A novel locus of resistance to severe malaria in a region of ancient balancing selection

66 authors, including:

Quang Le
University of Portsmouth
71 PUBLICATIONS **6,270** CITATIONS

Katja Kivinen
University of Cambridge
110 PUBLICATIONS **2,117** CITATIONS

Ellen Miranda Leffler
University of Chicago
21 PUBLICATIONS **792** CITATIONS

Victoria Cornelius
University of Oxford
27 PUBLICATIONS **372** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project Susceptibility to malaria View project

Project Community based rehabilitation and inclusion (CBR & I) View project

A novel locus of resistance to severe malaria in a region of ancient balancing selection

Malaria Genomic Epidemiology Network*

The high prevalence of sickle haemoglobin in Africa shows that malaria has been a major force for human evolutionary selection, but surprisingly few other polymorphisms have been proven to confer resistance to malaria in large epidemiological studies[1–3]. To address this problem, we conducted a multi-centre genome-wide association study (GWAS) of life-threatening *Plasmodium falciparum* infection (severe malaria) in over 11,000 African children, with replication data in a further 14,000 individuals. Here we report a novel malaria resistance locus close to a cluster of genes encoding glycophorins that are receptors for erythrocyte invasion by *P. falciparum*. We identify a haplotype at this locus that provides 33% protection against severe malaria (odds ratio = 0.67, 95% confidence interval = 0.60–0.76, *P* value = 9.5×10^{-11}) and is linked to polymorphisms that have previously been shown to have features of ancient balancing selection, on the basis of haplotype sharing between humans and chimpanzees[4]. Taken together with previous observations on the malaria-protective role of blood group O[1–3,5], these data reveal that two of the strongest GWAS signals for severe malaria lie in or close to genes encoding the glycosylated surface coat of the erythrocyte cell membrane, both within regions of the genome where it appears that evolution has maintained diversity for millions of years. These findings provide new insights into the host–parasite interactions that are critical in determining the outcome of malaria infection.

In the discovery phase of this study, we analysed GWAS data on 5,633 children with severe malaria and 5,919 population controls from The Gambia, Kenya and Malawi, and in the replication phase we analysed candidate single-nucleotide polymorphisms (SNPs) in a further 13,946 case–control samples from Burkina Faso, Cameroon, The Gambia, Ghana, Malawi, Mali and Tanzania (Extended Data Fig. 1). The majority of samples used in the discovery phase have been analysed previously by lower-resolution GWAS methods[3]. For this analysis we improved resolution by directly typing all samples at approximately 2.5 million SNPs using the Illumina Omni2.5M platform, followed by quality control (Extended Data Figs 1 and 2) and imputation of genotypes at over 10 million SNPs using haplotype data from the 1000 Genomes Project[6]. Imputation performance varied across populations, with accuracy varying as a function of the similarity between study and reference individuals (Extended Data Fig. 3). When testing for genetic association, principal components analysis was used to correct for population structure (Extended Data Fig. 4a–e), which reflected both geography and self-reported ethnicity. Similar results were obtained using a mixed-model approach (Extended Data Fig. 4f).

To assess the evidence for association in the discovery phase we used an approach that allows for heterogeneity in the protective effect of an allele across different study sites. This could be particularly important in our data, as high levels of genetic and ethnic diversity in Africa can result in variable patterns of linkage disequilibrium between study sites that can complicate GWAS analysis[7]. Other potential sources of heterogeneity include allelic heterogeneity and multiple independent origins of malaria resistance loci, as has been well documented at the *HBB* locus[1,3], as well as the high levels of genetic diversity in the

parasite[8]. Specifically, we used a Bayesian approach that combines evidence across multiple models of association by specifying a prior probability on the size and similarity of the genetic effect across populations, as well as the mode of inheritance[1]. A single statistical summary of the signal of association was obtained by averaging the evidence across models, weighting each by its prior probability, and comparing the evidence to the null model of no association (model-averaged Bayes factor (BF_{avg})). Having observed the data, a posterior probability was assigned to each model, conditional on it being a true association and the model assumptions, which are described in Methods and Extended Data Fig. 5. We replicated previously reported GWAS signals[2,3,9] at the *HBB* ($BF_{avg} = 5.8 \times 10^{24}$), *ABO* ($BF_{avg} = 6.7 \times 10^{9}$) and *ATP2B4* ($BF_{avg} = 4.4 \times 10^{5}$) loci, and a detailed analysis of key variants at these and other previously reported loci is presented elsewhere[1]. A previously reported association near the gene *MARVELD3* (ref. 2) did not replicate in this data set (Supplementary Note 1). Genome-wide patterns of association with severe malaria at the 34 regions of the genome containing a variant with either a Bayes factor for the most probable model (BF_{max}) > 2.5×10^{4} or with a BF_{avg} > 2.5×10^{3} are summarized in Extended Data Fig. 6 and Supplementary Table 1. Details of the evidence for association in these regions can be viewed online at http://www.malariagen.net/resource/14.

These data provide a rich resource of new candidate loci for further investigation. Here we focus on a region of chromosome 4 shown in Fig. 1, where the strongest signal of association (at SNP rs184895969) is located between the gene *FREM3* and a cluster of three glycophorin genes (*GYPE*, *GYPB* and *GYPA*). Glycophorins are sialoglycoproteins that are abundantly expressed in the erythrocyte membrane, providing a hydrophilic surface coat that is necessary for erythrocytes to flow freely in the circulation. A complex system of single-nucleotide and structural variants in this region determine the MNS blood group system[10]. These genes have a functional role in invasion of erythrocytes by *P. falciparum*. Glycophorin A is the receptor for the *P. falciparum* erythrocyte-binding ligand EBA-175 (ref. 11), and glycophorin B is a receptor for the parasite ligand EBL-1 (ref. 12). To follow up this observation, selected SNPs at this locus were genotyped by Sequenom iPLEX MassArray in the discovery and replication sample sets outlined earlier (Fig. 1 and Extended Data Fig. 7a). The combined data set of 25,498 samples provided convincing evidence of association at rs186873296 by standard fixed-effect meta-analysis ($P = 9.5 \times 10^{-11}$) as well as by the Bayesian approach described earlier ($BF_{overall} = 1.3 \times 10^{8}$; Fig. 2 and Methods). The derived (non-ancestral) allele of rs186873296 was at higher frequency in East Africa than in West Africa, and the greatest evidence of association was seen in Kenya, where the allele was most common with a frequency of approximately 10%. Using only replication data to avoid winner's curse, and assuming an additive model, we estimate that carrying one copy of the derived allele reduces the risk of severe malaria by about 40% in Kenya (odds ratio (OR) = 0.60, 95% confidence interval (CI) = 0.46–0.79), with a slightly smaller effect across all populations (OR = 0.67, 95% CI 0.56–0.80 in frequentist fixed-effect meta-analysis). Further details are given in Supplementary Note 2.

Figure 1 | Signal of association with severe malaria across the *FREM3/GYPE* region. a, Evidence for association ($\log_{10} BF_{avg}$) in the discovery data. Black plus signs denote SNPs that were directly typed, and black triangles denote SNPs selected for typing on the Sequenom platform. Dotted red vertical lines indicate a region of 0.25 cM ± 25 kb centred at the lead SNP (rs184895969). Coloured circles denote the correlation (outer circles) and |D'| (inner circles)

with rs184895969 in controls, computed from imputed haplotypes. **b**, Polymorphisms shared between humans and chimpanzees, and previously reported associations with other phenotypes. NHGRI, National Human Genome Research Institute. **c, d**, Genes in the region and the HapMap combined recombination rate.

The glycophorin gene cluster has a complex pattern of gene conversion and structural variation that has been previously noted; indeed, it has been proposed that selective pressure due to pathogens, including malaria, has contributed to shaping diversity in this region[10,13–16]. Using the human reference sequence and mapped sequence read data from the 1000 Genomes Project, we identified the boundaries of a 350 kb region of sequence homology surrounding these genes as well as a set of segregating gene deletions (Extended Data Fig. 8). The lead imputed marker (rs184895969) is located within 10 kb of this complex region. Imputation accuracy within the region is low using current reference data, so it is possible that the causal variant lies within the glycophorin genes themselves but that this is obscured in the current imputed data set. With this caveat, we computed a credible set of putatively causal variants in the region; this set includes both the lead imputed marker and a linked missense mutation (rs181620317) in *FREM3* (Extended Data Fig. 7b). We note that the protective allele at rs184895969 was associated with increased *GYPA* transcription in published gene expression data for HapMap lymphoblastoid cell lines[17] ($P = 0.016$; Extended Data Fig. 9); other regional analyses are described in

Supplementary Notes 1 and 2. Improved African genome variation reference panels[7,18] are needed to understand the complex patterns of variation in this region so that the causal variant of malaria resistance can be fine-mapped with greater confidence.

A striking feature of these data is that all of the loci that reach conventional criteria for genome-wide significance ($P < 5 \times 10^{-8}$ using a fixed-effects model) are in or near genes that are important for erythrocyte function (*HBB*, *ABO*, *ATP2B4*, *FREM3/GYPE* region), the primary host cell of *P. falciparum*. Other erythrocyte-related genes are identified in the discovery phase analysis but do not reach genome-wide significance, including *EPB41*, which encodes erythrocyte membrane protein band 4.1 and has been implicated as a possible receptor for *P. falciparum* invasion[19] (rs2985337: $BF_{avg} = 3{,}443$; fixed-effect meta-analysis OR = 1.16, 95% CI = 1.09–1.23, $P = 1.2 \times 10^{-6}$; see Extended Data Fig. 6).

The *ABO* locus contains a number of polymorphisms that are shared between humans and other primates, and recent analyses of sequence variation across species indicated that some of these are ancient polymorphisms that have been maintained by balancing

selection over millions of years[20]. The current findings are of particular interest since the *FREM3/GYPE* region is one of the most prominent examples of putative ancient balancing selection in a genome-wide analysis of haplotype sharing between humans and chimpanzees[4]. The peak GWAS signal at this locus is less than 45 kb away from the shared human–chimpanzee haplotypes, which fall within the region covered by the credible set (Extended Data Fig. 7b), although they do not exhibit a strong association with severe malaria (Supplementary Note 3). To explore the genealogical relationship between the putative ancient balanced polymorphisms (ABPs) and SNPs associated with severe malaria in the *FREM3/GYPE* region, we inferred an ancestral tree[21] from the African (YRI + LWK) part of the 1000 Genomes Project data and used it to order haplotypes in the region, labelled with the positions of ABPs, malaria-associated SNPs, and other variants of interest (Fig. 3). All three haplotypes carrying the protective allele at the directly typed marker with most evidence of association (rs186873296) carry the minor allele at the ABP markers ($D' = 1$, $P = 0.017$). By inferring the positions of putative causal mutations on the estimated genealogical tree[21] at the lead imputed marker (rs184895969) we found evidence for a single protective mutation in Kenya ($\log_{10} \mathrm{BF} = 3.09$; OR = 0.6) estimated to lie on the branch ancestral to the protective allele at the lead marker. Although the most likely position for an additional mutation was on the branch ancestral to the ABPs, a single haplotype explains most of the signal of association in this region.

These observations raise the question of whether malaria resistance loci are more likely to be found in regions of the genome that show evidence of ancient balancing selection. We therefore analysed the relationship between the regions of association in our GWAS and 125 regions of the genome found previously[4] to contain haplotypes shared between humans and chimpanzees. The SNPs defining these haplotypes (ABPs) were not themselves enriched for association with severe malaria ($P > 0.1$, Methods). We used a simulation approach to assess the physical proximity of ABPs to the peak of association within the 34 strongest regions of association (tier 1) and 73 weaker signals (tier 2), and observed a significant relationship with tier 1 over a range of length scales (Extended Data Fig. 10a, d). We also identified the nearest gene to the lead marker within each association region as well as the gene nearest to each ABP haplotype, and performed a gene-based test for genome-wide enrichment. Strong evidence for enrichment was seen at tier 1 loci (OR 41.0; $P = 4 \times 10^{-7}$ by Fisher's exact test, $P_{\mathrm{sim}} = 5 \times 10^{-4}$ using a simulation approach described in Methods, and $P = 1 \times 10^{-4}$ using the INRICH algorithm[22]; Extended Data Fig. 10b–d) and a weaker trend at tier 2 loci (OR 7.7, $P_{\mathrm{sim}} = 0.15$). Apart from *ABO* and *FREM3/GYPE*, there were six other GWAS loci (four in tier 1, two in tier 2) where the nearest gene to the lead marker was also the nearest gene to an ABP haplotype (*DSCAM*, *NRG1*, *CNTNAP5*, *TMEM132C*, *CACNA2D1*, *RYR2*). Although the current association evidence at these loci does not satisfy conventional criteria for genome-wide significance and they should be regarded as putative until convincingly replicated, it is noteworthy that they are all involved in key aspects of membrane biology (Supplementary Note 3).

In the largest genetic association study of malaria to date, we have identified a new locus of resistance to severe malaria that lies next to a cluster of glycophorin genes involved in erythrocyte invasion by *P. falciparum*, and that also overlaps a locus of putative ancient balancing selection identified by analysis of haplotype sharing between humans and chimpanzees. It is possible that malaria is not the cause of the ancient balancing selection, or that it is just one of a number of opposing evolutionary driving forces, as at *ABO*, where blood group O reduces the risk of severe malaria but increases the risk of severe cholera[23]. Nonetheless, these new findings raise the intriguing question of whether natural selection on malaria susceptibility has been shaping genetic diversity in humans and their ancestors for millions of years. *P. falciparum* is closely related to the chimpanzee parasite *P. reichenowi* and other parasites of African great apes[24–26].
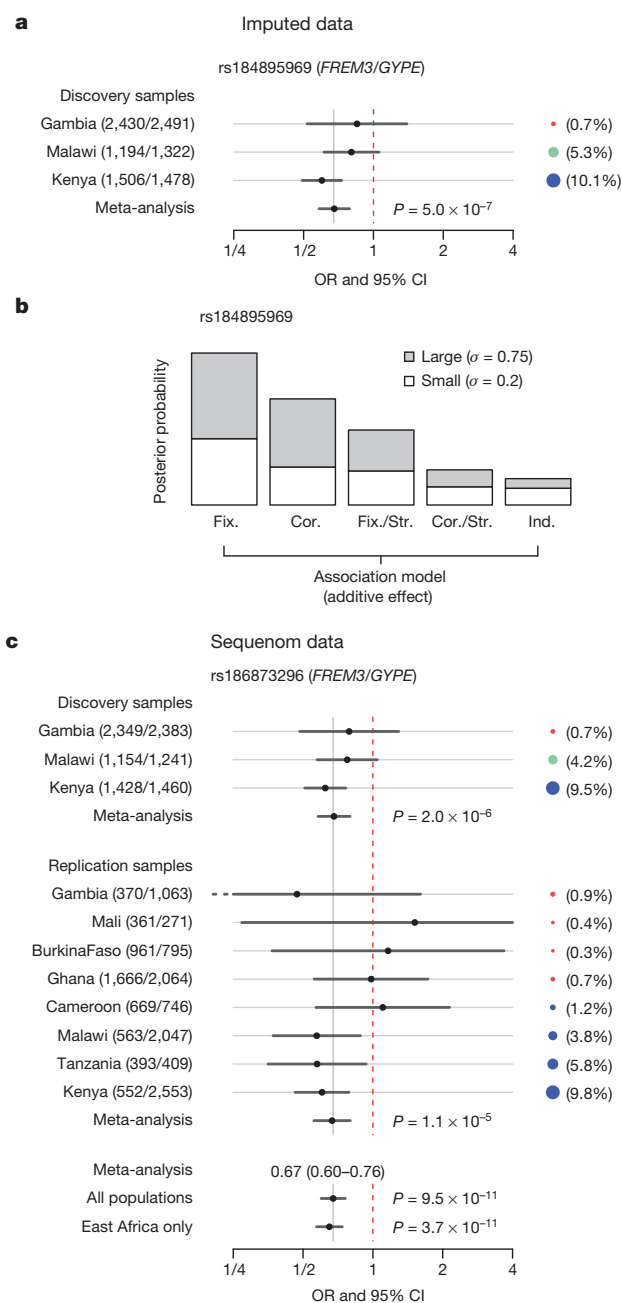

Figure 2 | Evidence for association at SNPs in the *FREM3/GYPE* region, assuming an additive model of association. a, Forest plot showing sample size, estimated OR and 95% CI for the lead imputed SNP in each population and under fixed-effect meta-analysis. The frequency of the protective allele in controls in each population is shown to the right. **b**, The posterior weight on different models of heterogeneity at rs184895969 under the prior used in the GWAS. Cor., correlated effects model; Cor./Str., correlated-structured-effect model; Fix., fixed effects model; Fix./Str., fixed-structured-effect model; Ind., independent effects model. Models are described in Methods. **c**, Forest plot for the Sequenom-typed SNP rs186873296 in discovery and replication samples, with fixed-effect meta-analysis across all populations and across East African populations (here taken as Kenya, Malawi, Tanzania and Cameroon.)

It has been proposed that *P. falciparum* was introduced into human populations from chimpanzees or gorillas in the recent past, but this remains a matter of intense debate[25–27]. Population genetic data are consistent with an ancient origin followed by a marked expansion of the parasite population approximately 10,000 years ago, coincident with the introduction of agriculture[28]. The *P. falciparum* genome
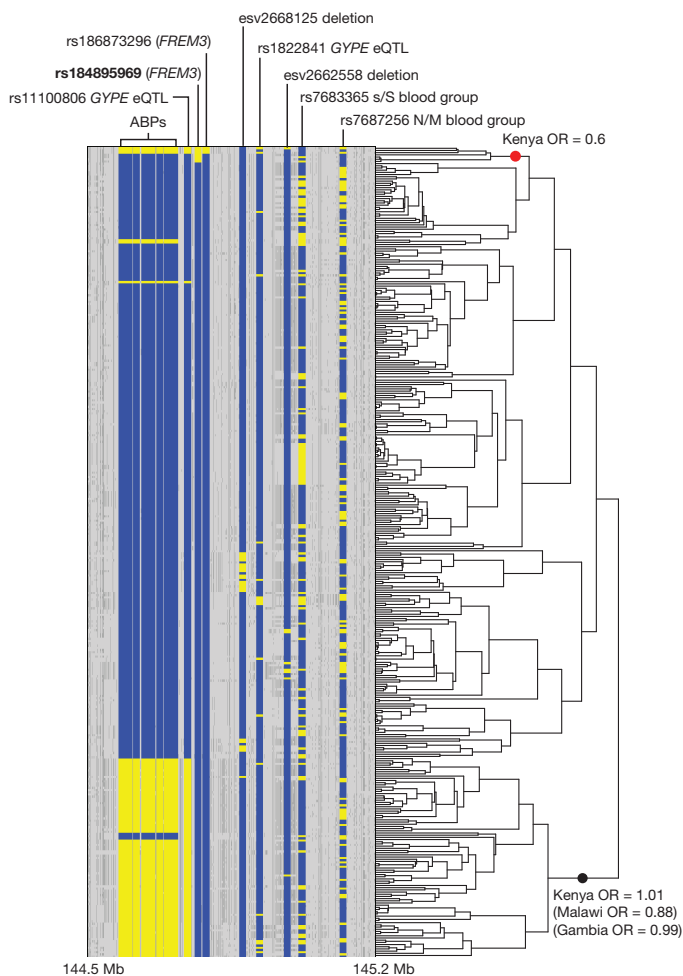
Figure 3 | Haplotype analysis across the *FREM3/GYPE* region. Left, haplotypes at 7,321 polymorphic SNPs between 144.5 Mb and 145.2 Mb on chromosome 4 in the LWK and YRI samples of the 1000 Genomes Project reference panel. Key variants (Methods and Supplementary Note 2) are enlarged for clarity and labelled, with reference and non-reference alleles coloured blue and yellow, respectively. Right, the estimated topology of the genealogical tree at rs184895969. Dots indicate the position of the inferred protective mutation in Kenya and the branch ancestral to the ABPs, and are labelled with the estimated OR.

possesses a huge repertoire of polymorphism[8], and it is possible that the host and parasite genomes are engaged in a longstanding evolutionary arms race, each maintaining diversity to try to outflank the other[29]. Intriguingly, the parasite surface receptor EBA-175, which directly binds glycophorin A during erythrocyte invasion, also contains structural polymorphisms with features of ancient dimorphism[30]. The present findings provide new leads both to investigate these evolutionary mechanisms and to discover further genetic determinants of resistance to severe malaria in African children.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genet.* **46,** 1197–1204 (2014).
2. Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* **489,** 443–446 (2012).
3. Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.* **9,** e1003509 (2013).
4. Leffler, E. M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339,** 1578–1582 (2013).
5. 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).
6. The 1000 Genomes Project Consortium.. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
7. Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature Rev. Genet.* **11,** 149–160 (2010).
8. Manske, M. *et al.* Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487,** 375–379 (2012).
9. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genet.* **41,** 657–665 (2009).
10. Blumenfeld, O. O. & Huang, C. H. Molecular genetics of the glycophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification. *Hum. Mutat.* **6,** 199–209 (1995).
11. Sim, B. K., Chitnis, C. E., Wasniowska, K., Hadley, T. J. & Miller, L. H. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science* **264,** 1941–1944 (1994).
12. Mayer, D. C. *et al.* Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proc. Natl Acad. Sci. USA* **106,** 5348–5352 (2009).
13. Baum, J., Ward, R. H. & Conway, D. J. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.* **19,** 223–229 (2002).
14. Ko, W. Y. *et al.* Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *Am. J. Hum. Genet.* **88,** 741–754 (2011).
15. Tarazona-Santos, E. *et al.* Population genetics of GYPB and association study between GYPB*S/s polymorphism and susceptibility to *P. falciparum* infection in the Brazilian Amazon. *PLoS ONE* **6,** e16123 (2011).
16. Wang, H. Y., Tang, H., Shen, C. K. & Wu, C. I. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. *Mol. Biol. Evol.* **20,** 1795–1804 (2003).
17. Stranger, B. E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8,** e1002639 (2012).
18. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517,** 327–332 (2015).
19. Lanzillotti, R. & Coetzer, T. L. The 10 kDa domain of human erythrocyte protein 4.1 binds the *Plasmodium falciparum* EBA-181 protein. *Malar. J.* **5,** 100 (2006).
20. Ségurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109,** 18493–18498 (2012).
21. Su, Z., Cardin, N., The Wellcome Trust Case Control Consortium, Donnelly P. & Marchini, J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.* **24,** 430–450 (2009).
22. Lee, P. H., O'Dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28,** 1797–1799 (2012).
23. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nature Rev. Genet.* **15,** 379–393 (2014).
24. Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature Commun.* **5,** 4754 (2014).
25. Liu, W. *et al.* Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467,** 420–425 (2010).
26. Prugnolle, F. *et al.* A fresh look at the origin of *Plasmodium falciparum*, the most malignant malaria agent. *PLoS Pathog.* **7,** e1001283 (2011).
27. Rich, S. M. *et al.* The origin of malignant malaria. *Proc. Natl Acad. Sci. USA* **106,** 14902–14907 (2009).
28. Joy, D. A. *et al.* Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300,** 318–321 (2003).
29. Gilbert, S. C. *et al.* Association of malaria parasite population structure, HLA, and immunological antagonism. *Science* **279,** 1173–1177 (1998).
30. Binks, R. H. *et al.* Population genetic analysis of the *Plasmodium falciparum* erythrocyte binding antigen-175 (*eba-175*) gene. *Mol. Biochem. Parasitol.* **114,** 63–70 (2001).

Malaria Genomic Epidemiology Network

Writing group Gavin Band[1], Kirk A. Rockett[1,2], Chris C. A. Spencer[1], Dominic P. Kwiatkowski[1,2]; **Data analysis** Gavin Band[1], Quang Si Le[1], Geraldine M. Clarke[1], Katja Kivinen[2], Ellen M. Leffler[1], Kirk A. Rockett[1,2], Dominic P. Kwiatkowski[1,2], Chris C. A. Spencer[1]; **Project management** Kirk A. Rockett[1,2], Chris C. A. Spencer[1], Victoria Cornelius[1], David J. Conway[3,4], Thomas N. Williams[5,6], Terrie Taylor[7,8], Dominic P. Kwiatkowski[1,2]; **Study site lead investigators** David J. Conway[3,4], Kalifa A. Bojang[3], Ogobara Doumbo[9], Mahamadou A. Thera[9], David Modiano[10], Sodiomon B. Sirima[11], Michael D. Wilson[12], Kwadwo A. Koram[12], Tsiri Agbenyega[13,14], Eric Achidi[15], Thomas N. Williams[5,6], Kevin Marsh[5], Hugh Reyburn[16,17], Chris Drakeley[16,17], Eleanor Riley[17], Terrie Taylor[7,8], Malcolm Molyneux[18]; **Clinical data and sample collection** Muminatou Jallow[3,19], Kalifa A. Bojang[3], David J. Conway[3,4], Margaret Pinder[3], Ogobara Doumbo[9], Mahamadou A. Thera[9], Ousmane B. Toure[9], Salimata Konate[9], Sibiri Sissoko[9], Edith C. Bougouma[11], Valentina D. Mangano[10], David Modiano[10], Sodiomon B. Sirima[11], Lucas N. Amenga-Etego[20], Anita K. Ghansah[12], Abraham V. O. Hodgson[20], Kwadwo A. Koram[12], Michael D. Wilson[12], Tsiri Agbenyega[13,14], Daniel Ansong[13,14], Anthony Enimil[13], Jennifer Evans[21,22], Eric Achidi[15], Tobias O. Apinjoh[23], Alexander Macharia[5], Kevin Marsh[5], Carolyne M. Ndila[5], Charles Newton[5], Norbert Peshu[5], Sophie Uyoga[5], Thomas N. Williams[5,6], Chris Drakeley[16,17], Alphaxard Manjurano[16,17], Hugh Reyburn[16,17], Eleanor Riley[17], David Kachala[18], Malcolm Molyneux[18], Vysaul Nyirongo[18], Terrie Taylor[7,8]; **Sample processing, genotyping, data management and project coordination** Kirk A. Rockett[1,2], Katja Kivinen[2], Daniel Mead[2], Eleanor Drury[2], Sarah Auburn[1], Susana G. Campino[2], Bronwyn MacInnis[2], Jim Stalker[2], Emma Gray[2], Christina Hubbart[1], Anna E. Jeffreys[1], Kate Rowlands[1], Alieu Mendy[1], Rachel Craik[1], Kathryn Fitzpatrick[1], Sile Molloy[1], Lee Hart[1], Robert Hutton[1], Angeliki Kerasidou[1,24], Kimberly J. Johnson[1], Victoria Cornelius[1]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [3]Medical Research Council Unit, Atlantic Boulevard, Fajara, PO Box 273, The Gambia. [4]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [5]KEMRI-Wellcome Trust Research Programme CGMRC, PO Box 230-80108, Kilifi, Kenya. [6]Faculty of Medicine, Department of Medicine, Imperial College, Exhibition Road, London SW7 2AZ, UK. [7]Blantyre Malaria Project, Queen Elizabeth Central Hospital, College of Medicine, P.O. Box 30096, Chichiri, Blantyre 3, Malawi. [8]College of Osteopathic Medicine, Michigan State University, East Lansing, Michigan 48824, USA [9]Malaria Research and Training Centre, Faculty of Medicine University of Bamako, Bamako, Mali. [10]University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy. [11]Centre National de Recherche et de Formation sur le Paludisme (CNRFP), 01 BP 2208 Ouagadougou 01, Burkina Faso. [12]Noguchi Memorial Institute for Medical Research, University of Ghana, P.O. Box LG 25, Legon, Accra, Ghana. [13]Komfo Anokye Teaching Hospital, PO Box 1934, Kumasi, Ghana [14]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [15]Department of Medical Laboratory Sciences, University of Buea, P.O. Box 63, Buea, South West Region, Cameroon. [16]Joint Malaria Programme, Kilimanjaro Christian Medical Centre, PO box 2228, Moshi, Tanzania. [17]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [18]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Queen Elizabeth Central Hospital, College of Medicine, P.O. Box 30096, Chichiri, Blantyre 3, Malawi. [19]Royal Victoria Teaching Hospital, Independence Drive, PO Box 1515, Banjul, The Gambia. [20]Navrongo Health Research Centre, PO Box 114, Navrongo, Ghana. [21]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Postfach 30 41 2, D-20324 Hamburg, Germany. [22]Kumasi Centre for Collaborative Research, School of Medical Sciences, KNUST, Kumasi, Ghana. [23]Department of Biochemistry & Molecular Biology, University of Buea, P.O. Box 63 Buea, South West Region, Cameroon. [24]The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

METHODS

Samples, ethics and clinical information. Samples were collected from nine partner projects from across sub-Saharan Africa (Extended Data Fig. 1a) as described previously[1,3]. The studies and sample sets described in this manuscript form part of a larger ongoing project within the Malaria Genomic Epidemiology Network (MalariaGEN; http://www.malariagen.net). We used the World Health Organization definition of severe malaria, which comprises a broad spectrum of life-threatening clinical complications of *Plasmodium falciparum* infection[29,31]. Investigators from study sites worked together to agree on principles for sharing data and standardizing clinical definitions, and to define best ethical practices across different local settings including the development of guidelines for informed consent. Relevant ethics committees are listed in Extended Data Fig. 1a. Further information on policies, research and the consent process may be found on the MalariaGEN website (http://www.malariagen.net/community/ethics-governance). No statistical methods were used to predetermine sample size.

DNA extraction and Sequenom typing. As described previously[1,3], all samples submitted to the MalariaGEN Resource Centre underwent a standard set of procedures that included quantification using picogreen, genotyping of 65 polymorphisms (including the polymorphism rs334 causing sickle haemoglobin, and three gender-typing SNPs) on the Sequenom iPLEX MassArray platform and matching to baseline clinical data (for example, gender, ethnic group and case–control status).

High-density genotyping. Three cohorts (Gambia, Malawi and Kenya) were genotyped on the Illumina HumanOmni2.5-4 (Kenya) and Illumina HumanOmni2.5-8 (Gambia, Malawi) platforms. As described previously[3] we used three different calling algorithms (Illuminus[32], Illumina's Gencall algorithm as provided in BeadStudio, and GenoSNP[33]), each of which uses slightly different information in the data. We formed final genotype calls by taking consensus between the three algorithms. Genotypes where any two of the three calling algorithms were discordant, and genotypes where fewer than two algorithms were confident enough to make a call were treated as missing. This process showed improved calling, evaluated using Mendelian error counts in a subset of Kenyan samples, relative to each of the three algorithms separately (data not shown).

After genotype calling, we aligned genotypes to the forward strand of the human reference sequence (GRCh37), using both the Illumina-supplied manifest and publicly available strand files (http://www.well.ox.ac.uk/~wrayner/strand) obtained by mapping allele probes to the reference by BLAT[34]. We removed SNPs whose position or strand mismatched between the Illumina manifest and the strand file. To simplify analyses, we restricted attention to the set of SNPs having the same name, chromosome, position, strand, and probe sequences across the two genotyping platforms. Because the Omni2.5M contains multiple probes for some variants, we further removed SNPs to ensure positions were unique. In total we were left with 2,322,985 SNPs in each cohort across the autosomes and the X and Y chromosomes. We note that SNPs annotated as lying on the pseudo-autosomal region (PAR) of the X chromosome were not included, as these had position equal to 0 in the manifest for the HumanOmni2.5-4 array. Finally, we flipped alleles where necessary so that in downstream analyses the first allele always corresponds to the reference allele of the human genome sequence.

Sample quality control. We performed sample quality control separately on each cohort by computing autosome-wide averages of normalized X channel intensity, normalized Y channel intensity, and heterozygosity based on the consensus call. To identify outlying samples we applied ABERRANT[35], adjusting the λ parameter per cohort to account for differences in genotyping quality (Extended Data Fig. 1d–f). In Gambia, a tail of samples showing low heterozygosity but otherwise appearing to be well typed was apparent. We explicitly included these samples in downstream analyses.

To estimate genome-wide relatedness between samples, we selected a list of 178,775 high-quality SNPs satisfying the criteria missingness <1%, minor allele frequency (MAF) > 1% and thinned to be at least 0.005 cM apart and to have pairwise $r^2 < 0.3$. Treating each cohort separately, we used SHELLFISH (http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php) to compute a matrix of pairwise relatedness values $R = (r_{ij})$ between samples, where r_{ij} denotes the genome-wide average covariance of frequency-normalized genotypes in individuals i and j. In samples with few close relationships, the value of r_{ij} can be thought of as an estimate of kinship[36] with, for example, values close to 1 representing identity between samples, and values close to zero reflecting a lack of close relatedness (relative to the rest of the sample). To remove samples with duplicate typing and close relationships, we excluded one of each pair of samples with $r_{ij} > 0.2$, taking all remaining samples through to phasing and imputation. Extended Data Figure 1c lists the number of samples before and after quality control, and the number removed by each quality control step.

We used SHELLFISH to compute principal components on the post-quality-control sample set. Consistent with our removal of poor quality, duplicate and closely related samples, principal components plots show no substantially outlying samples (see Extended Data Fig. 4a–c). The top few principal components in each cohort reflect substantial population structure, as evidenced by colouring by reported ethnicity. As found previously[3,9], the top principal components are also significantly correlated with case/control status (Extended Data Fig. 4d), indicating that population structure may act as a confounding factor in association analyses if not controlled.

Genotyping quality control. Treating each cohort separately, we used SNPTEST (http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) to test for association at each autosomal SNP using a genotypic model of association that allows for different effects at heterozygote and homozygote genotypes, including the leading five principal components as covariates. Genotypic model association tests are particularly sensitive to confounding by genotyping error[37], and we used the resulting P values as a guide to finding appropriate quality control criteria. To detect potential spurious genotypes due to batch effects, we modelled genotypes as predicted by the leading five principal components and case/control status in a linear regression framework in R[38], and tested whether including an indicator of the plate on which each sample was genotyped contributed significantly to model fit. We refer to this as the 'plate test'. For downstream analyses we excluded SNPs with MAF < 1%, missing data proportion >5% (in Gambia and Malawi) or >2.5% (in Kenya, which had fewer missing genotypes overall), Hardy–Weinberg $P < 1 \times 10^{-20}$ in controls, or plate-test $P < 0.01$. We inspected cluster plots of all remaining SNPs showing association test $P < 1 \times 10^{-5}$ (Gambia) or $P < 1 \times 10^{-4}$ (Kenya, Malawi) and excluded those with clear genotyping problems. Extended Data Figure 2a shows the number of SNPs excluded by each criterion. The post-exclusion genome-wide association analysis is shown in Extended Data Fig. 2c.

SNP quality control on the X chromosome was performed as on the autosomes, with a few differences as follows. We treated males and females separately, using a genotypic model of association in females and an allelic model in males (who have only one copy of the X chromosome). Because genotype calling was performed blind to the gender of samples, some males appear to be heterozygous at some X chromosome SNPs. We treated all such heterozygous calls as missing. We computed missing data proportion and plate-test P values in males and females separately, and tested for departure from Hardy–Weinberg in female controls and for differences in frequency between males and females. We excluded SNPs with MAF < 1%, missingness >5% (Gambia, Malawi) or 2.5% (Kenya) in males or females, or plate-test $P < 0.01$ in males or females. We further excluded SNPs with Hardy–Weinberg $P < 1 \times 10^{-20}$ in female controls or showing significant difference in allele frequencies between males and females ($P < 1 \times 10^{-20}$). Extended Data Figure 2b shows the number of SNPs excluded by each criterion.

Some regions of the X chromosome showed an elevated number of male heterozygote calls, contributing to the high number of SNPs excluded due to missingness in males. These included SNPs in the pseudo-autosomal regions at either end of the chromosome[39] (indicating that XY designation in the chip manifests does not adequately cover these regions) as well as SNPs within the X transposed region near the centromere.

Phasing and imputation. We phased genotype data in each cohort separately using SHAPEIT v2.r644 (ref. 40), specifying 200 hidden states and an effective population size of 17,469, as recommended for African populations by the SHAPEIT documentation, and phasing each chromosome separately. We used IMPUTE v.2.3.0 (refs 41, 42) to impute phased genotypes into the 1000 Genomes[6] Phase I integrated variant set (version of 24th August 2012, as downloaded from the IMPUTE website, which we refer to here as the 1000 Genomes reference panel) in 5Mb chunks with a buffer region of 500kb. For phasing and imputation we used the combined HapMap recombination map in build 37 coordinates included with the 1000 Genomes reference panel. Unless otherwise stated, downstream analyses included only SNPs with MAF > 0.5% and IMPUTE info measure >0.75.

Assessment of imputation. At each genotyped SNP, IMPUTE computes squared correlation (referred to here as accuracy) and concordance between typed genotypes and genotypes obtained by masking the SNP and re-imputing. To assess imputation performance, we plotted the cumulative distribution of concordance (Extended Data Fig. 3a) and accuracy (Extended Data Fig. 3b), accuracy by frequency (Extended Data Fig. 3d) as well as average per-sample accuracy (Extended Data Fig. 3c). We also assessed accuracy relative to direct typing on the Sequenom platform at variants typed in the *FREM3/GYPE*, *ARL14* and *INPP4B* regions of association as described later.

Association testing. We used SNPTEST to test for association at approximately 38 million variants obtained through imputation, including five principal components as covariates to control for population structure separately in each cohort. SNPTEST uses a missing data likelihood to account for the uncertainty in genotypes at imputed SNPs. We fit additive, dominant, recessive, and heterozygote models of association and ran SNPTEST separately in each imputation chunk.

Below, we refer to the estimated effect size for population i and mode of inheritance m as $\beta_{i,m}$ and its estimated standard error as $SE_{i,m}$.

Frequentist meta-analysis. For each SNP and each mode of inheritance (additive, dominant, recessive, heterozygote) we computed the fixed-effect inverse variance-weighted meta-analysis effect size, standard error, and P value. In this context, fixed-effect meta-analysis assumes a single true effect size that is identical between the three cohorts, and can be thought of as finding the maximum likelihood estimate of the effect size under the assumption that the likelihood in cohort i is proportional to the density of the normal distribution with mean $\beta_{i,m}$ and standard error $SE_{i,m}$.

Bayesian meta-analysis. We have previously[1,3] used Bayesian meta-analysis techniques to allow for between-population heterogeneity of effect sizes in these three cohorts. Here, we applied this method to compute Bayes factors for association under four modes of inheritance and six different models of correlation between cohorts. In this framework, true effect sizes are modelled by a multivariate normal distribution centred on zero and with a given prior covariance matrix that can be written as $\sigma^2 P$, where σ^2 is a prior variance controlling the magnitude of plausible effects and P is a prior correlation matrix.

We used the following correlation models.

Fixed effects (all elements of P equal to 1): as with frequentist fixed-effect meta-analysis, this assumes effect sizes are equal across the three cohorts.

Correlated effects (all off-diagonal elements of P equal to 0.96): this assumes effect sizes are similar but allows for some variability.

Independent effects (all off-diagonal elements of P equal to 0.1): this assumes there is little similarity between effect sizes in different cohorts.

Structured effects: we also considered models where effects in the two East African populations (Kenya and Malawi) were more similar to each other than to that in The Gambia. We assumed either effects were fixed between Kenya and Malawi and correlated ($\rho = 0.96$) between East Africa and The Gambia (we refer to this as the fixed-structured-effect model), or that effects were correlated ($\rho = 0.96$) between Kenya and Malawi and largely independent with The Gambia ($\rho = 0.1$) (referred to as the correlated-structured-effect model).

We used prior variance parameters of $\sigma^2 = 0.2^2$, reflecting a belief in relatively small effects (OR <1.5 with 95% probability)[37], and $\sigma^2 = 0.75^2$, reflecting a belief in larger effects (OR <4.5 with 95% probability).

To form a single summary measure of evidence for association we formed a model-averaged Bayes factor (BF), referred to as the mean BF and denoted BF_{avg}, as a weighted average of model-specific Bayes factors using the following weighting scheme. We assigned weights of 0.4 for additive mode of inheritance and 0.2 for dominant, recessive, or heterozygote modes of inheritance, reflecting a belief that variants which tag causal variants by linkage disequilibrium are more likely to display additive effects. We assigned a weight of 0.4, 0.2, and 0.1 to fixed-, correlated-, and independent-effect models, respectively, and 0.2 and 0.1 to fixed-structured and correlated-structured effects models. Finally we assigned weight of 0.5 to small-effect models ($\sigma^2 = 0.2^2$ and large-effect models ($\sigma^2 = 0.75^2$). Overall prior weights were assigned by multiplying across these categories; for example, the model representing small effect size distribution, fixed-effect across cohorts with additive mode of inheritance was assigned prior weight equal to $0.5 \times 0.4 \times 0.4 = 0.08$, while the model representing small, correlated-structured, dominant effects was assigned a prior weight of $0.5 \times 0.1 \times 0.2 = 0.01$. For each SNP we also recorded the model having the highest posterior weight, and refer to its Bayes factor as the maximum BF (denoted BF_{max}). Extended Data Figure 5a depicts slices through the combined prior on effect sizes across three cohorts for additive effect models. The mean BF behaves similarly to a minimum over all four fixed-effect meta-analysis P values, but additionally captures effects that vary between cohorts (Extended Data Fig. 5b).

As described previously[1,3], to compute model-specific Bayes factors efficiently we used an approximation of the likelihood by the density of a normal distribution with the estimated mean ($\beta_{i,m}$) and standard error ($SE_{i,m}$) in each cohort. Thus, overall, observed effect sizes are modelled as normally distributed around zero with covariance that depends on the prior covariance in true effect sizes and on model standard errors

$$\left(\beta_{i,m} \right) \sim N\left(0, \sigma^2 P + V \right)$$

where V is a diagonal matrix with ith diagonal entry equal to the squared standard error, $SE_{i,m}^2$. The approximate or asymptotic Bayes factor can then be computed by evaluating a ratio of two normal densities. In the univariate case this method is the same as described previously[43]. To facilitate working with genome-wide meta-analysis results, we wrote custom software to compute frequentist and Bayesian meta-analyses, and stored details of genotype counts, association model fit, and meta-analysis results directly in a SQLite database file.

Further discussion of the Bayesian approach can be found in Supplementary Note 4.

X chromosome association testing and meta-analysis. Association testing on the X chromosome was performed as for the autosomes with a few differences as follows. We ran SNPTEST separately in males and females, estimating effect sizes under additive, dominant, recessive and heterozygote modes of inheritance in females. In this usage, SNPTEST assumes a model of complete inactivation so encodes male genotypes as 0/1 and females as 0/0.5/1 for an additive mode of inheritance. We then meta-analysed the six gender-specific association analyses to produce frequentist fixed-effect and model-averaged Bayesian meta-analyses.

For Bayesian analysis, in addition to summing over models of between-population heterogeneity, we adopted the view that differences in sex might lead to heterogeneity in effect. We therefore included models of heterogeneity between males and females as follows. Let ρ_{sex} denote the correlation between effects in male and female samples within a single population. We included models where males and females have the same effect ($\rho_{sex} = 1$, termed fixed-sex model and given prior weight 0.45), correlated effects ($\rho_{sex} = 0.96$, termed correlated-sex model and given prior weight 0.225) or essentially independent effects ($\rho_{sex} = 0.1$, termed independent-sex model and given prior weight 0.225). Because some parts of the X chromosome escape inactivation[39,44], we also included a model where the effect in females is twice that in males (given prior weight 0.1).

To fully specify prior correlation for each model, for each pair of populations A, B we also need to specify the correlation (denoted ρ_{cross}) between males in A and females in B. We set $\rho_{cross} = \rho_{pop} \times \rho_{sex}$ where ρ_{pop} is the chosen prior correlation between same-sex samples in A and B (that is, $\rho_{pop} = 1$, 0.96 or 0.1 as defined earlier).

As earlier, we formed overall weights by multiplying across categories, so that, for example, the model of small, additive effects that are fixed across populations and across sexes had prior weight $0.5 \times 0.4 \times 0.4 \times 0.45 = 0.036$.

Linear mixed model analysis. To compare association test results for logistic regression as implemented in SNPTEST with the use of a linear mixed model, we reran association test scans in each discovery cohort using the program MMM[45]. We used the same genome-wide relatedness matrix as used to compute principal components (see above) and assumed an additive model of association. We plotted the $-\log_{10}(P$ value) for MMM against the corresponding $-\log_{10}(P$ value) based on the SNPTEST scan (which used five principal components as described above), for each discovery population and for fixed-effect meta-analysis (Extended Data Fig. 4f). P values under both methods at tier 1 loci are listed in Supplementary Table 1.

Lead SNPs, region and tier definitions. We formed a list of lead SNPs within approximately independent regions of interest as follows. We restricted to variants with IMPUTE info measure at least 0.75 across all three populations and ranked variants by the model-averaged Bayes factor (highest to lowest). We iteratively picked lead SNPs from the top of the list and excluded other variants within a recombination interval of 0.25 cM ± 25 kb centred at the lead SNP (referred to later as the association region), continuing until no more SNPs remained with $BF_{avg} > 250$ or $BF_{max} > 2,500$.

We grouped lead SNPs into two tiers as follows: tier 1 containing all lead SNPs with $BF_{avg} > 2,500$ or $BF_{max} > 25,000$, and tier 2 containing all lead SNPs not in tier 1 with $BF_{avg} > 1,000$ or $BF_{max} > 10,000$.

In total, across the autosomes and the X chromosome there were 34 regions in tier 1 and 73 in tier 2, with association regions covering approximately 13 Mb and 26 Mb of the genome, respectively.

Regional association analysis. For each tier 1 and 2 region, we examined the pattern of association in the region, generating a regional association plot for the region annotated with details of the meta-analysis for the lead SNP as follows.

In each region we computed pairwise linkage disequilibrium statistics between the lead SNP and surrounding SNPs using best-guess imputed haplotypes for control samples across three populations. To facilitate this computation, we stored imputed haplotypes in a SQLite-format database allowing us effective random access to haplotype data. We used R to compute both Pearson correlation coefficient (r^2) and Lewontin's $|D'|$ between the lead SNP and all other SNPs in the region.

For each SNP in tier 1 and 2 we plotted $\log_{10}(BF_{avg})$ in the association region around the lead SNP plus 1 Mb on either side, colouring points according to linkage disequilibrium, with outer circles representing r^2 and inner circles representing $|D'|$. We further annotated SNPs that were typed in at least one cohort (using black plus signs) and SNPs that had Sequenom genotype data available (with black triangles).

We plotted all ABPs (supplementary tables 4 and 5 from ref. 4), expression quantitative trait loci (eQTLs) from the Genotype-Tissue Expression project[46] (GTEx), and previously reported[47] GWAS loci. Where SNPs matched an imputed or typed variant, we computed linkage disequilibrium statistics and coloured these points as described earlier. We plotted all RefSeq genes (as downloaded from the UCSC Genome Browser MySQL database[48] on 18 March 2013) in the region,

annotating the direction of transcription. We further plotted local recombination rate estimates from the HapMap combined recombination map included with the IMPUTE haplotypes described earlier.

We annotated plots with effect sizes and confidence intervals for the lead SNP for each mode of inheritance (additive, dominant, recessive and heterozygote) that informed the model averaging. We also produced barplots showing the posterior distribution on models of association using the prior weights described earlier. Finally, for each region, we produced and inspected cluster plots for all typed SNPs with $BF_{avg} \geq 10$ that were not excluded by quality control. As phasing fills in missing genotypes and potentially improves genotype calling based on linkage disequilibrium with surrounding SNPs, we coloured plots based on genotype calls taken from the output of phasing.

Website. Regional association plots and cluster plots can be viewed online at http://www.malariagen.net/resource/14.

Validation and replication typing. On the basis of a preliminary version of the data presented here we selected SNPs for typing on the Sequenom platform across the whole MalariaGEN Consortium Project 1 sample set, which includes the discovery samples, further cases and controls in the same populations that were not included in the GWAS, and further large sample sets from five other populations from sub-Saharan Africa (see Extended Data Fig. 1a). Data were available for SNPs tagging the lead markers in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions ($r^2 > 0.5$ in controls, as estimated using the EM algorithm) as well as other regions not represented in tier 1.

Replication analysis. Replication analysis using Sequenom data was restricted to the set of samples with less than 10% missingness as measured across the set of 70 SNPs chosen for replication typing. In each population we conducted logistic regression in R, including five principal components (discovery samples) or reported ethnicity (replication samples) as covariates to control for population structure. For each GWAS lead SNP, we examined each Sequenom SNP in the region and computed r^2 and $|D'|$ with the lead SNP. To allow for the effects of incomplete linkage disequilibrium on the Bayesian model fitting, we recomputed the Bayesian analysis in the discovery samples based on Sequenom genotypes at each SNP, to obtain a Sequenom-based mean Bayes factor and a 'best model' Bayes factor at the model with highest posterior weight. Where linkage disequilibrium is incomplete or where imputation is imperfect, this model may differ from the best model for imputed data.

For each SNP we computed fixed-effect meta-analysis across discovery samples, across replication samples, and across all samples. We also computed the Bayes factor for replication samples ($BF_{replication}$) for the model with highest posterior weight in the discovery samples. To compute an overall Bayes factor for association, we combined the discovery BF_{avg} computed at the imputed lead marker with the replication Bayes factor at the Sequenom SNP, as $BF_{overall} = BF_{avg} \times BF_{replication}$. We use the imputed lead marker here because the number of discovery samples directly typed was smaller than the number of imputed samples. Conditional on the lead marker and replication markers reflecting the same signal of association, this $BF_{overall}$ represents an overall measure of the evidence for association at the locus that reflects all the samples in our study.

A discussion of the replication evidence in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions can be found in Supplementary Note 1.

MalariaGEN encourages individual study sites to perform more detailed analyses of local patterns of disease association, and a Tanzanian-focused analysis of *FREM3* and other candidate SNPs that were genotyped as part of this study is reported elsewhere[49].

Credible interval analysis. In a given region of the genome, under the assumption that exactly one variant (that is accessible to our typing or imputation) is causal, the posterior probability that each variant is the causal variant can be computed by a simple reweighting of Bayes factors[50]. Using this approach we computed 95% and 99% credible intervals (that is, the smallest set of SNPs accounting for 95% or 99% of the posterior mass) for variants in the association region around rs184895969, plus a margin of 50 kb at either end (Extended Data Fig. 7b). We noted that rs181620317, which is annotated as a missense mutation for the gene *FREM3*, is within the 95% confidence interval in our data. We note that while this analysis is simple and appealing, its interpretation depends on assumptions about the true disease model, and on the behaviour of imputation in the region[50]; in particular, the difficulty of imputing variants around the three glycophorin genes (Extended Data Fig. 8) might make this analysis fail to capture putatively important variation within or around *GYPA*, *GYPB*, or *GYPE*.

Sequence homology and alignability in the glycophorin region. To investigate the location of our GWAS signal with respect to the pattern of sequence homology around the three glycophorin genes, we generated a dot plot (Extended Data Fig. 8a) showing co-occurrence of short segments of DNA (k-mers; $k = 25$ or 100 bases) in the human reference sequence[51] in the region. Considerable sequence

homology was observed over a region of about 350 kb covering the three genes. Our lead GWAS marker and the ABPs lie just outside this region.

The high level of homology should affect our ability to align probes or sequences to this part of the genome. To confirm this, we also plotted the UCSC alignability track ('CRG alignability 100'; Extended Data Fig. 8c), which shows the degree to which the 100-mer starting at each position in the reference sequence is alignable (with a value of $1/n$ indicating that the 100-mer aligns to n positions across the genome, allowing up to two mismatches). As expected, even short regions of shared sequence affect alignability considerably, with alignability dropping to an average of about 0.7 within the large region of homology. Similarly, we plotted imputation performance (as measured by the IMPUTE info measure) and observed a marked drop within the region of sequence homology.

Structural variation in the glycophorin region. We attempted to identify structural variation in the glycophorin region by examining sequence read data generated by the 1000 Genomes Project, using the set of BAM files available from the 1000 Genomes Project in October 2014, downloaded from ftp://ftp.1000genomes. ebi.ac.uk/vol1/ftp/data. These data contain a mixture of read lengths, with most reads of at least 90 bp and some reads of 76 bp. We considered only reads with mapping quality at least 20. For each sample we computed coverage across the glycophorin region, normalized by average coverage computed across chromosome 1, and refer to these values here as 'coverage'.

Coverage values across the region were correlated with genome alignability (as defined earlier), with coverage dropping substantially in regions of low alignability. We therefore restricted attention to the set of perfectly alignable positions, here defined as positions such that every 100-mer overlapping the position in the reference sequence aligns uniquely, allowing up to two mismatches (and computed using the CRG alignability 100 track described earlier). As expected, coverage at perfectly alignable positions showed little or none of the variation present in genome alignability (Extended Data Fig. 8e–g).

Two large structural variants, present at frequencies of at least 1% in LWK+YRI, were evident in coverage data. Genotypes for both these variants were called by the 1000 Genomes Phase I and are referred to as esv2668125 (also called MERGED_DEL_2_26708) and esv2662558 (also called MERGED_DEL_2_26722). Both deletions putatively represent deletions of all or part of *GYPB*. We noted some uncertainty as to the location of the breakpoints of both deletions— with breakpoints identified by the 1000 Genomes Project differing from the breakpoints as they appeared in coverage data by at least 10 kb. We further noted that three samples (NA18519, NA19185, NA19222) that were called as heterozygote for one or both deletions by the 1000 Genomes Project appeared to have homozygous genotypes (Extended Data Fig. 8e–g, red blue and green lines).

eQTL analysis. To investigate the effect of associated SNPs in the *FREM3* region on expression of nearby genes, we downloaded publicly available data on RNA expression levels[17] in HapMap samples, and plotted expression levels of genes in the region (Fig. 1) against genotypes at the lead SNP and other SNPs of interest in African samples (YRI+LWK, Extended Data Fig. 9). Data was available for genes *INPP4B*, *USP38*, *GAB1*, *SMARCA5*, *GYPE*, *GYPB*, *GYPA*, *HHIP*, *ANAPC10* and *ABCE1*, with three probes available for *GYPE*, two for *GAB1* and one probe for each of the other genes. Five samples in this data set (NA19318, NA19324, NA19377, NA19429, NA19190) carry the protective allele at the lead marker rs184895969, while only two carry the protective allele at the directly typed rs186873296. For each gene and SNP we tested for a trend of genotype on expression using linear regression. Extended Data Fig. 9 shows all probes for the glycophorins as well as other regional genes for which a *P* value less than 0.05 was observed.

GENECLUSTER analysis. We used the program TREESIM[21] to estimate an ancestral recombination graph for the LWK and YRI individuals in the 1000 Genomes Project haplotype data in a region from 144.6 Mb to 144.8 Mb on chromosome 4 centred on the lead SNP in the *FREM3/GYPE* region. We then ran GENECLUSTER[21] in the region, allowing it to assign either one or two causal mutations in each marginal genealogy. GENECLUSTER works by probabilistically assigning study individuals to the tips of the tree estimated by TREESIM, and attempts to explain case/control status by assigning causal mutations to the branches of the tree in a Bayesian framework. This analysis is somewhat similar in spirit to our marginal SNP analysis, but has important differences. In particular, GENECLUSTER may detect associated mutations anywhere on the tree (which may not correspond directly to any typed or imputed variant), and can assess models of association involving more than a single causal variant. However, GENECLUSTER does not take into account principal components or otherwise control for population structure, and for computational reasons we only included the Luhyan (LWK) and Yoruban (YRI) populations from the 1000 Genomes reference panel in the analysis.

To investigate the relationship between ancestry and variants of interest, we plotted IMPUTE reference panel haplotypes in the region ordered by the marginal tree at the position of the imputed lead marker (rs184895969), annotating the lead
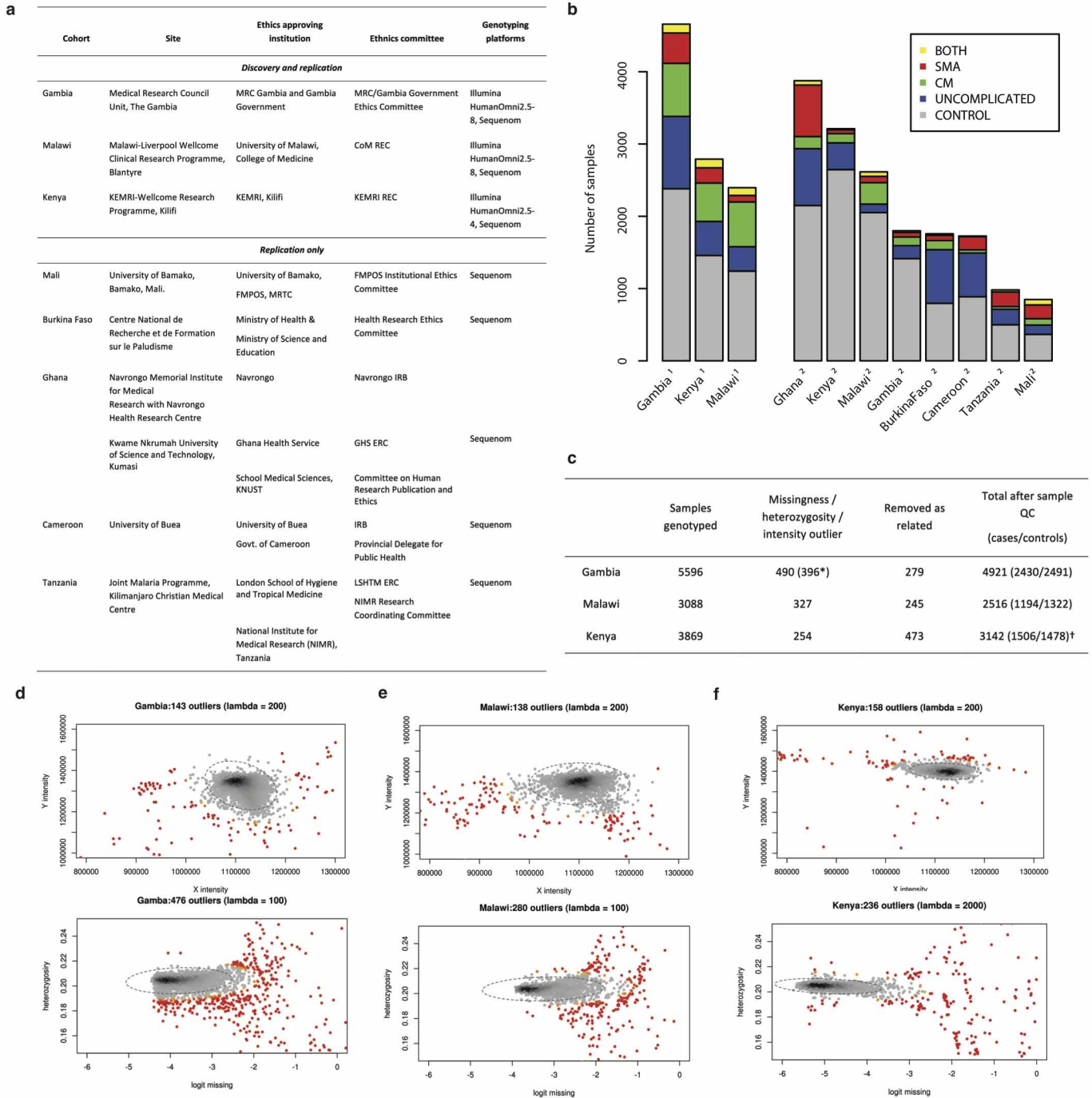
imputed and Sequenom-typed markers, ABPs, previously reported *GYPE* eQTLs[52], common deletions, and variants determining the M/N and S/s blood groups as described in Supplementary Note 2 (Fig. 3).

Analysis of enrichment of malaria-associated loci in functional categories. Full details of enrichment analyses are provided in Supplementary Note 3.

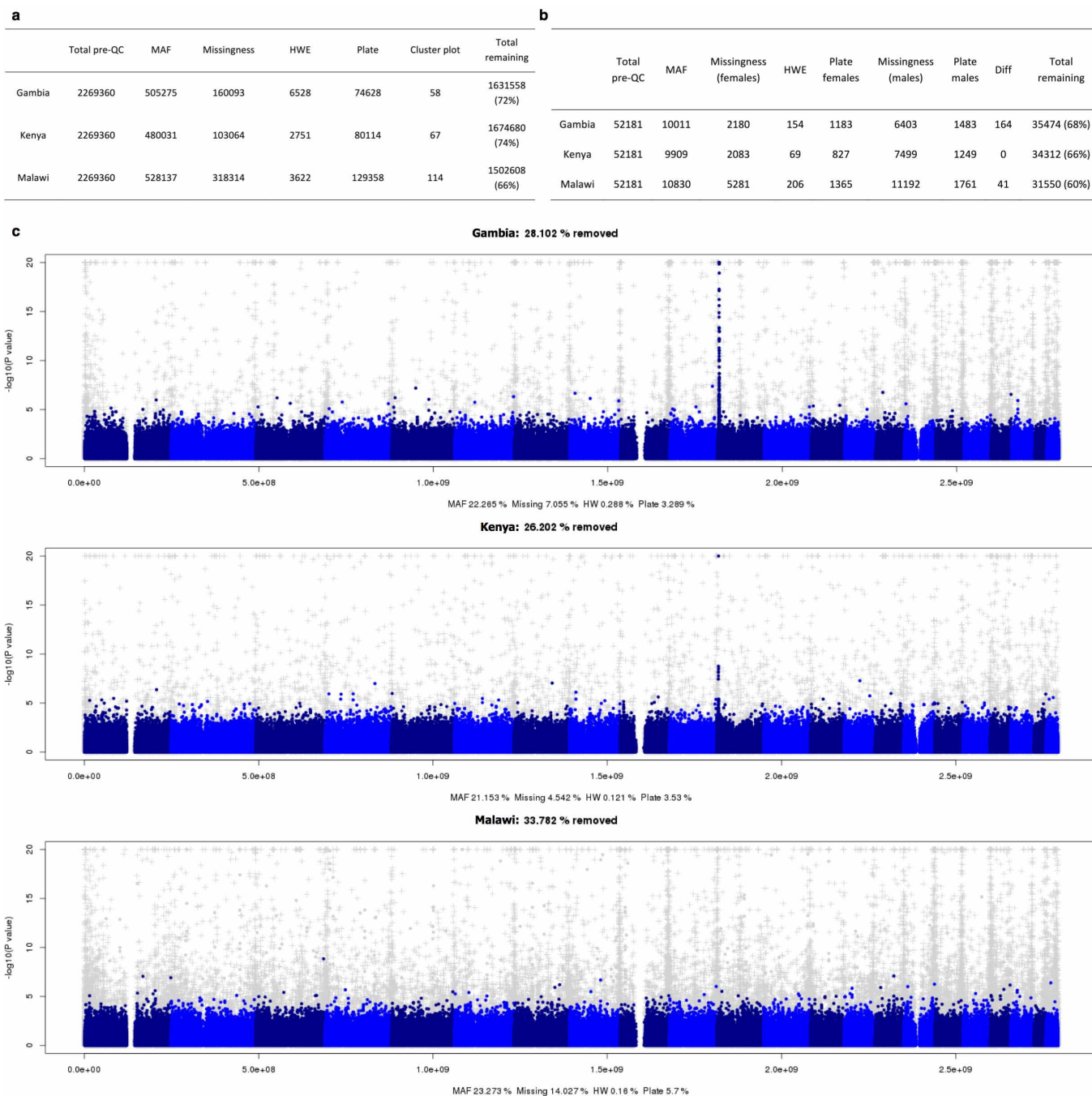**Code availability.** Executables and source code for inthinnerator are available at http://www.well.ox.ac.uk/~gav/inthinnerator.

Further software for generating key results for this paper will be made available at http://www.malariagen.net/resource/14.

31. World Health Organization. Severe falciparum malaria. *Trans. R. Soc. Trop. Med. Hyg.* **94** (suppl. 1), S1–S90 (2000).
32. Teo, Y. Y. Genotype calling for the Illumina platform. *Methods Mol. Biol.* **850**, 525–538 (2012).
33. Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J. & Holmes, C. C. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics* **24**, 2209–2214 (2008).
34. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
35. Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
36. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
37. 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).
38. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
39. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
40. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).
41. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genet.* **44**, 955–959 (2012).
42. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
43. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
44. Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl Acad. Sci. USA* **96**, 14440–14444 (1999).
45. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2013).
46. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
47. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
48. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
49. Manjurano, A. *et al.* *USP38, FREM3, SDC1, DDC,* and *LOC727982* gene polymorphisms and differential susceptibility to severe malaria in Tanzania. *J. Infect. Dis.* **212**, 1129–1139 (2015).
50. The Wellcome Trust Case Control Consortium et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genet.* **44**, 1294–1301 (2012).
51. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724–731 (2005).
52. Zeller, T. *et al.* Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5**, e10693 (2010).

a

Cohort	Site	Ethics approving institution	Ethnics committee	Genotyping platforms
Discovery and replication				
Gambia	Medical Research Council Unit, The Gambia	MRC Gambia and Gambia Government	MRC/Gambia Government Ethics Committee	Illumina HumanOmni2.5-8, Sequenom
Malawi	Malawi-Liverpool Wellcome Clinical Research Programme, Blantyre	University of Malawi, College of Medicine	CoM REC	Illumina HumanOmni2.5-8, Sequenom
Kenya	KEMRI-Wellcome Research Programme, Kilifi	KEMRI, Kilifi	KEMRI REC	Illumina HumanOmni2.5-4, Sequenom
Replication only				
Mali	University of Bamako, Bamako, Mali.	University of Bamako, FMPOS, MRTC	FMPOS Institutional Ethics Committee	Sequenom
Burkina Faso	Centre National de Recherche et de Formation sur le Paludisme	Ministry of Health & Ministry of Science and Education	Health Research Ethics Committee	Sequenom
Ghana	Navrongo Memorial Institute for Medical Research with Navrongo Health Research Centre	Navrongo	Navrongo IRB	
	Kwame Nkrumah University of Science and Technology, Kumasi	Ghana Health Service	GHS ERC	Sequenom
		School Medical Sciences, KNUST	Committee on Human Research Publication and Ethics	
Cameroon	University of Buea	University of Buea	IRB	Sequenom
		Govt. of Cameroon	Provincial Delegate for Public Health	
Tanzania	Joint Malaria Programme, Kilimanjaro Christian Medical Centre	London School of Hygiene and Tropical Medicine	LSHTM ERC	Sequenom
		National Institute for Medical Research (NIMR), Tanzania	NIMR Research Coordinating Committee	

b

c

	Samples genotyped	Missingness / heterozygosity / intensity outlier	Removed as related	Total after sample QC (cases/controls)
Gambia	5596	490 (396*)	279	4921 (2430/2491)
Malawi	3088	327	245	2516 (1194/1322)
Kenya	3869	254	473	3142 (1506/1478)†

d

e

f

Extended Data Figure 1 | Sample collections included in the study. a, Study sites and ethics approving institutions. **b,** Phenotypic makeup of discovery and replication samples from each site. 'Uncomplicated' refers to case individuals who were not identified as cerebral malaria (CM) or severe malarial anaemia (SMA) cases. 'Both' refers to individuals who have both CM and SMA phenotypes. **c,** Overall sample counts and number of samples excluded by each quality control (QC) criterion. Asterisk denotes the number of samples removed after explicitly including samples with low heterozygosity in Gambia. Dagger: the Kenyan cohort included parents of a subset of case samples; these were not used in subsequent analyses **d–f,** Plots of average genome-wide heterozygosity and missingness with outliers coloured, as outputted by the ABERRANT algorithm.
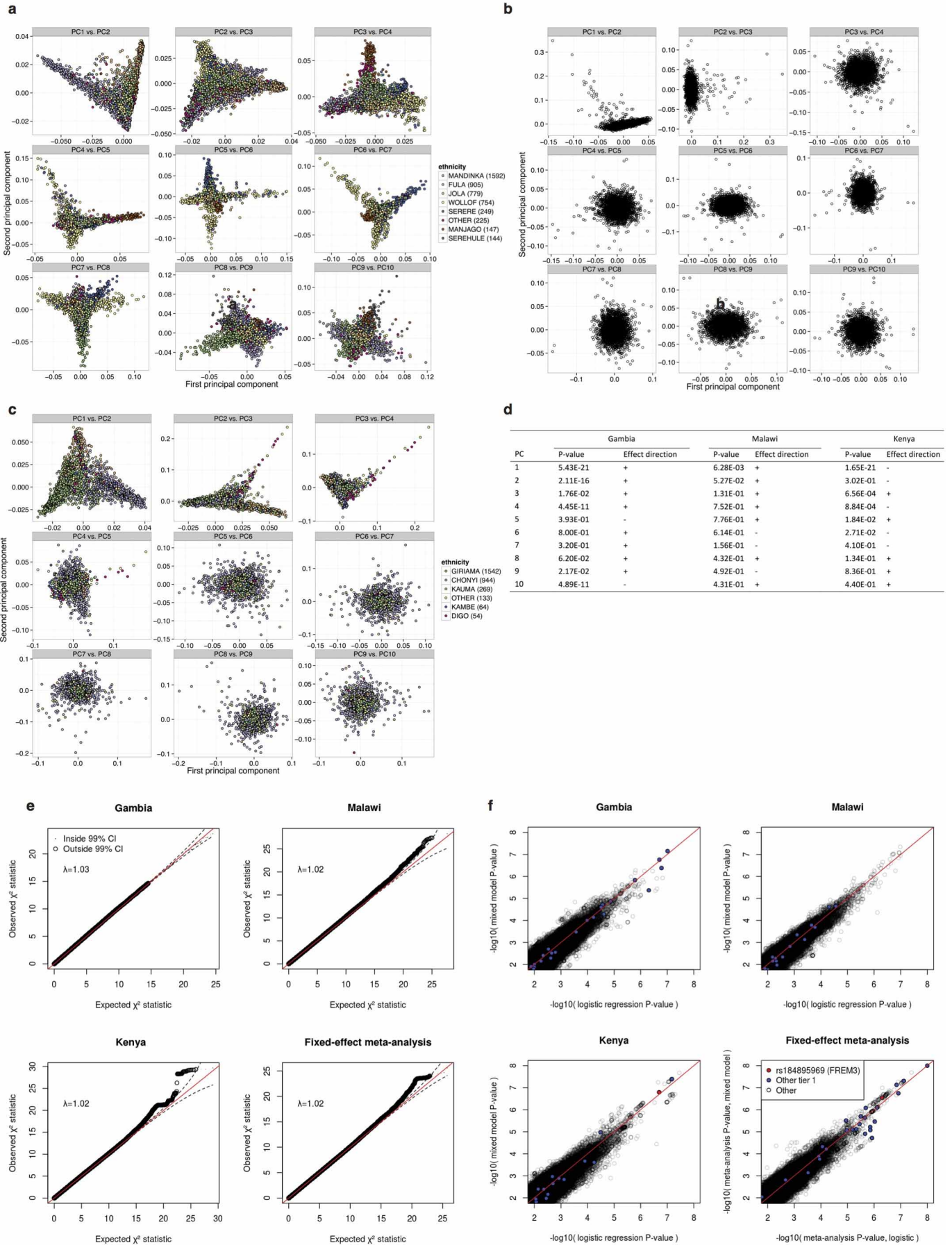
a

	Total pre-QC	MAF	Missingness	HWE	Plate	Cluster plot	Total remaining
Gambia	2269360	505275	160093	6528	74628	58	1631558 (72%)
Kenya	2269360	480031	103064	2751	80114	67	1674680 (74%)
Malawi	2269360	528137	318314	3622	129358	114	1502608 (66%)

b

	Total pre-QC	MAF	Missingness (females)	HWE	Plate females	Missingness (males)	Plate males	Diff	Total remaining
Gambia	52181	10011	2180	154	1183	6403	1483	164	35474 (68%)
Kenya	52181	9909	2083	69	827	7499	1249	0	34312 (66%)
Malawi	52181	10830	5281	206	1365	11192	1761	41	31550 (60%)

c

Extended Data Figure 2 | Genotyped SNP quality control for the three discovery cohorts. a, b, Total numbers of pre- and post-quality control SNPs on the autosomes (**a**) and the X chromosome (**b**), and numbers of SNPs excluded by each quality control (QC) criterion. Diff, the test of difference in frequency between males and females; HWE, Hardy–Weinberg equilibrium; Plate, plate test of association. Details of quality control are given in Methods. **c,** Plot showing the $-\log_{10}(P$ values) for the genotypic association test in the discovery data including the first five principal components as covariates. Grey dots show SNPs that are removed due to the quality control as defined in Methods. The total fraction of SNPs removed from each cohort is given at the top of the plot.
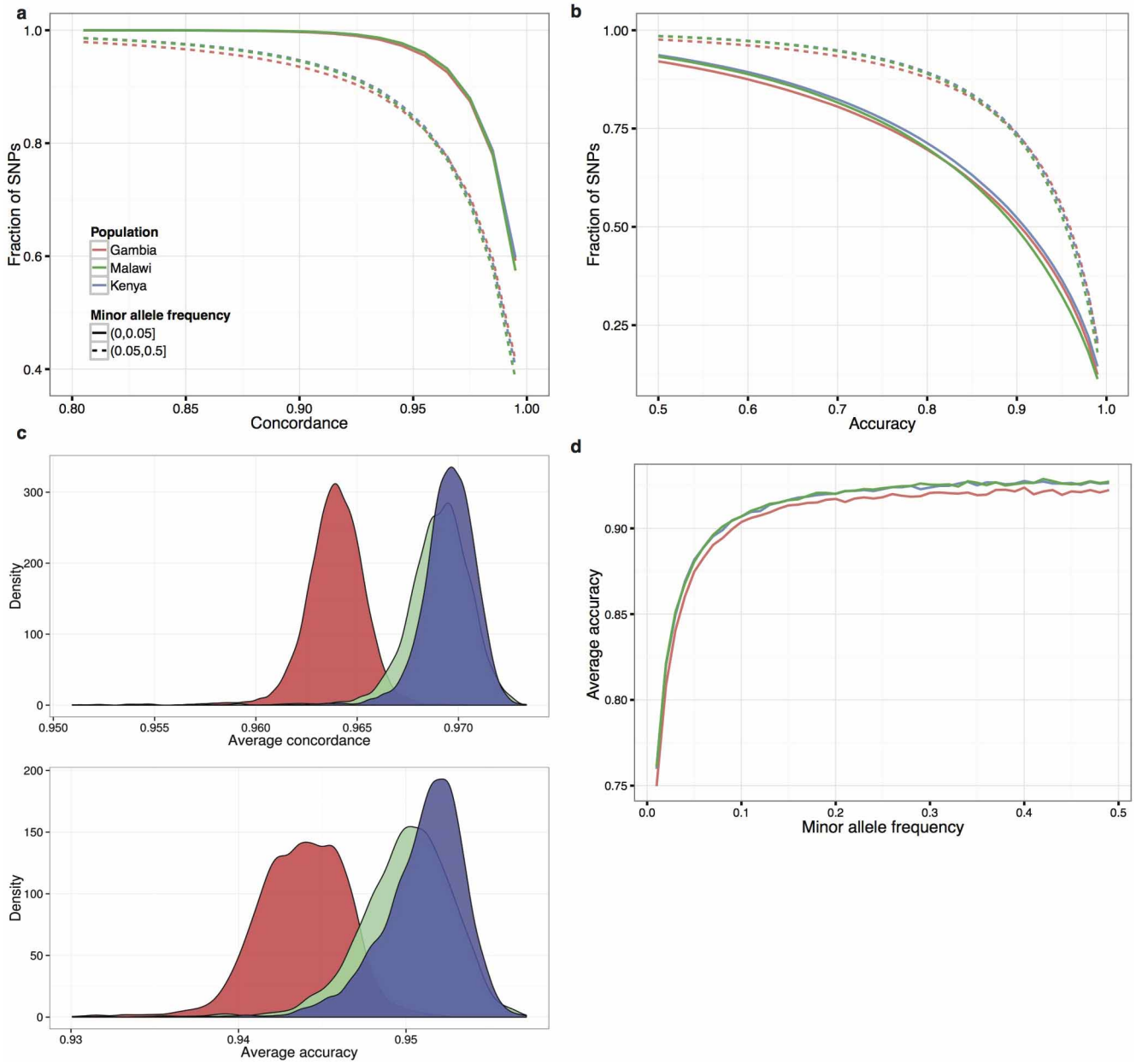
Extended Data Figure 3 | Imputation performance. a, b, Empirical distribution of concordance and accuracy (r^2) between typed and re-imputed SNPs in the three discovery cohorts. Solid lines represent SNPs with frequency below 5% and dashed lines represent SNPs with frequency of at least 5%. **c**, Per-sample concordance and accuracy (type 0 r^2) across the whole genome, as estimated by re-imputing genotyped SNPs. Values are averaged over imputation chunks. **d**, Average accuracy between genotype and re-imputed SNPs in each cohort, plotted against frequency, in 1% frequency bins.

a

b

c

d

PC	Gambia		Malawi		Kenya	
	P-value	Effect direction	P-value	Effect direction	P-value	Effect direction
1	5.43E-21	+	6.28E-03	+	1.65E-21	-
2	2.11E-16	+	5.27E-02	+	3.02E-01	-
3	1.76E-02	+	1.31E-01	+	6.56E-04	+
4	4.45E-11	+	7.52E-01	+	8.84E-04	-
5	3.93E-01	-	7.76E-01	+	1.84E-02	+
6	8.00E-02	+	6.14E-01	-	2.71E-02	+
7	3.20E-01	+	1.56E-01	-	4.10E-01	-
8	6.20E-02	+	4.32E-01	+	1.34E-01	+
9	2.17E-02	+	4.92E-01	-	8.36E-01	+
10	4.89E-11	-	4.31E-01	+	4.40E-01	+

e

f

Extended Data Figure 4 | Detail of population structure. **a–c**, Top ten principal components (PCs) in Gambia (**a**), Malawi (**b**) and Kenya (**c**). Where ethnicity was reported, points are coloured by ethnicity for ethnicities with at least 50 samples. **d**, Logistic regression P values and direction of effect for the top ten PCs on severe malaria status in each cohort. **e**, q–q plots for additive model association test P values in Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Dashed lines represent the 99% CI computed marginally at each variant. Circles and points represent points lying respectively outside and inside the 99% CI. **f**, Comparison of association test P values for logistic regression (SNPTEST, x-axis) and linear mixed model (MMM, y-axis) for Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Variants in tier 1 are coloured blue, with the lead marker at the $FREM3/GYPE$ region coloured red.
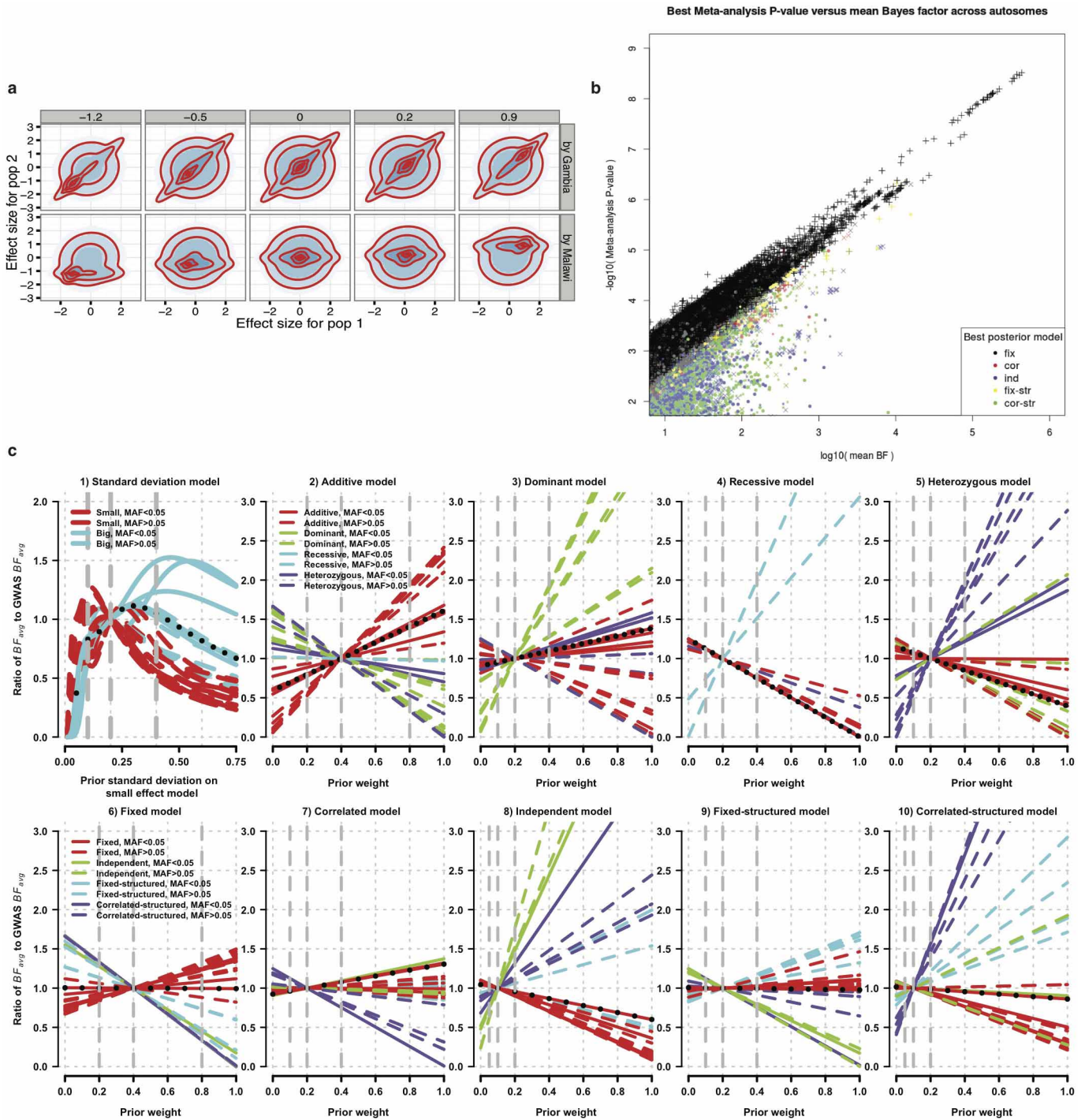
Extended Data Figure 5 | Detail of Bayesian analysis of discovery cohorts.
a, Visualization of slices through the combined prior on effect sizes in three cohorts for mode-of-inheritance-specific models. Top row: slices through the prior effect size on Kenya (x-axis) and Malawi (y-axis) for constant effect size in The Gambia (panels). Bottom row: slices through the prior effect size on Kenya (x-axis) and The Gambia (y-axis) for constant effect size in Malawi (panels). Red lines represent a factor of 10 in the prior density. **b**, Comparison of BF_{avg} (x-axis) with the minimum fixed-effect meta-analysis P value minimized across additive, dominant, recessive or heterozygote modes of inheritance (y-axis). Values are plotted on $\log_{10}$ and $-\log_{10}$ scales. Colour indicates the heterogeneity model of the model with the highest posterior weight. **c**, Sensitivity of BF_{avg} to changes in prior. Plots show BF_{avg} ratio (y-axis) plotted against one-dimensional parameterisations of the prior (x-axis), for the 32 autosomal SNPs in tier 1. Solid lines represent variants with MAF < 5%

averaged across populations, and dashed lines represent variants with MAF ≥ 5%. Black dots indicate the lead marker at the *FREM3/GYPE* locus. Colour indicates the effect size, mode of inheritance, or heterogeneity model for the model with highest posterior weight under the GWAS prior. Dashed grey vertical lines indicate the x-axis value corresponding to the prior used in the GWAS, and one-half and twice that value. Plots are parameterized by (1) the prior standard deviation of the small-effect model keeping the prior standard deviation of the large and small-effect models in the ratio 0.75:0.2; (2–5) the prior weight on additive, dominant, recessive or heterozygote modes of inheritance; (6–10) the prior weight on fixed, correlated, independent, fixed-structured and correlated-structured models. For each parameterization prior, weights on other models are kept in the same relative proportion. For further details see Supplementary Note 4.
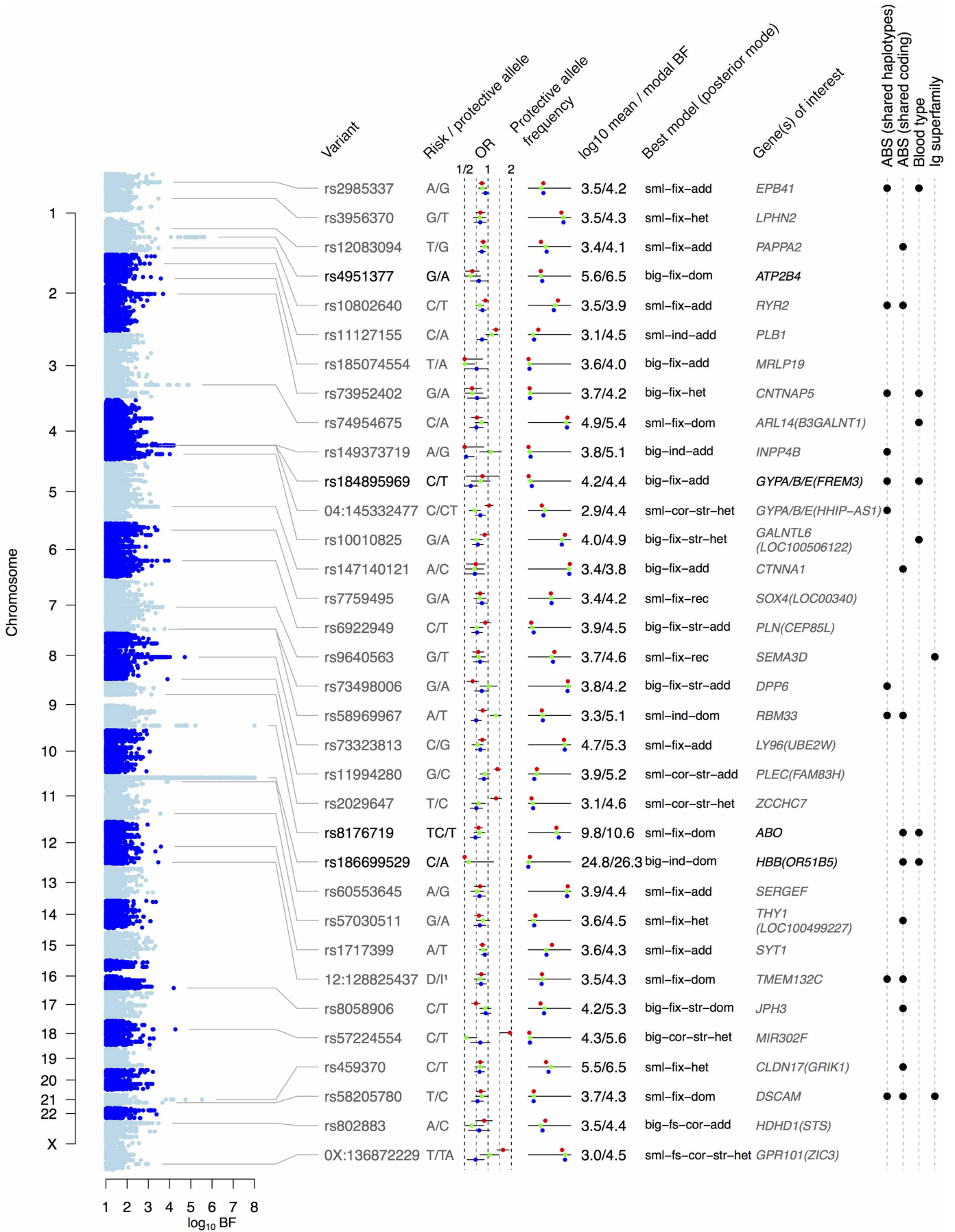
Extended Data Figure 6 | Strongest regions of association in the Bayesian analysis of the three discovery cohorts. Plot on left shows the $\log_{10}$ model-averaged Bayes Factor (BF_{avg}). Table shows the SNP with the highest BF_{avg} in each region (lead SNP), gene(s) of interest in the region, the model with the highest posterior weight at the lead SNP and its BF. Models are described in Methods; sml, big: small- and large-effects models; fix, cor, ind, fix-str, cor-str: fixed, correlated, independent, fixed-struc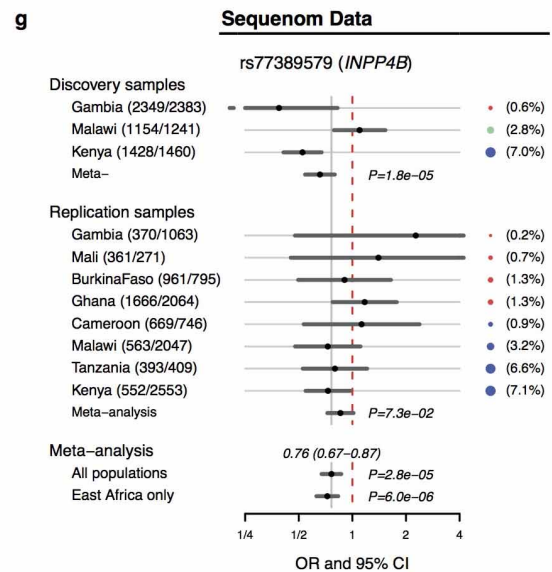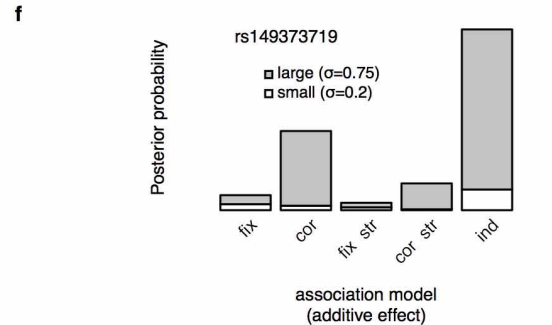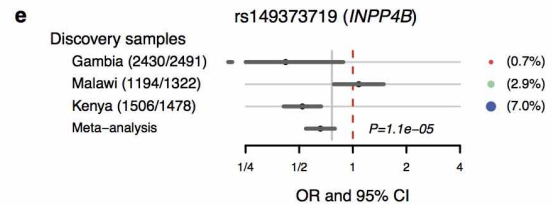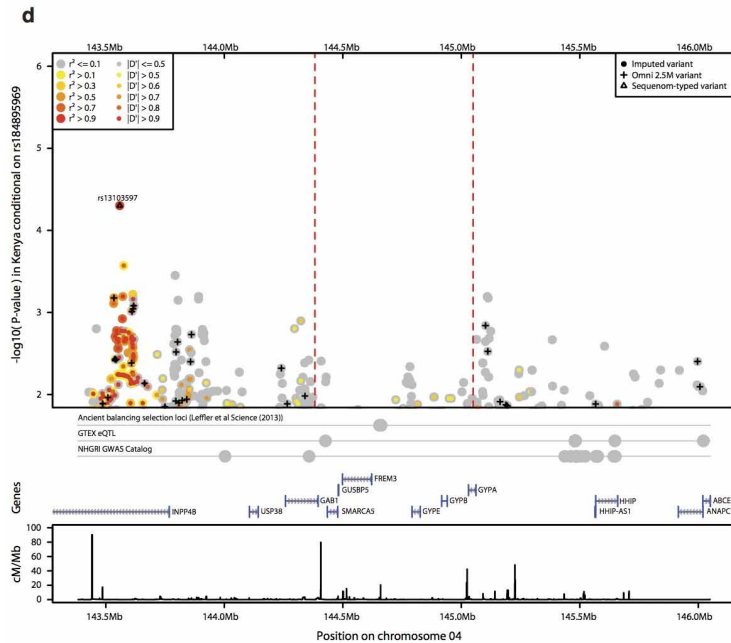tured and correlated-structured eff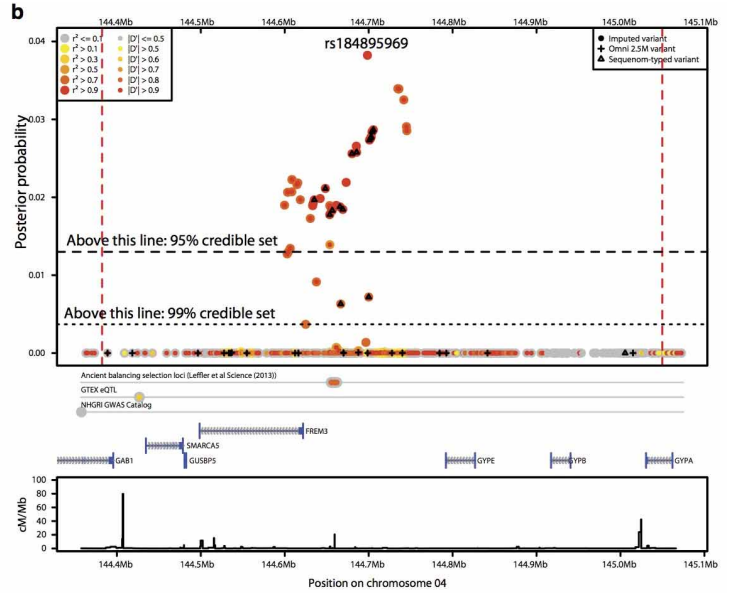ect models; add, dom, rec, het: additive, dominant, recessive, heterozygote models; fs: fixed-between-sex model. Coloured points indicate the OR and the protective allele frequency in Gambia (red), Malawi (green) and Kenya (blue). The right-hand columns indicate regions containing shared chimpanzee–human haplotypes or coding SNPs[4] (ABPs), blood group genes, or immunoglobulin superfamily genes.

a

| Region	SNP	r^2 with lead imputed marker	Two-sided fixed-effect P (discovery)	One-sided fixed-effect P (replication)	Two-sided fixed-effect P (discovery + replication)
FREM3 region					
	rs18673296	0.93	2.0×10^{-6}	5.4×10^{-6}	9.5×10^{-11}
	rs186790584	0.94	2.4×10^{-6}	1.4×10^{-5}	2.9×10^{-10}
	rs149914432	0.87	8.0×10^{-7}	6.6×10^{-5}	5.0×10^{-10}
INPP4B region					
	rs77389579	1.0	1.8×10^{-5}	3.7×10^{-2}	2.8×10^{-5}
ARL14 region					
	rs74954675	0.94	1.4×10^{-6}	0.63	0.0022
	rs75731597	0.79	9.1×10^{-6}	0.52	0.0016
	rs76033371	0.81	4.6×10^{-7}	0.71	2×10^{-4}

b

c

RS id	Chromosome and position (NCBI b37)	Ancestral allele	Ref/non-ref allele (NCBI b37)	Ref / non-ref amino acid (NCBI b37, reverse strand)	Ref / non-ref blood group (NCBI b37)
rs7658293	4:145041707	C	C/A	**GAG** (Glu) / **GGT** (Gly)	N / M
rs7687256	4:145041708	T	T/C		
rs7682260	4:145041720	G	A/G	**TTA** (Leu) / **TCA** (Ser)	
rs7683365	4:144920956	C	G/A	A**CG** (Thr) / A**TG** (Met)	s / S

d

Imputed Data

e

rs149373719 (*INPP4B*)

f

rs149373719

Sequenom Data

g

rs77389579 (*INPP4B*)

Extended Data Figure 7 | Evidence for association in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions. a, Evidence for association at directly typed SNPs. **b**, Posterior probability that variants in the *FREM3/GYPE* region are causal, assuming a single variant in the region is causal[50], based on the BF_{avg} for typed and imputed variants. Dashed lines indicate the 95% and 99% credible sets. See Fig. 1 legend for further details. **c**, Details of SNPs encoding the common MNS blood groups. Coordinates and alleles are with respect to the NCBI b37 human reference sequence. **d**, Evidence for possible independence of effects at the *FREM3* and *INPP4B* loci in Kenya by conditional analysis. *y*-Axis represents $-\log_{10}$(association *P* value) conditional on the imputed dosage at rs184895969. Points are coloured by linkage disequilibrium with the top SNP rs13103597. **e**, Forest plot showing sample size, estimated OR and 95% CI for the lead imputed SNP (rs149373719) in *INPP4B* under an additive model of association. **f**, Bar plot showing the posterior weight on different models of heterogeneity at rs149373719 under the prior used in the GWAS, assuming an additive model of association. **g**, Forest plot showing evidence in both discovery and replication samples in the Sequenom data at rs77389579 in *INPP4B*. See Fig. 2 legend for further details.
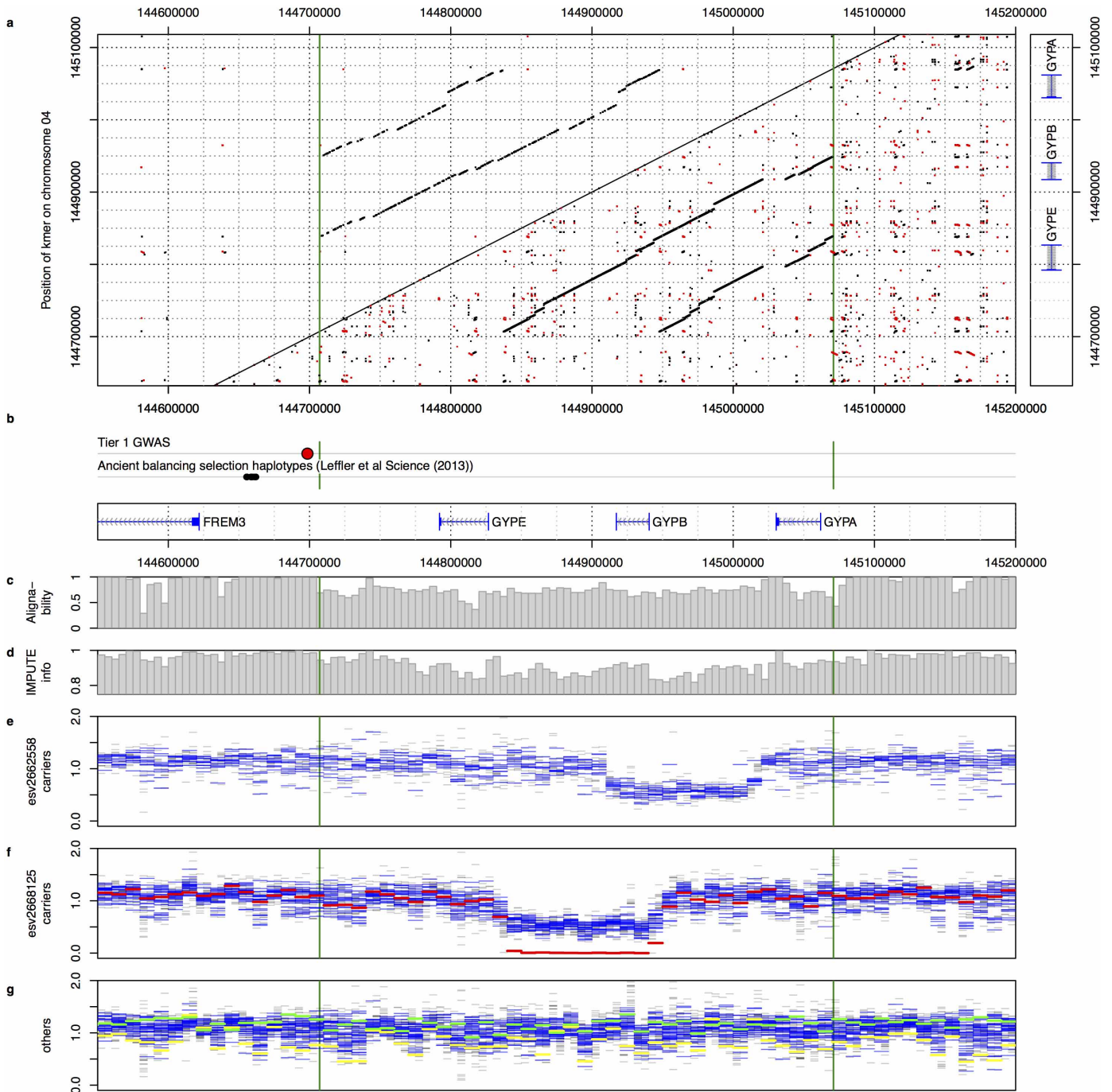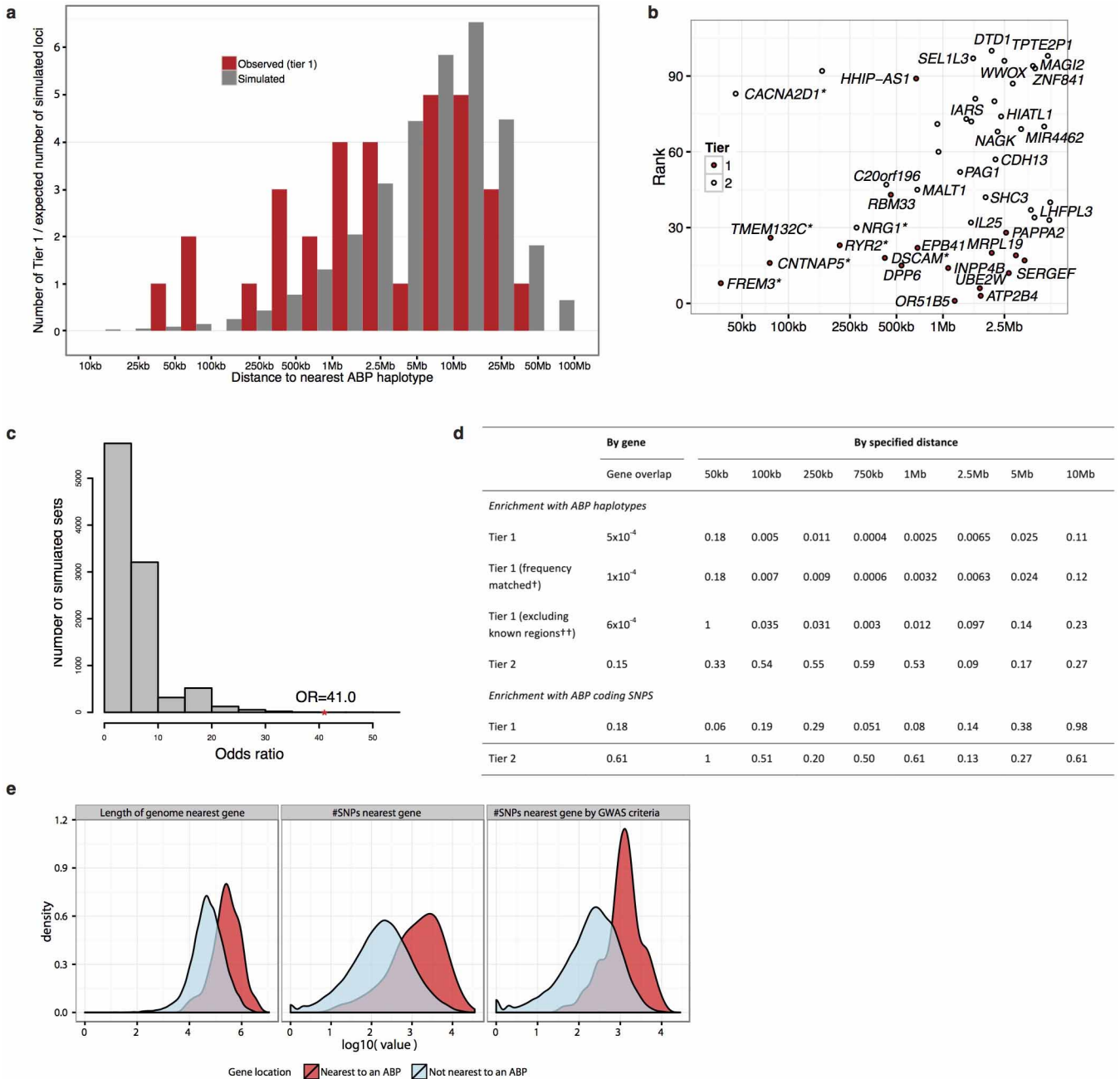
Extended Data Figure 8 | Sequence homology, alignability and structural variation in the glycophorin region. a, Co-occurrence of short segments of DNA (k-mers; top triangle: $k = 100$, bottom triangle: $k = 25$) in the human reference sequence Each point represents a k-mer that maps to the locations indicated by the x- and y-axis positions, either on the same strand (black points) or opposite strands (red points). Green vertical lines in this and subsequent panels delineate the region of high homology surrounding the three glycophorins. **b**, The location of the lead GWAS marker, ABPs, and protein-coding genes in the region. **c**, Alignability of the 100-mer at each position of the reference, up to two mismatches. Values are taken from the UCSC genome browser mappability track and averaged over 5 kb bins. **d**, IMPUTE info measure in Kenya for variants with frequency at least 5%, averaged over 5 kb bins. **e–g**, Coverage for samples from YRI and LWK in 1000 Genomes Project Phase I carrying esv2662558, carrying esv2668125, or not carrying either deletion, respectively. Coverage for each individual is normalized by the mean coverage for that individual across chromosome 1, and only computed at positions with alignability = 1 for all 100-mers overlapping the position, and for reads with mapping quality at least 20. Values are averaged over 5 kb (grey) and 10 kb (blue) bins. Three samples with apparently erroneous calls in the 1000 Genomes Phase I genotype release are coloured (NA18519, red; NA19185, yellow; NA19222, green) and assigned to the status indicated by their coverage profile. Panel **g** represents a random sample of 30 individuals not carrying the deletion in addition to the two with erroneous genotype calls. Coverage computation was performed using the BAM files available from the 1000 Genomes Project in October 2014, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data. Four African samples in the Phase I release were not assessed because they are not included in this directory.
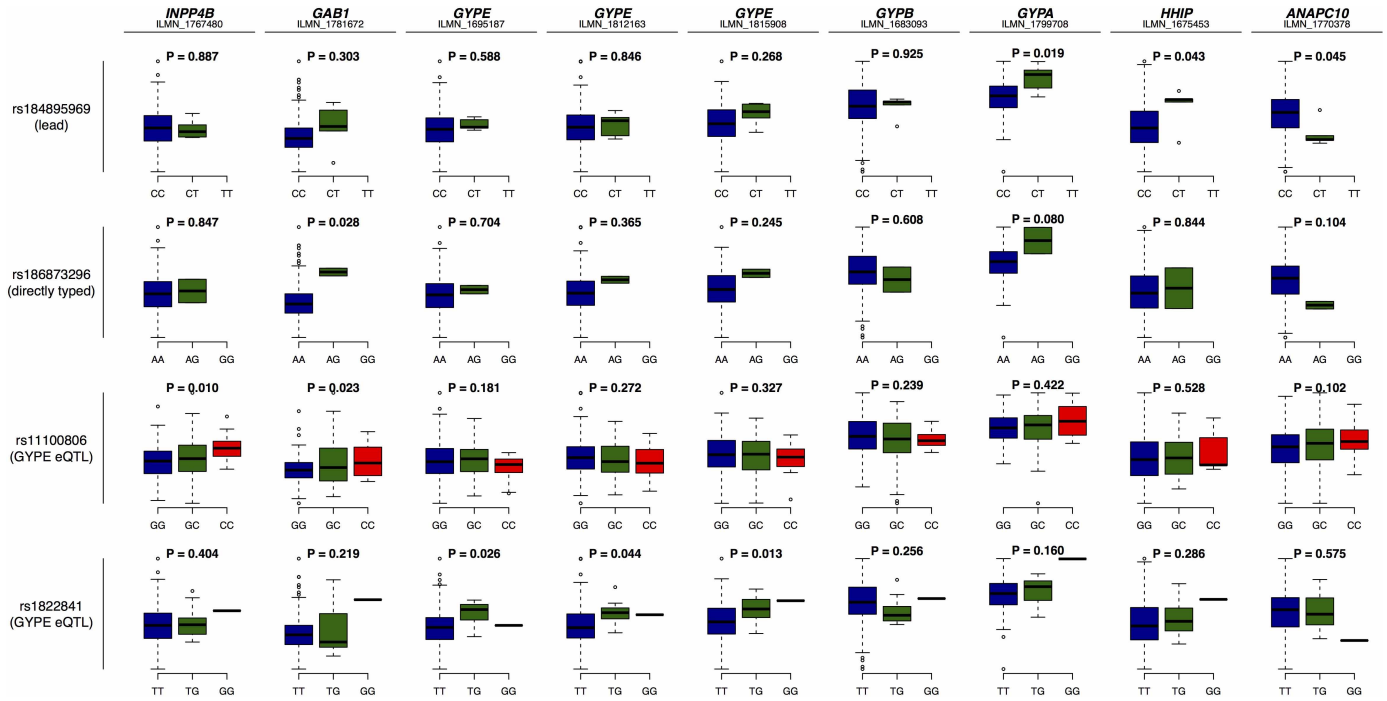
Extended Data Figure 9 | Correlation between the genotypes at SNPs of interest within the *GYPE/A/B* locus and reported gene transcription levels in samples from the YRI and LWK HapMap cohorts[17]. *P* values are for a trend test of association where more than one genotype class is present. Only assays targeting the glycophorins, and those with a *P* value below 0.05 are shown.

a

b

c

d

	By gene	By specified distance							
	Gene overlap	50kb	100kb	250kb	750kb	1Mb	2.5Mb	5Mb	10Mb
Enrichment with ABP haplotypes									
Tier 1	5x10⁻⁴	0.18	0.005	0.011	0.0004	0.0025	0.0065	0.025	0.11
Tier 1 (frequency matched†)	1x10⁻⁴	0.18	0.007	0.009	0.0006	0.0032	0.0063	0.024	0.12
Tier 1 (excluding known regions††)	6x10⁻⁴	1	0.035	0.031	0.003	0.012	0.097	0.14	0.23
Tier 2	0.15	0.33	0.54	0.55	0.59	0.53	0.09	0.17	0.27
Enrichment with ABP coding SNPS									
Tier 1	0.18	0.06	0.19	0.29	0.051	0.08	0.14	0.38	0.98
Tier 2	0.61	1	0.51	0.20	0.50	0.61	0.13	0.27	0.61

e

Gene location — Nearest to an ABP / Not nearest to an ABP

Extended Data Figure 10 | Detail of enrichment analysis. a, Red histogram: the empirical distribution of the log10 distance of observed tier 1 loci to the nearest ABP haplotype. Grey histogram, distribution of distances for 10,000 simulated tier 1 sets. b, The log10 distance of tier 1 (filled red circles) and tier 2 (empty circles) loci to the nearest ABP, plotted against their rank in BFavg order (stronger signals have lower rank). Loci are annotated with the nearest gene where a gene exists within the association region. Asterisks denote nearest genes that are also the nearest gene to an ABP shared haplotype. c, Empirical null distribution of the odds ratio for the enrichment of tier 1 loci in the set of genes closest to an ABP shared haplotype, based on 10,000 simulated SNP sets. The red asterisk and text indicate the odds ratio for the observed tier 1 loci. d, Distribution of the proportion of the genome which identifies a given gene as nearest, for genes in or not in the set annotated as nearest an ABP haplotype. Left, distribution of the length of the genome for which the given gene is unambiguously the closest gene. Middle, distribution of the number of SNPs in our study for which the given gene is the closest gene. Right, distribution of the number of SNPs in our study for which the given gene is the nearest gene within a recombination interval of 2.5 cM ± 25 kb around the SNP, as used to determine nearest genes to GWAS lead SNPs. e, Empirical *P* values for enrichment of ABP haplotypes and coding SNPs in tier 1 and tier 2 GWAS regions. Second column, *P* values for enrichment by gene overlap. Third to tenth column, P values for enrichment by proximity at different length scales. †Results for simulations using SNPs frequency-matched to GWAS tier 1 loci in 1% frequency bins. ††Results for simulations excluding the regions of *ABO, HBB, ATP2B4, FREM3, INPP4B,* and *HHIP-AS1.*