Draft Manuscript for Review

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

Journal:	<i>Human Genetics</i>
Manuscript ID:	HumGen-07-0503.R2
Manuscript Type:	Original Investigations
Date Submitted by the Author:	n/a
Complete List of Authors:	sikora, martin; universitat pompeu fabra, unitat de biologia evolutiva, departament de ciències experimentals i de la salut ferrer-admetlla, anna; universitat pompeu fabra, unitat de biologia evolutiva, departament de ciències experimentals i de la salut mayor, alfredo; Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Universitat de Barcelona, Barcelona Centre for International Health Research (CRESIB) Bertranpetit, Jaume; Universitat Pompeu Fabra, Evolutionary Biology casals, ferran; universitat pompeu fabra, unitat de biologia evolutiva, departament de ciències experimentals i de la salut
Keywords:	malaria, candidate gene analysis, positive selection, functional pathways

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

Martin Sikora¹, Anna Ferrer-Admetlla¹, Alfredo Mayor^{2,3}, Jaume Bertranpetit^{1,4} and Ferran Casals^{1§}

¹ Evolutionary Biology Unit, Department of Experimental Sciences and Health, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, Carrer Dr Aiguader 88, 08003 Barcelona, Catalonia, Spain

² Barcelona Center for International Health Research (CRESIB), Hospital Clinic, Institut d'Investigacions Biomedicas August Pi i Sunyer (IDIBAPS), Universitat de Barcelona, Catalonia, Spain

³ Centro de Investigação em Saúde da Manhiça (CISM), Maputo, Mozambique

⁴ CIBER en Epidemiología y Salud Pública (CIBERESP, Spain)

§Corresponding author:

Phone: +34 933160845

Fax: +34 933160901

E-mail: ferran.casals@upf.edu

Keywords: Malaria; Positive selection; Candidate gene analysis; Functional pathways

Abstract

Placental malaria is a special form of malaria that causes up to 200,000 maternal and infant deaths every year. Previous studies show that two receptor molecules, hyaluronic acid and chondroitin sulphate A, are mediating the adhesion of parasite-infected erythrocytes in the placenta of patients, which is thought to be a key step in the pathogenesis of the disease. In this study, we aimed at identifying sites of malaria-induced adaptation by scanning for signatures of natural selection in 24 genes in the complete biosynthesis pathway of these two receptor molecules. We analyzed a total of 24 Mb of publicly available polymorphism data from the International HapMap project for three human populations with European, Asian and African ancestry, with the African population from a region of presently and historically high malaria prevalence. Using methods based on allele frequency distributions, genetic differentiation between populations, and on long range haplotype structure, we found only limited evidence for malaria-induced genetic adaptation in this set of genes in the African population; however, we identified one candidate gene with clear evidence of selection in the Asian population. Although historical exposure to malaria in this population cannot be ruled out, we speculate that it might be caused by other pathogens, as there is growing evidence that these molecules are important receptors in a variety of host-pathogen interactions. We propose to use the present methods in a systematic way to help identify candidate regions under positive selection as a consequence of malaria.

Introduction

Historically, studies of human genetic variations have focused mostly on demography, in the form of studies of population history (Cavalli-Sforza et al. 1994). In the last few years however, a new interest in natural selection as the object of study has arisen (Sabeti et al. 2006), partly due to the availability of large, publicly available genotype data sets, such as the one produced by the International HapMap project (The International HapMap Consortium 2005, 2007). It has been proposed that searching for the footprint of selection is a good strategy to find functional variation (Bamshad and Wooding 2003). A variety of methods exist for these kinds of analyses, differing in their applicability to different types of genetic data as well as in the evolutionary timeframe they have power to detect selection (for thorough reviews see Bamshad and Wooding (2003); Nielsen (2005); Sabeti et al. (2006)).

The malaria parasite *Plasmodium falciparum* causes an estimated 350 – 500 million clinical episodes of malaria worldwide every year, with a death toll of over one million annually, most of which are children (World Health Organization 2005). Not surprisingly, it is thought to be the strongest selective force that humans have experienced in recent history (Kwiatkowski 2005), and a variety of genetic variants have been described that are generally assumed to be a direct consequence of that (Ohashi et al. 2004; Sabeti et al. 2002; Tishkoff et al. 2001; Wood et al. 2005). As some well-known examples, they are the underlying cause of erythrocyte deficiencies like sickle-cell disease (caused by the hemoglobin S variant), thalassemias or glucose-6-phosphate (G6PD) deficiency. The fact that these variants are maintained at appreciable frequencies in affected populations, despite most of them being almost always lethal in the homozygous state, are clear signs of the selective force exerted by the parasite (Min-Oo and Gros 2005). Given this strong pressure, it has been recently

suggested that, in the case of malaria, incorporating evidence of selection is a good strategy to improve power to detect protective loci in genetic association studies (Ayodo et al. 2007).

Plasmodium falciparum infection in adults living in areas of high malaria transmission rarely causes severe malaria, due to their acquired immunity to infection. Pregnant women however are an exception, especially during their first pregnancy (McGregor et al. 1983). Pregnancy-associated malaria can lead to a feto-maternal syndrome characterized by maternal anemia, fetal growth retardation and low birth weight (Steketee et al. 1996). These conditions are associated with substantial morbidity and mortality in mothers and especially newborns, with an estimated death toll of up to 100,000 infants related to placental malaria (Brabin 1983; Desai et al. 2007; Menendez et al. 2000; Okoko et al. 2002). It not only results in lower birthweight in newborns, leading to higher mortality, but also in a significant increase in malaria susceptibility of newborns (Mutabingwa et al. 2005). While there have been numerous studies to date on natural selection related to malaria (see above), none of them has been specifically aimed at placental malaria infection. The present study is a first attempt to detect signatures of natural selection related to this type of malaria, which is thought to involve different susceptibility factors than malaria in children. There is growing evidence that placental parasites express host cell surface antigens in the infected erythrocytes that are functionally distinct from those of non-placental parasites. Various studies indicate that two glycosaminoglycans, hyaluronic acid (HA) and especially chondroitin sulfate (CS) - A, serve as placenta-specific receptors for infected erythrocytes in pregnancy-associated malaria (Beeson et al. 2000; Fried et al. 2006; Rogerson et al. 2007), making them interesting targets for an in-depth study of signatures of selection.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

This article presents the result of a study aimed at using methods of evolutionary genetic analysis to identify functionally important variation in a set of candidate genes related to placental malaria. Candidate gene studies are usually done using genes with prior evidence of being involved in the disease (e.g. genes with signals of association), but not necessarily with a clear functional relation among them. Given the increasing amount of information about protein interactions in functional pathways, including biosynthesis, available in public databases, it seems logical to incorporate this information in the selection of candidate genes, in order to include all possible sources of variation related to a given phenotype. To this end, we employ a pathway-centric approach, examining the complete set of genes in the biosynthesis pathways of HA and CS. We analyzed publicly available single nucleotide polymorphism (SNP) genotype data of the 24 genes in these pathways in three populations with different exposure to malaria. We propose this systematic approach as a useful tool for the identification of possible targets of natural selection.

Materials & Methods

Candidate gene regions

A list of the complete set of genes for CS biosynthesis was obtained by querying the KEGG PATHWAY database for this pathway (http://www.genome.jp/dbget-bin/www_bget?path:hsa00532). Genes for HA biosynthesis were obtained from Varki (1999). The complete list included a total of 24 genes, 21 genes for CS biosynthesis and 3 genes for HA biosynthesis (Table 1). Chromosomal positions for all genes (NCBI human genome build 35 / HG17) were retrieved by querying the *knownGenes* table of the UCSC Genome Browser (<http://genome.ucsc.edu>), whereas information regarding the molecular function of the genes were obtained by batch query of the PANTHER database (<http://www.pantherdb.org>).

Genotype data

Genotype data for the genes of interest was obtained from the HapMap website (<http://www.hapmap.org>). We downloaded SNP genotype data of the HapMap release 21 (phase 2, Jul 06) for all unrelated samples in the four HapMap populations (CEU – CEPH North Americans with European ancestry; YRI – Yorubans from Ibadan, Nigeria; JPT – Japanese from Tokyo; CHB – Chinese from Beijing). Due to the high degree of genetic similarity between the two Asian populations we analyzed them as a single population (JC). Thus, our final analysis panel consisted of 60 CEU, 60 YRI and 90 JC individuals. For our 24 genes of interest, SNP genotype data and phased haplotypes of all SNPs that are polymorphic in at least one of the study populations were downloaded for 1 Mb regions centered on the gene of interest. This size was chosen to adequately cover even the largest gene in the list (*XYLT1*, 360 kb) as well as to obtain sufficient data for background

distributions for our analysis. In no case more than one candidate gene was found in the same regions, thus all regions are completely independent. Functional status information for each SNP was retrieved by querying the *snp125* table of the UCSC Genome Browser.

Data handling & analysis

Unless otherwise stated, all data analysis was carried out using the R statistical software package (version 2.4.0, <http://www.r-project.org>) as well as the SNPator web application package (<http://bioinformatica.cegen.upf.es>). Since we only worked with released HapMap data that had already passed stringent quality control, none of the usual tests like departures from Hardy-Weinberg equilibrium were necessary to be performed.

Allele frequency analysis

For each window across every gene region in each population, we analyzed the distribution of the minor allele frequencies (MAF analysis) and the derived allele frequencies (DAF analysis) of the corresponding SNPs. For both MAF and DAF analysis, we calculated the proportions of SNPs with allele frequencies higher or lower than a defined threshold within a sliding window of 100 kb size. Thresholds were chosen to maximize sensitivity to selection as suggested by Walsh et al. (2006). For each region, we calculated the proportions for each window that contains a minimum of 20 SNPs, with a 10 kb step size for the sliding window, in order to obtain an overall background distribution from the data. For the MAF analysis, we calculated the proportion of SNPs with a $MAF < 0.1$ versus the proportion of SNPs with $MAF > 0.4$. For the DAF analysis, we calculated the proportion of SNPs with $DAF < 0.2$ versus the proportion of SNPs with $DAF > 0.8$. We then plotted all windows that overlap with our studied genes against this background distribution to identify

potentially interesting outliers. Ancestral states for all analyzed SNPs were obtained by querying the *snpl26orthoPanTro2RheMac2* table of the UCSC Genome Browser and assigning the chimpanzee allele as the ancestral state. In cases where neither of the human alleles corresponded to the chimp sequence (244 SNPs) we excluded these SNPs from the DAF analysis. We defined interesting outliers as regions where multiple windows were found in the top 5% of the distribution of the respective proportions.

F_{ST} calculation

We calculated molecular fixation indices (F_{ST}) per locus using the ARLEQUIN software package (version 3.11) (Excoffier et al. 2005). Genetic structure of the study populations was inferred using the Analysis of Molecular Variance (AMOVA) framework (Excoffier et al. 1992). To treat missing data accordingly, we carried out a locus-by-locus AMOVA. Empirical percentiles of F_{ST} values were calculated based on the empirical distribution of all F_{ST} values over all analyzed regions. Markers on the X chromosome generally show elevated levels of F_{ST} compared to autosomes due to different population genetic properties (e.g. lower effective population size). We therefore excluded the *CHST7* region from this analysis.

Long range haplotype (LRH) analysis

The analysis was carried out using the SWEEP software package (version 1.1, <http://www.broad.mit.edu/mpg/sweep/index.html>) as well as our own implementation of the test. We employed two different approaches to look for selection, both based on the extended haplotype homozygosity (EHH) measure introduced by Sabeti et al. (2002). In order to obtain enough data for a background distribution of EHH values as well as being able to examine regions with possible low recombination rates and

therefore long extended haplotypes, we analyzed phased haplotype data for 1 Mb regions centered on the genes of interest, excluding the *CHST7* region on the X chromosome to only compare autosomal data. One of the analyzed regions (*B4GALT7*) showed a very low SNP density and gaps >100 kb between consecutive SNPs around the gene of interest. Since this can lead to inflated EHH values at long distances, we also excluded this region from our analysis.

The first approach was the EHH analysis method suggested by Sabeti et al. (2002), and we defined core haplotypes according to the same criteria used by the authors. For each identified core haplotype in each of the regions, we calculated EHH and the relative EHH (REHH) at a genetic distance of 0.3 cM in both directions and plotted them against the frequency of the core haplotype. Core haplotypes were then placed in 5% frequency bins and the respective EHH and REHH values were log-transformed for each bin to obtain approximately normally distributed values.

The second approach we used is a method similar to the integrated haplotype score (iHS) introduced by Voight et al. (2006). In this approach, we start by treating each SNP in turn as a “core haplotype”. We then find the physical distance in both directions at which the EHH falls to 0.25, to obtain a total extended haplotype length for both the minor and the major allele at the core SNP analyzed. The test score HLS (haplotype length score) is then given by:

$$HLS = \log\left(\frac{EHL_m}{EHL_M}\right)$$

where EHL_m is the extended haplotype length at the minor allele and EHL_M at the major allele. To account for differences in the values of the score across different frequency bins, we normalize it by:

$$nHLS = \frac{HLS - E_{bin}(HLS)}{SD_{bin}(HLS)}$$

where $E_{\text{bin}}(\text{HLS})$ is the mean and $SD_{\text{bin}}(\text{HLS})$ the standard deviation of HLS for all core SNPs in the same frequency bin as the tested SNP. The nHLS was calculated for each SNP in the data that reached an EHH of 0.25 or lower in both directions for both alleles. Furthermore, in order to identify regions along a chromosome that contain multiple SNPs with high scores we calculated the average score, as well as the proportion of SNPs with an absolute value of $\text{nHLS} > 2.5$, over sliding windows of 15 adjacent SNPs for each SNP in each of the regions.

Statistical significance and multiple testing

Empirical p-values for the LRH analyses were obtained by using the mean and standard deviation of the empirical distribution of the respective scores. To account for multiple testing, we estimated the positive false discovery rate (pFDR) (Storey and Tibshirani 2003) and calculated the q-value for the scores within each frequency bin using the package qvalue (version 1.1) for R. The q-value for a particular p-value is defined as the expected proportion of false positives among all significant p-values when calling that p-value significant. We used a q-value cutoff of 0.05 for assigning significance.

Results

HapMap data

We analyzed SNP polymorphism data from Phase 2 of the International HapMap Project (The International HapMap Consortium 2005, 2007) for 24 gene regions involved in the biosynthesis of CS and HA (see Methods and Table S1). Table 1 gives an overview of the analyzed regions. The data consist of a total number of 26,282 SNPs in 210 individuals of three populations, covering roughly 24 Mb of the human genome. We restricted our analysis to SNPs that were polymorphic in at least one of the populations, including SNPs that were not typed in all three populations. Thus, the total numbers of SNPs analyzed for each population were 25,380 for Europeans (CEU), 25,716 for Asians (JC) and 25,125 for Africans (YRI). Out of the respective totals, 20,389 (80.3%) were polymorphic in CEU, 19,909 (77.4%) in JC and 22,674 in YRI (90.2%), consistent with previous observations that African populations show generally higher levels of diversity. The average SNP density over all regions was 1 SNP per ~900 base pairs, with the lowest average density at 1 SNP per 3.8 kb in the *B3GALT6* region.

MAF – DAF analysis

We next carried out two types of analysis of the allele frequency distributions, the minor allele frequency (MAF) and the derived allele frequency (DAF) analysis, in a similar way as previously suggested by Walsh et al. (2006; see Materials & Methods). This method aims at detecting selective sweeps due to the characteristic allele frequency spectrum caused by selection. An excess of rare alleles and/or a high proportion of high frequency derived alleles could indicate positive selection, whereas an excess of intermediate frequency alleles could be produced by balancing selection. In the MAF analysis, outliers with high proportion of rare MAFs could indicate

positive selection (due to the local reduction of variation caused by a selective sweep), whereas outliers with a high proportion of high frequency MAFs could point towards balancing selection. Similarly, regions with high proportion of high frequency derived alleles in the DAF analysis could also indicate positive selection.

It is important to point out three caveats of this method that have to be taken into account when interpreting the results. First, it is well-known that allele-frequency based methods are not powerful for detecting recent selective sweeps (Hanchard et al. 2006). However, we expect that a strong sweep would still be leaving a detectable signature in the allele frequency spectrum, as demonstrated by the example of *LCT* (see Supplementary information). Second, due to the ascertainment scheme of the HapMap, which was aimed mainly at genotyping common SNPs ($MAF > 0.05$), the allele frequency spectrum of the data is skewed towards common alleles. This bias makes it more difficult to detect regions with excess of rare alleles or an excess of high frequency derived alleles that are near fixation, and is also the reason why it is not possible to apply statistical tests for resequencing data like Tajima's *D* to SNP data (see the case shown in Soldevila et al. (2005)). The extent of the bias is also likely to be different between populations; however, since we only do comparisons within populations, we do not expect this to influence our results significantly. Third, the method is not a statistical test, but merely a crude way of identifying regions that show interesting patterns in the distribution of allele frequencies. When analyzing 24 Mb of data we expect that some windows will be outliers just by chance. Despite these caveats, in combination with the results of the other analyses undertaken in this study, the method will help to identify interesting candidate regions.

Figure 1 shows the results of this analysis. As a first observation we note that for both the MAF - (Figure 1a) and DAF - analysis (Figure 1b) the distributions are

more spread for the CEU and JC populations compared to the YRI, indicating that there is more fluctuation in the allele frequencies in these populations. This is easily explained by the different demographic histories of these populations (Cavalli-Sforza and Feldman 2003). Due to these, non-African populations generally show stronger effects of genetic drift, leading to the more varied allele frequency spectrum observed in Figure 1. In all three populations we identified potentially interesting outlier regions, which are summarized in Table 2. As a general observation, it is interesting to note that genes detected as outliers in the two types of analyses greatly overlap. In the CEU, a 130 kb region including the *B3GALT6* gene shows high proportions of rare MAFs (0.66 – 0.72) with the same, albeit slightly shorter region also showing a high proportions of high frequency derived alleles (0.38 – 0.50). In the JC, we identified three different regions showing very high proportions of rare alleles, with two of these regions additionally exhibiting a high proportion of high frequency derived alleles. In the *B3GAT3* region a stretch of 200 kb shows proportions of rare MAFs that range from 0.70 to 0.96, but only one of these 100 kb windows also is an outlier in the DAF analysis. Another interesting region is the one containing *XYLT1*, with a 200 kb large region showing high proportion of rare MAFs as well as high frequency DAFs. Examining the distribution of MAF and DAF for each SNP along the chromosome in the coding region of this gene, compared to the same gene in the CEU population, we found that the region covering exons 1 – 3 (roughly from 17250 kb – 17500 kb) shows a very distinct pattern of allele frequencies, with almost complete lack of intermediate frequency alleles, as opposed to the CEU SNPs (see Figure S1). The region between exon 1 and exon 2 has a particularly striking distribution, with basically all SNPs being either rare or near fixation, and a significant proportion of the near-fixation alleles being derived alleles. Lastly, we also

identified two outlier regions in the YRI sample. As mentioned previously, the YRI show a much more compact distribution with less extreme proportions. In the MAF analysis, a 130 kb region containing *GALNACT-2* shows a relative excess of rare alleles. Interestingly, the same region in the DAF analysis exhibits a relative excess of low frequency derived alleles, which would indicate possible purifying selection in the region. Additionally, a 130 kb stretch in the *B3GATI* shows a relatively high fraction of high frequency derived alleles (0.21 – 0.26), making it another candidate for further investigation.

Population differentiation

We calculated F_{ST} for each SNP in the full regions to identify loci that show a high population differentiation, which can be an indicator of local adaptation in a population (Akey et al. 2002; Sabeti et al. 2007). The analysis was carried out globally as well as for all three pairwise comparisons (CEU vs. JC, CEU vs. YRI, JC vs. YRI). Mean F_{ST} was in agreement with overall HapMap data at 0.13 (Weir et al. 2005), and the 0.95 and 0.99 percentile were 0.37 and 0.54, respectively. We then analyzed our candidate genes compared to the background distribution obtained from the full regions. Table 3 shows the results for each of the analyzed genes. The gene with the highest global F_{ST} is *D4ST1* (0.259), followed by *CHSY1* (0.255) and *XYLT1* (0.253). It has been shown that F_{ST} values show substantial variation even among SNPs that are very close to each other (Weir et al. 2005), and that individual extreme F_{ST} values do not necessarily imply positive selection (Gardner et al. 2007). Therefore we focused on regions that show high values among multiple markers. *CHSY1* has the second highest average F_{ST} of the data, and nine out of 87 SNPs in the region have values above the 0.95 percentile. Examining the F_{ST} profile over the whole region, we found that the YRI are very differentiated from either JC or CEU in the region with

1
2
3 elevated F_{ST} values, which could be an indicator of local adaptive pressure in YRI
4
5
6 (see Figure S2).
7

8 The most interesting result is again in the *XYLT1* gene, which also contains the
9
10 SNP with the highest F_{ST} of all our candidate genes (0.66). Out of the 484 SNPs, 62
11
12 (13%) have values exceeding the 0.95 percentile, and 22 (5%) have values above the
13
14 0.99 percentile. Looking at the F_{ST} profile of the region, we found that high values
15
16 occur all over the gene, with multiple SNPs showing extreme values in a stretch of
17
18 around 150 kb from positions 17250 kb up to 17400 kb (see Fig. 2a). In the pairwise
19
20 comparisons, the differentiation between JC and each of the other two populations is
21
22 generally higher than between the two (Fig. 2b), suggesting that this region might be
23
24 under positive selection in JC. Notably, this region shows extensive overlap with the
25
26 region identified as a clear outlier in JC in the MAF – and DAF – analysis above (see
27
28 also Fig. 2c), adding further support to this hypothesis.
29
30
31
32
33

34
35 **Long range haplotype analysis**

36 As a last approach for detecting selection in our data we used two different
37
38 methods based on the EHH measure (see Materials and Methods). Both methods aim
39
40 at detecting selection from local haplotype structure. A selective sweep causes the
41
42 selected allele to spread rapidly through the population, leading to more extended
43
44 linkage disequilibrium compared to the neutral expectation.
45
46
47

48 In the first approach we use the methodology introduced by Sabeti et al.
49
50 (2002), by looking at the relative EHH at a genetic distance of 0.3 cM in both
51
52 directions from a particular core haplotype. As before, we obtained a background
53
54 distribution to compare our genes to by analyzing the full 1 Mb regions. Figure 3
55
56 shows the distributions of REHH versus frequency for the three studied populations,
57
58 and Table 4 shows the list of significant core haplotypes after correction for multiple
59
60

testing. The most immediate result that can be seen is five cores with ~60% frequency which are clear outliers in the JC sample. Four of those remain significant after correcting for multiple testing (see Methods), and all of those cores are again found in *XYLT1*. Interestingly, all four core haplotypes only show significant extended haplotypes in the plus direction, indicating the possibility of a recombination hotspot in the minus direction of the cores (see below). The strongest signal comes from a 5 SNP core haplotype with an allele frequency of 0.622 (core XYLT1-1), which exhibits an extremely high REHH value of 80.7 (q-value 0.002). The signal extends beyond 0.5 cM, corresponding to roughly 200 kb in this region. The other significant result in our data comes from two adjacent core haplotypes around ~18 kb upstream of *UST* in the YRI (q-value 0.045). In addition, two SNPs within the cores also show F_{ST} values greater than the 0.99 percentile (rs4130936, $F_{ST} = 0.55$ and rs6938228, $F_{ST} = 0.54$). Haplotype bifurcation plots and REHH decay plots for the top core in *XYLT1* (XYLT1-1) as well as for one of the two significant cores in *UST* (UST-1) are given in Figure S3.

The second approach we employed is a modified version of the iHS score introduced by Voight et al. (2006). In our own implementation, instead of calculating the integral under the EHH decay plot we simply measure the total length of the haplotype between the two points where EHH falls to 0.25. We then calculate a normalized score with mean of zero and standard deviation of one, based on the ratio of the total haplotype lengths for each SNP (nHLS, see Materials & Methods for a more detailed description). Large positive or negative values indicate unusual extended haplotype homozygosity of haplotypes carrying the minor or major allele, respectively. The advantage of this method over the original approach of Sabeti et al. (2002) is that we now have a single value statistic that can be plotted along the

chromosome to look for trends in the values and identify regions of interest, e.g. that contain multiple SNPs with high scores. Furthermore, since we use a ratio of values at the two alleles, local differences in recombination rates should cancel each other out, making scores more comparable among regions. To identify interesting regions, we calculate and plot the average score as well as the proportion of SNPs with an absolute score above a given threshold within sliding windows of 15 consecutive SNPs. We validated our method by analyzing the Lactase region in CEU, and the Hemoglobin B region in the YRI, two well-established examples of positive selection in humans. Full details are given in the Supplementary information.

The total number of SNPs for which scores could be obtained in our datasets was 15,817 in CEU, 15,014 in JC and 18,307 in YRI. Table 5 gives a summary of all SNPs with high scores in our genes. Although none of the SNPs by themselves reached significance after multiple testing correction, we still identified two interesting candidate regions that merit further investigation, namely *ChGn* in the YRI and again *XYLT1* in the JC sample.

Three out of the top five scoring SNPs in the YRI are within *ChGn*, and are concentrated in a small stretch in the centre of the region (see Figure S4). Furthermore, this cluster contains a non-synonymous SNP (rs17128518, MAF = 0.075) which, although having a score slightly below our threshold for calling SNPs with high scores, shows the highest score of all non-synonymous SNPs in the YRI sample (nHLS = 2.702). In the JC sample, 8 out of the top 10 scores are found in *XYLT1*. The SNP with the highest score (rs7200407, nHLS = -3.988) has the second highest score of all analyzed SNPs in JC. Analyzing the distribution of scores along the region, it can be clearly seen that multiple SNPs with high scores cluster within the gene, particularly between position 17250 kb and 17450 kb (see Fig. 2d and e).

This region again coincides with the regions already identified in the other analyses. For example, the two top scoring SNPs are also both outliers in the 99th percentile of the F_{ST} distribution, with values of 0.59 (rs7200407) and 0.57 (rs4780701).

Figure 4 shows the haplotype bifurcation and EHH decay plots for two of the SNPs mentioned above. At the non-synonymous SNP in *ChGn* in the YRI (rs17128518, Figure 4a), the low frequent haplotype 2, we note that a stretch of sequence around 100 kb long surrounding the SNP is made up by only a single haplotype. In the JC, the highest scoring SNP in *XYLT1* (rs7200407, Figure 4b) shows a very suggestive extended haplotype pattern, with the EHH of the more frequent allele decaying much slower than that of the minor allele. Considering also the results of the previous methods, these results are strong indicators of positive selection in *XYLT1*.

Discussion

In this article we present the result of a study aimed at identifying candidate regions that could be involved in host-pathogen interaction in humans. We took advantage of the wealth of publicly available data, in two different forms: large databases of human variation data (the HapMap), as well as curated databases of metabolic pathways, that allow the quick retrieval of the complete set of genes known to be involved in a given biological process.

We utilized this pathway-centric approach to analyze all the genes in the biosynthesis of chondroitin sulfate and hyaluronic acid, two glycosaminoglycans with prior evidence for interaction with the malaria parasite *Plasmodium falciparum* in placental malaria (Beeson et al. 2000; Fried et al. 2006; Rogerson et al. 2007). Moreover, we also took advantage of the geographic distribution of the three HapMap populations, with very different pathogen environments and the African population (YRI) having been strongly affected by malaria.

We used three different methods to analyse our data, based on allele frequencies, population differentiation and long-range haplotypes. Given the recent origin of selection related to malaria, we expect the LRH analysis to be the most powerful (for a comparison of the different timeframes for each method, see Sabeti et al. (2006)). Another advantage of this method is that it is not significantly influenced by the HapMap ascertainment bias (see Voight et al. (2006)). Both allele frequency and population differentiation methods are more susceptible to this bias.

Nonetheless, each of these methods can give us additional information on the possibility of selection at a locus, and strong signals in each analysis will give more confidence in a detected signal (see also Supplementary information). In fact, a recent paper by Sabeti et al. (2007) used a similar strategy to filter results obtained from

LRH based analyses. Based on our analysis we identified three candidate genes that show evidence of positive selection: *UST* and *ChGn* in YRI and *XYLT1* in JC.

UST codes for a sulfotransferase that catalyzes the transfer of sulfate onto uronyl-residues of the polysaccharide backbone of the CS molecule. Two adjacent core haplotypes located 18 kb upstream of the gene showed a significant result for the LRH analysis in the YRI, with two SNPs within the cores also exhibiting large F_{ST} . Due to the 5' – location and lack of other genes located in this region, we speculate that this region could be involved in the transcriptional regulation of *UST*. Despite the quite large distance from the gene, important regulatory regions may very well be found tens of kilobases away from the actual coding region. An example of this is *LCT*, where two SNPs found to be associated with lactase persistence are 14 kb and 22 kb upstream of the gene, within two introns of a neighboring gene (Enattah et al. 2002).

Both *ChGn* and *XYLT1* code for enzymes involved in the synthesis of the polysaccharide backbone of the CS molecule. *ChGn* is interesting due to the fact that it contains the non-synonymous SNP with the highest nHL score of all analyzed regions in the YRI, within a small cluster of the highest scores in our genes of interest. It serves as an example that it is important to include functional information of the particular SNP in the analysis, as the score by itself, although high, is not found within the extreme outliers of the distribution. The SNP rs17128518 is only polymorphic in the YRI, with an allele frequency of 0.075. The cytosine to thymine transition in the first exon of the gene causes the substitution of a valin by an isoleucin at amino acid position 137 in the protein. Although this is a conservative substitution (Grantham 1974), it occurs within the catalytic domain of the protein, which could indicate a possible effect on protein function. However, since low frequency SNPs are

assumed to be of very recent origin and therefore generally occur on long extended haplotypes, this result has to be interpreted with caution. Tests for selection based on extended haplotypes have the highest power to detect partial sweeps for allele frequencies greater than 10% (Sabeti et al. 2006), which makes it difficult to distinguish a real selection event from an extreme case within neutral variation. Even so, the fact that our test is normalized using all scores calculated in the same frequency bin as well as the complete lack of variation in a 100 kb region surrounding the SNP could indicate that we are observing a very recent selective sweep in the region.

The most promising candidate of our study is *XYLTI*, which showed strong signatures of selection in the JC sample in all our analyses. Figure 2 gives a summary of the results in the region. The region from exons 1 – 3 is particularly interesting, where a very suggestive pattern of allele frequencies coincides with multiple extreme F_{ST} values as well as high nHL scores. Over most of the region we find that the CEU and YRI are more closely related than any of the two populations are with the JC, particularly upstream of exon two and around exon three. The most extreme nHLS values cluster on both sides of exon three. Interestingly, the significant core haplotypes observed in the REHH approach of the LRH analysis do not map to this region, but to the region from exons nine to eleven, which does not exhibit nHLS values as high (see Figure S5). On the other hand, we did not find any significant core haplotypes in the REHH approach in the region that showed the highest nHL scores. For a possible explanation of the different results we note that the region of the significant cores is flanked by two predicted hotspots of recombination, one of them immediately upstream and the other one roughly 50 kb downstream. Closer examination of the EHH decay around the cores revealed that EHH falls to below

0.25 at roughly the positions of the hotspots (data not shown), resulting in those haplotypes not being detected as unusually extended by our score method due to our measuring of the haplotype length at an EHH of 0.25. Despite the apparent discrepancy, the combined evidence presented above strongly suggests that the target of selection is more likely to be found in the region surrounding exons two and three.

Our primary aim was to identify variants related to placental malaria, which should lead to signatures of selection detectable in the African population, YRI. Indeed we found some evidence for selection in two regions in the YRI sample, *UST* and *ChGn*, but it is not totally conclusive. Apart from the possibility of a lack of power to detect very recent selective sweeps like in the case of *ChGn* (see above), the most parsimonious explanation would be that no variants with a strong protective effect exist in this set of genes in the YRI population. However, this does not rule out the possibility of adaptation related to placental malaria in these genes in other populations. In fact, one particular feature of protective variants in malaria is that different populations have evolved different genetic variants to combat the disease (Kwiatkowski 2005). It would be interesting to examine variation in these genes in other exposed populations to investigate this possibility. Another explanation would be that changes in the ability of infected erythrocytes to bind to HA and CS in the placenta have no significant impact on disease outcome in pregnant women. This explanation would challenge the idea that adhesion in the placenta is a key step in the pathogenesis, but given the evidence on the contrary (see Rogerson et al. (2007) for a review), this explanation seems unlikely.

The most promising candidate was found in the JC sample, which is a combined sample of two East Asian populations from Beijing, China and Tokyo, Japan. Although it cannot be ruled out that malaria transmission occurred in the past,

these regions are currently not endemic regions (World Health Organization 2005). Several studies have shown that the various forms of chondroitin sulfate serve as receptors for a variety of pathogens (Sugahara et al. 2003; Trowbridge and Gallo 2002), indicating the possibility that one of those pathogens might be the causative agent of the strong signal observed in our data. Encouragingly, *XYLT1* was also identified as candidate region for selection in two recent genome-wide studies using HapMap data (Tang et al. 2007; The International HapMap Consortium 2007).

In conclusion, our results show that an in-depth analysis of functional pathways can assist the identification of functionally important genetic variation. Mining of the publicly available human variability datasets allows this to be done in a rapid and cheap manner. The increasing amount of data on metabolic pathways and protein interactions provides additional information that should be incorporated to give a more detailed view of natural selection acting on functional pathways. We emphasize the broader applications of this approach to the rapidly emerging field of pathway genetics.

Acknowledgements

This study funded by Ministerio de Educación y Ciencia, Spain (SAF2007-63171 to JB, and BFU2005-00243), Direcció General de Recerca, Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608) and the National Institute for Bioinformatics (www.inab.org), a platform 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).

We are grateful to Pedro Alonso and Clara Menéndez for fruitful discussion about the pathogenesis of placental malaria. We also would like to thank Elena Bosch, Francesc Calafell, David Comas and Arcadi Navarro for providing very useful comments on the draft version of the manuscript and Andrés Moreno for helpful discussion. We also thank four anonymous reviewers for their helpful comments.

References

- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12: 1805-14
- Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, Patterson N, Reich D (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am. J. Hum. Genet.* 81: 234-42
- Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nat. Rev. Genet.* 4: 99-111
- Beeson JG, Rogerson SJ, Cooke BM, Reeder JC, Chai W, Lawson AM, Molyneux ME, Brown GV (2000) Adhesion of *Plasmodium falciparum*-infected erythrocytes to hyaluronic acid in placental malaria. *Nat. Med.* 6: 86-90
- Brabin BJ (1983) An analysis of malaria in pregnancy in Africa. *Bull. World Health Organ.* 61: 1005-16
- Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* 33 Suppl: 266-75
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, N.J.
- Desai M, ter Kuile FO, Nosten F, McGready R, Asamo K, Brabin B, Newman RD (2007) Epidemiology and burden of malaria in pregnancy. *Lancet Infect. Dis.* 7: 93-104
- Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30: 233-7
- Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1: 47-50
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479-91
- 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. *Exp. Parasitol.* 113: 36-42
- Gardner M, Williamson S, Casals F, Bosch E, Navarro A, Calafell F, Bertranpetit J, Comas D (2007) Extreme individual marker F(ST) values do not imply population-specific selection in humans: the NRG1 example. *Hum. Genet.* 121: 759-62
- Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185: 862-4
- Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP (2006) Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* 78: 153-9
- Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* 77: 171-92
- 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. *Trans. R. Soc. Trop. Med. Hyg.* 77: 232-44

- Menendez C, Ordi J, Ismail MR, Ventura PJ, Aponte JJ, Kahigwa E, Font F, Alonso PL (2000) The impact of placental malaria on gestational age and birth weight. *J. Infect. Dis.* 181: 1740-5
- Min-Oo G, Gros P (2005) Erythrocyte variants and the nature of their malaria protective effect. *Cell. Microbiol.* 7: 753-63
- 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 Med.* 2: e407
- Nielsen R (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* 39: 197-218
- Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am. J. Hum. Genet.* 74: 1198-208
- Okoko BJ, Ota MO, Yamuah LK, Idiong D, Mkpanam SN, Avieka A, Banya WA, Osinusi K (2002) Influence of placental malaria infection on foetal outcome in the Gambia: twenty years after Ian McGregor. *J. Health Popul. Nutr.* 20: 4-11
- Rogerson SJ, Hviid L, Duffy PE, Leke RF, Taylor DW (2007) Malaria in pregnancy: pathogenesis and immunity. *Lancet Infect. Dis.* 7: 105-17
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-7
- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. *Science* 312: 1614-20
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KAPI, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RAPI, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang HPI, Zeng CPI, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SBPL, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler DPI, Shen YPI, Yao Z, Huang WPI, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Wayne MMPI, Tsui SK, Xue H, Tze-Fei Wong J, Galver LMPL, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MSPI, Montpetit APL, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJPI, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-918
- Soldevila M, Calafell F, Helgason A, Stefansson K, Bertranpetit J (2005) Assessing the signatures of selection in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. *Trends Genet.* 21: 389-91
- Steketee RW, Wirima JJ, Slutsker L, Heymann DL, Breman JG (1996) The problem of malaria and malaria control in pregnancy in sub-Saharan Africa. *Am. J. Trop. Med. Hyg.* 55: 2-7

- Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100: 9440-5
- Sugahara K, Mikami T, Uyama T, Mizuguchi S, Nomura K, Kitagawa H (2003) Recent advances in the structural biology of chondroitin sulfate and dermatan sulfate. *Curr. Opin. Struct. Biol.* 13: 612-20
- Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biol.* 5: e171
- The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437: 1299-320
- The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861
- Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drouiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293: 455-62
- Trowbridge JM, Gallo RL (2002) Dermatan sulfate: new functions from an old glycosaminoglycan. *Glycobiology* 12: 117R-25R
- Varki A (1999) *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4: e72
- Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, de Bakker PI, Varilly P, Palma AA, Roy J, Cooper R, Winkler C, Zeng Y, de The G, Lander ES, O'Brien S, Altshuler D (2006) Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum. Genet.* 119: 92-102
- Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15: 1468-76
- 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: 637-42
- World Health Organization (2005) *World Malaria Report 2005*.

Figures

Figure 1 - Scatter plots of minor allele frequency (MAF) – and derived allele frequency (DAF) – analysis.

In all plots 100 kb windows that show overlap with a region ± 5 kb of the candidate genes (colored diamonds) are plotted against the background distribution of the remaining windows (grey dots). Grey shading and dotted lines indicate 95th and 99th percentile of the corresponding proportions based on the full distribution. Outliers are marked according to the legend in the respective figure (a) MAF – Analysis. Plots the proportion of SNPs with $MAF < 0.1$ versus the proportion of SNPs with $MAF > 0.4$. (b) DAF – Analysis. Plots the proportion of SNPs with $DAF < 0.2$ versus the proportion of SNPs with $DAF > 0.8$.

Figure 2 - Summary of signatures of selection in *XYLT1* in the JC sample

Depicted is a summary of the results of the different types of analyses in *XYLT1*. (a) F_{ST} for each SNP (black bars) and 15 SNP sliding window average of F_{ST} (grey line) over the region. Horizontal lines indicate mean (grey dashed line). (b) 15 SNP sliding window average of F_{ST} pairwise population comparisons CEU vs. JC (blue), CEU vs. YRI (yellow) and JC vs. YRI (green), 95th percentile (red dotted line), and 99th percentile (red dashed line) of the overall F_{ST} distribution. (c) DAF for each SNP in the region (yellow dots), compared to DAF in the CEU (grey dots). (d, e) nHLS profile over the region. Dashed line indicates 15 SNP sliding window average of nHLS. The bordered triangle indicates the only coding SNP (synonymous) in the region. The region from exons 1 to 3 shows interesting signatures in all types of analyses conducted.

Figure 3 - Distribution of REHH versus frequency

Core haplotypes within ± 5 kb of the candidate genes (colored diamonds) are plotted over the background distribution of cores from the full regions (grey dots) for the

study populations. Dashed lines indicate 0.95, 0.99 and 0.999 percentiles of REHH. Significant cores are marked according to the legend in the respective figure

Figure 4 - Haplotype bifurcation and REHH decay plots for two core SNPs
Shown are haplotype bifurcation plots (left) as well as EHH decay (right) over physical distance for two core SNPs. (a) Plots for rs17128518 (MAF = 0.075), the high score non-synonymous SNP in *ChGn*. To note is the complete homozygosity of the haplotypes carrying the minor allele (red) for a stretch of 100 kb, as seen by an EHH of 1 (right) and a lack of bifurcations (left). (b) Plot for rs7200407 (MAF = 0.183), the highest scoring SNP in *XYLT1*. It is clearly visible that the EHH decays much slower at the major allele (orange), despite the very high allele frequency (0.817), a clear indicator of selection leading to the high scores observed.

Table 1 - Summary of analyzed regions

Region	Chr	Gene ^a					Full region ^b				
		Start ^c	End ^c	Size	SNPs	kb /	Start ^c	End ^c	Size (kb)	SNPs	kb /
		(HG17)	(HG17)	(kb)	total	SNP ^d	(HG17)	(HG17)		total	SNP ^d
<i>B3GALT6</i>	1	1207568	1210341	2.773	1	2.773	761153	1685834	924.681	244	3.790
<i>B3GAT1</i>	11	133753612	133787022	33.410	27	1.237	133270337	134270132	999.795	1445	0.692
<i>B3GAT2</i>	6	71628087	71722853	94.766	105	0.903	71175614	72175437	999.823	1192	0.839
<i>B3GAT3</i>	11	62139345	62146024	6.679	2	3.340	61643248	62641929	998.681	880	1.135
<i>B4GALT7</i>	5	176959778	176969936	10.158	2	5.079	176467029	177456843	989.814	339	2.920
<i>ChGn</i>	8	19305952	19584374	278.422	511	0.545	18945192	19943977	998.785	1903	0.525
<i>CHPF</i>	2	220229183	220233943	4.760	2	2.380	219732070	220730466	998.396	1138	0.877
<i>CHST11</i>	12	103353271	103654156	300.885	585	0.514	103004424	104002857	998.433	1386	0.720
<i>CHST12</i>	7	2216464	2247455	30.991	25	1.240	1732172	2731745	999.573	971	1.029
<i>CHST13</i>	3	127725874	127744831	18.957	11	1.723	127236567	128232857	996.290	1188	0.839

<i>CHST3</i>	10	73394126	73443318	49.192	48	1.025	72919193	73918300	999.107	1129	0.885
<i>CHST7</i>	X	46189446	46214091	24.645	13	1.896	45739304	46699792	960.488	459	2.093
<i>CHSY1</i>	15	99533456	99609649	76.193	79	0.964	99073148	100070375	997.227	1402	0.711
<i>CSGlcA-T</i>	7	150367233	150373553	6.320	3	2.107	149870692	150870222	999.530	1115	0.896
<i>CSS3</i>	5	129268064	129550225	282.161	234	1.206	128910761	129904944	994.183	1070	0.929
<i>D4ST1</i>	15	38550505	38552645	2.140	2	1.070	38052602	39050730	998.128	881	1.133
<i>GALNAC4S-6ST</i>	10	125757176	125796231	39.055	16	2.441	125276809	126276656	999.847	1232	0.812
<i>GALNACT-2</i>	10	42953940	43000753	46.813	43	1.089	42479041	43475789	996.748	1065	0.936
<i>HAS1</i>	19	56908177	56919033	10.856	5	2.171	56414394	57413597	999.203	1416	0.706
<i>HAS2</i>	8	122694721	122722811	28.090	36	0.780	122210622	123206591	995.969	1409	0.707
<i>HAS3</i>	16	67698944	67709072	10.128	6	1.688	67207308	68203623	996.315	638	1.562
<i>UST</i>	6	149110157	149439818	329.661	548	0.602	148775867	149774212	998.345	1541	0.648
<i>XYLT1</i>	16	17108840	17472239	363.399	514	0.707	16791812	17790173	998.361	1274	0.784
<i>XYLT2</i>	17	45778392	45793511	15.119	9	1.680	45293520	46284502	990.982	965	1.027

total	23828.704	26282	0.907
--------------	------------------	--------------	--------------

- ^a Gene of interest
- ^b Full 1 Mb region centered on the respective gene
- ^c Start and end positions (HG17 assembly of UCSC Genome Browser) of the first and last SNP in the respective region
- ^d SNP Density, measured as average number of kilobases per SNP over the region

Table 2 - Outlier regions of MAF - DAF analysis

Region	Population	Position (kb) ^a	Proportions	
			<i>MAF</i> < 0.1 ^b	<i>MAF</i> > 0.4 ^b
<i>B3GALT6</i>	CEU	1171 – 1311	0.66 – 0.77	0.00
<i>B3GAT3</i>	JC	62043 – 62243	0.70 – 0.96	0.00
<i>HAS3</i>	JC	67657 – 67807	0.73 – 0.97	0.00 – 0.03
<i>XYLT1</i>	JC	17342 – 17572	0.66 – 0.90	0.00 – 0.02
<i>GALNACT-2</i>	YRI	42889 – 43019	0.55 – 0.62	0.02 – 0.04
			<i>DAF</i> < 0.2 ^b	<i>DAF</i> > 0.8 ^b
<i>B3GALT6</i>	CEU	1201 – 1311	0.48 – 0.50	0.42 – 0.43
<i>B3GAT3</i>	JC	62063 – 62163	0.48	0.48
<i>ChGn</i>	JC	19205 – 19425	0.38 – 0.54	0.40 – 0.48
<i>XYLT1</i>	JC	17352 – 17572*	0.44 – 0.58	0.41 – 0.51
<i>B3GAT1</i>	YRI	133610 – 133740	0.34 – 0.36	0.21 – 0.26
<i>GALNACT-2</i>	YRI	42909 – 43079	0.75 – 0.81	0.03 – 0.12

^a Start and end position of the window(s). Every single window covering the region shows the range of proportions given in columns (^b). The exception is the DAF – proportions of the XYLT1 – region in the JC sample (*) , where only 3 out of 12 overlapping windows in the region show the range of DAF proportions given

Table 3 - F_{ST} for the candidate genes

Gene	SNPs region ^a	Average F_{ST}				Max F_{ST}	$F_{ST} > 0.95$ percentile ^b	$F_{ST} > 0.99$ percentile ^b
		Global	CEU - JC	CEU - YRI	JC - YRI			
<i>B3GALT6</i>	5	0.024	0.020	0.028	0.096	0.093	0 (0)	0 (0)
<i>B3GAT1</i>	28	0.194	0.116	0.137	0.295	0.422	1 (0.036)	0 (0)
<i>B3GAT2</i>	113	0.197	0.169	0.316	0.130	0.527	17 (0.150)	0 (0)
<i>B3GAT3</i>	6	0.118	0.158	0.094	0.113	0.252	0 (0)	0 (0)
<i>B4GALT7</i>	10	0.086	0.011	0.114	0.133	0.167	0 (0)	0 (0)
<i>ChGn</i>	513	0.202	0.165	0.109	0.298	0.532	37 (0.072)	0 (0)
<i>CHPF</i>	9	0.104	0.069	0.108	0.131	0.165	0 (0)	0 (0)
<i>CHST11</i>	539	0.133	0.134	0.076	0.184	0.569	14 (0.026)	1 (0.002)
<i>CHST12</i>	26	0.070	0.054	0.052	0.099	0.201	0 (0)	0 (0)
<i>CHST13</i>	15	0.148	0.008	0.186	0.241	0.294	0 (0)	0 (0)
<i>CHST3</i>	54	0.181	0.183	0.177	0.184	0.381	1 (0.019)	0 (0)
<i>CHSY1</i>	87	0.255	0.170	0.254	0.338	0.471	9 (0.103)	0 (0)

<i>CSGlcA-T</i>	7	0.090	0.057	0.035	0.155	0.112	0 (0)	0 (0)
<i>CSS3</i>	234	0.069	0.037	0.078	0.093	0.365	0 (0)	0 (0)
<i>D4ST1</i>	9	0.259	0.107	0.292	0.349	0.478	1 (0.111)	0 (0)
<i>GALNAC4S-6ST</i>	17	0.099	0.100	0.075	0.119	0.288	0 (0)	0 (0)
<i>GALNACT-2</i>	47	0.199	0.178	0.065	0.275	0.360	0 (0)	0 (0)
<i>HAS1</i>	19	0.126	0.127	0.046	0.185	0.280	0 (0)	0 (0)
<i>HAS2</i>	48	0.224	0.281	0.142	0.232	0.449	6 (0.125)	0 (0)
<i>HAS3</i>	5	0.142	0.014	0.156	0.249	0.193	0 (0)	0 (0)
<i>UST</i>	543	0.133	0.122	0.078	0.176	0.578	23 (0.042)	2 (0.004)
<i>XYLT1</i>	484	0.253	0.196	0.178	0.356	0.663	62 (0.128)	22 (0.045)
<i>XYLT2</i>	19	0.204	0.248	0.075	0.283	0.380	1 (0.053)	0 (0)

^a All SNPs within ± 5 kb of the respective gene are considered

^b Number (proportion) of SNPs with F_{ST} greater than respective percentile of full F_{ST} distribution

Table 4 - Core haplotypes with significant REHH values in CS - HAS regions

Gene	Pop	Core ID	Start SNP ^a	End SNP ^a	Dir ^b	Sequence	Frequency	REHH	p-value	q-value
<i>XYLT1</i>	JC	XYLT1-1	438	442	+	AGCAT	0.622	80.704	9.42×10^{-06}	0.002
<i>XYLT1</i>	JC	XYLT1-2	418	437	+	GGTCGGTGTGATAATCTCTG	0.633	35.316	2.14×10^{-04}	0.020
<i>XYLT1</i>	JC	XYLT1-3	444	463	+	GTGGAAGCGTGGCTCGGACA	0.639	25.022	6.69×10^{-04}	0.043
<i>XYLT1</i>	JC	XYLT1-4	464	483	+	TTAAGCCAGGCCCTGCCCAG	0.644	21.868	1.02×10^{-03}	0.049
<i>UST</i>	YRI	UST-1	417	429	-	GTGCTGAGGGTGT	0.558	6.620	6.31×10^{-04}	0.045
<i>UST</i>	YRI	UST-1	417	429	+	GTGCTGAGGGTGT	0.558	5.452	1.27×10^{-03}	0.045
<i>UST</i>	YRI	UST-2	412	416	-	GAGTT	0.558	5.789	1.03×10^{-03}	0.045
<i>UST</i>	YRI	UST-2	412	416	+	GAGTT	0.558	5.571	1.18×10^{-03}	0.045

XYLT1 cores are in or within 5kb of the gene; *UST* cores are 17kb upstream

^a Index number of the SNP within the region analyzed

^b (+) indicates upstream, (-) downstream direction

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Table 5 - SNPs with | nHLS | > 3 in CSA-HAS genes

SNP ID	Gene	Pop	SNP No. ^a	Minor Allele ^b	Other Allele ^b	MAF	nHLS	p-value	q-value	Rank in Population	SNP function
rs7515488	<i>B3GALT6</i>	CEU	121	T	C	0.150	-3.672	2.40 x 10 ⁻⁰⁴	0.136	27	intron
rs6586856	<i>ChGn</i>	CEU	1196	G	A	0.108	3.494	4.76 x 10 ⁻⁰⁴	0.154	37	unknown
rs4148914	<i>CHST3</i>	CEU	674	A	G	0.208	3.374	7.41 x 10 ⁻⁰⁴	0.191	49	untranslated
rs12195841	<i>UST</i>	CEU	483	G	A	0.333	3.358	7.84 x 10 ⁻⁰⁴	0.114	50	intron
rs687087	<i>B3GAT2</i>	CEU	480	C	G	0.333	-3.257	1.13 x 10 ⁻⁰³	0.135	69	intron
rs1552777	<i>ChGn</i>	CEU	1170	T	G	0.075	3.064	2.18 x 10 ⁻⁰³	0.364	96	untranslated
rs11155595	<i>UST</i>	CEU	488	A	C	0.325	3.047	2.31 x 10 ⁻⁰³	0.230	98	intron
rs6985674	<i>ChGn</i>	CEU	1184	T	C	0.258	3.020	2.53 x 10 ⁻⁰³	0.482	105	untranslated
rs12208914	<i>UST</i>	CEU	480	T	C	0.042	3.015	2.57 x 10 ⁻⁰³	0.387	106	intron
rs7200407	<i>XYLT1</i>	JC	706	G	A	0.183	-3.988	6.68 x 10 ⁻⁰⁵	0.093	2	intron
rs4780701	<i>XYLT1</i>	JC	703	A	G	0.189	-3.865	1.11 x 10 ⁻⁰⁴	0.093	7	intron
rs7196324	<i>XYLT1</i>	JC	643	T	A	0.183	-3.621	2.93 x 10 ⁻⁰⁴	0.127	14	intron

rs7750238	<i>B3GAT2</i>	JC	523	T	C	0.050	3.582	3.41×10^{-04}	0.111	18	intron
rs2310850	<i>XYLT1</i>	JC	697	A	G	0.211	-3.532	4.12×10^{-04}	0.299	22	intron
rs6498687	<i>XYLT1</i>	JC	779	C	A	0.228	-3.442	5.77×10^{-04}	0.299	28	intron
rs8061042	<i>XYLT1</i>	JC	780	G	A	0.228	-3.442	5.77×10^{-04}	0.299	28	intron
rs6498703	<i>XYLT1</i>	JC	850	A	G	0.078	-3.379	7.28×10^{-04}	0.310	34	intron
rs7206651	<i>XYLT1</i>	JC	711	T	C	0.183	-3.207	1.34×10^{-03}	0.300	52	intron
rs2694406	<i>CHST11</i>	JC	779	C	G	0.433	-3.180	1.47×10^{-03}	0.610	56	intron
rs7500021	<i>XYLT1</i>	JC	793	C	T	0.122	-3.071	2.13×10^{-03}	0.410	73	intron
rs4782018	<i>XYLT1</i>	JC	654	C	T	0.172	-3.041	2.36×10^{-03}	0.300	78	intron
rs2435385	<i>GALNACT-2</i>	JC	557	G	A	0.450	-3.032	2.43×10^{-03}	0.610	80	intron
rs10111704	<i>ChGn</i>	YRI	897	T	C	0.108	3.196	1.39×10^{-03}	0.835	19	untranslated
rs7823526	<i>ChGn</i>	YRI	864	G	A	0.242	3.159	1.58×10^{-03}	0.385	22	intron
rs11112170	<i>CHST11</i>	YRI	820	A	C	0.033	-3.114	1.85×10^{-03}	0.463	31	intron
rs17128520	<i>ChGn</i>	YRI	869	C	T	0.042	3.073	2.12×10^{-03}	0.463	36	untranslated
rs1047577	<i>B4GALT7</i>	YRI	176	C	T	0.483	3.010	2.61×10^{-03}	0.663	46	unknown

All SNPs within ± 5 kb of the respective gene are considered

^a Index number of the SNP within the region analyzed

^b Alleles with the unusual extended haplotype are shown in bold italic typeface

For Peer Review

304x203mm (600 x 600 DPI)

254x304mm (600 x 600 DPI)

342x114mm (600 x 600 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

562x237mm (600 x 600 DPI)

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 (Bersaglieri et al. 2004; Voight et al. 2006), 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 (Enattah et al. 2002), 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 (Kwiatkowski 2005; Tishkoff and Williams 2002), due to the roughly 10-fold reduction of risk of severe malaria for heterozygote carriers (Ackerman et al. 2005). 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.

Results of allele frequency analysis and F_{ST} analysis for *LCT* & *HBB*

We also applied both the allele frequency- and the F_{ST} analysis to *LCT* and *HBB* in order to assess the performance of the other two methods in the positive controls.

In case of *LCT*, we find strong evidence of selection with both methods. In the allele-frequency methods, a 1 Mb region containing *LCT* and the putative functional SNP rs4988235 at the 5' region show an excess of windows with allele-frequency

proportions greater than the 95th percentile of the empirical background distribution (Figure S7). Also worth noting is the complete absence of intermediate frequency alleles as shown in Figure S7 (B). This region also corresponds very well with the region of high nHLS scores discussed above (see Figure S6). Results of the F_{ST} analysis are shown in Table S2 and Figure S8. Looking at the average F_{ST} of the gene, *LCT* does not stand out exceptionally, although in the pairwise comparison of CEU vs YRI the average F_{ST} is high with 0.379. However, over the whole region we find multiple SNPs with extreme F_{ST} , and the putatively associated SNP rs4988235 shows an F_{ST} of 0.769, which is the 6th highest value in the whole 2 Mb region (Figure S8).

In case of *HBB*, we do not expect to find strong signals with these two methods, given the strong deleterious effect in homozygote individuals. Due to that, *HBB* should be considered to be under balancing selection, because the selected allele cannot rise above a certain frequency. Our results confirm this notion. In the allele-frequency methods, we do not find any outlier regions (Figure S9). It is worth noting that we also do not detect it as an outlier with an excess of intermediate frequencies, which would be interpreted as balancing selection. This is explained by the particular characteristics of selection at the *HBB* locus, which is balanced at a very low frequency (0.12 in the HapMap YRI). This also explains why we do not observe an accumulation of high F_{ST} in the region (Figure S10 and Table S2).

References

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. *Ann. Hum. Genet.* 69: 559-65

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74: 1111-20

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30: 233-7

Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* 77: 171-92

Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3: 611-21

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

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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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 shows 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.

Figure S7 – MAF – and DAF - analysis for *LCT* region

Results of MAF – and DAF – analysis for a 5 Mb region centered on *LCT* in the CEU. Plotted are proportions of SNPs with MAF < 0.1 (A) or MAF > 0.4 (B) as well as DAF > 0.8 for each 100 kb window over the region. Dashed lines indicate 0.95 percentile of empirical background distribution.

Figure S8 – Distribution of F_{ST} along *LCT* region

(A) 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; (B) 15 SNP sliding window average F_{ST} for the three pairwise populations comparisons. Shown are results over a 2 Mb region centered on *LCT*, with the region of the gene ± 5 kb indicated through darker bars (A) or grey background (B). Asterisk indicates the very high F_{ST} value for rs4988235.

Figure S9 – MAF – and DAF - analysis for *HBB* region

Results of MAF – and DAF – analysis for a 1 Mb region centered on *HBB* in the YRI. Plotted are proportions of SNPs with MAF < 0.1 (A) or MAF > 0.4 (B) as well as DAF > 0.8 for each 100 kb window over the region. Dashed lines indicate 0.95 percentile of empirical background distribution.

Figure S10 – Distribution of F_{ST} along *HBB* region

(A) 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; (B) 15 SNP sliding window average F_{ST} for the three pairwise populations comparisons. Shown are results over a 1 Mb region centered on *HBB*, with the region of the gene ± 5 kb indicated through darker bars (top) or grey background (bottom).

Supplementary tables

Table S1. List of candidate genes and their function

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

^a UCSC Genome Browser known Gene ID^b from PANTHER Database

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Table S2. F_{ST} for *LCT* and *HBB*

Gene	SNPs region ^a	Average F_{ST}				Max F_{ST}	$F_{ST} > 0.95$ percentile ^b	$F_{ST} > 0.99$ percentile ^b
		Global	CEU - JC	CEU - YRI	JC - YRI			
<i>LCT</i>	61	0.199	0.118	0.379	0.129	0.446	5 (0.081)	0 (0)
<i>HBB</i>	25	0.157	0.137	0.146	0.171	0.408	2 (0.082)	0 (0)

^a All SNPs within ± 5 kb of the respective gene are considered
^b Number (proportion) of SNPs with F_{ST} greater than respective percentile of full F_{ST} distribution