Combining Evidence of Natural Selection with Association Analysis Increases Power to Detect Malaria-Resistance Variants

George Ayodo, Alkes L. Price, Alon Keinan, Arthur Ajwang, Michael F. Otieno, Alloys S. S. Orago, Nick Patterson, and David Reich

Statistical power to detect disease variants can be increased by weighting candidates by their evidence of natural selection. To demonstrate that this theoretical idea works in practice, we performed an association study of 10 putative resistance variants in 471 severe malaria cases and 474 controls from the Luo in Kenya. We replicated associations at *HBB* (P = .0008) and *CD36* (P = .03) but also showed that the same variants are unusually differentiated in frequency between the Luo and Yoruba (who historically have been exposed to malaria) and the Masai and Kikuyu (who have not been exposed). This empirically demonstrates that combining association analysis with evidence of natural selection can increase power to detect risk variants by orders of magnitude—up to P = .000018 for *HBB* and P = .00043 for *CD36*.

Malaria infection (MIM 248310) has exerted severe pressure on the human genome within the past 10,000 years,[1–3] and there are more cases today than ever before, with an estimated 300–660 million new episodes of clinical *Plasmodium falciparum* malaria every year.[4] Despite high infection rates, only 1%–2% of patients develop life-threatening complications, such as cerebral malaria and profound anemia,[5] so natural selection has likely operated, to a large extent, on severity. In the context of high infection rates, the genetics of host response are likely to play an important role.[6] In sub-Saharan Africa, the populations in which malaria is endemic generally have a lower proportion of cases with severe disease.[5,7] This suggests that there exist genetic variants that have risen to higher frequency in malaria-endemic populations because they modulate risk of *P. falciparum* malaria, similar to the case of the Duffy-null variant that protects against *P. vivax* malaria.[8]

A handful of genetic variants have already been associated with risk of or protection against severe malaria infection.[5] Our first objective in this study was to test variants of β-globin (*HbAS*[9,10]), intercellular adhesion molecule (*ICAM* TT[11]), *CD36* (*CD36* GT[12]), nitric oxide synthase (*NOS2A* 1659 AA[13]), tumor necrosis factor (*TNF* 238 A[14] and *TNF* 308 A[14–16]), Fc γ-receptor IIA (*CD32* AA[17,18]), interferon-α receptor-1 (*IFNARI* LI168V CC[19] and *IFNARI* 17470 CC[19]), and Toll-like receptor (*TLR4*[20]), which had previously been associated with malaria susceptibility. The particular phenotype we focused on was high levels of parasitemia in young children due to malaria infection.

Second, we compared the frequency differentiation in populations in which malaria is endemic and in closely related populations in which it is not endemic, searching for the differences that would be expected if natural selection had affected those alleles in one population but not in the other, because malaria began to affect only one group. Finally, we formally combined the evidence of association from case-control studies with evidence of natural selection in populations that have been exposed to malaria infection. We note that there has been discussion elsewhere of how one could formally combine case-control association studies with statistical weights obtained on the basis of evidence of natural selection.[21] Our goal in this study was to empirically demonstrate the power of this approach.

Material and Methods

Human Subjects

We collected 471 severe malaria cases and 474 controls from the Luo ethnic group, a population that speaks a Nilotic language and lives in a malaria-endemic region in western Kenya. All the severe malaria cases were collected from the Bondo District Hospital's children's emergency ward or from its outpatient clinic between May 2004 and August 2005. The average age of the cases was 2.6 years (table 1), reflecting our focus on individuals with no previous immunological protection against malaria. The controls were randomly collected from volunteers at nearby secondary schools, with an average age of 16.9 years (table 1). We focused on older controls, because we knew that they had survived to an older age. Thus, the control samples selected for this study may be slightly enriched for variants protecting against severe malaria, which should make it slightly easier to detect associations.

For the selection study, we assembled population control samples from the Masai, Kikuyu, and Yoruba ethnic groups. We collected samples from the Masai and Kikuyu from secondary schools in Narok and Nyeri, Kenya, respectively (table 1). The Yoruba samples were from the International Haplotype Map project[22]; we analyzed data from unrelated men and women, the parents in HapMap mother-father-child trios.

From the Department of Genetics, Harvard Medical School, and Broad Institute of Harvard and MIT (G.A; A.L.P; A.K; N.P; D.R), Boston; and Department of Pre-Clinical Sciences, Kenyatta University, Nairobi (G.A; A.A; M.F.O; A.S.S.O)

Received March 6, 2007; accepted for publication April 25, 2007; electronically published June 15, 2007.

Address for correspondence and reprints: Dr. David Reich, Harvard Medical School, Department of Genetics, 77 Avenue Louis Pasteur, New Research Building, Boston, MA 02115. E-mail: reich@genetics.med.harvard.edu

Am. J. Hum. Genet. 2007;81:000–000. 0002-9297/2007/8102-00XX$15.00
DOI: 10.1086/519221

Table 1. Characteristics of the Populations Included in This Study

Population	Mean Age (Range)	No. in Sample (Male/Female)	Source	Malaria Endemicity	Altitude above Sea Level (m)
Luo cases	2.6 (1.5–10.0)	471 (232/239)	Bondo District Hospital, Kenya	Endemic	~1,240
Luo controls	16.9 (14–20)	474 (290/184)	Bondo schools, Kenya	Endemic	~1,240
Masai controls	16.9 (13–21)	97 (42/55)	Narok schools, Kenya	Nonendemic	~1,880
Kikuyu controls	17.1 (15–19)	110 (46/64)	Nyeri schools, Kenya	Nonendemic	~1,950
Yoruba controls	NA[a]	55 (27/28)	International Haplotype Map	Endemic	~700

[a] NA = not available.

About 2 ml of blood was obtained by venipuncture for all the samples we collected in Kenya. We extracted DNA within 10 h of blood collection, using a Qiagen DNA Blood mini kit, and then stored it at −20°C. All the participants provided informed consent, and, for children, informed consent was obtained from the parents and/or guardians. The study was reviewed and approved by the Harvard Medical School and Kenyatta University ethical review boards and by the Kenyan government.

Clinical Identification of Human Subjects with Severe Malaria

We identified human subjects who had severe malaria according to World Health Organization criteria. Blood smears and Giemsa staining were used to determine the asexual parasite count (parasitemia level). We identified cases as young children with >12 parasites per 200 red blood cells. All cases were also required to have overlapping clinical manifestations at the time of hospitalization, such as respiratory distress, convulsions, prostration, and hyperthermia (>39°C).

Genotyping at Candidate Genetic Variants

We genotyped all human subjects for 13 candidate malaria SNPs, using mass spectrometry (Sequenom).[23] We discarded SNPs with minor-allele frequency averaging <5% across the four ethnic groups, leaving 10 SNPs for subsequent analysis (table 2). Although the X-linked *G6PD* and *CD40* genes are important candidates for malaria-resistance genes,[3] we excluded them from this study because we wished to focus on autosomal SNPs that we could compare with an empirical panel of autosomal variants in the genome.

As an assessment of genotyping quality, we observed that, for 85 genotypes obtained in duplicate, there were 2 discrepancies, for a discordance rate of 2.4%. After removing samples with <80% genotyping completeness, we found that the average completeness of genotypes was 97.8%. We also compared genotyping of 10 SNPs in Yoruba samples with data from HapMap.[22] Of 30 genotypes obtained in duplicate, there was 1 discrepancy (3.33%). All SNPs were in Hardy-Weinberg equilibrium in all the ethnic groups we studied ($P > .05$).

Genotyping at 1,454 Random SNPs

For the assessment of allele-frequency differentiation at random SNPs, we used the Illumina Bead Lab System to genotype 1,536 random SNPs from the Illumina linkage panel (covering chromosomes 1, 2, 3, and 22) in 45 of the Luo controls, 47 Masai controls, and 37 Kikuyu controls. We also obtained genotypes for these SNPs in 55 Yoruba samples from the HapMap database.[22] Of these SNPs, 1,454 passed standard quality checks and had been genotyped in all four populations.

Case-Control Association Analysis

We assessed the statistical significance of allele-frequency differences between Luo cases and Luo controls, using a χ^2 test with 1 df. We used a one-tailed test of statistical significance, since our interest was in assessing whether a genotype or allele previously associated with malaria is more common in cases than in controls. We computed odds ratios (ORs) as $A = (f_{case}/1 - f_{case})/(f_{control}/1 - f_{control})$, where f_{case} is the frequency in cases and $f_{control}$ is the frequency in controls. We also computed a 95% CI as the range of ORs that produced a likelihood ratio consistent with the data ($P > .05$). Specifically, we estimated the SE of the log OR as

$$B = \left(\frac{1}{n_{\text{case-ref}}} + \frac{1}{n_{\text{case-var}}} + \frac{1}{n_{\text{control-ref}}} + \frac{1}{n_{\text{control-var}}}\right)^{0.5},$$

where $n_{\text{case-ref}}$ and $n_{\text{case-var}}$ are the counts of the reference and variant genotypes in cases, and $n_{\text{control-ref}}$ and $n_{\text{control-var}}$ are the analagous quantities in controls. The 95% CI is quoted as the range ($e^{\ln(A)-1.65B}$ to $e^{\ln(A)+1.65B}$).

Epistasis Testing

To test for possible epistasis between any two SNPs, we used logistic regression. We compared the fit of three models with the data (case-control status for all the Luo samples): (1) genotype at the first SNP, (2) genotype at the second SNP, and (3) genotype at both SNPs.[24] We performed a one-tailed test for association with the genotypes previously associated with malaria. We calculated a Wald statistic and assessed significance for the epistatic interaction by a χ^2 test with 1 df.

Statistical Test for Natural Selection

The model of allele-frequency differentiation between two populations that we used to test for selection is that the difference in population frequencies at a given polymorphism is normally distributed with mean 0 and variance $cp(1-p)$, where p is the ancestral frequency. This model is similar to that of Nicholson et al.,[25] who showed that, for populations with modest genetic divergence times, it is a good approximation for allele-frequency differentiation. Under certain assumptions, the c parameter is expected to equal $2 \times F_{ST}$. From a population genetics perspective, c can be viewed as measuring genetic drift between populations.

To estimate c empirically, we used data from the 1,454 randomly chosen markers. For a given pair of populations, we estimated c as the empirical variance of the difference in population

Table 2. Replication Analysis for 10 Genotypes or Alleles Previously Associated with Malaria Susceptibility

Genotype or Allele	Reference SNP	Direction of Previous Association	Frequency in Controls (%)	No. of Cases/Controls Genotyped	OR (95% CI)	*P*
HbAS[10]	*rs334*	Protection[a,b]	25	447/454	.57 (.41–.79)	.0004
CD36 GT[12]	*rs3211938*	Risk[c]	12	456/457	1.50 (1.03–2.18)	.015
ICAM TT[11]	*rs5491*	Protection[a]	7	460/455	.71 (.42–1.21)	.10
NOS2A 1659 AA[13]	*rs8078340*	Risk[c,d]	6	450/455	.42 (.21–.83)	.99
TNF 238 A[14,15]	*rs361525*	Risk[c]	9	459/457	1.00 (.73–1.39)	.49
CD32 AA[17,18]	*rs1801274*	Protection[d,e]	25	455/447	.95 (.71–1.29)	.38
IFNARI LI168V CC[19]	*rs2257167*	Protection[c]	3	455/457	1.18 (.54–2.07)	.76
TNF 308 A[14,16]	*rs1800629*	Risk[c]	9	450/433	1.13 (.82–1.56)	.21
IFNARI 17470 CC[19]	*rs1012335*	Protection[c]	3	455/452	.85 (.53–1.36)	.34
TLR4 AG[20]	*rs4986790*	Risk[a]	10	407/303	1.36 (.85–2.17)	.10

[a] Previously published association with severe malaria.
[b] Previously published association with mild malaria.
[c] Previously published association with cerebral malaria.
[d] Previously published association with severe malarial anemia.
[e] Previously published association with parasitemia.

frequencies, after normalizing by $p(1-p)$ and accounting for sampling noise, which has variance $p(1-p)(1/N_1 + 1/N_2)$, where N_1 and N_2 are total allele counts for the two populations at a given marker. We approximated the normalization term $p(1-p)$ by setting p equal to the average of observed frequencies of the two populations, and we approximated binomial sampling noise as normally distributed. The same approximations were applied both to our estimation of c and to our subsequent analysis of individual markers. SNPs with average minor-allele frequency <5% for the two populations being compared were omitted from all computations, since the normal approximation becomes less reliable (table 3).

To test whether an individual marker was more differentiated than expected between two populations, we compared the observed difference in frequency with the expected distribution $N[0,p(1-p)(c + 1/N_1 + 1/N_2)]$, using the value of c estimated above, and computed a χ^2 statistic with 1 df. A feature of this test is that the χ^2 statistic has a mean value of 1 across the set of markers used to infer c. The test appropriately handles different sample sizes for candidate markers versus random markers used to infer c. A detailed statistical treatment will appear elsewhere (A.L.P, N.P., and D.R., unpublished data).

Combining Case-Control Association and the Test for Differentiating Selection

The combined test formally evaluates whether the observed data are consistent with the model of no case-control association and no selection. The test is performed by summing the association χ^2 statistic and the differentiation χ^2 statistic, forming a χ^2 statistic with 2 df. We note that the association χ^2 statistic used in this test is, by definition, a two-tailed statistic. We computed this sum for each pair of populations, using the same association statistic in each case. When one of the two populations being compared was the Luo population, we used the summed counts of Luo cases and Luo controls in the combined statistics reported in table 4. This generally leads to less significant *P* values than does using Luo controls only (and so is conservative). Using summed counts of Luo cases and Luo controls is appropriate under the null assumption of no association and ensures that the association statistic and differentiation statistic are independent. However, for the selection-only statistics reported in table 3, we used Luo controls only, since we wished to evaluate the evidence of selection in the control population, without regard to evidence of case-control association.

Results

Case-Control Association

We tested each of the 10 variants for association with malaria, comparing Luo cases with Luo controls. Two of the variants showed nominally statistically significant associations by one-tailed tests that searched for an association with the genotype or allele previously proposed to affect malaria resistance (table 3). We replicated the well-known association in which heterozygotes for the sickle-cell trait HbAS (*HbAS* T) are protected against severe malaria ($P = .0004$; OR 0.57 [95% CI 0.41–0.79]) (see the "Material and Methods" section). Although the OR of 0.57 is less strong than that observed in some previous studies,[9] it is in the same range as the OR of 0.45 (0.24–0.84), which was observed in another study of young children with a similar phenotype of severe malaria.[10] Different case-control studies focus on different phenotypes, and the protection of *HbAS* against severe malaria is known to vary with age,[26] so it is not surprising that the estimated ORs are heterogeneous across studies. We also replicated the association in which heterozygotes for *CD36* GT are at increased risk for severe malaria ($P < .015$; OR 1.50 [95% CI 1.03–2.18]).[12]

We note in passing that *NOSA* (*rs8078340*) gives a nominally significant *P* value (by a two-tailed test), but the association is in the opposite direction to previous reports ($P = .99$) (table 2). Our null findings at the other variants do not necessarily mean that they are unassociated; the CIs for the ORs are broad (table 2) and are often consistent with substantial association. We also note that our study included only individuals with parasitemia; we had no power to detect associations that were specific to cerebral malaria, a phenotype that was the focus of some previous studies.[13,27,28]

Table 3. Tests for Differentiating Selection between Malaria-Endemic and -Nonendemic Populations

Allele	Reference SNP	Allele Frequency (%) (No. of Alleles Used in Assessment)				*P*[a]			
		Luo	Yoruba	Masai	Kikuyu	Luo vs. Masai	Luo vs. Kikuyu	Yoruba vs. Masai	Yoruba vs. Kikuyu
HbAS T	*rs334*	13 (908)	11 (102)	0 (194)	0 (200)	**.00149**	**.00036**	**.044**	**.025**
CD36 G	*rs3211938*	6 (914)	22 (100)	1 (186)	0 (202)	NA	NA	**.00590**	**.00096**
ICAM T	*rs5491*	25 (910)	24 (100)	16 (186)	18 (206)	.19	.25	.41	.48
NOS2A 1659 A	*rs8078340*	21 (910)	19 (98)	25 (188)	21 (204)	.62	.86	.57	.89
TNF 238 A	*rs361525*	9 (914)	1 (100)	21 (192)	16 (202)	**.04**	.13	**.00741**	**.010**
CD32 A	*rs1801274*	50 (900)	50 (98)	50 (190)	44 (210)	1.0	.37	1.0	.56
IFNARI L168V C	*rs2257167*	16 (914)	16 (98)	25 (192)	19 (208)	.20	.61	.37	.72
TNF 308 A	*rs1800629*	9 (866)	6 (96)	6 (188)	7 (204)	.60	.74	1.0	.84
IFNARI 17470 C	*rs1012335*	32 (904)	22 (108)	33 (190)	35 (202)	.92	.64	.32	.18
TLR4 G	*rs4986790*	5 (633)	4 (114)	7 (178)	5 (170)	.60	1.0	.6	.84

[a] *P* values for selection are based on allele-frequency differentiation tests between malaria-endemic (Luo and Yoruba) and -nonendemic populations (Kikuyu and Masai). Values in bold are significant. Statistics for SNPs with average minor-allele frequency <5% for the two populations analyzed are denoted as NA (not available).

Finally, we tested for epistatic interactions between each pair of variants,[29] but no pair showed a statistically significant interaction by a Wald test (not shown). We also tested for different strengths of association by sex but found no evidence of this (table A1).

Allele-Frequency Differentiation and Tests for Natural Selection

To test for differentiating natural selection, we compared the frequencies of the putative susceptibility variants between populations in which malaria is endemic and nonendemic (tables 3 and A2). We observed the most-significant frequency differentiation at the two SNPs that also showed the strongest associations (table 3). The sickle-cell allele *HbAS* T is present at appreciable frequency in the Luo (13%) and Yoruba (11%) but is absent in the malaria-nonendemic Masai and Kikuyu. The *CD36* G allele is present at 22% in the Yoruba and at 6% in the Luo but occurs at only ~1% frequency in the populations in which malaria is nonendemic.

To test whether these allele-frequency differences are greater than what could be explained in the absence of selection, we compared them with a panel of 1,454 random SNPs[30] for which we obtained genotypes in 45 Luo, 47 Masai, 37 Kikuyu, and 59 Yoruba. (We first assessed whether there was evidence of population substructure in the Luo,[31] which could, in principle, confound our case-control tests of association. No structure was detected, indicating that population stratification is not likely to cause false-positive or false-negative results in the association analysis.) We also used the data to assess the genetic relationships among the populations; understanding this is crucial to the tests for differentiating selection.

The genetic differentiation among populations ranges from $F_{ST} = 0.0012$ between Masai and Kikuyu (lowest differentiation) to $F_{ST} = 0.021$ between Yoruba and Masai (highest differentiation). We found that the Luo and Masai do not cluster genetically, despite the fact that they both speak Nilotic languages, whereas the Masai and Kikuyu are closely related (despite the fact that the Kikuyu speak a Bantu language) (fig. 1). These results show that the linguistic patterns in Kenya do not correlate with the genetic patterns, which is at odds with what has been suggested elsewhere.[32] Sampling of more populations should elucidate the relationships between genetic and linguistic groups in East Africa.[33]

To formally test for differentiating selection, we computed a χ^2 statistic for frequency differentiation at each tested SNP, assuming it was drawn from the empirical distribution defined by 1,454 random SNPs (see the "Material and Methods" section and table 3). Allele-frequency differentiation between malaria-endemic and -nonendemic populations is significant at *HbAS* T ($P = .00036$ for the most extreme Luo-Kikuyu comparison) and *CD36* G ($P = .00096$ for Yoruba-Kikuyu), with the results significant even after use of a Bonferroni correction for testing 40 comparisons of malaria-endemic and -nonendemic populations at 10 SNPs (this essentially involves multiplying the nominal *P* values by a factor of 40). By contrast, the eight SNPs that do not give positive case-control association show no evidence of differentiating selection ($P > .25$ for each SNP and pair of populations after correction for multiple hypotheses tested; $P = .39$ for the sum of 32 χ^2 statistics at these eight SNPs). We further evaluated the robustness of our selection test by computing χ^2 statistics for each of the 1,454 random SNPs for each of the four pairs of populations. If the test is robust, we would expect to achieve a χ^2 value >3.84, with probability 0.05. Restricting the analysis to SNPs in which the average allele frequency across the two populations tested was at least 5%, we observed a χ^2 value >3.84 in 255 (5%) of 5,090 of tests performed. Similarly, only 4 of 5,090 tests produced a *P* value <.001, and the lowest *P* value was not statistically significant after correction for 5,090 hypotheses tested ($P > .16$). These results show that our test for differenti-

Table 4. Formal Combination of Case-Control Association Analysis and Tests of Natural Selection

Allele	Reference SNP	P[a] Luo vs. Masai	Luo vs. Kikuyu	Yoruba vs. Masai	Yoruba vs. Kikuyu
HbAS T	*rs334*	**.000056**	**.000018**	**.00048**	**.00029**
CD36 G	*rs3211938*	NA	NA	**.0023**	**.00043**
ICAM T	*rs5491*	.19	.23	.32	.36
NOS2A 1659 A	*rs8078340*	.033	.038	.032	.038
TNF 238 A	*rs361525*	.12	.30	.028	.038
CD32 A	*rs1801274*	.96	.65	.96	.81
IFNARI L168V C	*rs2257167*	.34	.68	.52	.73
TNF 308 A	*rs1800629*	.63	.69	.76	.75
IFNARI 17470 C	*rs1012335*	.91	.79	.56	.38
TLR4 G	*rs4986790*	.42	.41	.38	.42

[a] *P* values from combining case-control association studies with the test for differentiating selection between malaria-endemic (Luo and Yoruba) and -nonendemic (Masai and Kikuyu) populations. Values in bold are significant. NA = not available.

ating natural selection is not prone to false-positive results in a large selection of randomly chosen SNPs.

We note that both *HbAS* T and *CD36* G have been identified elsewhere as targets of recent positive natural selection.[22,34–36] However, the long-range haplotype test used to detect selection at these alleles detects evidence of selection from any cause and thus is not specific to a particular type of selection (e.g., for malaria resistance). The tests of allele-frequency differentiation we present here are much more specific to malaria. By comparing malaria-endemic and -nonendemic populations, we increase the probability that the loci detected as being affected by selection are specifically associated with malaria resistance. Of all the SNPs we tested for population differentiation—1,454 random SNPs and 10 candidates for malaria susceptibility—2 of those that achieve a nominal P value <.001 for at least one pair of populations were among the candidate malaria-resistance SNPs.

Combined Analysis of Case-Control Association and Selection

Finally, we formally combined the evidence of association with the evidence from the selection test (see the "Material and Methods" section). The combined test evaluates whether the observed data are consistent with the model of no case-control association and no selection. Whereas the evidence of association at *HbAS* T and *CD36* G is only moderate by the association analysis alone (see *P* values in table A1), significance is greatly increased when the association and selection evidence is combined: P = .000018–.00029 for *HbAS* T and P = .00043–.017 for *CD36* G, depending on which populations are compared (table 4). These results remain statistically significant after correction for 40 hypotheses tested (P = .00072 for *HbAS* T and P = .017 for *CD36* G).

Discussion

We performed a case-control association study of malaria resistance in the Luo, an East African population, analyzing 10 previously implicated variants. We replicated associations at *HbAS* (OR 0.57 [95% CI 0.41–0.79]) and *CD36* (OR 1.50 [95% CI 1.05–2.18]). Our OR for *CD36* is in agreement with the results published elsewhere.[12] Similarly, the OR for *HbAS* is in agreement with the previously reported longitudinal study in the same population (OR 0.45 [95% CI 0.24–0.84]; P = .0001).[10] For *HbAS,* the protective effect that we observed is smaller than in some previous reports, which is potentially due to the fact that the cases we studied were young (average age 2.6 years) and thus lacked an immune basis for *HbAS* protection. (Williams and colleagues showed that *HbAS* has a more protective effect for older individuals.[9,10,26]) A possible reason why we did not replicate all the previous associations is that, in our study, the phenotype was parasitemia, whereas previous studies sometimes focused on cerebral malaria (table 2). We also show in table 2 that the CIs for the ORs are broad; thus, many of the variants we tested are consistent with an effect on malaria susceptibility, even if we could not reject the hypothesis of no association.

A particularly striking observation is that, at *CD36,* where we observe significant case-control association and highly significant allele-frequency differentiation, the variant increasing susceptibility actually has higher frequency in malaria-endemic populations. A possible historical explanation is that the selection pressures on this variant may have changed over time because of host-parasite genetic interactions. For example, the variant may have historically reduced susceptibility to malaria, and then, as the parasite evolved to adapt to the human immune system, the allelic association might have reversed. This hypothesis would be consistent with the known temporal and geographical heterogeneity in *CD36* binding

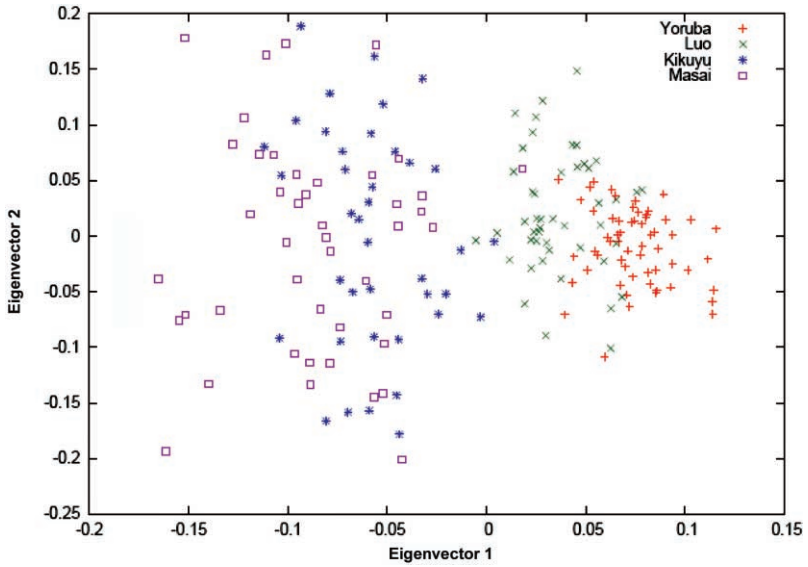

Figure 1. Principal-components analysis of samples from four different populations genotyped for 1,454 SNPs. The first eigenvector clusters the Yoruba, Luo, Kikuyu, and Masai. This is contrary to the expectation based on linguistics (the Luo and Masai both speak Nilotic languages) or geography. The F_{ST} values estimated between pairs of populations are as follows: Yoruba-Luo, 0.008; Yoruba-Masai, 0.021; Yoruba-Kikuyu, 0.015; Luo-Kikuyu, 0.008; Luo-Masai, 0.011; and Kikuyu-Masai, 0.0012. The top eigenvector is highly statistically significant by principal-components analysis[31] ($P \ll 10^{-12}$). The second eigenvector is not significant ($P = .09$).

and pathogenicity.[12] For example, genetic variation at the *PfEMP1* gene in the malaria parasite has been shown elsewhere to be associated with the pathogenicity,[37,38] and parasite *PfEMP1* and human *CD36* are known to interact.[39–41] In future studies, it will be interesting to explore whether human variants at *CD36* have different interactions with genetically different malaria parasite strains.[42]

These results finally provide empirical validation for a long-standing idea.[21] The idea is that, to increase power in case-control studies, one can combine the evidence of association with that from tests of natural selection. Previous studies have prioritized SNPs by natural selection on the basis of a combination of the alleles being frequent and being surrounded by a long-range haplotype[3,22,43]; the present study adds to this in several ways. First, we provide a formal χ^2 test of statistical significance, which can be combined with a case-control statistic to provide evidence that a SNP is a statistical outlier and, thus, a strong candidate for being associated with malaria. Second, our selection evidence is more specific to our phenotype of interest, since we are comparing frequency variants in populations differentiated by whether malaria has been historically endemic or nonendemic. Tishkoff et al.[33] recently applied a similar strategy to the phenotype of lactase persistence. They compared pastoral and nonpastoral populations in East Africa that have been differently exposed to diets including cow's milk. This analysis demonstrated high allele-frequency differences at variants near the lactase gene *LCT* and simultaneously showed that these highly differentiated variants also conferred the phenotype of lactase persistence.

We conclude that, in future whole-genome association scans, evidence from case-control comparisons can be combined with allele-frequency differentiation between differently exposed populations—and, potentially, other sources of evidence about recent selection[22,43]—to provide increased sensitivity and power in tests to detect disease-related genetic variants. It has been suggested that the identification of targets of selection may soon become a mainstream approach to finding genetic variants affecting human disease; our results provide empirical validation for this idea.[44] In our study, *P* values for *HbAS* and *CD36* were enhanced by several orders of magnitude with the use of <60 samples from each population analyzed, suggesting that this strategy may be cost effective relative to the number of additional samples needed to obtain a similar increase in power within the conventional case-control paradigm.

Acknowledgments

We are grateful to the human subjects and their families who participated in this study. We thank the medical officers and nursing staff in Bondo District Hospital, as well as the staff and students of the secondary schools where controls were collected. We are grateful to Pardis Sabeti, for thoughtful comments, and to Gavin McDonald, Alicja Waliszewska, Julie Neubauer, and Christine Schirmer, for technical advice. This work was supported by funds from Harvard Medical School, National Institutes of Health (NIH) grant R21-AI64519, and a Burroughs Wellcome Career Development Award in the Biomedical Sciences to D.R. A.L.P. is supported by a Ruth Kirschstein K-08 award from the NIH. We declare that we have no competing financial interests.

Appendix A

Table A1. Statistical Tests for Association

Genotype/Allele, Reference SNP, and Sex	Frequency in Luo Controls (%)	No. of Cases/ Controls	OR (95% CI)	*P* in Two-Tailed (One-Tailed) Tests of Previously Associated		*P* in Tests Based on 3 × 2 Genotype Table
				Genotype	Allele	
HbAS AT:						
rs334:						
All	25	447/454	.57 (.41–.79)	**.0008 (.0004)**	**.006 (.003)**	**.0031**
Female	25	232/175	.55 (.33–.90)	**.02 (.01)**		
Male	25	215/279	.59 (.38–.93)	**.02 (.01)**		
CD36 GT:						
rs3211938:						
All	12	456/457	1.50 (1.03–2.18)	**.03 (.015)**	**.06 (.03)**	.061
Female	12	232/174	1.40 (.85–2.33)	**1.19 (.60)**		
Male	13	224/283	1.57 (.90–2.74)	.11 (0.06)		
ICAM TT:						
rs5491:						
All	7	460/455	.71 (.42–1.21)	.20 (.10)	.89 (.45)	.22
Female	8	235/176	.43 (.18–1.00)	**.04 (.02)**		
Male	7	225/279	1.05 (.53–2.09)	.90 (.45)		
NOS2A 1659 AA:						
rs8078340:						
All	6	450/455	.42 (.21–.83)	.01 (.99)	.69 (.5)	**.015**
Female	6	229/179	.55 (.22–1.40)	.21 (.11)		
Male	6	221/276	.28 (.09–.85)	.02 (.01)		
TNF 238 A:						
rs361525:						
All	9	459/457	1.00 (.73–1.39)	.28 (.14)	.97 (.49)	.47
Female	8	233/173	1.33 (.82–2.17)	.24 (.12)		
Male	9	226/284	.74 (.47–1.16)	.20 (.10)		
CD32 AA:						
rs1801274:						
All	25	455/447	.95 (.71–1.29)	.76 (.38)	.81 (.45)	.95
Female	24	229/176	1.06 (.67–1.67)	.37 (.19)		
Male	25	226/271	.45 (.30–.70)	.66 (.33)		
IFNARI LI168V CC:						
rs2257167:						
All	3	455/457	1.18 (.54–2.07)	.48 (.76)	.93 (.47)	.86
Female	2	231/180	1.37 (.45–4.17)	.37 (.19)		
Male	3	224/277	.37 (.11–1.24)	.66 (.33)		
TNF 308 A:						
rs1800629:						
All	9	450/433	1.13 (.82–1.56)	.21 (.11)	.42 (.21)	.25
Female	6	234/166	1.33 (.75–2.37)	.33 (.17)		
Male	11	216/267	1.17 (.78–1.74)	.45 (.23)		
IFNARI 17470 CC:						
rs1012335:						
All	3	455/452	.85 (.53–1.36)	.68 (.34)	.70 (.35)	.78
Female	2	234/180	1.73 (.52–5.70)	.34 (.17)		
Male	3	221/272	.46 (.15–1.43)	.66 (.33)		
TLR4 AG:						
rs4986790:						
All	10	407/299	1.36 (.86–2.17)	.20 (.10)	.067 (.02)	.20
Female	9	201/123	1.48 (.70–3.16)	.67 (.34)		
Male	11	206/176	1.33 (.74–2.40)	.17 (.09)		

NOTE.—This table is an expansion of table 2. Values in bold are significant.

Table A2. Allele Counts in Cases and Controls for the 10 Polymorphisms

Allele	Reference SNP	Luo Controls		Luo Cases		Masai		Kikuyu		Yoruba	
		Ref	Var	Ref	Var	Ref	Var	Ref	Var	Ref	Var
HbAS T	*rs334*	789	119	819	75	194	0	200	0	91	11
CD36 G	*rs3211938*	856	58	883	79	184	2	202	0	78	22
ICAM T	*rs5491*	683	227	688	232	156	30	169	37	76	24
NOS2A 1659 A	*rs8078340*	714	196	713	187	141	47	162	42	79	19
TNF 238 G	*rs361525*	833	81	837	81	152	40	169	33	99	1
CD32 A	*rs1801274*	454	446	453	455	95	95	92	118	49	49
IFNAR1 LI168V C	*rs2257167*	764	150	762	148	144	48	168	40	98	16
TNF 308 A	*rs1800629*	791	75	813	87	176	12	189	15	90	6
IFNAR1 17470 C	*rs1012335*	616	288	622	280	128	62	131	71	84	24
TLR4 G	*rs4986790*	571	31	755	59	165	13	161	9	109	5

NOTE.—Ref = reference allele; Var = variant allele.

Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for malaria infection)

References

1. Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111
2. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, et al (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293:455–462
3. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837
4. Guerra CA, Snow RW, Hay SI (2006) Defining the global spatial limits of malaria transmission in 2005. Adv Parasitol 62: 157–179
5. Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77:171–192
6. Mackinnon MJ, Mwangi TW, Snow RW, Marsh K, Williams TN (2005) Heritability of malaria in Africa. PLoS Med 2:e340
7. Clarke SE, Brooker S, Njagi JK, Njau E, Estambale B, Muchiri E, Magnussen P (2004) Malaria morbidity among school children living in two areas of contrasting transmission in western Kenya. Am J Trop Med Hyg 71:732–738
8. Miller LH, Mason SJ, Clyde DF, McGinniss MH (1976) The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. N Engl J Med 295:302–304
9. Modiano D, Luoni G, Sirima BS, Simpore J, Verra F, Konate A, Rastrelli E, Olivieri A, Calissano C, Paganotti GM, et al (2001) Haemoglobin C protects against clinical Plasmodium falciparum malaria. Nature 414:305–308
10. Aidoo M, Terlouw DJ, Kolczak MS, McElroy PD, ter Kuile FO, Kariuki S, Nahlen BL, Lal AA, Udhayakumar V (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet 359:1311–1312
11. Kun JF, Klabunde J, Lell B, Luckner D, Alpers M, May J, Meyer C, Kremsner PG (1999) Association of the ICAM-1Kilifi mutation with protection against severe malaria in Lambarene, Gabon. Am J Trop Med Hyg 61:776–779
12. Aitman TJ, Cooper LD, Norsworthy PJ, Wahid FN, Gray JK, Curtis BR, McKeigue PM, Kwiatkowski D, Greenwood BM, Snow RW, et al (2000) Malaria susceptibility and CD36 mutation. Nature 405:1015–1016
13. Burgner D, Usen S, Rockett K, Jallow M, Ackerman H, Cervino A, Pinder M, Kwiatkowski DP (2003) Nucleotide and haplotypic diversity of the NOS2A promoter region and its relationship to cerebral malaria. Hum Genet 112:379–386
14. Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, Marsh K, Kwiatkowski D (1999) A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. Nat Genet 22:145–150
15. 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: 508–510
16. 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 6:472–480
17. Shi YP, Nahlen BL, Kariuki S, Urdahl KB, McElroy PD, Roberts JM, Lal AA (2001) Fcγ receptor IIa (CD32) polymorphism is associated with protection of infants against high-density *Plasmodium falciparum* infection. VII. Asembo Bay Cohort Project. J Infect Dis 184:107–111
18. Cooke GS, Aucan C, Walley AJ, Segal S, Greenwood BM, Kwiatkowski DP, Hill AV (2003) Association of Fcgamma receptor IIa (CD32) polymorphism with severe malaria in West Africa. Am J Trop Med Hyg 69:565–568
19. Aucan C, Walley AJ, Hennig BJ, Fitness J, Frodsham A, Zhang L, Kwiatkowski D, Hill AV (2003) Interferon-alpha receptor-1 (IFNAR1) variants are associated with protection against cerebral malaria in the Gambia. Genes Immun 4:275–282
20. Mockenhaupt FP, Cramer JP, Hamann L, Stegemann MS, Eckert J, Oh NR, Otchwemah RN, Dietz E, Ehrhardt S, Schroder NW, et al (2006) Toll-like receptor (TLR) polymorphisms in African children: common TLR-4 variants predispose to severe malaria. Proc Natl Acad Sci USA 103:177–182
21. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using

linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78:243–252
22. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
23. Tang K, Fu DJ, Julien D, Braun A, Cantor CR, Koster H (1999) Chip-based genotyping by mass spectrometry. Proc Natl Acad Sci USA 96:10016–10020
24. Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York
25. Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P (2002) Assessing population differentiation and isolation from single nucleotide polymorphism data. J R Stat Soc 64:695–715
26. Williams TN, Mwangi TW, Roberts DJ, Alexander ND, Weatherall DJ, Wambua S, Kortok M, Snow RW, Marsh K (2005) An immune basis for malaria protection by the sickle cell trait. PLoS Med 2:e128
27. Burgner D, Rockett K, Kwiatkowski D (1999) Nitric oxide and infectious diseases. Arch Dis Child 81:185–188
28. Xu W, Humphries S, Tomita M, Okuyama T, Matsuki M, Burgner D, Kwiatkowski D, Liu L, Charles IG (2000) Survey of the allelic frequency of a NOS2A promoter microsatellite in human populations: assessment of the NOS2A gene and predisposition to infectious disease. Nitric Oxide 4:379–383
29. Williams TN, Mwangi TW, Wambua S, Peto TE, Weatherall DJ, Gupta S, Recker M, Penman BS, Uyoga S, Macharia A, et al (2005) Negative epistasis between the malaria-protective effects of α^+-thalassemia and the sickle cell trait. Nat Genet 37:1253–1257
30. Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan JB, Chee MS, et al (2004) A highly informative SNP linkage panel for human genetic studies. Nat Methods 1:113–117
31. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190
32. Cavalli-Sforza LL (2005) The human genome diversity project: past, present and future. Nat Rev Genet 6:333–340
33. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40
34. 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 78:153–159
35. 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 312:1614–1620
36. Moormann AM, Embury PE, Opondo J, Sumba OP, Ouma JH, Kazura JW, John CC (2003) Frequencies of sickle cell trait and glucose-6-phosphate dehydrogenase deficiency differ in highland and nearby lowland malaria-endemic areas of Kenya. Trans R Soc Trop Med Hyg 97:513–514
37. Kraemer SM, Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. Mol Microbiol 50:1527–1538
38. Rottmann M, Lavstsen T, Mugasa JP, Kaestli M, Jensen AT, Muller D, Theander T, Beck HP (2006) Differential expression of *var* gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children. Infect Immun 74:3904–3911
39. Ndungu FM, Sanni L, Urban B, Stephens R, Newbold CI, Marsh K, Langhorne J (2006) CD4 T cells from malaria-non-exposed individuals respond to the CD36-binding domain of Plasmodium falciparum erythrocyte membrane protein-1 via an MHC class II-TCR-independent pathway. J Immunol 176: 5504–5512
40. Urban BC, Cordery D, Shafi MJ, Bull PC, Newbold CI, Williams TN, Marsh K (2006) The frequency of BDCA3-positive dendritic cells is increased in the peripheral circulation of Kenyan children with severe malaria. Infect Immun 74:6700–6706
41. Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, Ingle CE, Thomas A, Quail MA, Siebenthall K, Uhlemann AC, et al (2007) Genome variation and evolution of the malaria parasite Plasmodium falciparum. Nat Genet 39:120–125
42. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner DA Jr, Daily JP, Sarr O, Ndiaye D, Ndir O, et al (2007) A genome-wide map of diversity in Plasmodium falciparum. Nat Genet 39:113–119
43. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72
44. Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet 13:R245–R254