nature
genetics

Genome-wide and fine-resolution association analysis of malaria in West Africa

Muminatou Jallow[1,34], Yik Ying Teo[2,3,34], Kerrin S Small[2,3,34], Kirk A Rockett[2,3], Panos Deloukas[3], Taane G Clark[2,3], Katja Kivinen[3], Kalifa A Bojang[1], David J Conway[1], Margaret Pinder[1], Giorgio Sirugo[1], Fatou Sisay-Joof[1], Stanley Usen[1], Sarah Auburn[2,3], Suzannah J Bumpstead[3], Susana Campino[2,3], Alison Coffey[3], Andrew Dunham[3], Andrew E Fry[2], Angela Green[2], Rhian Gwilliam[3], Sarah E Hunt[3], Michael Inouye[3], Anna E Jeffreys[2], Alieu Mendy[2], Aarno Palotie[3], Simon Potter[3], Jiannis Ragoussis[2], Jane Rogers[3], Kate Rowlands[2], Elilan Somaskantharajah[3], Pamela Whittaker[3], Claire Widden[3], Peter Donnelly[2,4], Bryan Howie[4], Jonathan Marchini[2,4], Andrew Morris[2], Miguel SanJoaquin[2,5], Eric Akum Achidi[6], Tsiri Agbenyega[7], Angela Allen[8,9], Olukemi Amodu[10], Patrick Corran[11], Abdoulaye Djimde[12], Amagana Dolo[12], Ogobara K Doumbo[12], Chris Drakeley[13,14], Sarah Dunstan[15], Jennifer Evans[7,16], Jeremy Farrar[15], Deepika Fernando[17], Tran Tinh Hien[15], Rolf D Horstmann[16], Muntaser Ibrahim[18], Nadira Karunaweera[17], Gilbert Kokwaro[19], Kwadwo A Koram[20], Martha Lemnge[21], Julie Makani[22], Kevin Marsh[19], Pascal Michon[8], David Modiano[23], Malcolm E Molyneux[5], Ivo Mueller[8], Michael Parker[24], Norbert Peshu[19], Christopher V Plowe[25,26], Odile Puijalon[27], John Reeder[8], Hugh Reyburn[13,14], Eleanor M Riley[13,14], Anavaj Sakuntabhai[27], Pratap Singhasivanon[28], Sodiomon Sirima[29], Adama Tall[30], Terrie E Taylor[25,31], Mahamadou Thera[12], Marita Troye-Blomberg[32], Thomas N Williams[19], Michael Wilson[20] & Dominic P Kwiatkowski[2,3], Wellcome Trust Case Control Consortium[33] & Malaria Genomic Epidemiology Network[33]

We report a genome-wide association (GWA) study of severe malaria in The Gambia. The initial GWA scan included 2,500 children genotyped on the Affymetrix 500K GeneChip, and a replication study included 3,400 children. We used this to examine the performance of GWA methods in Africa. We found considerable population stratification, and also that signals of association at known malaria resistance loci were greatly attenuated owing to weak linkage disequilibrium (LD). To investigate possible solutions to the problem of low LD, we focused on the HbS locus, sequencing this region of the genome in 62 Gambian individuals and then using these data to conduct multipoint imputation in the GWA samples. This increased the signal of association, from $P = 4 \times 10^{-7}$ to $P = 4 \times 10^{-14}$, with the peak of the signal located precisely at the HbS causal variant. Our findings provide proof of principle that fine-resolution multipoint imputation, based on population-specific sequencing data, can substantially boost authentic GWA signals and enable fine mapping of causal variants in African populations.

[1]MRC Laboratories, Fajara, Banjul, Gambia. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [4]Department of Statistics, Oxford University, Oxford, UK. [5]Malawi–Liverpool–Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Chichiri, Blantyre, Malawi. [6]The University of Buea, Buea, South West Province, Cameroon. [7]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [8]Papua New Guinea Institute of Medical Research, Madang, Papua New Guinea. [9]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [10]Institute of Child Health, College of Medicine, University of Ibadan, Ibadan, Nigeria. [11]National Institute for Biological Standards and Control, Hertfordshire, UK. [12]The Malaria Research & Training Centre, University of Bamako, Bamako, Mali. [13]London School of Hygiene & Tropical Medicine, London, UK. [14]Joint Malaria Programme, Kilimanjaro Christian Medical Centre, Moshi, Tanzania. [15]Oxford University Clinical Research Unit, The Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam. [16]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany. [17]Faculty of Medicine, University of Colombo, Colombo, Sri Lanka. [18]Institute for Endemic Diseases, University of Khartoum, Medical Service Science Campus, Khartoum, Sudan. [19]Kenyan Medical Research Institute (KEMRI)–Wellcome Trust Programme, Kilifi, Kenya. [20]Noguchi Memorial Institute for Medical Research, University of Ghana, Accra, Ghana. [21]National Institute for Medical Research, Dar es Salaam, Tanzania. [22]Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania. [23]University of Rome 'La Sapienza', Rome, Italy. [24]The Ethox Centre, Department of Public Health and Primary Health Care, University of Oxford, Headington, Oxford, UK. [25]Blantyre Malaria Project, Chichiri, Blantyre 3, Malawi. [26]Howard Hughes Medical Institute/University of Maryland School of Medicine, Baltimore, Maryland, USA. [27]Institut Pasteur, Unité d'Immunologie Moléculaire des Parasites, Paris, France. [28]Faculty of Tropical Medicine, Mahidol University, Ratchathewi, Bangkok, Thailand. [29]Centre National de Recherche et Formation sur le Paludisme, Ouagadougou, Burkina Faso. [30]Institut Pasteur de Dakar, Dakar, Senegal. [31]Michigan State University, Department of Internal Medicine, College of Osteopathic Medicine, East Lansing, Michigan, USA. [32]The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden. [33]A full list of members is provided in the **Supplementary Note** online. [34]These authors contributed equally to this work. Correspondence should be addressed to D.P.K. (dominic@sanger.ac.uk).
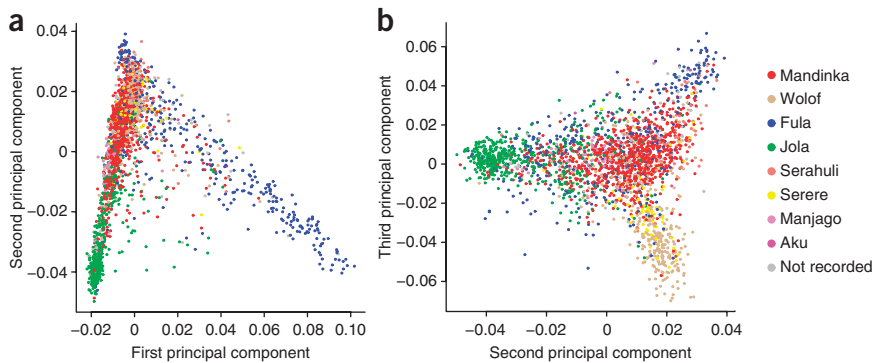
Figure 1 Principal components analysis of population structure within The Gambia. Plots of the first three principal components from EIGENSTRAT using 100,715 SNPs selected to minimize intermarker LD. Each solid circle represents an individual, and the color is assigned according to self-reported ethnicity. (**a**) Plot of the first two principal components for all Gambian samples. (**b**) Plot of the second and third principal component for all Gambian samples.

The malaria parasite *Plasmodium falciparum* kills on the order of a million African children each year[1], and this is a small fraction of the number of infected individuals in the population[1–3]. In communities where everyone is repeatedly infected with *P. falciparum*, host genetic factors account for $\sim$25% of the risk of severe malaria, that is, life-threatening forms of the disease[3]. The strongest known determinant of risk, hemoglobin S (HbS), accounts for $\sim$2% of the total variation, implying that only a small fraction of genetic resistance factors have so far been discovered[3]. Identifying the genetic basis of protective immunity against severe malaria may provide important insights for vaccine development.

Here we examine the possibility of approaching this problem by genome-wide association (GWA) analysis. There are many unsolved methodological questions about how to conduct an effective GWA study in Africa[4]. High levels of ethnic diversity may result in false-positive associations owing to population structure. Variations in haplotype structure between different ethnic groups may reduce power to detect GWA signals, particularly when data are amalgamated across multiple study sites. Low LD implies the need for denser genotyping arrays than are currently available: a crude estimate is that an African GWA study with 1.5 million SNPs would have approximately the same statistical power as a European study with 0.6 million SNPs[5], but this is based on HapMap data from a single ethnic group and a larger number of SNPs may be needed to achieve adequate power across different ethnic groups.

We carried out an initial GWA study in Gambian children that explores these methodological questions. Genotyping of $\sim$500,000 SNPs was conducted on 1,060 cases of severe malaria and 1,500 population controls using the Affymetrix GeneChip 500K Mapping Array Set. The results reported here are based on a set of 402,814 SNPs in 958 cases and 1,382 controls that passed stringent quality control procedures. Access to these data may be requested online (see URLs section in Methods).

RESULTS

Examining population structure

Subjects were recruited from an area of approximately 400 square miles in the Kombos region of The Gambia. Four ethnic groups, Mandinka, Jola, Wolof and Fula, accounted for 89% cases and 86% controls (**Supplementary Table 1** online). Using Wright's F_{ST} across all autosomal SNPs, we found that differences between ethnic groups accounted for a small fraction of genetic variation within the population as a whole ($F_{ST} = 0.004$). The greatest differentiation was seen between Fula and Jola ($F_{ST} = 0.007$) and the least between Mandinka and Wolof ($F_{ST} = 0.002$) (**Supplementary Table 2** online).

To investigate the relationship between population structure and self-reported ethnicity, we carried out principal components analysis (PCA) of 100,715 SNPs, selected to reduce LD between markers (**Fig. 1**)[6]. The first two principal components distinguished Fula and Jola, and the third principal component separated the Mandinka and Wolof from others. Some individuals could be confidently assigned to a specific ethnic group, whereas others seemed to have a more complex ancestry. These findings were verified using the STRUC-TURE[7] program on 8,000 SNPs, which gave an optimal model of population structure with four genetic subpopulations corresponding to the four most common ethnic groups (**Supplementary Fig. 1** online).

To place these findings in the context of global population structure, we compared the Gambian sample with populations studied by the HapMap project[5,8]. The Gambian sample can be clearly distinguished by PCA from the Yoruba people of Ibadan, Nigeria (a different part of West Africa) but is much closer to Yoruba than to European, Han Chinese or Japanese samples (**Fig. 2a**). Individual ethnic groups within The Gambia seem to have greater genetic diversity than the HapMap Yoruba sample (compare **Fig. 1a** with **Fig. 2b**, **Supplementary Fig. 2** online). This may reflect the fact that Gambian samples were recruited from the general population, whereas the HapMap Yoruba samples were collected in a particular community from individuals with four Yoruba grandparents.

Genome-wide association of severe malaria

To evaluate the likelihood of false-positive GWA findings due to population structure, we conducted a trend test of association in
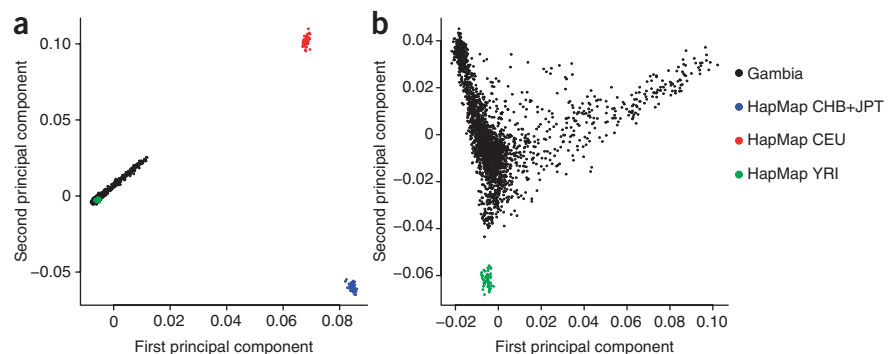

Figure 2 Principal components analysis of population structure for the Gambian study sample in relation to HapMap reference panels. Plots of the first two principal components from EIGENSTRAT using 100,715 SNPs selected to minimize intermarker LD. Each solid circle represents an individual. (**a**) Plot of the first two principal components for HapMap and Gambian samples. (**b**) Plot of the first two principal components for HapMap YRI and Gambian samples.
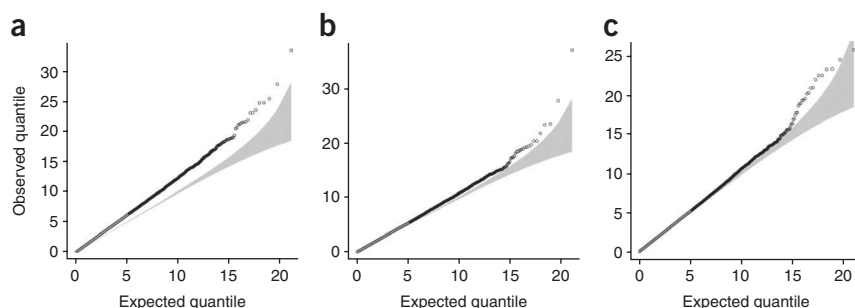
Figure 3 Quantile-quantile plots of association test statistic. (**a**–**c**) Quantile-quantile plots of the trend test statistic for the unstratified analysis, which uses all 958 cases and 1,382 controls (**a**); the ethnic-stratified analysis, which tests 854 cases, 1,195 controls from the four major ethnic groups (**b**); and the PCA analysis, which corrects for the first three principal components from EIGENSTRAT and uses all 958 cases and 1,382 controls (**c**). The shaded region in gray represents the lower and upper 95% probability bounds for the expected quantiles.

cases versus controls on all SNPs, and compared observed χ^2 values to expected values under the null hypothesis in a quantile-quantile plot (**Fig. 3**). The overdispersion of association test statistics ($\lambda = 1.23$) implies a high number of false-positive associations in the raw data, but this was greatly reduced by correction for self-reported ethnicity ($\lambda = 1.07$), and became negligible when the first three principal components from the eigenanalysis of population structure were entered as covariates in logistic regression analysis ($\lambda = 1.02$). For comparison, λ has been estimated to be 1.03–1.11 in case-control studies in the British population, a range considered acceptable for GWA analysis[9]. Thus, with appropriate statistical correction, false-positive GWA findings arising from the Gambian population structure can be reduced to a very low level.

After PCA correction for population structure, we tested each SNP for disease association using an unguided genotypic test with 2 degrees of freedom (d.f.), as well as tests with 1 d.f. for models of dominance, recessiveness, heterozygous advantage and trend. Cluster plots were visually inspected on all potentially significant results, which yielded 139 SNPs with unequivocal genotype results in 100 independent regions of the genome with $P < 10^{-4}$ (**Supplementary Table 3** online), including 6 with $P < 10^{-6}$ (**Fig. 4**). The strongest signal of association was close to the *HBB* gene on chromosome 11p15, where the HbS polymorphism is located, with 13 SNPs at $P < 10^{-4}$ and a minimum of $P = 3.9 \times 10^{-7}$ by trend test. In the following sections, we examine the signal of association around the HbS polymorphism, evaluate other known and putative malaria resistance–associated genes and describe newly identified signals of association.

Fine-resolution association mapping at the HbS locus

HbS provides a benchmark for evaluating GWA methods, as the causal polymorphism responsible for the malaria-protective effect is known: it is a SNP (rs334) in the coding region of *HBB* on chromosome 11p15.4 which results in replacement of glutamic acid with valine at amino-acid residue 6 of the β-globin chain. When we genotyped rs334 on the same samples used in the GWA study, using the Sequenom iPlex platform, we found a much stronger signal of association ($P = 1.3 \times 10^{-28}$). This raises several questions: why was the GWA signal ($P = 3.9 \times 10^{-7}$) much weaker than the true effect; is there an effective way to increase the GWA signal; and is there an effective way to get from the GWA signal to identification of the causal variant?

To investigate these questions, we sequenced 111 kb in the center of the GWA signal on chromosome 11p15 in a reference panel of 62 randomly selected Gambian controls (see Methods). These reference data were used to impute genotypes for all ∼2,500 individuals in the GWA study with the IMPUTE program[10], and a trend test of association was conducted at each imputed SNP. Out of 202 SNPs examined across this 111 kb region, three imputed SNPs had stronger signals of association than any of the SNPs genotyped on the initial GWA scan (**Fig. 5**). The HbS causal polymorphism (rs334) stands out

as the imputed SNP with the strongest association ($P = 4.5 \times 10^{-14}$), several orders of magnitude more significant than the strongest signal from SNPs that were directly genotyped ($P = 3.9 \times 10^{-7}$).

This result provides proof of principle that it is possible to identify the causal polymorphism within a GWA signal by regional sequencing followed by multipoint association mapping using model-based imputation, provided the appropriate reference panel is used. We observed two features of LD in this region of the genome in this population, which may together be favorable for fine mapping. First, over a 1-Mb region we identified 55 SNPs with $D' = 1$ in relation to the HbS causal polymorphism rs334 (**Fig. 6**): this is consistent with previous evidence that the HbS allele is associated with an extended haplotype a result of recent positive selection[5,11] Second, the region as a whole has weak LD with a well-known recombination hot spot[12], and the correlation between rs334 and neighboring SNPs does not exceed an r^2 of 0.36 (**Fig. 6**). In other words, there are no neighboring SNPs that are sufficiently strongly correlated with rs334 to imitate the true signal of association generated from the causal variant. We are still at an early stage of understanding how the process of fine mapping is affected by different patterns of natural variation in the human genome, and this example of extended haplotype within a region of generally low LD provides an interesting case study.

In general, the performance of imputation strategies depends on the overall information content the genotyped SNPs carry for the untyped SNPs in the region, which was estimated at only 40% for rs334 for our data. This may explain, in part, why the imputed association signal ($P = 4.5 \times 10^{-14}$) was weaker than the value obtained when we genotyped rs334 directly on the same samples ($P = 1.3 \times 10^{-28}$).
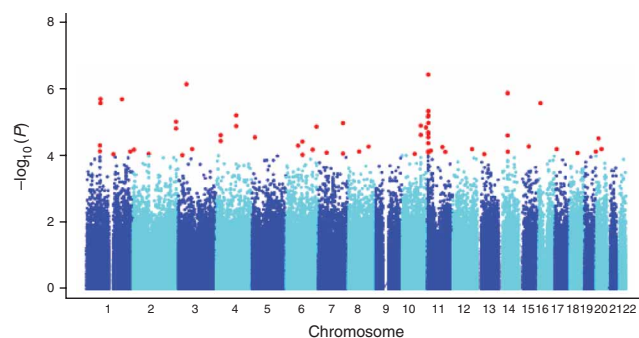

Figure 4 Genome-wide signals of association with severe malaria. Plot of the $-\log_{10} P$ values for the trend test correcting for the first three principal components from EIGENSTRAT. Each point represents a SNP from the 402,814 remaining after quality control filters were applied. Different bands of blue are used to differentiate SNPs on consecutive autosomal chromosomes. SNPs with P values less than 10^{-4} are represented by red points.

The HapMap Yoruba sample has been used as the basis for designing GWA genotyping arrays intended for African populations in general, and these data provide an example of where this approach may fail. When viewed at a macroscopic level, patterns of LD in The Gambia and the HapMap Yoruba sample are similar, both for the genome as a whole (**Supplementary Fig. 3** online) and for the genomic region around the HbS locus (**Supplementary Fig. 4** online). However, when we attempted to impute rs334 genotypes in our Gambian data using the HapMap Yoruba as the reference panel, we failed to identify any association signal ($P = 0.06$). This may be explained by the fact that the SNP on the Affymetrix array with the strongest LD with rs334 in The Gambia (rs11036238, $r^2 = 0.32$) has negligible LD in the HapMap Yoruba samples ($r^2 = 0.009$); conversely, the SNP in strongest LD with rs334 in the HapMap Yoruba samples (rs7936221, $r^2 = 0.35$) is not in LD with rs334 in The Gambia ($r^2 = 0.005$). This is consistent with evidence that the HbS allele has arisen independently in different African populations[11,13,14]. Although the HbS allele may not be representative of genomic variation as a whole, it highlights the possibility of local anomalies particularly in regions under strong selective pressure, and thus raises important questions about the design of an optimal SNP tagging strategy for African populations in general.

Taken together, these findings support the view that low LD in African populations can help to distinguish the causal polymorphism from neighboring polymorphisms. But they also highlight the importance of understanding regional variations in haplotype structure when designing and interpreting GWA studies in African populations, particularly for loci that are under selective pressure.

Signals at other known loci

The GWA analysis did not identify any of the well-known erythrocyte variants that have been selected by malaria, other than HbS. This can partly be explained by population genetic factors; for example, the Duffy FY*O allele has reached fixation in The Gambia, whereas other variants such those affecting hemoglobin C and southeast Asian ovalocytosis are rare or absent in this population. We might have expected associations at *G6PD* and *HBA1-HBA2*, the loci for glucose-6-phosphate deficiency[15–17] and α+thalassaemia[18–20], respectively, but our GWA dataset had no SNP within 100 kb of *G6PD* and only one SNP within 50 kb of *HBA1-HBA2*.

To investigate *G6PD* in more detail, we used the Sequenom iPlex platform to genotype rs1050828, a coding polymorphism that has received considerable attention as a marker of protection against severe malaria[15,17], although there are other polymorphisms associated with reduced G6PD enzyme activity that have been less well studied in malaria and could possibly also be involved[21]. The minor allele frequency of rs1050828 in the Gambian control sample was 0.03, considerably lower than samples from Kenya (0.18) and Malawi (0.19) that we genotyped by the same method. Power to detect association with rs1050828 in The Gambia is affected by this low allele frequency, and the results were consistent with a modest protective effect but were not statistically significant: odds ratio (OR) for male hemizygotes 0.71 (95% CI = 0.34–1.49) and for female heterozygotes 0.79 (95% CI = 0.43–1.46). Even if it had been a strong effect, it would not have given a GWA signal because the best tagging SNP for rs1050828 on the Affymetrix 500K array had $r^2 = 0.06$.

We also examined the *ABO* locus, where the functional variant is known and an effect has been conclusively replicated across different populations[22]. A previous study combining case-control and family-based analyses of ~9,000 individuals in three populations found that individuals who are not of blood group O (as defined by the functional variant rs8176719, a splice-site insertion in the *ABO* gene) have ~1.2-fold increased risk of severe malaria with a combined P value of 2×10^{-7} (ref. 22). We genotyped rs8176719 in our GWA sample and found an association that was entirely consistent with previous data (OR = 1.26, 95% CI = 1.11–1.44, $P = 5 \times 10^{-4}$) but which would not have passed our initial GWA significance threshold of
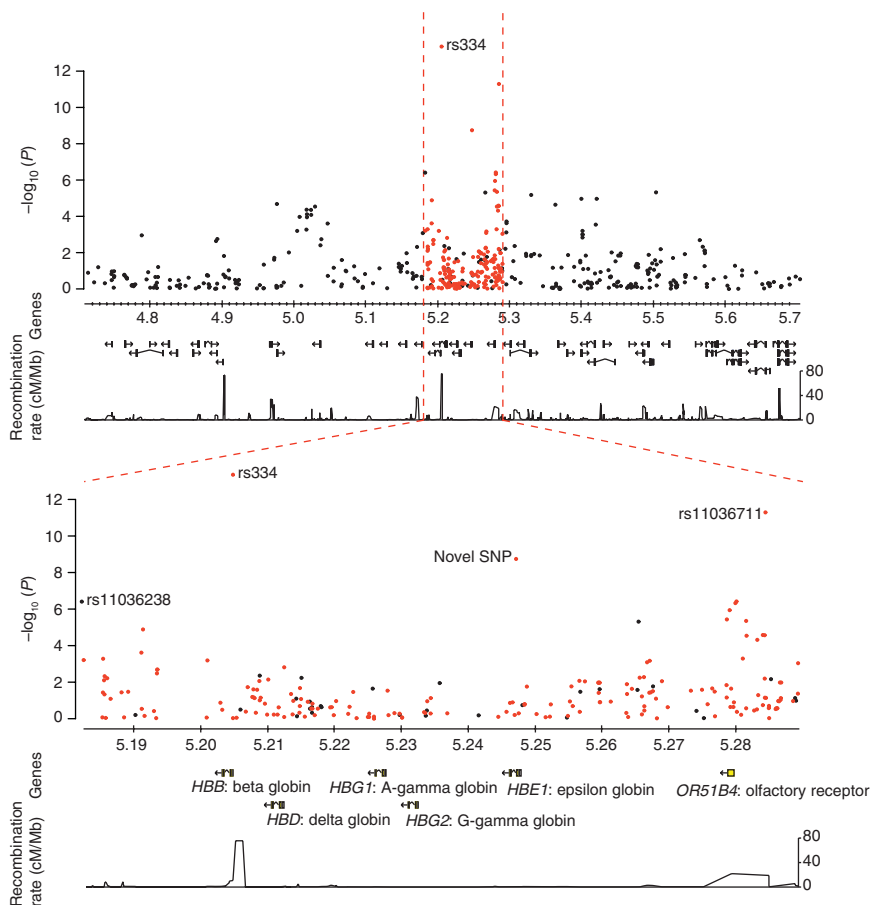

Figure 5 Association signal at the *HBB* locus. The top panel shows the association signals across a 1-Mb region on chromosome 11 centering on rs334, with the vertical axis representing the $-\log_{10}$ *P* values from the Armitage trend test. Points in black represent SNPs that are found on the Affymetrix array, and points in red represent SNPs imputed with the resequenced Gambian reference panel. The dashed lines in red indicate the start and end of the sequenced region. The bottom panels focus on the 110-kb sequenced region, together with a map of the recombination rates and genes found in the region. Recombination rates and genes were extracted from the HapMap Genome Browser.
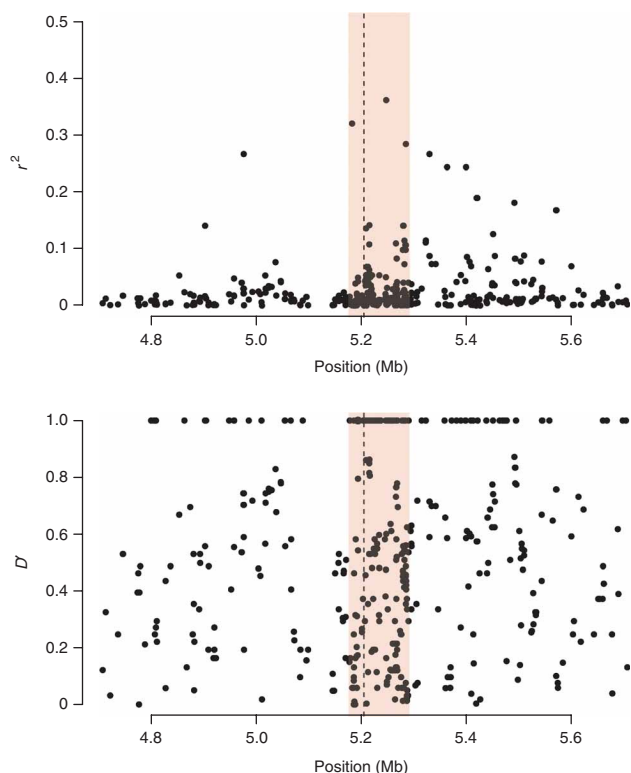
Figure 6 Extent of LD surrounding HbS. Each point show r^2 (top panel) and D' (bottom panel) between the HbS SNP (rs334) and SNPs in the Gambian reference panel. The shaded pink region indicates the boundaries of the resequenced region. The dashed vertical line indicates the position of rs334.

$P < 10^{-4}$. The lack of a GWA signal can be explained by the fact that the best tagging SNP had $r^2 = 0.15$.

Other SNP associations have been reported for malaria but have not been conclusively replicated in large studies across different populations, and are mostly thought to be markers rather than true causal variants. At seven loci (*CD36, CD40LG, CR1, ICAM1, IL22, NOS2, TNF*) we genotyped nine candidate SNPs previously reported to show association with malaria (**Supplementary Table 4** online). A weak association was identified at *TNF* for rs2516486 ($P = 0.02$) but this did not result in a GWA signal, as the best tagging SNP had $r^2 = 0.51$. Other SNPs tested showed no significant association, but had they done so it might have been missed by GWA analysis, as all candidate SNPs were poorly tagged by the Affymetrix 500K array (median $r^2 = 0.45$; range 0.01–0.61) (**Supplementary Table 4**).

In summary, the lack of GWA signals corresponding to previously reported malaria associations can at least in part be explained by low tagging efficiency of the Affymetrix 500K array in this population and other causes of low statistical power, particularly low allele frequencies. However these data also raise the question of how many previously reported associations may have been false positives. In some cases an authentic association may fail to replicate because the effect size was overestimated in initial reports ('winner's curse'); because the frequency of the causal variant varies between populations; because LD between the marker SNP and the causal variant varies between populations; or because the effect is complex, for example, due to allelic heterogeneity or epistasis. These issues are currently being addressed by the MalariaGEN consortium in a multicenter study across 11 different malaria-endemic populations[4].

Loci identified by genome-wide association analysis

From the above analyses it is clear that in the Gambian population the Affymetrix 500K array may fail to detect authentic resistance loci with weak effects, and that even strong genetic determinants may give relatively weak GWA signals. In the following analysis we focus primarily on GWA signals with P-values $<10^{-4}$, although it will be important to follow up weaker GWA signals in future work.

Table 1 Regions of the genome showing association

Chr.	Region (Mb)	No. of SNPs	SNP	Minor allele	Model	Case MAF	Control MAF	P value	Odds ratio	Nearby gene
1p34.1	46.0–46.08	4	rs10890361	A	Het	0.37	0.36	9.4×10^{-6}	1.46 (1.24–1.73)	*MAST2*
1p31.1	72.9–72.96	2	rs10889990	A	Trend	0.39	0.34	5.1×10^{-5}	1.29 (1.14–1.46)	–
1p31.1	76.1–76.15	2	rs12405994	T	Dom	0.01	0.03	8.2×10^{-7}	0.33 (0.20–0.53)	*ASB17*
2q37.1	231.7–231.71	5	rs10192428	G	Dom	0.33	0.38	5.1×10^{-7}	0.65 (0.55–0.77)	*SPATA3*
3p22.1	43.10–43.11	1	rs488069	C	Trend	0.27	0.21	7.6×10^{-7}	1.42 (1.24–1.64)	*C3orf39*
4p15.2	26.1–26.19	2	rs2046784	G	Trend	0.27	0.33	2.5×10^{-5}	0.76 (0.66–0.86)	*CCKAR*
4q24	107.6–107.67	2	rs2949632	G	Trend	0.27	0.33	6.5×10^{-6}	0.74 (0.65–0.85)	*SCYE1*
5p12	43.0–43.13	3	rs316414	A	Het	0.16	0.18	4.5×10^{-7}	0.66 (0.49–0.88)	*ZNF131*
7p12.2	50.39–50.40	3	rs10249420	A	Het	0.18	0.21	6.8×10^{-5}	0.69 (0.57–0.83)	*DDC*
7q32.3	131.5–131.53	2	rs10269601	A	Trend	0.28	0.34	1.1×10^{-5}	0.75 (0.66–0.85)	*PLXNA4*
8p22	13.6–13.61	2	rs1384057	A	Rec	0.16	0.20	2.4×10^{-5}	0.32 (0.18–0.57)	*DLC1*
10p12.2	23.1–23.14	2	rs11013140	T	Rec	0.31	0.36	5.6×10^{-6}	0.53 (0.40–0.70)	*PIP4K2A*
10q24.2	101.0–101.1	2	rs11190062	A	Trend	0.12	0.09	1.3×10^{-5}	0.65 (0.54–0.79)	*CNNM1*
10q26.13	126.2–126.4	2	rs7076268	C	Het	0.22	0.25	1.6×10^{-5}	0.68 (0.56–0.81)	*FAM53B*
11p15.4	4.6–5.6	13	rs11036238	C	Trend	0.09	0.14	3.9×10^{-7}	0.61 (0.50–0.74)	*HBB*
13q31.3	92.49–92.52	2	rs1444227	C	Het	0.15	0.11	3.6×10^{-5}	1.54 (1.25–1.89)	*GPC6*
14q21.2	45.43–45.54	3	rs17728971	G	Rec	0.16	0.12	6.5×10^{-7}	1.63 (1.34–1.98)	–
16q13	56.0–56.1	2	rs16957051	C	Dom	0.15	0.20	3.4×10^{-5}	0.68 (0.56–0.81)	*CIAPIN1*
19p13.3	6.8–6.9	2	rs460375	G	Dom	0.45	0.48	1.2×10^{-5}	0.66 (0.55–0.79)	*EMR1*

Shown are regions meeting either of two criteria: (i) at least two SNPs with P value $< 10^{-4}$ for the PC-corrected analysis within 250 kb of each other or (ii) a SNP with P value $< 10^{-6}$. For each region, we report the SNP with the strongest signal and the model which gave the signal. The minor allele is defined with respect to the controls and each odds ratio is defined for the minor alleles. The most significant model is reported, either trend, dominant (dom), recessive (rec) or heterozygous advantage (het). Nearby gene is defined as the closest gene within 200 kb of the region.

661

Table 2 Combined results from the genome-wide association and replication experiments

Chr.	SNP	Minor allele	Model	GWA experiment				Replication experiment				Combined			Nearby gene
				Case MAF	Control MAF	P value	Odds ratio	Case MAF	Control MAF	P value	Odds ratio	P value	Odds ratio	95% CI	
1p34.1	rs6429582	T	Het	0.37	0.36	6.0×10^{-5}	1.51	0.36	0.38	8.0×10^{-1}	0.98	2.5×10^{-4}	1.24	1.11–1.40	MAST2
1p31.1	rs10889990	A	Trend	0.39	0.34	5.1×10^{-5}	1.29	0.38	0.36	2.4×10^{-1}	1.07	2.0×10^{-4}	1.20	1.09–1.31	–
1p31.1	rs12405994	T	Dom	0.01	0.03	8.2×10^{-7}	3.03	0.02	0.03	6.1×10^{-1}	1.10	1.1×10^{-3}	1.67	1.23–2.27	ASB17
1p22.3	rs1523993	T	Trend	0.33	0.29	7.8×10^{-4}	1.72	0.34	0.30	1.3×10^{-2}	1.18	2.4×10^{-4}	1.26	1.12–1.41	DDAH
2q37.1	rs6750230	G	Dom	0.37	0.43	1.6×10^{-5}	0.65	0.38	0.40	1.0×10^{-1}	0.83	3.2×10^{-5}	0.74	0.64–0.85	SPATA3
3p22.1	rs488069	C	Trend	0.27	0.21	7.6×10^{-7}	1.42	0.24	0.23	8.4×10^{-1}	1.01	1.3×10^{-4}	1.22	1.10–1.34	C3orf39
5p12	rs316414	A	Het	0.16	0.18	4.5×10^{-7}	0.66	0.17	0.17	6.4×10^{-1}	1.04	1.3×10^{-3}	0.79	0.69–0.91	ZNF131
7p12.2	rs1451375	A	Dom	0.18	0.21	3.0×10^{-4}	0.69	0.19	0.22	1.4×10^{-2}	0.81	6.1×10^{-6}	0.75	0.66–0.85	DDC
7q32.3	rs10269601	A	Trend	0.28	0.34	1.1×10^{-5}	0.75	0.29	0.30	2.4×10^{-1}	0.93	2.2×10^{-4}	0.82	0.74–0.90	PLXNA4
10p12.2	rs11013140	T	Rec	0.31	0.36	5.6×10^{-6}	0.53	0.33	0.35	5.4×10^{-1}	0.93	2.2×10^{-4}	0.63	0.49–0.80	PIP4K2A
10q24.2	rs11190062	A	Trend	0.12	0.09	1.3×10^{-5}	0.65	0.10	0.10	7.8×10^{-1}	0.97	3.9×10^{-4}	0.78	0.67–0.89	CNMM1
10q26.13	rs7076268	C	Het	0.22	0.25	1.6×10^{-5}	0.68	0.26	0.25	4.7×10^{-2}	1.19	1.1×10^{-4}	0.71	0.60–0.85	FAM53B
11p15.4	rs11036238	C	Trend	0.09	0.14	3.9×10^{-7}	0.61	0.10	0.14	6.8×10^{-6}	0.65	3.7×10^{-11}	0.63	0.55–0.72	HBB
13q31.3	rs1444227	C	Het	0.15	0.11	3.6×10^{-5}	1.54	0.12	0.11	9.5×10^{-1}	1.01	1.4×10^{-3}	1.27	1.10–1.48	GPC6
14q21.2	rs6572335	G	Dom	0.17	0.12	9.3×10^{-7}	1.64	0.12	0.14	1.9×10^{-1}	0.88	4.4×10^{-3}	1.22	1.06–1.40	–
16q13	rs2161647	A	Dom	0.17	0.23	4.1×10^{-5}	0.69	0.21	0.21	8.8×10^{-1}	1.01	5.3×10^{-3}	0.84	0.74–0.95	C1APIN1
17p13.1	rs6503319	T	Trend	0.51	0.45	6.6×10^{-5}	1.28	0.49	0.45	2.1×10^{-2}	1.14	7.2×10^{-7}	1.21	1.12–1.31	SCO1
19p13.3	rs460375	G	Dom	0.45	0.48	1.2×10^{-5}	0.66	0.45	0.45	5.1×10^{-1}	0.94	1.4×10^{-4}	0.76	0.66–0.88	EMR1

SNPs showing association in the GWA study of 958 cases and 1,382 controls that were subsequently typed in an independent sample of 1,087 cases and 2,376 controls. The GWA analysis is PCA-corrected and the replication analysis is stratified by ethnic group classified as Mandinka, Wollof, Fula, Jola or other. The combined results are obtained by a weighted linear combination of odds ratios. Because our replication assay method was unable to type certain SNPs, we do not have replication data on the 4p15.2, 4q24 and 8p22 loci, and at some other loci the SNPs in the replication experiment are not those with the strongest signals in the GWA experiment.

We identified 19 regions of the genome in which there were ≥2 SNPs with $P < 10^{-4}$ or a single SNP with $P < 10^{-6}$ (**Table 1**). Three regions other than *HBB* had signals of $P < 10^{-6}$. Chromosome 2q37 had five SNPs at $P < 10^{-4}$ with a minimum of $P = 5.1 \times 10^{-7}$ in a recessive model: the closest genes are *SPATA3*, encoding a spermatogenesis-associated protein; *LOC257407*, encoding a hypothetical protein; *PSMD1*, encoding a proteasome subunit; and *GPR55*, encoding a G protein–coupled receptor. Chromosome 5p12 had three SNPs at $P < 10^{-4}$ with a minimum of $P = 4.5 \times 10^{-7}$ in a heterozygous advantage model: this region has a number of genes encoding proteins of unknown function. Chromosome 14q21 had three SNPs at $P < 10^{-4}$ with a minimum of $P = 6.5 \times 10^{-7}$ in a dominant model: this area seems to be a gene desert.

Signal plots for each of the regions listed in **Table 1** are shown in **Supplementary Figure 5** online. In addition to SNPs that were directly genotyped, **Supplementary Figure 6** online shows imputed genotypes in each of these regions, with the caveat that imputation is based on the HapMap Yoruba reference panel and this can give erroneous results when applied to the Gambian population, as we show above for HbS.

Severe malaria is composed of several overlapping clinical entities, notably cerebral malaria and severe malaria anemia[23,24], and there may be genetic effects that are specific for a particular subphenotype. We carried out separate principal component–corrected analyses on the 758 cases with cerebral malaria and 297 cases with severe malaria anemia (97 subjects had both cerebral malaria and severe malaria anemia). Results of the subphenotype association tests are included in **Supplementary Table 3**. Acknowledging the limitation of low sample size, there is an indication that some GWA signals may be specific for cerebral malaria. They include two signals of $P \sim 10^{-5}$, one close to *CAMTA1* (encoding calmodulin-binding transcription activator) and another within *RYR2* (encoding a ryanodine receptor involved in calcium-dependent signaling). These loci are particularly notable in view of the importance of calcium signaling in cerebral function.

Replication study of severe malaria

We conducted a replication study in an independent sample of 1,087 severe malaria cases and 2,376 controls for a number of loci containing genes of interest (**Table 2**) and for all of the listed in **Table 1**. The replication study used the Sequenom iPlex platform: at some loci it was not possible to type the SNP with the strongest association in the GWA study by this method. Analysis of the replication data was stratified by self-reported ethnicity and was integrated with the PCA-corrected GWA analysis using a weighted combination of linear odds ratios. The combined sample after application of quality control filters to the GWA data comprised 2,045 cases and 3,758 controls.

At the *HBB* region on chromosome 11p15.4, a trend test for rs11036238 had OR of 0.61 ($P = 3.9 \times 10^{-7}$) in the GWA study, 0.65 ($P = 6.8 \times 10^{-6}$) in the replication study and 0.63 (95% CI = 0.55–0.72, $P = 3.7 \times 10^{-11}$) in the combined sample.

A number of newly identified loci that showed notable associations in the combined sample are described in the **Supplementary Note** online. The strongest effects were observed on chromosome 17p13 (rs6503319: trend test OR = 1.21, 95% CI = 1.12–1.31, $P = 7.2 \times 10^{-7}$) close to *SCO1*, which encodes a protein involved in cytochrome oxidase function; and on chromosome 7p12.2 (rs1451375: dominant model OR = 0.75, 0.66–0.85, $P = 6 \times 10^{-6}$; and rs7803788, dominant model OR = 0.76, 0.68–0.85, $P = 2.4 \times 10^{-6}$) intronic to *DDC*, encoding dopa decarboxylase, which is involved in dopamine and serotonin synthesis. Notably, *DDC* has also been linked to malaria refractoriness in mosquitoes[25]. Although these findings are of

considerable interest, they cannot be regarded as conclusive until they have been replicated at multiple study sites. Below we discuss the challenges of multicenter replication studies in African populations.

DISCUSSION

The application of GWA analysis to populations in Africa could provide fundamental insights into resistance to infectious disease and the genetic origins of common diseases[26]. The main conclusion of the current study is that this will require a different methodological approach than that used for GWA analysis in European or Asian populations. GWA analysis proceeds in three stages: first, discovery of regions of the genome with significant associations; second, multicenter replication studies; and third, fine mapping of causal variants. Here we consider each of these stages of analysis, and how they might be effectively implemented in Africa.

At the first stage of GWA analysis, screening many SNPs across the genome, a stringent threshold for statistical significance is used to reduce false-positive rates. There is theoretical debate about where to set this threshold but take, for example, the level of $P < 5 \times 10^{-7}$ used by the Wellcome Trust Case Control Consortium in the British population[9,27,28]. In practice it is difficult to achieve this threshold in Africa, because of weak LD between the marker SNPs that are genotyped and causal variants. We found, for example, that the GWA signal at the HbS locus barely reaches this level of significance, although the causal variant is common and confers tenfold reduction in risk of severe malaria. It is likely that there are many authentic loci with modest genetic effects that are several orders of magnitude below this level of significance, as we have shown here for the ABO locus.

Multipoint imputation offers a potential solution to this problem. We found that the GWA signal around the HbS variant can be boosted by several orders of magnitude by imputation, from $P = 3.9 \times 10^{-7}$ to 4.5×10^{-14}. Although GWA signals are often increased by imputation, this is a much larger effect than is commonly observed in European populations. It reflects the fact that no individual marker SNP is strongly correlated with the HbS causal variant but that it is possible to gather much additional information about the causal variant from haplotypic combinations. The ability of multipoint imputation to boost a GWA signal is potentially greatest in situations where there is no single marker SNP that is strongly correlated with the causal variant, and imputation could therefore be of particular importance in Africa.

Multipoint imputation cannot work effectively without accurate data on sequence variation and haplotype structure. This is critical in situations of low LD, where a significant proportion of common variants have no strong marker SNP, as is the case for HbS, and important genetic effects might be missed if these variants are not genotyped, or accurately imputed, at the first stage of GWA analysis. The genetic diversity found across Africa increases the imperative for the data underpinning imputation to be population specific. Population-specific data are needed because sequence variants and haplotypes can be highly localized within Africa; for example, the malaria resistance factor hemoglobin C (HbC) is found in some parts of West Africa but not others[29,30].

Development of an optimal genome-wide SNP genotyping platform for use in Africa would help to strengthen the signals of association that are directly observed at the first stage of GWA analysis, as well as increase the accuracy of imputation. Although there have been steady improvements in genotyping platforms since the start of the present study, there is no platform that achieves genome coverage in Africa that is close to that in Europe[31–34]; indeed, the number of SNPs needed to achieve this cannot be precisely estimated until more

complete sequencing information from African populations is available[35]. From initial HapMap data it was estimated that 1 SNP per 2 kb in Africa would give approximately the same tagging efficiency as 1 SNP per 5 kb in Europe, but as more information emerges about sequence variation in Africa, the estimated number of SNPs required to tag common variants is tending to increase[5,35]. The challenge is to determine the optimal number of genotyped SNPs that, when combined with genome-wide resequencing data from a representative sample of the same population, would allow accurate imputation of all common variants.

The second stage of GWA analysis is to replicate signals of association in large multicenter studies. The problem is that replication of association at multiple locations depends on the allele frequency of the marker SNP and the causal variant, as well as the LD between the marker SNP and the causal variant, being relatively constant across locations. The high extent of genetic diversity across Africa therefore creates uncertainty about whether associations with marker SNPs will replicate at different locations, even if there is a true causal variant. For example, we find that patterns of LD between the HbS causal variant and surrounding SNPs vary to such an extent between the Gambian and Yoruba populations that none of the GWA associations found at the HbS locus in the present study would be expected to replicate across West Africa. Associations with marker SNPs may also fail to replicate when the causal variant differs in frequency between locations.

It is therefore difficult to design effective multicenter replication studies without information about sequence variation and haplotype structure in the relevant African populations. However, because the main problem is the variable relationship between marker SNPs and causal variants, the problem of replication becomes greatly simplified once a shortlist of potential causal variants has been identified. This is an additional reason to carry out high-resolution multipoint imputation at the first stage of GWA analysis, as it allows putative causal variants to be tested directly in different populations.

The third stage of GWA analysis is to identify causal variants. In European populations it can be extremely difficult to distinguish causal variants from nonfunctional polymorphisms that are in strong LD with causal variants. This study provides proof of principle that, under the conditions of low LD found in Africa, it is possible to uniquely identify a causal variant at the first stage of GWA analysis by multipoint imputation, based on deep sequencing data that are population specific. Thus, although a considerable amount of work is needed to overcome the problem of low LD at the initial stages of GWA analysis in Africa, a byproduct of this work is that it might be possible to proceed relatively rapidly from a GWA scan and replication studies to the identification of causal variants. The major limiting factor, at all stages of GWA analysis in Africa, is the need for population-specific data on genome sequence variation. In the near future, this limiting factor will be overcome by advances in genome sequencing technologies, through initiatives such as the 1000 Genomes Project.

METHODS

Study sample. The study sample comprised 1,060 cases and 1,500 controls from a mixed urban and rural area of approximately 400 square miles in the Kombos region of The Gambia. Cases were children admitted to hospital with severe malaria: they had a median age of 4.3 years and 18% had a fatal outcome. Controls were newborns recruited from routine births at local health clinics. The control data were shared with a GWA study of tuberculosis (unpublished data). We could theoretically achieve a modest increase in power by selecting controls that had gone through

childhood without developing severe malaria, but past medical history is difficult to ascertain with confidence in this population.

Severe malaria is made up of several overlapping clinical entities[23,24,36]. The cases analyzed here included 82% with cerebral malaria, 30% with severe malarial anemia and 11% with respiratory distress. By estimating the protective effect of the HbAS genotype in this study sample we can exclude high rates of diagnostic misclassification, which can arise when other severe diseases mimic the clinical features of severe malaria: we found ORs of 0.12 (95% CI = 0.07–0.21) for cerebral malaria, 0.10 (0.04–0.24) for severe malaria anemia, 0.08 (0.02–0.38) for respiratory distress and 0.09 (0.05–0.16) for severe malaria in general. These analyses are stratified for self-reported ethnic group using the Mantel-Haentzel test. Because the control group was chosen to represent all births, and because only a minority of children develop severe malaria, these ORs can be viewed as an estimate of relative risk in the general population.

Sample inclusion criteria. The initial sample set consisted of 1,060 cases of severe malaria and 1,500 controls from the Kombos region of The Gambia, West Africa. Cases of severe malaria, defined essentially according to WHO criteria[36], were recruited on admission to the Paediatric Department of the Royal Victoria Hospital in Banjul after obtaining informed consent. The main forms of severe malaria in Gambian children are cerebral malaria, severe malaria anemia and respiratory distress[23,24]. In this study we define cerebral malaria as a Blantyre coma score[37] of ≤3, persisting for >30 min after cessation of a transient seizure or after correction of hypoglycaemia, in a child with asexual forms of *P. falciparum* on blood film and no other evident cause of coma. Severe malaria anemia is here defined as packed cell volume ≤15%, or hemoglobin ≤5 g/dl with asexual forms of *P. falciparum* on blood film. Controls were cord blood samples obtained from routine births at local health clinics in the Kombos region after obtaining informed consent. The study was approved by The Gambia Government/Medical Research Council Joint Ethics Committee and by the Oxford Tropical Research Ethics Committee. Ethics approval documents and informed consent forms are available upon request.

Imputation. The program IMPUTE[10] was used to infer the genotypes of SNPs in the Yoruba panel of the International HapMap Project. To avoid SNPs with genotyping errors in the Gambian study, we used only SNPs on the Affymetrix platform for the Gambian data with minor allele frequency >1% and missingness <5% in both case and control cohorts, and with Hardy-Weinberg equilibrium $P \geq 10^{-7}$ in controls as the input for IMPUTE. The results from IMPUTE were used in our association study in two ways: (i) the same spectrum of association tests was conducted at each imputed SNP where, for each individual, the genotype corresponding to the maximum posterior probability is assigned unless this probability is <0.9 and a missing genotype is assigned instead; (ii) the same spectrum of association tests was conducted at each imputed SNP where the posterior probabilities of the calls are used to average over the uncertainty in the inference of the genotypes. In the former analysis, only imputed SNPs that satisfy the genome-wide SNP quality control criteria (see 'SNP filters' in **Supplementary Methods** online) were considered; in the latter analysis, only imputed SNPs with relative statistical information >0.5 were considered.

Sequencing in the beta-globin region. The beta-globin region was sequenced in 62 randomly selected Gambian control individuals who had previously been genotyped on the Affymetrix 500K array. The sequenced region spans 110 kb from 5,179,297 to 5,289,530, and encompasses all five beta-globin genes (*HBB*, *HBD*, *HBE1*, *HBG1*, *HBG2*) and an olfactory receptor (*OR51B1*). To avoid false positives introduced by sequencing errors and as the underlying interest here is in detecting malarial association with common polymorphisms, we only define a polymorphic marker if the minor allele frequency of this marker exceeds 5% or, equivalently, if at least 4 minor alleles out of the 124 chromosomes are observed. On the basis of this definition we identified 202 SNPs in the sequenced region (66 of which were previously unknown SNPs), although we expect the true number to be higher.

To avoid edge effects in haplotype phasing and imputation, the data for each sequenced sample was extended by including SNPs from the Affymetrix array flanking both ends of the sequenced region, creating a 1-Mb region centered on

rs334 (at 5,204,808 in build 35) from 4,705,000 to 5,705,000 spanning 453 SNPs in total. Haplotypes for this 1-Mb region were constructed using fastPHASE[38], and were subsequently used as the reference panel for imputation using IMPUTE[10].

Genotypes for all 453 SNPs in this 1-Mb region were imputed across the remaining 1,325 controls and all 958 cases. We carried out tests for malaria association on all the imputed SNPs and the results from the trend test are shown in **Figure 5** as red dots. Association results for SNPs on the Affy 500K array are shown in black. Rather than threshold the imputed calls with a missingness filter, the association test averages over the imputation uncertainty by using the genotype posterior probabilities from the imputation. As before, the first three principal components estimated from the array genotypes using EIGENSTRAT were included as covariates in the test for association.

Additional description of the DNA handling protocol, data quality control and analytical methodologies can be found in the **Supplementary Methods** and **Supplementary Figures 7–10** online.

URLs. MalariaGEN, http://www.malariagen.net.

Note: Supplementary information is available on the Nature Genetics website.

AUTHOR CONTRIBUTIONS
The clinical study in The Gambia was designed and carried out by M.J., K.A.B., D.J.C., D.P.K., M. Pinder, G.S., F.S-J. and S.U. Genotyping and sequencing studies were designed and performed by K.A.R., P. Deloukas, S.A., S.J.B., S.C., A.C., A. Dunham, A.E.F., A.G., R.G., S.E.H., M. Inouye, A.E.J., K.K., A. Mendy, A.P., S.P., J. Ragoussis, J. Rogers, K.R., E.S., P.W. and C.W. The analysis group comprised Y.Y.T., K.S.S., T.G.C., P. Donnelly, J. Marchini, A. Morris, M.S. and D.P.K. This study forms part of a multicenter investigation of genetic resistance to malaria led by E.A.A., T.A., S.A., O.A., K.A.B., D.J.C., P.C., P. Deloukas, A. Djimde, A. Dolo, O.K.D., C.D., S.D., J.E., J.F., D.F., T.T.H., R.D.H., M. Ibrahim, N.K., G.K., K.A.K., D.P.K., M.L., J. Makani, K.M., P.M., D.M., M.E.M., I.M., M. Parker, N.P., C.V.P., O.P., J. Ragoussis, J. Rogers, J. Reeder, H.R., E.M.R., A.S., P.S., S.S., G.S., A.T., T.E.T., M.T., M.T-B., T.N.W. and M.W. The manuscript was written by D.P.K., Y.Y.T. and K.S.S.

1. Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y. & Hay, S.I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434**, 214–217 (2005).
2. Greenwood, B.M. *et al.* Mortality and morbidity from malaria among children in a rural area of The Gambia, West Africa. *Trans. R. Soc. Trop. Med. Hyg.* **81**, 478–486 (1987).
3. Mackinnon, M.J., Mwangi, T.W., Snow, R.W., Marsh, K. & Williams, T.N. Heritability of Malaria in Africa. *PLoS Med.* **2**, e340 (2005).
4. Malaria Genomic Epidemiology Network. A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732–737 (2008.).
5. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
7. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
8. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
9. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
10. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

11. Hanchard, N. *et al.* Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC Genet.* **8**, 52 (2007).

12. Chakravarti, A. *et al.* Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).

13. Pagnier, J. *et al.* Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl. Acad. Sci. USA* **81**, 1771–1773 (1984).

14. Chebloune, Y. *et al.* Structural analysis of the 5′ flanking region of the beta-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. *Proc. Natl. Acad. Sci. USA* **85**, 4431–4435 (1988).

15. Ruwende, C. *et al.* Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**, 246–249 (1995).

16. Tishkoff, S.A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).

17. Guindo, A., Fairhurst, R.M., Doumbo, O.K., Wellems, T.E. & Diallo, D.A. X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. *PLoS Med.* **4**, e66 (2007).

18. Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986).

19. Mockenhaupt, F.P. *et al.* Alpha(+)-thalassemia protects African children from severe malaria. *Blood* **104**, 2003–2006 (2004).

20. Williams, T.N. *et al.* Both heterozygous and homozygous \{alpha\}+thalassemia protect against severe and fatal *Plasmodium falciparum* malaria on the coast of Kenya. *Blood* **106**, 368–371 (2005).

21. Cappellini, M.D. & Fiorelli, G. Glucose-6-phosphate dehydrogenase deficiency. *Lancet* **371**, 64–74 (2008).

22. Fry, A.E. *et al.* Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Hum. Mol. Genet.* **17**, 567–576 (2008).

23. Marsh, K. *et al.* Indicators of life-threatening malaria in African children. *N. Engl. J. Med.* **332**, 1399–1404 (1995).

24. Taylor, T. *et al.* Standardized data collection for multi-center clinical studies of severe malaria in African children: establishing the SMAC network. *Trans. R. Soc. Trop. Med. Hyg.* **100**, 615–622 (2006).

25. Romans, P., Black, W.C., Sakai, R.K. & Gwadz, R.W. Linkage of a gene causing malaria refractoriness to Diphenol oxidase-A2 on chromosome 3 of *Anopheles gambiae. Am. J. Trop. Med. Hyg.* **60**, 22–29 (1999).

26. Tishkoff, S.A. & Williams, S.M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3**, 611–621 (2002).

27. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).

28. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

29. Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96**, 2358–2363 (2000).

30. Modiano, D. *et al.* Haemoglobin S and haemoglobin C: 'quick but costly' versus 'slow but gratis' genetic adaptations to *Plasmodium falciparum* malaria. *Hum. Mol. Genet.* **17**, 789–799 (2008).

31. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).

32. Conrad, D.F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 1251–1260 (2006).

33. de Bakker, P.I. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).

34. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38**, 663–667 (2006).

35. Bhangale, T.R., Rieder, M.J. & Nickerson, D.A. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.* **40**, 841–843 (2008).

36. World Health Organization. Communicable Diseases Cluster. Severe falciparum malaria. *Trans. R. Soc. Trop. Med. Hyg* **94**(Suppl. 1), S1–S90 (2000).

37. Molyneux, M.E., Taylor, T.E., Wirima, J.J. & Borgstein, A. Clinical features and prognostic indicators in paediatric cerebral malaria: a study of 131 comatose Malawian children. *Q. J. Med.* **71**, 441–459 (1989).

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