doi:10.1038/nature15390

1. Assessment of association evidence in specific regions

Replication evidence in the *FREM3/GYPE* region

In the *FREM3/GYPE* region we typed several SNPs including the lead imputed marker rs184895969 on the Sequenom platform (Figure 1, Extended Data Figure 7a). However, agreement between the Sequenom and best-guess imputed genotypes at this SNP was poor (r^2<0.1, estimated using EM algorithm) both across three cohorts and in East African populations, where the frequency is highest. The best Sequenom markers by r^2 were rs186873296 (r^2=0.932), rs186790584 (r^2=0.943) and rs149914432 (r^2=0.87). Replication evidence at these three SNPs was similar and is presented in Figure 2 and Extended Data Figure 7a.

Association with severe malaria subtypes in the *FREM3/GYPE region*

To estimate the effect of the association in the *FREM3/GYPE* region on clinical subtypes of severe malaria, we fit a logistic regression model for cerebral malaria (CM) versus controls and for severe malarial anaemia (SMA) versus controls using Sequenom data for rs186873296 across discovery and replication samples in Kenya, including ethnicity as a covariate to control for population structure. In total there were 4118 controls, 659 CM cases and 263 SMA cases in Kenya, of which 136 cases were classified as both CM and SMA. A further 944 cases were not classified as either CM or SMA and were not used to estimate subtype-specific effects; see also Extended Data Figure 1b. We observed a similar protective effect for CM (OR = 0.58, 95% CI=0.45-0.74) and SMA (OR = 0.55, 95% CI=0.38 -0.81).

Evidence for association in *INPP4B*

We investigated the evidence for association at a second peak of association (rs149373719) located approximately 1Mb away from the lead marker in the *FREM3/GYPE* region (see Figure 1). Discovery phase evidence for association at this SNP was driven largely by Kenya, where a strong effect was observed, but an opposite direction of effect was observed in Malawi (Extended Data Figure 7e). Replication evidence at the directly-typed marker rs77389579 was more consistent across East African populations, but with little evidence in central or western Africa (Extended Data Figure 7g) where the frequency was low.

To determine if this signal was driven by LD with the lead imputed marker at the FREM3/GYPE region, we conducted association analysis in the region conditional on the additive dosage at rs184895969 in Kenya. SNPs in *INPP4B* remained associated ($P<1x10^{-4}$ at rs13103597, Extended Data Figure 7d), suggesting that this association may be somewhat independent of the *FREM3/GYPE* association signal.

Replication evidence in the *ARL14* region

We observed a strong signal of association in discovery populations near the gene *ARL14* (rs74954675, $BF_{avg}=7.8x10^4$, OR=1.34, 95% CI 1.21-1.49, $P=5.1x10^{-8}$ in fixed-effect meta-analysis for dominant model; see Extended Data Figure 6). We directly typed rs74954675 and two other regional SNPs on the Sequenom platform, with genotypes showing good agreement with best-guess imputed genotypes (r^2=0.94 between Sequenom-typed and best-guess imputed genotypes in discovery populations at rs74954675; Extended Data Figure 7a). However, replication evidence was poor (one-sided fixed-effect $P > 0.5$), with no replication population showing strong evidence, and the effect estimated as protective in three of seven replication populations.

We note two factors that might impede our ability to replicate signals of association. Firstly, the lack of a genome-wide data in replication samples may mean we do not perfectly control for population structure[1]. Secondly, the relative enrichment of CM and SMA subtype among discovery samples (see Extended Data Figure 1b) may indicate phenotypic differences between discovery and replication sets.

Evidence for association at *MARVELD3*

We examined evidence for association at rs2334880, previously reported as associated with severe malaria in a population from Ghana[2], in our discovery data but saw no evidence of association at this SNP ($BF_{avg} = 0.12$, $P > 0.18$ in fixed-effect meta-analysis or population-specific analyses) or for other SNPs in the region ($BF_{avg} < 20$ for SNPs within 500kb). A regional association plot detailing the evidence in this region can be found at http://www.malariagen.net/resource/14.

2. Analysis of putative functional and structural variants in the *FREM3/GYPE* region

Determination of MNS blood group variants

We used dbRBC[3] and the Ensembl Genome Browser[4] to identify mutations encoding the MNS blood group antigens, cross-checking with previously published protein/sequence alignments[5]. The M/N blood groups are encoded by three SNPs in the 20th and 24th amino acid encoded by *GYPA*, which is transcribed on the reverse strand of the human reference sequence. Comparison with the cDNA and reference sequence of the major *GYPA* transcript in Ensembl identifies these as rs7658293, rs7687256 and rs7687256 (Extended Data Figure 7c). The human reference sequence encodes Glutamine and Leucine at these codons and hence the N blood group. Of these SNPs, only one (rs7687256) was present in the 1000 Genomes reference panel.

The S/s blood groups are encoded by a single SNP in the 48th amino acid of *GYPB*, which is also transcribed on the reverse strand of the human reference sequence. As above we inspected the cDNA sequence of the major transcript of *GYPB* to identify this SNP as rs7683365, also present in the 1000 Genomes reference panel. At this amino acid, the human reference sequence encodes Threonine and hence the s blood group (Extended Data Figure 7c).

Reimputation of large structural variants

To improve imputation of the deletions esv2668125 and esv2662558 (see Methods), we generated a custom reference panel for each deletion in the region 140-145Mb on chromosome 4, plus a margin of 500kb at either end. These panels differed from the 1000 Genomes reference panel in two ways. Firstly, we corrected the haplotypes for three individuals NA18519, NA19185, NA19222 based on examination of the underlying read data, as described in Methods and Extended Data Figure 8e-g. Secondly, because neither the 1000 Genomes pipeline nor our genotype calling took into account ploidy within large structural variants such as these, typed variants within the deletions could lead to spurious imputation. To avoid this, for each deletion we removed all variants within the deletion breakpoints from the modified panel.

Reimputed calls for esv2668125, which was imputed with high confidence in the discovery imputation (IMPUTE info>0.93), showed good agreement with the original calls (r^2=0.98, concordance = 0.99 for the genotype with highest imputation probability; non-reference allele frequency = 5% across populations).

Deletion esv2662558 was reimputed with higher confidence (IMPUTE info > 0.59) than in the discovery scan (IMPUTE info < 0.4). Reimputed calls showed low correlation (r^2=0.13, concordance = 0.97) and an increase in allele frequency (2.1% versus 1.7%) relative to the discovery imputation.

Association analysis with regional variants

To investigate the association signal within the *FREM3/GYPE* region, we fit association models including the lead imputed marker (rs184895969) or the Sequenom-typed lead marker (rs186873296), and a set of potentially functional variants, including the ABPs (represented by rs67600034), reported regional eQTLs[6] (including rs1822841 and rs11100806), large deletions (esv2662558 and esv2668125 based on re-imputation as described above), the FREM3 missense mutation that is within the 95% confidence interval in our data (rs181620317), a SNP tagging the N/M blood group (rs7687256), and a SNP tagging the S/s blood group (rs7683365). We used logistic regression in R to test whether the putatively functional variants jointly improved model fit relative to the null model, which included 5 PCs and either the lead imputed marker rs184895969 or the Sequenom-typed rs186873296 included as an additive effect. We observed no significant improvement in model fit (likelihood ratio test P>0.05), either across all samples or within each population.

To investigate whether specific haplotypes might better explain observed patterns of risk, we included all haplotypes having frequency at least 1% at the above variants, based on best-guess haplotypes from imputation, in a joint model of association and asked whether the haplotypes improved model fit relative to the model including only the lead marker. We observed no evidence that haplotypes explained additional association signal (P>0.1) either across all samples or within each population.

For each variant, we also tested whether including the variant marginally under an additive, dominant, recessive or heterozygote model improved model fit. In Malawi, we saw nominally significant evidence for a protective effect of non-reference alleles at the two SNPs rs67600034 and rs11100806, which are in LD (r^2=0.93 in YRI+LWK), with the highest evidence for a recessive protective effect of the non-reference allele (P=0.005 at rs11100806). However, effects were inconsistent across populations, with the same allele estimated has having a mild risk effect in Kenya.

We also tested for interactions of each variant with the lead *FREM3/GYPE* marker and noted nominally significant interactions of rs186873296 with rs7683365 (P=0.004), though this effect was less pronounced when testing with the imputed rs184895969, and with the *GYPE* eQTL rs7686794 (P=0.006), in Gambia. No variant showed evidence for a conditional or interaction effect across all populations (P>0.01).

GENECLUSTER analysis

We ran GENECLUSTER[7] as described in Methods. In Kenya, using the TREESIM tree inferred at the position of the lead marker (rs184895969), GENECLUSTER placed a mutation subtending all haplotypes carrying the risk allele with log10(*BF*) = 3.09 (*OR* = 0.6; Figure 3). Allowing for a second mutation defines three classes of haplotype risk and marginally

increased the evidence for association *log10(BF)* = 3.23. The second mutation defines a sub-tree which tags the ABP haplotype and had marginally increased risk (*OR = 1.01*). In Gambia and Malawi, the same branch of the tree was inferred as the location of the best single mutation, but with no evidence of association (*BF* < 1) and decreased odds ratios (OR = 0.98 and 0.88 respectively). We observed considerable variation between estimated marginal trees in the region. In Kenya, the position with the highest posterior was at 144694105 with *log10*(*BF)*=4.4 under the two-mutation model.

3. Analysis of enrichment of malaria-associated loci in functional categories

Gene-based enrichment analysis

We tested for enrichment of our tier 1 and tier 2 regions in predefined collections of functionally related genes as follows. We tested for enrichment in four broad sets of genes:

1. HUGO Gene families. These were downloaded from the HUGO Gene Nomenclature Committee website. A total of 560 gene sets were used.
2. Custom gene sets. We also included two sets of genes previously listed as nearest to putative ancient balancing polymorphisms (Leffler *et al*[8], Supplementary Table S4 and S5), treating shared coding SNPs and shared haplotypes separately, and also formed the combined set of genes in the ISET and VSET HUGO Gene families as single ISET/VSET gene set.
3. Gene Ontology (GO) terms. We downloaded the "latest-lite" database dated 2013-11-23 from www.geneontology.org and used a SQL query to extract a transitive list of members of each gene family. A total of 17718 GO terms were included.
4. Sets of genes previously identified as associated with red blood cell traits[9].

We aimed to test for enrichment of these gene sets in the list of genes nearest to our lead SNPs in each tier. To form this list, we considered all RefSeq genes (downloaded from the UCSC Genome Browser MySQL database[10] on 2013-03-18). For each of our tier 1 and tier 2 regions, we determined the nearest gene to the lead SNP by physical distance, including only genes which intersected the association region, and considering all possible transcripts where several transcripts existed for a gene.

For each gene set we computed an odds ratio (OR) by forming the 2x2 table

	In gene set	Not in gene set
In GWAS association region	*a*	*b*
Not in GWAS association region	*c*	*d*

where *a, b, c* and *d* represent the counts of genes in the given categories, and computing the odds ratio as *OR=ad/bc*. To do so, it is necessary to specify a 'universe' set containing all genes that appear in the table (i.e. containing a total of *a+b+c+d* genes). For most analyses we took the universe to consist of all RefSeq genes that lie on the autosomes and the X chromosome (23584 genes in total), while for Gene Ontology analyses we further restricted to genes annotated to at least one GO term (17708 genes).

Fisher's exact test can be used to test the hypothesis OR=1. However, Fisher's exact test assumes an underlying model of binomial sampling in each row of the table, which is inappropriate here because some genes are more likely to be the nearest gene to a GWAS lead marker by chance. In particular this is true of the set of genes closest to ABP haplotypes, which is enriched for genes covering an unusually large proportion of the genome (as measured by the size of the of the genome or the number of SNPs in our study for which the gene is the closest gene; see Extended Data Figure 10e). To assess enrichment we therefore adopted a simulation scheme.

We simulated 10,000 sets of tier 1 lead SNPs by iteratively picking SNPs uniformly from the set of SNPs included in the GWAS analysis and, as for the GWAS SNPs, excluding other SNPs within a region of 2.5cM±25kb centred at the chosen SNP, to form a list of 34 SNPs. For each set of simulated SNPs we computed odds ratios as above and computed an empirical P-value, denoted P_{sim}, as the proportion of odds ratios as large or larger than that observed for the true data. These P-values thus reflect the null model in which each SNP included in our study has equal probability of being a lead SNP.

Potentially, SNPs close to or in LD with ABPs might have a frequency distribution that differs from the genome-wide distribution, affecting the power of our study at these SNPs. We therefore conducted further simulations matching the frequency of picked SNPs to the frequency of SNPs in GWAS tier 1. Specifically, we computed the 1% frequency bin of each GWAS tier 1 SNP in controls across all cohorts. For each simulated set we then picked the *i*th SNP from the set of SNPs having the same frequency bin as the *i*th GWAS SNP (taken in order of decreasing BF_{avg}).

Simulations for tier 2 were conducted as for tier 1 except that we excluded all SNPs in the GWAS tier 1 regions from the simulation, and any gene intersecting the GWAS tier 1 regions from the counts forming the 2x2 table. For the list of genes nearest to true GWAS tier 2 hits, this removed gene *PAX5* which also overlaps the tier 1 region near *ZCCHC7*. To assess the extent to which enrichment in tier 1 is driven by the previously identified loci and the new FREM3 locus, we also conducted additional tier 1 simulations removing regions of *HBB*, *ABO*, *ATP2B4*, *FREM3*, *INPP4B* and *HHIP*.

All simulations were conducted using the program inthinnerator, available at http://www.well.ox.ac.uk/~gav/inthinnerator. Empirical P-values for all tier 1 and tier 2 simulations are shown in Extended Data Figure 10d.

Distance-based enrichment analysis

Using the same set of simulated SNP sets we also evaluated enrichment of ABP haplotypes and coding SNPs within different length scales of GWAS lead SNPs. For each distance 50kb, 100kb, 250kb, 500kb, 750kb, 1Mb, 2.5Mb, 5Mb and 10Mb we computed an indicator variable encoding whether an ABP occurred within the given distance of each lead SNP in the GWAS set and each simulated set. We computed empirical P-values (again denoted P_{sim}) as the proportion of simulations for which the number of regions containing an ABP within

the given distance was greater than or equal to the number for the GWAS set. Empirical P-values for SNPs in tier 1 and 2 are shown in Extended Data Figure 10d.

Overlap between genes nearest to ABPs and to GWAS loci

Apart from *ABO* and *FREM3/GYPE*, we find six other putative signals of association with severe malaria (4 in tier 1, 2 in tier 2) where the nearest gene to the GWAS peak is also the nearest gene to an ABP. *DSCAM* is a member of the immunoglobulin superfamily whose *Anopheles gambiae* orthologue plays a role in innate immune recognition of *Plasmodium falciparum*[11]. *NRG1* is a glycoprotein with diverse signalling properties and a great variety of isoforms that include immunoglobulin I-set and epidermal growth factor-like domains. *CNTNAP5* belongs to a family of cell adhesion and receptor proteins. *TMEM132C* encodes a transmembrane protein of unknown function. *CACNA2D1* and *RYR2* are both part of calcium channel complexes which is interesting in view of the importance of calcium signalling in host cell invasion and the known association with *ATP2B4*[2] which encodes the main erythrocyte calcium channel.

4. Interpretation and sensitivity of the model-averaged Bayes factor

Interpretation of the Bayes factor

Overall evidence for association is measured by the posterior probability for association, which takes into account both the Bayes factor and the prior odds of association:

(Posterior odds that variant is associated) = *BF* x (prior odds that variant is associated)

For GWAS studies, a value of 1 in 100,000 for prior odds, based on assumptions about the number of loci plausibly involved in disease susceptibility and the extent of LD surrounding the associated variants, has commonly been used[12,13]. For the lead marker in the *FREM3/GYPE* region this gives a posterior odds of association of 0.156 and a posterior probability of 13.5% using the BF_{avg} computed in discovery samples, and a posterior probability of almost 100% using the Bayes factor computed across discovery and replication samples, while a variant with BF_{avg}=2500 has posterior probability of association of 2.5%. Across the 31 tier 1 loci not previously reported, we would expect about three to represent real associations. However, this computation relies crucially on the prior odds of association. If associated variants are tagged over more of the genome (say around 0.01cM) and more regions of the genome (say 50 or more) are associated, giving a prior odds of around one in 20,000, a variant with BF_{avg}=2500 has a posterior probability of association of 11%, and we would expect about 8 of the novel 31 loci to be true associations.

Prior sensitivity analysis

To investigate the sensitivity of the model-averaged Bayes factor to the choice of prior, we plotted the ratio of BF_{avg} computed under modified priors to its value under the prior used in the GWAS (referred to here as the BF_{avg} ratio) for the 32 tier 1 autosomal variants (Extended Data Figure 5c). We considered modifications of the prior along several one-dimensional axes of variation:

- The prior standard deviation on effect sizes. We varied the small-effect standard deviation from 0.01 to 0.75, keeping the large-effect and small-effect standard deviations in the proportion 0.75 : 0.2.
- The prior weight assigned to additive, dominant, recessive and heterozygote models. For each model we varied the prior weight from 0 to 1, keeping prior weights on the other three models in the same relative proportion. (Note that because the BF_{avg} is a linear combination of component models, the BF_{avg} ratio depends linearly on the prior weight.)
- The prior weight assigned to fixed, correlated, independent, fixed-structured and correlated-structured models. For each heterogeneity model, we varied the weight from 0 to 1, keeping the other four models in the same relative proportion.

We quantified the sensitivity to the prior by considering the relative change in BF_{avg} corresponding to a doubling of each prior parameter. Theoretically, a variant showing strong evidence for association under a single mode of inheritance or heterogeneity model, and strong evidence against association under the other component models, would attain a doubling of BF_{avg} for a doubling of the corresponding prior parameter. Some variants in our tier 1 list came close to this value. Ten variants, including the lead markers in the *GRIK1* (rs459370, which showed strong evidence for heterozygote model) and *ATP2B4* (rs4951377, dominant model) had BF_{avg} ratio greater than 1.9 for doubling the weight on the preferred mode of inheritance. Lead markers at all three previously reported loci (*ATP2B4*, *HBB* and *ABO*) had BF_{avg} ratio > 1.7, indicating that placing at least some weight on nonadditive models is helpful in detecting loci such as these. In contrast, the lead marker at *FREM3/GYPE* (black dots in Extended Data Figure 5c) was among a group of SNPs for which the evidence for association remained relatively constant across model specifications. This behaviour reflects the low frequency of rs184895969, particularly in West Africa, which leads to similar levels of evidence across models.

We note that the effect of relative changes in the Bayes factor on posterior probability diminishes as the overall evidence grows. This is because, while the posterior odds is proportional to the Bayes factor, probability as a function of odds becomes flatter as the odds becomes large.

References

1 Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature reviews. Genetics* **11**, 149-160, doi:10.1038/nrg2731 (2010).

2 Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, doi:10.1038/nature11334 (2012).

3 Patnaik, S. K., Helmberg, W. & Blumenfeld, O. O. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic acids research* **40**, D1023-1029, doi:10.1093/nar/gkr958 (2012).

4 Flicek, P. *et al.* Ensembl 2014. *Nucleic acids research* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).

5 Siebert, P. D. & Fukuda, M. Molecular cloning of a human glycophorin B cDNA: nucleotide sequence and genomic relationship to glycophorin A. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 6735-6739 (1987).

6 Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS genetics* **8**, e1002639, doi:10.1371/journal.pgen.1002639 (2012).

7 Su, Z., Cardin, N., the Wellcome Trust Case Control, C., Donnelly, P. & Marchini, J. A Bayesian Method for Detecting and Characterizing Allelic Heterogeneity and Boosting Signals in Genome-Wide Association Studies. 430-450, doi:10.1214/09-sts311 (2009).

8 Leffler, E. M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578-1582, doi:10.1126/science.1234070 (2013).

9 van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369-375, doi:10.1038/nature11677 (2012).

10 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).

11 Dong, Y., Taylor, H. E. & Dimopoulos, G. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. *PLoS biology* **4**, e229, doi:10.1371/journal.pbio.0040229 (2006).

12 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).

13 Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics* **10**, 681-690, doi:10.1038/nrg2615 (2009).