a large number of receptors (large "fan-in"), a central small number of adaptor genes (the bow tie knot) and a large number of effectors (large "fan-out"). The bow tie structure is already known to characterize many metabolic and immunity system networks and it is recognized as a topological feature which provides robustness to systems (Kitano 2004, Csete and Doyle 2004). It also has the advantageous property of making robustness compatible with evolvability, since the small knot provides robustness to the system by easily accommodating for perturbations, while a great variability is allowed at the edges of the bow tie making possible the generation of new functionalities (Csete and Doyle 2004). In agreement with results obtained for bow tie shaped metabolic networks, here we find that in the human immunity network the core of the bow tie structure is highly conserved, being the adaptor genes the only ones showing a significantly low nucleotide diversity. However, signatures of adaptive evolution are found at the bow tie edges. The amount of selection acting on each gene appeared therefore highly dependent on its position in the structure. This clearly shows the strength of the constraint represented by the topology of the network on the evolution of individual genes.

References

Alvarez-Ponce, D., M. Aguade and J. Rozas, 2009 Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res 19: 234-242.

Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, et al., 2009 Targets of balancing selection in the human genome. Mol Biol Evol 26(12):2755-64

Assenov, Y., F. Ramirez, S. E. Schelhorn, T. Lengauer and M. Albrecht, 2008 Computing topological parameters of biological networks. Bioinformatics 24: 282-284.

Barabási, A.-L., R. Albert, 1999 Emergence of scaling in random networks. Science 286, 509-512.

Barabasi, A. L., and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.

Barreiro, L. B., M. Ben-Ali, H. Quach, G. Laval, E. Patin et al., 2009 Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet 5: e1000562.

Barreiro, L. B., E. Patin, O. Neyrolles, H. M. Cann, B. Gicquel et al., 2005 The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am J Hum Genet 77: 869-886.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz et al., 2005 Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

Cagliani, R., M. Fumagalli, S. Riva, U. Pozzoli, G. P. Comi et al., 2008 The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol 9: R143.

Calafell, F., F. Roubinet, A. Ramirez-Soriano, N. Saitou, J. Bertranpetit et al., 2008 Evolutionary dynamics of the human ABO gene. Hum Genet 124: 123-135.

Csete M, Doyle J. 2004 Bow ties, metabolism and disease. Trends Biotechnol. 22(9):446-50.

Dinarello CA, 2009 Immunological and inflammatory functions of the interleukin-1 family. Annu Rev Immunol 27:519-550

Enard D, Depaulis F, Roest Crollius H. PLoS Genet. 2010 Feb 5;6(2):e1000840. Human and non-human primate genomes share hotspots of positive selection.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Ferrer-Admetlla, A., E. Bosch, M. Sikora, T. Marques-Bonet, A. Ramirez-Soriano et al., 2008 Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol 181: 1315-1322.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Fumagalli, M., R. Cagliani, U. Pozzoli, S. Riva, G. P. Comi et al., 2008 Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res.

Fumagalli, M., R. Cagliani, U. Pozzoli, S. Riva, G. P. Comi et al., 2009a A population genetics study of the Familial Mediterranean Fever gene: evidence of balancing selection under an overdominance regime. Genes Immun.

Fumagalli, M., U. Pozzoli, R. Cagliani, G. P. Comi, S. Riva et al., 2009b Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med 206: 1395-1408.

Gao, J., A. S. Ade, V. G. Tarcea, T. E. Weymouth, B. R. Mirel et al., 2009 Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics 25: 137-138.

Kim, P. M., J. O. Korbel and M. B. Gerstein, 2007 Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proceedings of the National Academy of Sciences of the United States of America 104: 20274-20279.

Kitano H., 2004 Biological robustness. Nat Rev Genet 5: 826–837

Krieg AM., 2002 CpG motifs in bacterial DNA and their immune effects. Annu Rev Immunol 20:709-60)

Lazarus, R., D. Vercelli, L. J. Palmer, W. J. Klimecki, E. K. Silverman et al., 2002 Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease. Immunol Rev 190: 9-25.

Lazzaro, B. P., 2008 Natural selection on the Drosophila antimicrobial immune system. Curr Opin Microbiol 11: 284-289.

Librado, P., and J. Rozas, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Ma HW, Zeng AP., 2003 The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19(11):1423-30.

Mukherjee, S., N. Sarkar-Roy, D. K. Wagener and P. P. Majumder, 2009 Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc Natl Acad Sci U S A 106: 7073-7078.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton et al., 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170.

Oda K, Kitano H., 2006 A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol. 2006;2:2006.0015.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly et al., 2006 Positive natural selection in the human lineage. Science 312: 1614-1620.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly et al., 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576-1583.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang et al., 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.

Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai et al., 2008 Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol 69: 443-464.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

Yuan, F. F., K. Marks, M. Wong, S. Watson, E. de Leon et al., 2008 Clinical relevance of TLR2, TLR4, CD14 and FcgammaRIIA gene polymorphisms in Streptococcus pneumoniae infection. Immunol Cell Biol 86: 268-270.

Figure legends

Figure 1. Schematic representation of the function on the innate immune system of the 132 genes analyzed in this work. The ligands of the receptors are also showed.

Figure 2. Nucleotide diversity in the different genes and functional classes in African and European individuals. Colours for each functional class are the same that in Figure 1.

Figure 3.

Network of interactions among innate immunity genes. Genes are grouped according to their classes.

Figure 4. A) Tajima's D in the different genes and functional classes in African and European individuals. B) Fu and Li's F in the different genes and functional classes in African and European individuals. Grey lines show the 95th upper and lower percentiles in each population. Colours for each functional class are the same that in Figure 1. Cytokynes, in green and: ■ = Acute phase, ▲ = Antiinflamatory Cytokines, ○ = Cellular Immunity, □ = Chemokynes, ◊ = Complement, ▲ = Extracellular, ▼ = Neutrophil.

Table 1. Nucleotide diversity in the different functional classes

Categorie	N	π AFR	π EUR
Receptors	18	0.00093	0.00060
Adaptor	10	0.00057	0.00047
Modulators	6	0.00080	0.00063
Cytokines	77	0.00080	0.00066
Acute Phase	20	0.00091	0.00081
Cellular Immunity	10	0.00073	0.00060
Antiinflamatory Cytokines	10	0.00079	0.00063
Neutrophil	9	0.00074	0.00061
Chemokines	14	0.00078	0.00062
Extracellular	6	0.00097	0.00072
Complement	8	0.00061	0.00051
Effector	15	0.00096	0.00073
TOTAL	127	0.00082	0.00065

π, nucleotide diversity (average of pairwise differences).

N, number of genes included in each categorie.

Table 2. Number of genes in each functional class with evidences of adaptive selection.

Functional Class		Africans		Europeans	
	N	Positive Selection	Balancing Selection	Positive Selection	Balancing Selection
Receptors	18	TLR2	LY96	LY64, TLR9, TLR10	CD14, LY96, TLR6
Adaptors	10	-	-	-	-
Modulators	6	SOCS2, TOLLIP	-	-	-
Cytokines	77				
Acute Phase	20	-	LTA, IL18RAP, IL1F5, IL1F7	-	IL1F5, IL1A, IL1F10, IL18RAP, IL1F6, TNF
Cellular Immunity	10	-	-	-	-
Antiinf. Citokines	10	-	-	TGFB2	-
Neutrophil	9	IL17B	-	-	-
Chemokines	14	IL8	-	CCR4	-
Extracellular	6	-	-	IL5, ADAM33	-
Complement	8	-	CFH, C5, C3	-	C3
Effector	15	MPO, NGFR	NTRK	-	-

N, number of genes in each functional categorie.

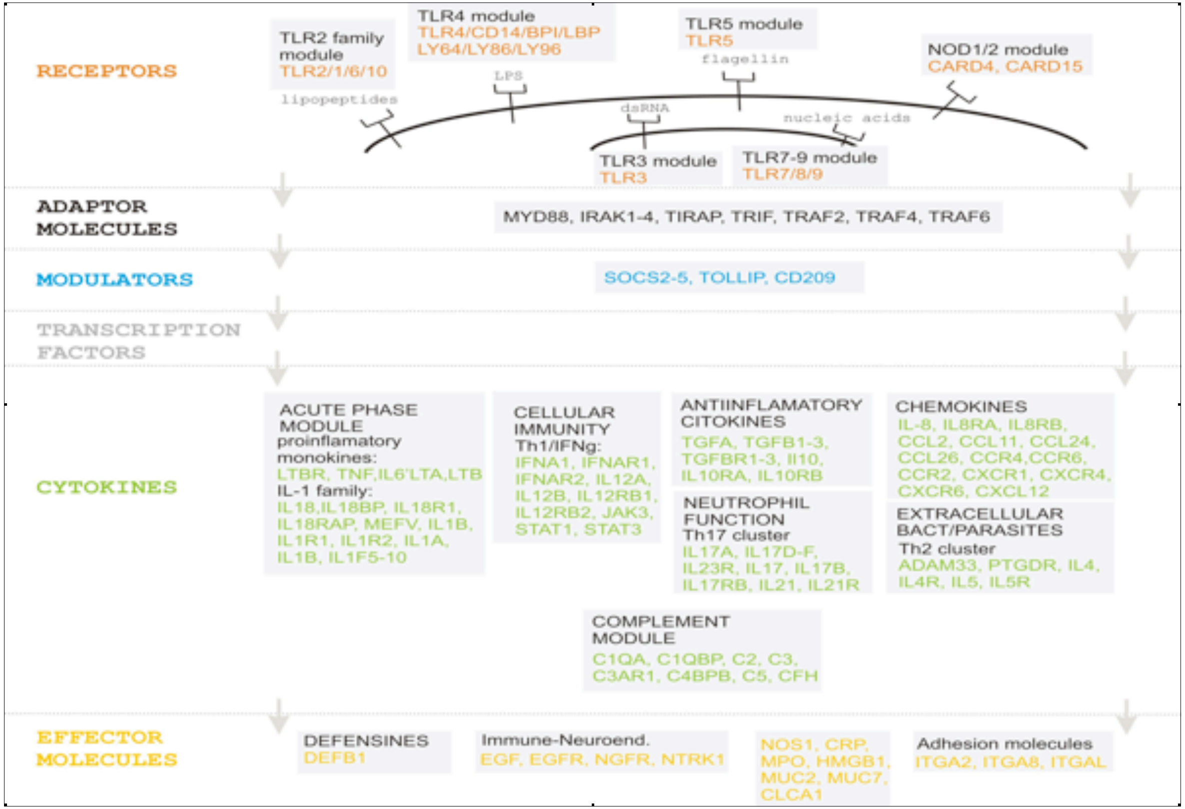

RECEPTORS
TLR2 family module
TLR2/1/6/10
lipopeptides
TLR4 module
TLR4/CD14/BPI/LBP
LY64/LY86/LY96
LPS
TLR5 module
TLR5
flagellin
NOD1/2 module
CARD4, CARD15
dsRNA
nucleic acids
TLR3 module
TLR3
TLR7-9 module
TLR7/8/9
ADAPTOR MOLECULES
MYD88, IRAK1-4, TIRAP, TRIF, TRAF2, TRAF4, TRAF6
MODULATORS
SOCS2-5, TOLLIP, CD209
TRANSCRIPTION FACTORS
CYTOKINES
ACUTE PHASE MODULE
proinflamatory monokines:
LTBR, TNF,IL6'LTA,LTB
IL-1 family:
IL18,IL18BP, IL18R1, IL18RAP, MEFV, IL1B, IL1R1, IL1R2, IL1A, IL1B, IL1F5-10
CELLULAR IMMUNITY
Th1/IFNg:
IFNA1, IFNAR1, IFNAR2, IL12A, IL12B, IL12RB1, IL12RB2, JAK3, STAT1, STAT3
ANTIINFLAMATORY CITOKINES
TGFA, TGFB1-3, TGFBR1-3, Il10, IL10RA, IL10RB
NEUTROPHIL FUNCTION
Th17 cluster
IL17A, IL17D-F, IL23R, IL17, IL17B, IL17RB, IL21, IL21R
CHEMOKINES
IL-8, IL8RA, IL8RB, CCL2, CCL11, CCL24, CCL26, CCR4,CCR6, CCR2, CXCR1, CXCR4, CXCR6, CXCL12
EXTRACELLULAR BACT/PARASITES
Th2 cluster
ADAM33, PTGDR, IL4, IL4R, IL5, IL5R
COMPLEMENT MODULE
C1QA, C1QBP, C2, C3, C3AR1, C4BPB, C5, CFH
EFFECTOR MOLECULES
DEFENSINES
DEFB1
Immune-Neuroend.
EGF, EGFR, NGFR, NTRK1
NOS1, CRP, MPO, HMGB1, MUC2, MUC7, CLCA1
Adhesion molecules
ITGA2, ITGA8, ITGAL

Figure 2

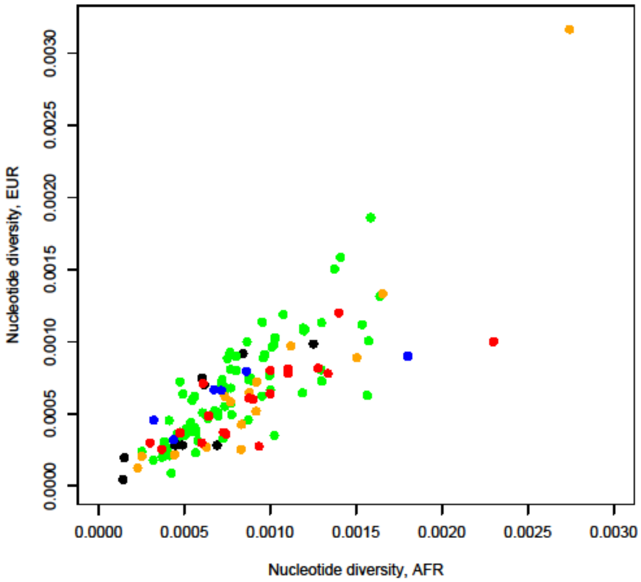

Nucleotide diversity, EUR
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
Nucleotide diversity, AFR
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030

Figure 3

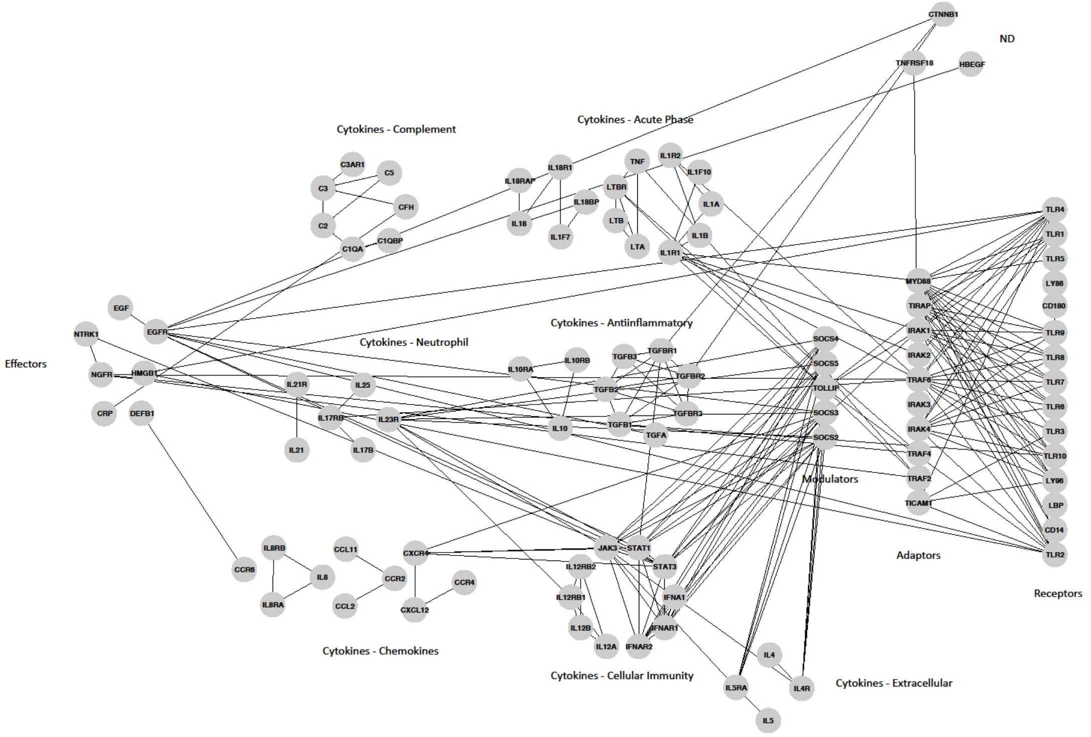
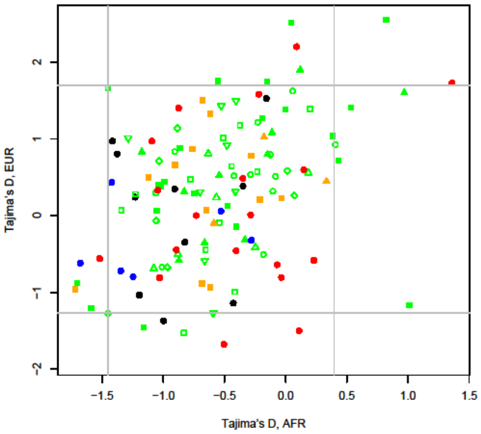
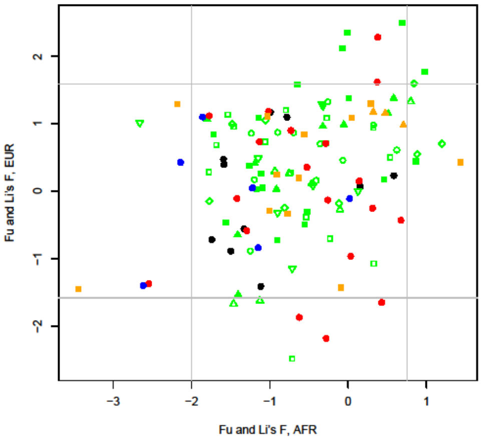

ND
CTNNB1
TNFRSF18
HBEGF
Cytokines - Complement
C3AR1
C5
C3
CFH
C2
C1QBP
C1QA
Cytokines - Acute Phase
IL18RAP
IL18R1
TNF
IL1R2
IL1F10
LTBR
IL18BP
IL1A
IL18
IL1F7
LTB
LTA
IL1B
IL1R1
TLR4
TLR1
TLR5
LY86
CD180
TLR9
TLR8
TLR7
TLR6
TLR3
TLR10
LY96
LBP
CD14
TLR2
Receptors
MYD88
TIRAP
IRAK1
IRAK2
TRAF6
IRAK3
IRAK4
TRAF4
TRAF2
TICAM1
Adaptors
EGF
NTRK1
EGFR
Effectors
NGFR
HMGB1
CRP
DEFB1
Cytokines - Neutrophil
IL21R
IL25
IL17RB
IL23R
IL21
IL17B
Cytokines - Antiinflammatory
IL10RA
IL10RB
TGFB3
TGFBR1
TGFBR2
TGFB2
TGFBR3
IL10
TGFB1
TGFA
SOCS4
SOCS5
TOLLIP
SOCS3
SOCS2
Modulators
IL8RB
CCL11
CXCR4
IL8
CCR2
CCR4
CCR6
IL8RA
CCL2
CXCL12
Cytokines - Chemokines
JAK3
STAT1
IL12RB2
STAT3
IL12RB1
IFNA1
IL12B
IFNAR1
IL12A
IFNAR2
Cytokines - Cellular Immunity
IL4
IL5RA
IL4R
IL5
Cytokines - Extracellular

Figure 4

A


Tajima's D, EUR
2
1
0
−1
−2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Tajima's D, AFR

B


Fu and Li's F, EUR
2
1
0
−1
−2
−3
−2
−1
0
1
Fu and Li's F, AFR

B.2 Balancing selection is the main force shaping the evolution of innate immunity genes

Anna Ferrer-Admetlla, Elena Bosch, **Martin Sikora**, Tomàs Marquès-Bonet, Anna Ramírez-Soriano, Aura Muntasell, Arcadi Navarro, Ross Lazarus, Francesc Calafell, Jaume Bertranpetit, Ferran Casals

Journal of Immunology 2008 Jul 15; 181(2): 1315-22

Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, et al. Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol. 2008; 181(2): 1315-22.

B.3 Human pseudogenes of the ABO family show a complex evolutionary dynamics and loss of function

Ferran Casals, Anna Ferrer-Admetlla, **Martin Sikora**, Anna Ramírez-Soriano, Tomàs Marquès-Bonet, Stéphanie Despiau, Francis Roubinet, Francesc Calafell, Jaume Bertranpetit, Antoine Blancher

Glycobiology 2009 Jun; 19(6): 583-91

Casals F, Ferrer-Admetlla A, Sikora M, Ramírez-Soriano A, Marquès-Bonet T, Despiau S, et al. Human pseudogenes of the ABO family show a complex evolutionary dynamics and loss of function. Glycobiology. 2009; 19(6): 583-91.

B.4 A natural history of *FUT2* polymorphism in humans

Anna Ferrer-Admetlla, **Martin Sikora**, Hafid Laayouni, Anna Esteve, Francis Roubinet, Antoine Blancher, Francesc Calafell, Jaume Bertranpetit, Ferran Casals

Molecular Biology and Evolution 2009 Sep; 26(9): 1993-2003

Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, et al. A natural history of FUT2 polymorphism in humans. Mol Biol Evol. 2009; 26(9): 1993-2003.

B.5 Interrogating 11 fast-evolving genes for signatures of recent positive selection in worldwide human populations

Andrés Moreno-Estrada, Kun Tang, **Martin Sikora**, Tomàs Marquès-Bonet, Ferran Casals, Arcadi Navarro, Francesc Calafell, Jaume Bertranpetit, Mark Stoneking, Elena Bosch

Molecular Biology and Evolution 2009 Oct; 26(10): 2285-97

Moreno-Estrada A, Tang K, Sikora M, Marquès-Bonet T, Casals F, Navarro A, et al. Interrogating 11fast-evolving genes for signatures of recent positive selection in worldwide human populations. Mol Biol Evol. 2009; 26(10): 2285-97.

B.6 Isolated populations as treasure troves in genetic epidemiology: the case of the Basques

Paolo Garagnani, Hafid Laayouni, Anna González-Neira, **Martin Sikora**, Donata Luiselli, Jaume Bertranpetit, Francesc Calafell

European Journal of Human Genetics 2009 Nov; 17(11): 1490-4

Garagnani P, Laayouni H, González-Neira A, Sikora M, Luiselli D, Bertranpetit J, et al. Isolated populations as treasure troves in genetic epidemiology: the case of the Basques. Eur J Hum Genet. 2009; 17(11): 1490-4.

B.7 Sequence variation and genetic evolution at the human F12 locus: mapping quantitative trait nucleotides that influence FXII plasma levels

Francesc Calafell, Laura Almasy, Maria Sabater-Lleal, Alfonso Buil, Carolina Mordillo, Anna Ramírez-Soriano, **Martin Sikora**, Juan Carlos Souto, John Blangero, Jordi Fontcuberta, Jose Manuel Soria

Human Molecular Genetics 2010 Feb 1; 19(3): 517-25.

Calafell F, Almasy L, Sabater-Lleal M, Buil A, Mordillo C, Ramírez-Soriano A, Sikora M, et al. Sequence variation and genetic evolution at the human F12 locus: mapping quantitative trait nucleotides that influence FXII plasma levels. Hum Mol Genet. 2010; 19(3): 517-25.