Research

Haplotypic analysis of the *TNF* locus by association efficiency and entropy

Hans Ackerman*†‡#, Stanley Usen§, Richard Mott*, Anna Richardson*‡, Fatoumatta Sisay-Joof§, Pauline Katundu¶, Terrie Taylor¶, Ryk Ward†¥, Malcolm Molyneux¶, Margaret Pinder§ and Dominic P Kwiatkowski*‡

Addresses: *Wellcome Trust for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. †Institute of Biological Anthropology, 58 Banbury Road, Oxford OX2 9QS, UK ‡University Department of Paediatrics, John Radcliffe Hospital, Oxford OX3 9DU, UK. §MRC Laboratories, Fajara, The Gambia. ¶Welcome Trust Research Laboratories and Malaria Project, College of Medicine, University of Malawi, Blantyre, Malawi. #Current address: Harvard Medical School, 64 Linnaean Street, Cambridge, MA 02138, USA. ¥Deceased.

Correspondence: Hans Ackerman. E-mail: ackerman@fas.harvard.edu

Published: 17 March 2003

Genome ***Biology*** 2003, **4**:R24

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/4/R24

Received: 12 July 2002
Revised: 8 December 2002
Accepted: 24 January 2003

Abstract

Background: To understand the causal basis of *TNF* associations with disease, it is necessary to understand the haplotypic structure of this locus. We genotyped 12 single-nucleotide polymorphisms (SNPs) distributed over 4.3 kilobases in 296 healthy, unrelated Gambian and Malawian adults. We generated 592 high-quality haplotypes by integrating family- and population-based reconstruction methods.

Results: We found 32 different haplotypes, of which 13 were shared between the two populations. Both populations were haplotypically diverse (gene diversity = 0.80, Gambia; 0.85, Malawi) and significantly differentiated ($p < 10^{-5}$ by exact test). More than a quarter of marker pairs showed evidence of intragenic recombination (29% Gambia; 27% Malawi). We applied two new methods of analyzing haplotypic data: association efficiency analysis (AEA), which describes the ability of each SNP to detect every other SNP in a case-control scenario; and the entropy maximization method (EMM), which selects the subset of SNPs that most effectively dissects the underlying haplotypic structure. AEA revealed that many SNPs in *TNF* are poor markers of each other. The EMM showed that 8 of 12 SNPs (Gambia) and 7 of 12 SNPs (Malawi) are required to describe 95% of the haplotypic diversity.

Conclusions: The *TNF* locus in the Gambian and Malawi sample is haplotypically diverse and has a rich history of intragenic recombination. As a consequence, a large proportion of *TNF* SNPs must be typed to detect a disease-modifying SNP at this locus. The most informative subset of SNPs to genotype differs between the two populations.

Background

The *TNF* locus (MIM *191160) has been associated with susceptibility to a wide range of infectious and inflammatory diseases, including malaria, typhoid, leishmaniasis, meningococcal sepsis, trachoma, asthma, multiple sclerosis, and inflammatory bowel disease [1-11]. Thus far, these associations have not been mapped in any detail, and as *TNF* lies in the central part of the major histocompatibility

complex (MHC), there are many candidate genes that could potentially be responsible. A strong candidate is *TNF* itself, as it encodes the potent pro-inflammatory cytokine TNF, and there is a significant body of clinical and experimental data suggesting a causal role for it in the pathogenesis of many of the diseases with which it has been associated. Furthermore, most of the reported associations are with polymorphisms located in the *TNF* promoter region, and cellular studies of gene regulation *in vitro* suggest that the molecular basis of the disease association may be, at least in some cases, a direct effect of the polymorphism on levels of gene expression [3,12].

To pursue the causal origin of these *TNF* disease associations we must begin with a detailed understanding of the allelic associations between different *TNF* SNPs. This is particularly important considering that within a few hundred base pairs, there are several potentially functional polymorphisms which show apparently independent disease associations with severe malaria [1-3]. Alternatively, these *TNF* SNPs may be serving as neutral markers of functional polymorphisms elsewhere in the central MHC. In order to understand, first, how the *TNF* SNPs relate to each other, and second, which SNPs are the best markers of the *TNF* locus in general, we applied two new analytical techniques to our haplotypic data. The first, association efficiency analysis (AEA), precisely defines the ability of one SNP to detect association at every other SNP in a case-control scenario. The second technique, entropy maximization method (EMM), selects those SNPs that most effectively dissect the underlying haplotypic structure of a locus. The results of these analyses allow us to prioritize SNPs for genotyping in future disease-association studies.

Results

Haplotyping the *TNF* locus

Twelve SNPs spanning 4.3 kb (Figure 1) were genotyped in 212 Gambian and 84 Malawian adults with no missing data. Allele frequencies for each SNP are listed in Table 1. The Gambian genotypic data had 354/2,544 (14%) sites where gametic phase was unknown, and the Malawian data had 188/1,008 (19%) sites where gametic phase was unknown. Where available, genotypes from offspring of the adults were used to phase these data (using the program PHAMILY), which reduced the number to 127/2,544 (5%) phase-unknown sites in the Gambian dataset and 75/1,008 (7%) phase-unknown sites in the Malawian dataset. The data were pooled and the program PHASE was then used to assign the remaining phase-unknown sites. After inferring haplotypes, only 6/2,544 (0.2%) phase assignments were less than 90% certain in the Gambian dataset, and only 2/1,008 (0.2%) phase assignments were less than 90% certain in the Malawian dataset. All other assignments (121/2,544 and 73/1,008) were greater than 90% certain. These 424 Gambian haplotypes and 168 Malawian haplotypes (Table 2) were the basis of subsequent analyses reported.

Haplotype distributions in two populations

In the Gambian sample, we observed 24 different haplotypes, and the distribution was dominated by one major haplotype (37.0%) and two others (16.7%, 16.5%) (Table 2, Figure 2). Thirteen haplotypes of low frequency made up another 28%, and 2% of alleles were singletons (haplotypes observed only once in the sample). The gene diversity, or probability that two haplotypes chosen at random from the sample are different, was 0.80 (standard deviation (SD) 0.01). There is one nonrecombinant haplotype that is found only in The Gambia (8).

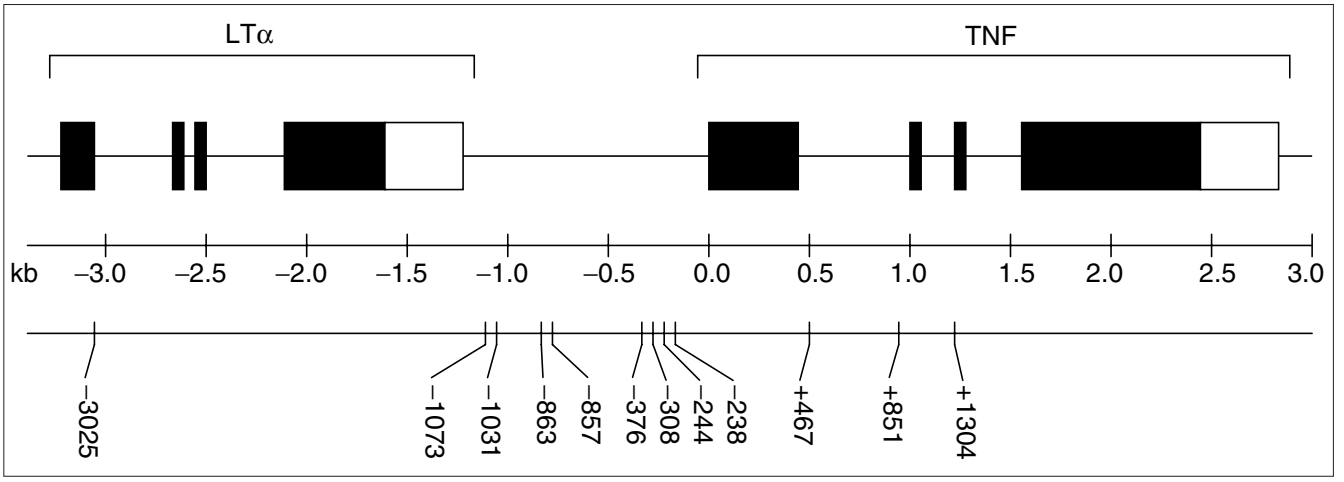

Figure 1
Diagram of the *TNF* locus drawn to scale with SNPs indicated. Filled boxes represent exons and the open boxes represent the 3′ untranslated region (3′ UTR). Positions are given in number of base pairs relative to the transcriptional start of *TNF*. The SNP at position -3025 is also referred to as an LTα *Nco*I restriction fragment length polymorphism in some references.

Table 1

Allele frequencies of 12 SNPs at the *TNF* locus in two populations

Polymorphism	Gambia (*n* = 424)	Malawi (*n* = 168)
TNF-3025 A/G	0.382	0.446
TNF-1073 T/C	0.002	0.030
TNF-1031 T/C	0.108	0.250
TNF-863 C/A	0.057	0.161
TNF-857 C/T	0.054	0.000
TNF-376 G/A	0.014	0.048
TNF-308 G/A	0.193	0.107
TNF-244 G/A	0.021	0.060
TNF-238 G/A	0.054	0.071
TNF+467 G/A	0.028	0.006
TNF+851 A/G	0.104	0.149
TNF+1304 A/G	0.085	0.131

The Malawian sample included 21 different haplotypes, 13 of which were shared with the Gambian population (Table 2, Figure 2). The predominant haplotype in Malawi was not as common, at 25.0%, and there were three other haplotypes with frequencies greater than 10% (23.2%, 16.1% and 10.1%). Haplotype 9, bearing the -1031C and -863A alleles, is much more common in Malawi than in The Gambia, at 16.1% versus 5.4%. Four percent of the Malawian alleles were present as singletons. The gene diversity in Malawi was slightly greater at 0.85 (SD 0.01).

Wright's *F* statistic reveals differences between populations by calculating the reduction in heterozygosity that is expected after random mating between the two populations. For the Gambian and Malawian populations the F_{ST} = 0.021. Permutations of the data under a null hypothesis of no differences between the populations gives a probability of 0.0001. An exact test of population differentiation finds the observed haplotype distributions to be a poor fit of a model of random mating between the Gambian and Malawian samples, with a probability less than 10^{-5}.

Recombination

Results of the four-gamete test on pairs of SNPs in each population reveal that in The Gambia, 19/66 (29%) pairs of loci have all four possible haplotypes present; while in Malawi, 15/55 (27%) pairs of SNPs have all four haplotypes present (Figure 3). Some high-frequency (and presumably older) SNPs show evidence of recombination in both populations (TNF-3025, TNF-1031). In The Gambia, the TNF-308 SNP shows evidence of recombination with six other SNPs. In Malawi, the TNF-244 shows evidence of recombination with many other SNPs. Interestingly, the TNF-863 SNP, of relatively high frequency in Malawi, shows no evidence of recombination with any other SNPs in the Malawian sample. Recombinant haplotypes of frequency greater than 1% are represented in Figure 2 with dotted lines. Although the

Table 2

TNF haplotype frequencies in two populations

		Gambia		Malawi	
Identifier	Haplotype	*n*	Frequency	*n*	Frequency
1	ATTCCGGGGGAA	157	0.370	42	0.250
2	.....GG	18	0.042	7	0.042
3	.....G	2	0.005	-	-
4	.....G.	-	-	3	0.018
5	.....A..	11	0.026	1	0.006
6	.....A...	2	0.005	-	-
7	.....A.....	4	0.009	-	-
8	....T.....	22	0.052	-	-
9	..CA.....	23	0.054	27	0.161
10	..C.....A.GG	13	0.031	2	0.012
11	..C.....	2	0.005	1	0.006
P	..C.....G	-	-	-	-
12	..C.....GG	1	0.002	-	-
13	..C....AA.GG	-	-	2	0.012
14	..C...A.....	1	0.002	-	-
15	..C..A..A.G.	6	0.014	1	0.006
16	..C..A..A.GG	-	-	4	0.024
17	..C..A.AA.G.	-	-	1	0.006
18	..C..A.AA.GG	-	-	2	0.012
19	G.....	70	0.165	39	0.232
20	G.....G	-	-	4	0.024
21	G.....G.	4	0.009	2	0.012
22	G.....GG	1	0.002	1	0.006
23	G.....A....	9	0.021	4	0.024
24	G...T.....	1	0.002	-	-
25	G..A.....	1	0.002	-	-
26	G.C.....	-	-	2	0.012
27	GC.....	1	0.002	5	0.030
28	G.....A.....	71	0.167	17	0.101
29	G.....A...GG	1	0.002	-	-
30	G.....A..A..	1	0.002	-	-
31	G.....A.A...	2	0.005	-	-
32	G.....AA....	-	-	1	0.006
		424		168	

SNPs are in genomic order from left to right as listed in Table 1. The haplotype labeled "P" was found in two chimpanzees, a gorilla, and two orang utans.

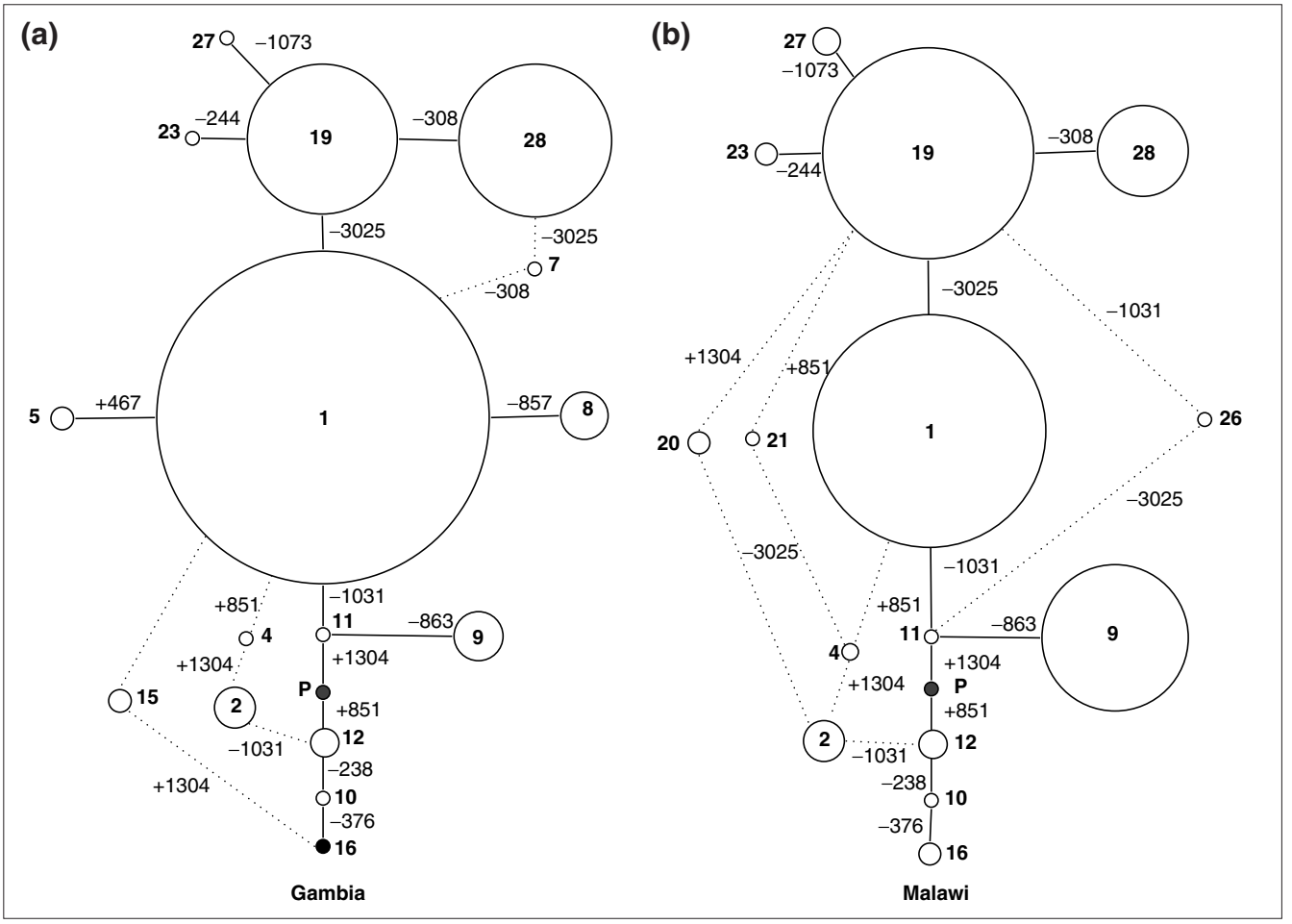

Figure 2
Minimum mutation networks of the *TNF* haplotypes. **(a)** The Gambian and **(b)** the Malawian population. Each circle represents a haplotype, and the size of the circle is proportional to the haplotype frequency. Each connecting line corresponds to a mutational event, and the position of the resulting SNP is indicated in base pairs from the transcriptional start of the gene. Broken lines indicate recombination. The numbers correspond to the haplotype indentifiers listed in Table 2. Filled circles represent haplotypes that were not found in that population. The haplotype labelled 'P' was found in two chimpanzees, a gorilla, and two orang utans. Haplotypes of frequency less than 1% are not shown.

branching structure of the *TNF* genealogy is similar in both populations, there are unique recombination events that have differentiated the two samples (haplotypes 17, 20 and 26). In Malawi, there appear to be more recombinant haplotypes of greater than 1% frequency.

Linkage disequilibrium

Consistent with the evidence for recombination, linkage disequilibrium is variable across the *TNF* locus (Figure 4c,d) and generally low. Values of pairwise linkage disequilibrium, r^2, are different between the two populations, though well correlated (r = 0.83). In The Gambia, 80% of SNP pairs have the minor alleles in the repulsion phase, while in Malawi, 65% of SNP pairs are in the repulsion phase. Most of the SNP pairs in the repulsion phase have very low linkage disequilibrium (r^2 = 0). There are six pairs of SNPs that are in the coupling phase in the Malawian sample, but are in the repulsion phase in The Gambian sample, reversing the allelic associations. Most of these occasions involve the TNF-244A allele, which is found on only one haplotype in the Gambia (in phase with the TNF-3025G allele), but on five different haplotypes in Malawi.

Association efficiency analysis

Using Equation (2) (see Materials and methods) we calculate the apparent relative risk at one marker given a relative risk of 10 at a second marker for all pairs of markers. As an example, we assign a relative risk of 10 to the TNF-376A allele (this is the relative risk of the A allele compared to the G allele) and calculate the relative risks at all other SNPs in the TNF gene using the haplotype frequencies from The Gambia (Figure 5). Log_{10} of the relative risk is plotted on the y-axis, so values above the x-axis correspond to disease risk while values below the x-axis correspond to disease protection. We

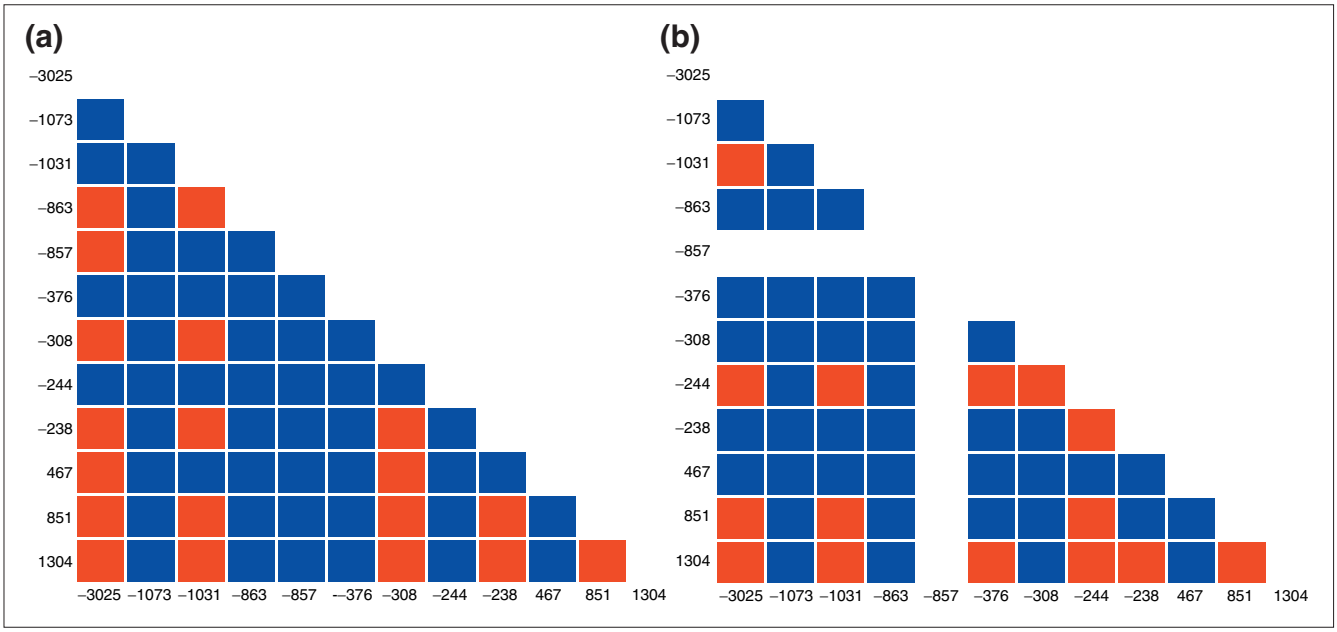

Figure 3
Recombination in the *TNF* gene region. **(a)** Four-gamete test, The Gambia; **(b)** four-gamete test, Malawi. Red squares indicate pairs of SNPs in which all four gametes were observed, which is evidence for recombination or recurrent mutation.

find that the relative risk varies greatly across the TNF gene region. The three SNPs that are in maximum positive linkage disequilibrium ($D' = 1$) with TNF-376A show relative risks of 2.2, 3.3, and 2.2 when TNF-376A gives a relative risk of 10. The remaining eight SNPs in the TNF gene would have a relative risk between 0.8 and 0.9 when the relative risk at TNF-376A is 10 (Figure 4).

Data for all pairs of markers are given in Figure 4a,b. Hypothetical disease alleles are given along the *x*-axis, and marker alleles are given on the *y*-axis. Each cell indicates the apparent relative risk (R) of a marker on the *y*-axis when the hypothetical disease allele indicated on the *x*-axis has a relative risk of 10. Using an arbitrary cutoff of $0.5 < R < 2.0$ (indicated by blue color), some markers perform poorly: for example in The Gambia TNF-308 can only detect TNF-3025 ($R = 2.9$). Interestingly, if the TNF-308 SNP were a disease allele, most of the other SNPs in TNF would show a moderate protective effect. TNF-1031 is good at detecting associations at seven other SNPs. In Malawi, the pattern of association efficiency is different, and generally stronger. Most hypothetical disease SNPs in Malawi would be detectable by other SNPs in TNF, with the exception of TNF+476 and TNF-1073 which have very low frequency and little prospect of being detected by other markers.

Association efficiency analysis compared to linkage disequilibrium

To illustrate the relationship between the association efficiency parameter and measures of linkage disequilibrium, we plot the $\log_{10}$ of relative risk for all pairs of SNPs (when one SNP is assigned a relative risk of 10) against r^2 (Figure 6a,b). Above the *x*-axis, r^2 appears to be correlated with the association efficiency parameter (correlation coefficients of $r = 0.77$ in The Gambia, and $r = 0.91$ in Malawi). Below the *x*-axis, where the marker SNP gives an apparently protective effect, the correlation is not as strong ($r = -0.61$ in The Gambia, $r = -0.60$ in Malawi). In this situation where minor alleles are in the repulsion phase, the values for r^2 remain low even when association efficiency parameter shows a strong protective effect.

The association efficiency parameter is correlated with D, the un-normalized linkage disequilibrium parameter (Figure 6c,d). When the absolute value of D is great, the association efficiency tends to be high; however, many SNP pairs also have high association efficiency even when the absolute value of D is low.

When we plot the association efficiency parameter against the normalized disequilibrium parameter, D', we see a strong positive correlation: $r = 0.89$ in the Gambia and $r = 0.91$ in Malawi (Figure 6e,f). Linkage disequilibrium is necessary to detect association; however, it is not sufficient. Despite $D' = 1.0$, some pairs of SNPs are poor markers of each other, showing relative risks between 0.5 and 2.0 when the linked hypothetical disease allele gives a relative risk of 10.

Entropy maximization method

Entropy (E), a measure of haplotypic diversity, was calculated for the full 12-SNP haplotypes in each population

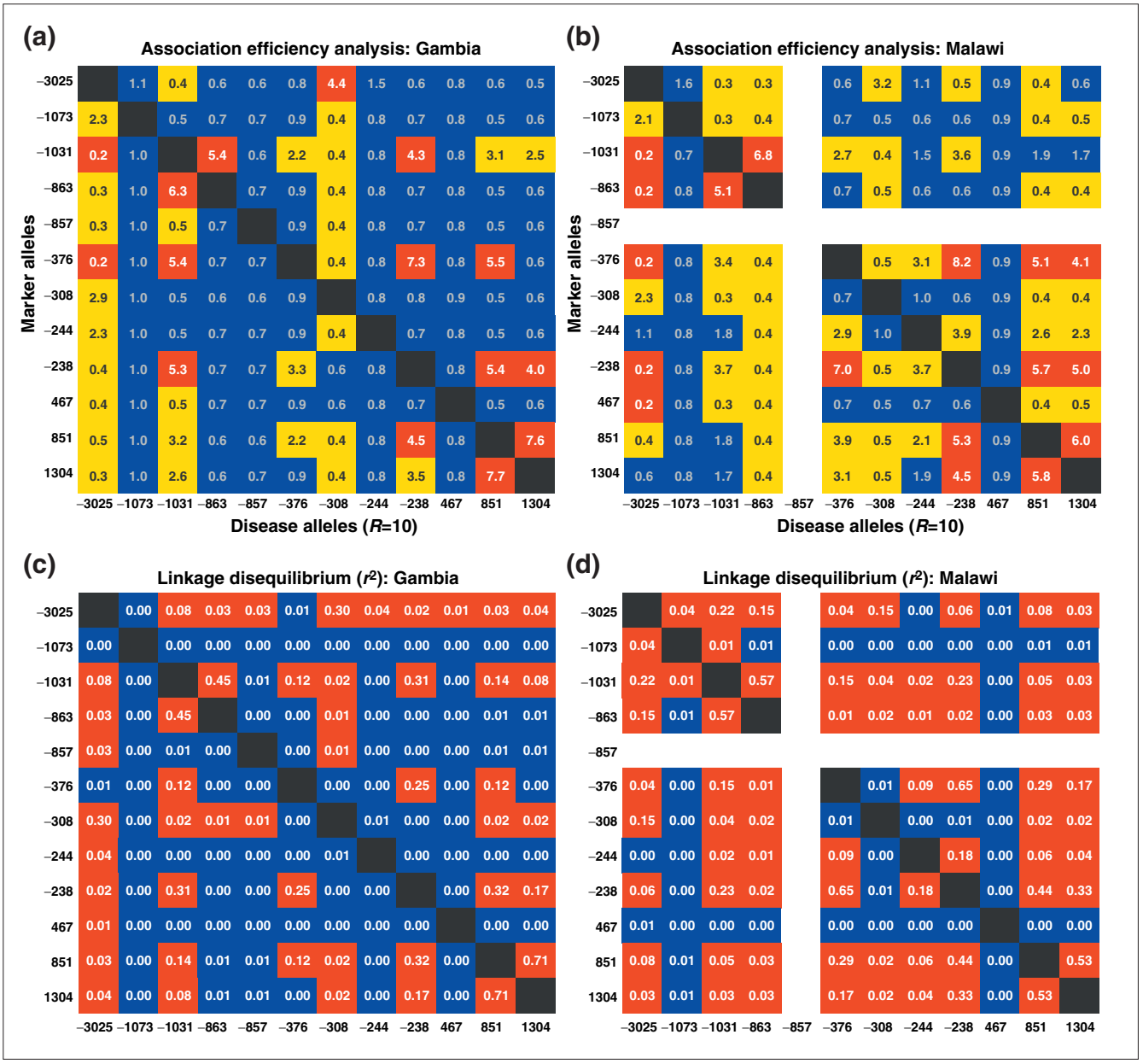

Figure 4
Association efficiency and linkage disequilibrium parameters for 12 SNPs at the *TNF* locus. **(a)** AEA, Gambia; **(b)** AEA, Malawi; **(c)** linkage disequilibrium, Gambia; **(d)** linkage disequilibrium, Malawi. The apparent relative risk at the marker allele indicated in the leftmost column is given when the hypothetical disease allele indicated in the bottom row is assigned a relative risk of 10. In (a,b) color indicates the magnitude of the apparent relative risk: blue indicates a relative risk of less than twofold, yellow between two- and fourfold, and red between fourfold and the maximum of 10-fold. In (c,d) red indicates $p < 0.05$ by the chi-squared test, uncorrected for multiple tests.

(Equation (4) in Materials and methods). In the Gambian sample, E = 2.95; in the Malawian sample, E = 3.23. To determine if a subset of these 12 SNPs could account for most of the haplotypic diversity in these samples, we calculated entropy for all possible subsets of SNPs, increasing the number of SNPs in a subset from 1 to 12. In Figure 7 we plot the subsets of markers with maximum entropy. We begin with a single SNP, and with each additional SNP we include, the entropy increases. The entropy plateaus when the additional markers contribute little additional information. To account for 95% of the haplotypic diversity of *TNF*, one must genotype 8 of 12 SNPs in The Gambian sample (in order of decreasing informativeness): -3025, -308, 851, -1031, -857, 467, -244, -238; and type 7 SNPs in the Malawian sample: -3025, -1031, 851, -308, 1304, -244, -1073.

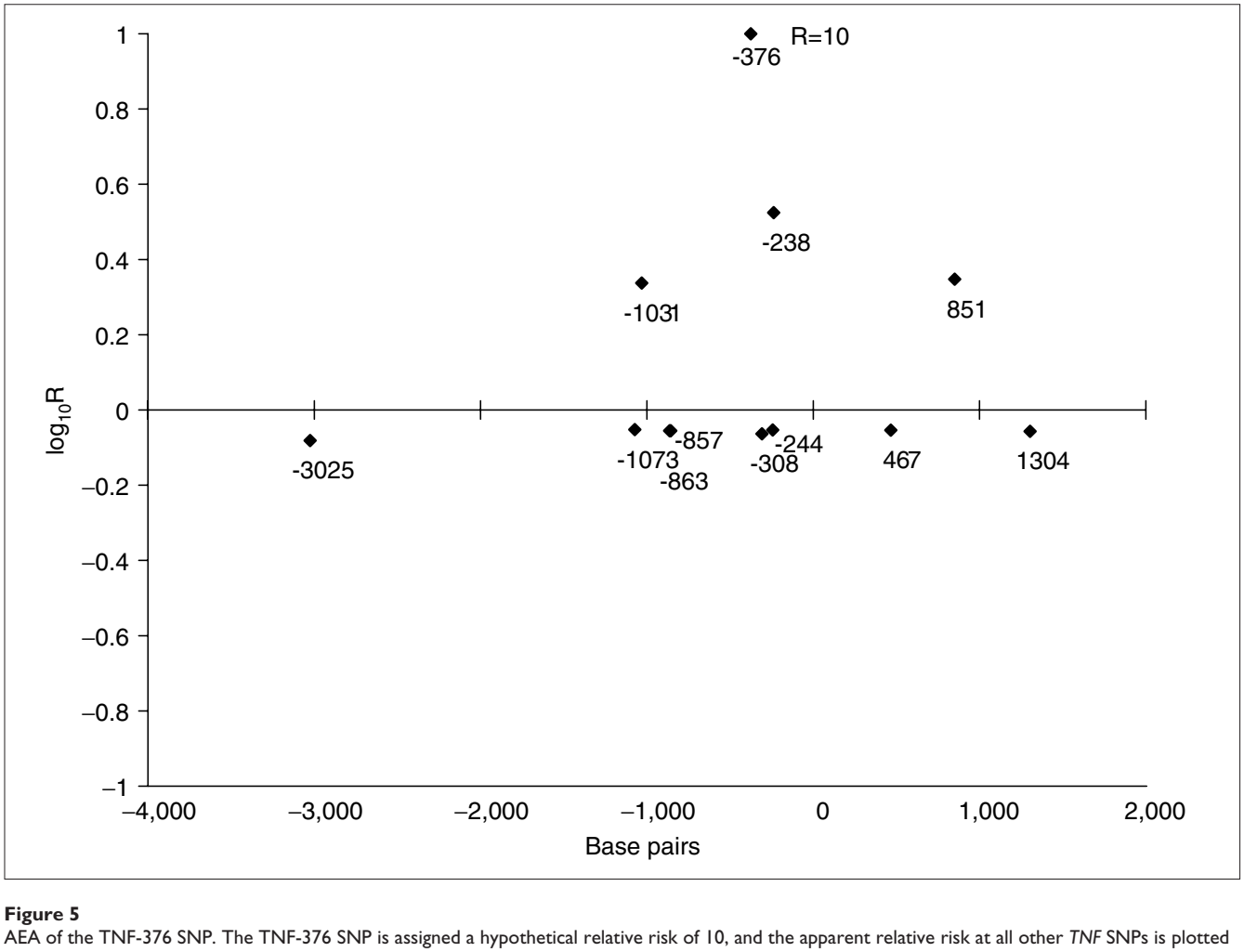

Figure 5
AEA of the TNF-376 SNP. The TNF-376 SNP is assigned a hypothetical relative risk of 10, and the apparent relative risk at all other *TNF* SNPs is plotted on a $\log_{10}$ scale ($\log_{10}$R). The positions of the other SNPs are indicated on the *x*-axis as base pairs from the start of transcription. Eight of the SNPs in the *TNF* gene would give a relative risk between 0.8 and 0.9 when the TNF-376 SNP gives a hypothetical relative risk of 10.

Discussion

The *TNF* locus has been associated with susceptibility to a wide range of infectious and inflammatory diseases. In African populations, three *TNF* promoter SNPs have been independently associated with severe malaria [1-3]. To better understand these complex genetic associations with malaria, we have defined the haplotypic structure of the *TNF* gene region using 12 SNPs genotyped in 212 Gambian and 84 Malawian adults. Gametic phase was determined by genotyping one offspring from each adult (when available), reducing the number of phase-unknown sites by 65%. The gametic phase of the remaining sites was determined by statistical inference. This integration of family- and population-based methods of haplotype reconstruction resulted in phase assignments of high certainty. This description of the pattern of DNA sequence variation provides insight into the genetic differentiation of these two African populations, the evolution of the TNF gene itself, and the suitability of these SNPs to serve as markers of disease association.

There are a number of interesting features that distinguish these two populations at the haplotypic level. Eleven haplotypes are unique to the Gambian sample, and eight are unique to the Malawian sample. The haplotype bearing the TNF-857T allele was not found in Malawi (this site is not polymorphic in the Malawian sample), whereas it has a frequency of 5.2% in The Gambia. When population differentiation is assessed by haplotype instead of by individual SNP frequencies, we observe a greater value for Wright's fixation index, F_{ST} = 0.021, compared to F_{ST} = 0.007 for individual TNF SNPs [13]. Intragenic recombination seems to have had an important role in diversifying the locus and in distinguishing the two populations, as evidenced by a number of population-specific recombinant haplotypes (7, 17, 20 and 26). In spite of the significant differences in haplotype distributions between the two African populations, it appears that most of the haplotypes have evolved over a period of time during which there was gene flow between the two sampled populations. That is, the basic branching structure of the

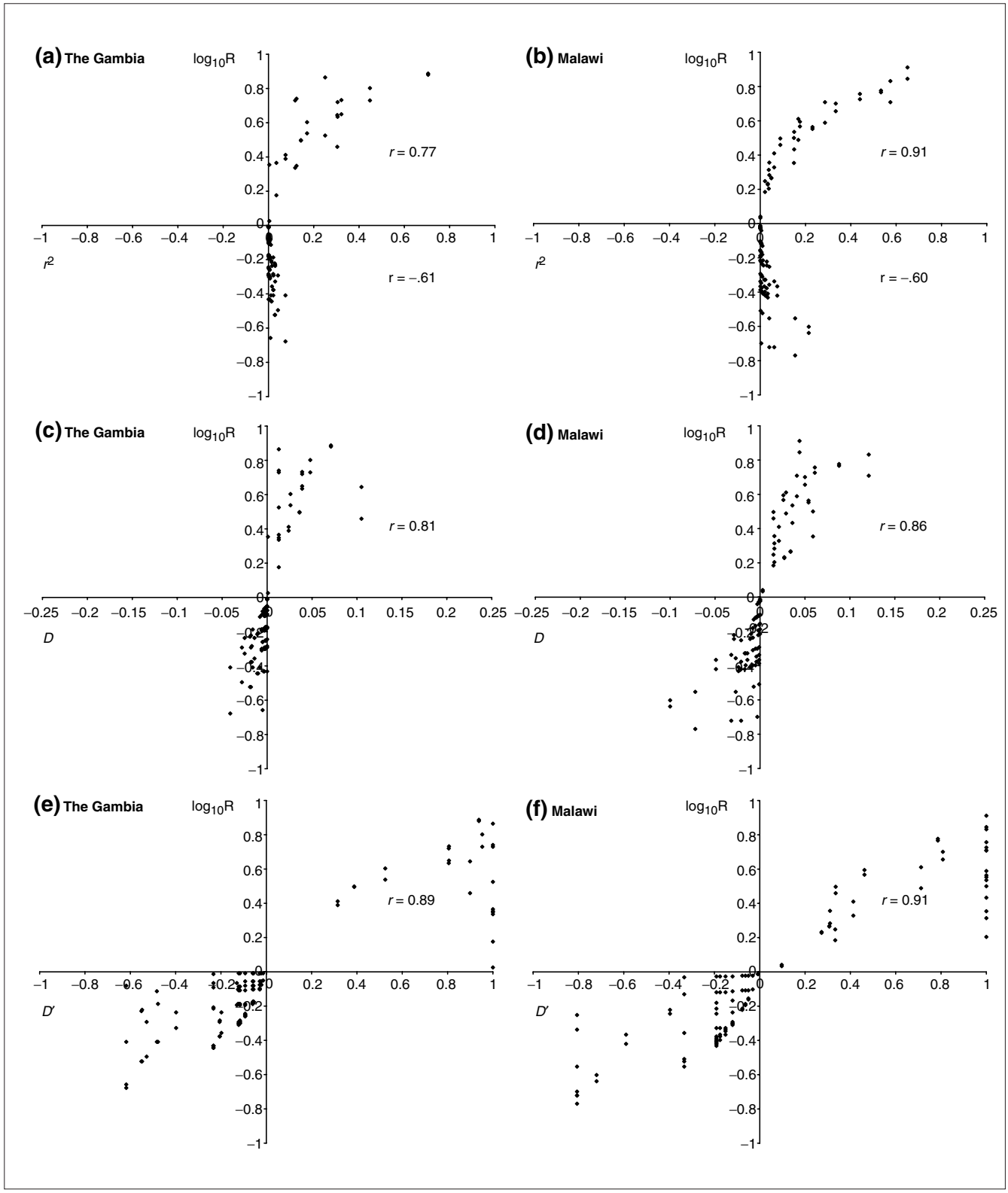

Figure 6
Plots of the AEA parameter versus linkage disequilibrium (LD) for all pairs of SNPs. The apparent relative risk for all marker SNPs (when paired with a hypothetical disease SNP of relative risk = 10) is plotted on a $\log_{10}$ scale against measures of LD in **(a,c,e)** the Gambian and **(b,d,f)** the Malawian populations. The AEA parameter is plotted against (a,b) r^2, (c,d) D, and (e,f) D'. The correlation coefficient (r) between the AEA parameter and the measure of LD is indicated on each graph.

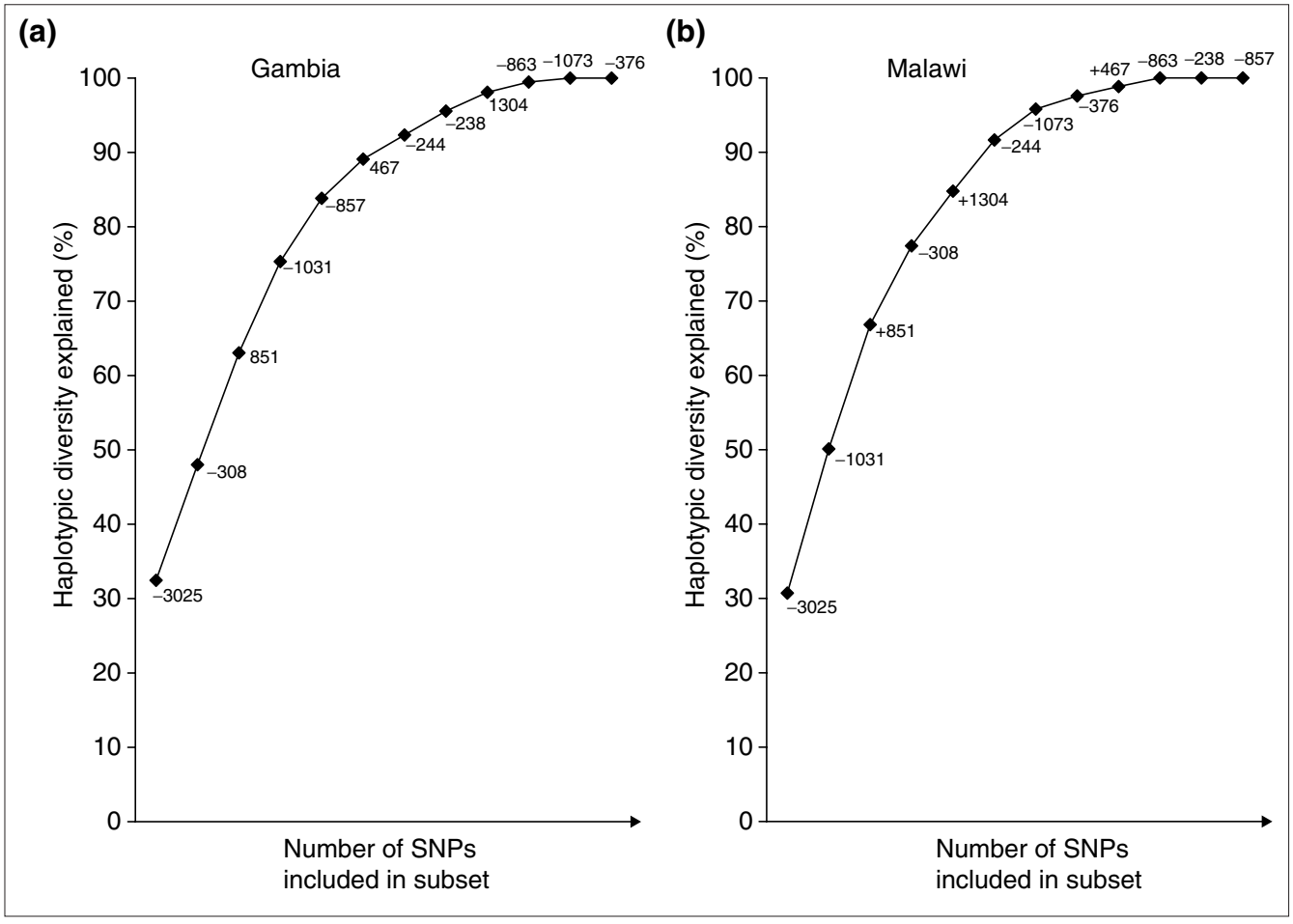

Figure 7
Results of the entropy maximization method. Plots of the percentage of haplotypic diversity accounted for by a subset of SNPs as the size of the subset goes from 1 SNP to 12 SNPs. The individual SNPs are labeled on the plot. The percentage is calculated as the entropy of the haplotype frequencies as defined by the best *n*-SNP subset divided by the entropy of the haplotype frequencies as defined by all 12 SNPs, as *n* goes from 1 to 12. Note that the order in which SNPs are included in an optimum subset differs between the two populations.

gene genealogy is the same in both populations, although frequencies are different and there appear to be population specific recombination events (Figure 2).

Our results have some similarities with a report of haplotype structure across 13 megabases (Mb) in a sample of Yorubans and African-Americans [14]. In a haplotype block defined by 12 SNPs, four haplotypes greater than 5% frequency were found; we find five in the Gambia, and four in Malawi with frequency greater than 5%. However, we find that a minimum of 8 and 10 haplotypes make up 90% of the sample in The Gambia and Malawi, respectively, while they report an average minimum of 5 haplotypes. Furthermore, they report values for D' that are constant across a haplotype block, while we find variability in linkage disequilibrium, even over short distances. Our data are more consistent with previous reports of haplotypic structure where SNPs were ascertained by resequencing in a large sample of chromosomes [15-17].

The increased frequency of haplotype 1 in the Gambian sample has a profound impact on the allelic associations between SNPs as evaluated by chi-square test or as defined by the pairwise association efficiency parameter. The 'star-shaped' genealogy of the Gambian *TNF* locus, dominated by one central haplotype with many derived haplotypes defined by single mutational events, produces negative disequilibria between most pairs of SNPs (the rare alleles are found in the repulsion phase). The chi-square test has low power to detect negative disequilibria [18]; this may explain why in the case where $p > 0.05$, 41/42 SNP pairs have negative disequilibria. The ability of one marker to detect another in a case-control scenario is also greatly reduced when alleles are in the repulsion phase. For example, in the Gambian sample, allele pairs in the repulsion phase are much more likely to be inefficient markers (where inefficient means the apparent relative risk at the marker SNP is betwen 0.5 and 2.0 when the relative risk of the hypothetical disease SNP is 10) than alleles in the coupling phase ($p < 10^{-7}$).

Although linkage disequilibrium between a disease susceptibility locus and a marker locus is required to detect an association with a complex disease, for a fixed sample size, it may not be sufficient. Even in the extreme case when linkage disequilibrium is at its theoretical maximum (D' = 100% and no recombination has occurred between the two SNPs), the relative risk at the marker locus will be some fraction of the relative risk at the disease susceptibility locus, unless the two alleles have the same frequency [19]. Using the relationship defined in Equation 2 (see Materials and methods), we can examine the haplotypic structure of a candidate locus and assess the ability of one marker to detect a disease association when a second marker contributes a given disease risk. We calculate two values of apparent relative risk for each pair of SNPs (assigning each SNP in turn a hypothetical risk of 10) and plot them against D' (Figure 6).

Pairs of SNPs where D' values are reversed between the populations also show a reversal of the relative risk (that is, relative risks greater than 1 become less than 1). For example, if the risk at SNP TNF+851 was 10, then the neighboring SNP, TNF+467, would give a relative risk of 3.7 in Malawi, but in the Gambian population, the effect would be protective with a relative risk of 0.8. This suggests that even with a marker in close physical proximity to a disease-susceptibility SNP, results may differ greatly and even give apparently contradictory results between two populations. This illustrates the importance of identifying population-specific recombinant chromosomes in the study populations. We identify six pairs of SNPs that would reverse relative risks (that is, reverse the sign of D) between the Gambian and Malawian populations.

Although the association efficiency parameter is correlated with D' (r = 0.99 for combined data), it is interesting to consider the situation where D' = 1.00 between the marker and hypothetical disease SNPs. In this situation, the apparent relative risk at a marker SNP may be considerably lower than at the disease SNP (Figure 6). This is especially evident in the Malawian population. These observations suggest that linkage-disequilibrium measures, as summarized by D', may not accurately reflect the ability of a marker to detect a disease association.

It is troubling to observe that even within a relatively small gene like *TNF* (4.3 kb in this study) the SNP markers are poorly correlated: if a true disease-susceptibility allele existed in the TNF gene, many of the other SNPs would be inefficient markers of that disease allele. Our calculations reveal that even if a hypothetical disease allele gave a 10-fold increased risk, out of the 132 possible pairs of SNPs, in 92 of them (70%), the marker would show a relative risk of less than twofold. Only a small minority of SNP pairs are efficient enough to detect a strong disease-modifying SNP in the *TNF* gene region.

Although it appears that individual SNPs in *TNF* are poor markers of each other, it is important to ask which SNPs serve as good markers of the *TNF* haplotypes. Depending on the haplotypic structure of a gene region, one may be able to choose a small subset of SNPs that capture most of the haplotypic diversity, thus reducing the cost of genotyping significantly with only modest reductions in information [20-22]. Two methods have been reported that select the best subset of SNPs [20,23]. In this paper we introduce a third, the entropy maximization method (EMM), which chooses the subset of SNPs that explains the largest proportion of the underlying haplotypic diversity as measured by entropy.

The single SNP with maximum entropy is the SNP with frequency closest to 0.50, in this case TNF-3025. In both populations, this SNP represents about one-third of the haplotypic diversity. With the addition of a second SNP to form a two-SNP haplotype, one-half of the haplotypic diversity is accounted for. In The Gambia, it is most informative to include the TNF-308 SNP as a second marker, whereas in Malawi, the TNF-1031 SNP contributes the most as a second marker. It is not surprising that the first few SNPs selected by the EMM are those that have frequencies closest to 0.50; however, the addition of a common SNP that is well correlated with a SNP already in the subset will contribute little additional information.

For example, in the Gambian sample, the four most common SNPs are selected first by the EMM (TNF-3025 (38%), TNF-308 (19%), TNF+851 (10%), TNF-1031 (11%)). The next most common SNPs are the TNF+1304 (9%) and the TNF-863 (6%); however, the EMM does not choose these SNPs next because the TNF+1304 tends to occur in phase with the TNF+851, and the TNF-863 tends to occur in phase with the TNF-1031, both of which are already in the subset of four SNPs. So the next two choices that maximize entropy are the TNF-857 and TNF+467 SNPs, which, although less common, represent haplotypes previously unaccounted for (5 and 8). These six-SNP haplotypes now represent about 90% of the haplotypic diversity. Two more SNPs must be added (for a total of eight) before 95% of the 12-SNP haplotypic diversity is accounted for.

In Malawi, the general features of marker selection are the same, though the optimal SNPs are different. Three of the four most common SNPs are selected first by the EMM (TNF-3025 (45%), TNF-1031 (25%), TNF+851 (15%)), followed by the TNF-308 (11%) and TNF+1304 (13%) SNPs. Interestingly, population-specific recombinant haplotypes make the TNF+1304 SNP more informative in Malawi; it is required to resolve haplotypes 2 and 4 and haplotypes 21 and 22, which together make up almost 10% of all the haplotypes. In Malawi seven SNPs must be included to represent 95% of the 12-SNP haplotypic diversity.

EMM is especially useful when combined with AEA. The latter can be used to determine how well the recommended haplotype-tagging SNPs detect association at a particular

candidate SNP. The results of AEA can then be used to adjust the power of an association study accordingly.

Analysis of the *TNF* locus in two African populations reveals a haplotypically diverse locus where markers are only weakly associated with each other. Although initial detection of a disease association in *TNF* may be difficult, weak allelic associations within the *TNF* gene region will make it possible to distinguish a disease-modifying SNP from its neighbors. Future studies may determine that this is a general feature of African genomes.

Materials and methods

Subjects

Healthy unrelated adults were recruited in Banjul, The Gambia and in Blantyre, Malawi. All were parents of children who presented to hospital with severe malaria. All subjects gave informed consent and the study was approved by the Gambia Government/Medical Research Council Joint Ethical Committee and the College of Medicine Research Committee of the University of Malawi.

After excluding erroneous pedigrees, the 187 parent-child pairs plus 45 other adults comprised the Gambian sample (a total of 212 adults), and 70 parent-child pairs plus 14 other adults (a total of 84 adults) comprised the Malawian sample. We also studied DNA from two chimpanzees, one gorilla and two orang utans.

SNP identification and genotyping

In a previous study, we described the nucleotide diversity of the *TNF* locus in 36 healthy, unrelated Gambian adults [13]. There we identified 11 SNPs by forward and reverse sequencing of the entire TNF gene from -1,389 bp relative to the start of transcription to +3,004 bp on 72 chromosomes. Here we have genotyped the 11 SNPs identified in *TNF*, plus one SNP in the neighboring gene, *LT*α, in a sample of 212 Gambian and 84 Malawian adults. The amplification refractory mutation system PCR (ARMS-PCR) was used to genotype the polymorphisms identified by sequencing. Methods are as previously described [13]. The accuracy of the ARMS-PCR method was tested on sequenced individuals before extending it to DNA samples of unknown genotype.

Haplotype construction

For most of these adults, genotypes were available from their children, raising the question of how to construct haplotypes most efficiently from a dataset where pedigree information is available for some but not all individuals. We developed a program PHAMILY, which uses information from two-generation pedigrees to construct parental haplotypes at all unambiguous sites. These partial haplotypes of unambiguous sites serve as input for PHASE [24,25], which uses the Stephens-Donnelly method of haplotype construction to assign the remaining phase-unknown sites among the unrelated parents. The two study populations were pooled before haplotype construction; after haplotype construction, individuals were assigned to their original populations on the basis of their sample identifiers.

Population genetic analyses

Estimating gene diversity

An estimate of gene diversity (H), the probability that two randomly chosen haplotypes are different in a sample, was calculated using Equation 8.5 from Nei [26],

$$H = \frac{n}{n\text{-}1}\left(1\text{-}\sum_{i=1}^{k} p_i^2\right)$$

where n is the number of gene copies in the sample, k is the number of different haplotypes, and p_i is the frequency of the ith haplotype. The sampling variance of this estimate was calculated according to Nei and Roychoudhury [27]. The computer software Arlequin was used to perform these calculations [28,29].

Construction of haplotype networks

Network2.0c was used to construct median-joining networks of the haplotypic data [30]. Because of the occurrence of many low-frequency recombinant haplotypes in the dataset, only haplotypes observed twice or more were represented in the network. Output from Network2.0c was critically evaluated and some modifications made to best fit a model of minimum mutation.

Four-gamete test

All pairs of biallelic SNPs were cross-tabulated to identify the number of unique two-locus haplotypes present. An observation of four different haplotypes from a pair of biallelic loci was considered evidence for recombination or recurrent mutation.

Population differentiation and pairwise F_{ST}

Pairwise F_{ST} was calculated as the genetic variance among populations divided by the genetic variance of the total population using haplotypic data [31]. An exact test of population differentiation was applied to the haplotypic data [32]. The computer software Arlequin was used to perform these calculations [28,29].

Linkage disequilibrium calculations

Pairwise linkage disequilibria were calculated using the r^2 statistic. Terms for allele and two-locus haplotype frequencies are given in Table 3. r^2 was calculated using

$$r^2 = \left(\frac{D}{\sqrt{p_1 p_2 q_1 q_2}}\right)^2$$

where $D = f_{11}f_{22} - f_{12}f_{21}$. Normalized linkage disequilibrium parameters D' were calculated using the method of Lewontin

[33] in Hartl and Clark [34]. Statistical significance of the standardized linkage disequilibrium parameter was calculated using the formula

$$\chi^2 = N\left(\frac{D}{\sqrt{p_1 p_2 q_1 q_2}}\right)^2$$

with one degree of freedom [34], where N is the number of chromosomes.

Association efficiency analysis (AEA)

To assess the ability of one marker to detect a given disease association at a second marker, we derived a measure we call association efficiency, which is the relative risk at one SNP by virtue of its association with a second SNP. Consider a disease-modifying SNP A that is in linkage disequilibrium with a neighboring nonfunctional SNP B. We define R_A as the relative risk associated with allele A_2 compared to allele A_1. Under a multiplicative mode of inheritance, the frequencies of the different haplotypes (A_1B_1, A_1B_2, A_2B_1, A_2B_2) found in diseased individuals is expected to be $f_{11}, f_{12}, R_A f_{21}$, and $R_A f_{22}$ (see Table 3). Given the risk associated with A_2 and the frequency of each haplotype, we can derive the frequency of allele B_2 among diseased individuals (q_2').

$$q_2' = \frac{f_{12} + R_A f_{22}}{f_{11} + f_{12} + R_A f_{21} + R_A f_{22}} \quad (1)$$

The frequency of allele B_2 in the general population is defined as q_2. Thus, in a case-control study where diseased individuals are compared to individuals drawn from the general population, we would obtain an odds ratio for allele B_2 of

$$R_B = \frac{(1 - q_2)(R_A f_{22} + f_{12})}{q_2 (R_A f_{21} + f_{11})} \quad (2)$$

This odds ratio provides an approximation of the relative risk that is associated with the non-functional allele B_2 by virtue of its association with the disease allele A_2.

Table 3

Terms for linkage disequilibrium and association efficiency calculations

Allele		A_1	A_2
	Frequency	p_1	p_2
B_1	q_1	A_1B_1	A_2B_1
		f_{11}	f_{21}
B_2	q_2	A_1B_2	A_2B_2
		f_{12}	f_{22}

When locus B is considered to be the disease-modifying locus, with R_B as the relative risk associated with allele B_2 compared to B_1, it can be shown that the relative risk at locus A is

$$R_A = \frac{(1 - p_2)(R_B f_{22} + f_{21})}{p_2 (R_B f_{12} + f_{11})} \quad (3)$$

In this paper, we consider the possibility that the rarer allele gives a true relative risk of 10 for a given disease, and calculate the apparent relative risk at all other loci. We do this for all pairs of loci. Because this property does not commute (that is, the apparent relative risk at locus B given a true relative risk of 10 at locus A is not necessarily equal to the apparent relative risk at locus A given a true relative risk of 10 at locus B) we calculate two values for each pair of SNPs.

Entropy maximization method (EMM)

A measure of haplotypic diversity, entropy (E), was calculated using

$$E = -\sum_{i=1}^{k} p_i \log p_i \quad (4)$$

where p_i is the frequency of the ith haplotype, and k is the number of unique haplotypes in the sample. Entropy is maximized when all haplotypes have the same frequency. We developed an algorithim that chooses SNPs, first individually, then in multi-SNP haplotypes, that maximize entropy [35]. Thus we can define a subset of SNPs that represent the greatest proportion of the full 12-SNP haplotypic diversity. This method provides for the selection of optimal haplotype-tagging SNPs even in the absence of a clear haplotypic block structure.

Additional data files

The algorithm that chooses SNPs, first individually, then in multi-SNP haplotypes that maximize entropy, is available with the online version of this article and from [35].

Acknowledgements

We thank the families and our colleagues at the Royal Victoria Hospital, Banjul and the Queen Elizabeth Central Hospital, Blantyre, who made this study possible. The authors would also like to thank Matthew Stephens for providing software for haplotype reconstruction. This work was funded by the Medical Research Council, UK. This work was funded by the Medical Research Council, UK. This article is dedicated to the memory of Ryk Ward, who passed away on February 14, 2003.

References

1. McGuire W, Hill AV, Allsopp CE, Greenwood BM, Kwiatkowski D: **Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria.** *Nature* 1994, **371**:508-510.
2. McGuire W, Knight JC, Hill AV, Allsopp CE, Greenwood BM, Kwiatkowski D: **Severe malarial anemia and cerebral malaria**

are associated with different tumor necrosis factor promoter alleles. *J Infect Dis* 1999, **179:**287-290.
3. Knight JC, Udalova I, Hill AV, Greenwood BM, Peshu N, Marsh K, Kwiatkowski D: **A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria.** *Nat Genet* 1999, **22:**145-150.
4. Knight JC, Kwiatkowski D: **Inherited variability of tumor necrosis factor production and susceptibility to infectious disease.** *Proc Assoc Am Physicians* 1999, **111:**290-298.
5. Dunstan SJ, Stephens HA, Blackwell JM, Duc CM, Lanh MN, Dudbridge F, Phuong CX, Luxemburger C, Wain J, Ho VA, *et al.*: **Genes of the class II and class III major histocompatibility complex are associated with typhoid fever in Vietnam.** *J Infect Dis* 2001, **183:**261-268.
6. Cabrera M, Shaw MA, Sharples C, Williams H, Castes M, Convit J, Blackwell JM: **Polymorphism in tumor necrosis factor genes associated with mucocutaneous leishmaniasis.** *J Exp Med* 1995, **182:**1259-1264.
7. Nadel S, Newport MJ, Booy R, Levin M: **Variation in the tumor necrosis factor-alpha gene promoter region may be associated with death from meningococcal disease.** *J Infect Dis* 1996, **174:**878-880.
8. Conway DJ, Holland MJ, Bailey RL, Campbell AE, Mahdi OS, Jennings R, Mbena E, Mabey DC: **Scarring trachoma is associated with polymorphism in the tumor necrosis factor alpha (TNF-alpha) gene promoter and with elevated TNF-alpha levels in tear fluid.** *Infect Immun* 1997, **65:**1003-1006.
9. Moffatt MF, Cookson WO: **Tumour necrosis factor haplotypes and asthma.** *Hum Mol Genet* 1997, **6:**551-554.
10. Fernandez-Arquero M, Arroyo R, Rubio A, Martin C, Vigil P, Conejero L, Figueredo MA, de la Concha EG: **Primary association of a TNF gene polymorphism with susceptibility to multiple sclerosis.** *Neurology* 1999, **53:**1361-1363.
11. Negoro K, Kinouchi Y, Hiwatashi N, Takahashi S, Takagi S, Satoh J, Shimosegawa T, Toyota T: **Crohn's disease is associated with novel polymorphisms in the 5′-flanking region of the tumor necrosis factor gene.** *Gastroenterology* 1999, **117:**1062-1068.
12. Udalova IA, Richardson A, Denys A, Smith C, Ackerman H, Foxwell B, Kwiatkowski D: **Functional consequences of a polymorphism affecting NF-kappaB p50-p50 binding to the TNF promoter region.** *Mol Cell Biol* 2000, **20:**9113-9119.
13. Richardson A, Sisay-Joof F, Ackerman H, Usen S, Katundu P, Taylor T, Molyneux M, Pinder M, Kwiatkowski D: **Nucleotide diversity of the TNF gene region in an African village.** *Genes Immun* 2001, **2:**343-348.
14. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al.*: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.
15. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, *et al.*: **Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, **63:**595-612.
16. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, *et al.*: **Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism.** *Am J Hum Genet* 2000, **67:**881-900.
17. Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF: **Recombinational and mutational hotspots within the human lipoprotein lipase gene.** *Am J Hum Genet* 2000, **66:**69-83.
18. Thompson EA, Deeb S, Walker D, Motulsky AG: **The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes.** *Am J Hum Genet* 1988, **42:**113-124.
19. Muller-Myhsok B, Abel L: **Genetic analysis of complex diseases.** *Science* 1997, **275:**1328-1330.
20. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, *et al.*: **Haplotype tagging for the identification of common disease genes.** *Nat Genet* 2001, **29:**233-237.
21. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: **High-resolution haplotype structure in the human genome.** *Nat Genet* 2001, **29:**229-232.
22. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, *et al.*: **Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease.** *Nat Genet* 2001, **29:**223-228.
23. Zhang K, Deng M, Chen T, Waterman MS, Sun F: **A dynamic programming algorithm for haplotype block partitioning.** *Proc Natl Acad Sci USA* 2002, **99:**7335-7339.
24. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68:**978-989.
25. **Mathematical Genetics Group: Software** [http://www.stats.ox.ac.uk/mathgen/software.html]
26. Nei M: *Molecular Evolutionary Genetics*. New York: Columbia University Press; 1987.
27. Nei M, Roychoudhury AK: **Sampling variances of heterozygosity and genetic distance.** *Genetics* 1974, **76:**379-390.
28. Schneider S, Roessli D, Excoffier L: *Arlequin ver. 2.000: a software for population genetics analysis.* Geneva: Genetics and Biometry Laboratory, University of Geneva; 2000.
29. **Arlequin** [http://lgb.unige.ch/arlequin/]
30. **Free Phylogenetic Network Analysis Shareware Software** [http://www.fluxus-engineering.com/sharenet.htm]
31. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16:**97-159.
32. Rousset F, Raymond M: **Testing heterozygote excess and deficiency.** *Genetics* 1995, **140:**1413-1419.
33. Lewontin R: **The interaction of selection and linkage. I. General considerations; heterotic models.** *Genetics* 1964, **49:**49-67.
34. Hartl DL, Clark AG: *Principles of Population Genetics* 3rd edn. Sunderland, MA: Sinauer; 1997.
35. **SNP selection page** [http://www.well.ox.ac.uk/~rmott/SNPS/]