A structural variant encoding hybrid glycophorins is associated with resistance to severe malaria

Ellen M. Leffler[1,2], Gavin Band[1,2], George B.J. Busby[1], Katja Kivinen[2], Quang Si Le[1], Geraldine M. Clarke[1], Kalifa A. Bojang[3], David J. Conway[3,4], Muminatou Jallow[3,5], Fatoumatta Sisay-Joof[3], Edith C. Bougouma[6], Valentina D. Mangano[7], David Modiano[7], Sodiomon B. Sirima[6], Eric Achidi[8], Tobias O. Apinjoh[9], Kevin Marsh[10,11], Carolyne M. Ndila[10], Norbert Peshu[10], Thomas N. Williams[10,12], Chris Drakeley[13,14], Alphaxard Manjurano[13,14], Hugh Reyburn[13,14], Eleanor Riley[14], David Kachala[15], Malcolm Molyneux[15,16], Vysaul Nyirongo[15], Terrie Taylor[17,18], Nicole Thornton[19], Louise Tilley[19], Shane Grimsley[19], Eleanor Drury[2], Jim Stalker[2], Victoria Cornelius[1], Christina Hubbart[1], Anna E. Jeffreys[1], Kate Rowlands[1], Kirk A. Rockett[1,2], Chris C.A. Spencer[1+], Dominic P. Kwiatkowski[1,2+], Malaria Genomic Epidemiology Network[1,2]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.
[2] 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] Royal Victoria Teaching Hospital, Independence Drive, PO Box 1515, Banjul, The Gambia.
[6] Centre National de Recherche et de Formation sur le Paludisme (CNRFP), 01 BP 2208 Ouagadougou 01, Burkina Faso.
[7] University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy.
[8] Department of Medical Laboratory Sciences, University of Buea, PO Box 63, Buea, South West Region, Cameroon.
[9] Department of Biochemistry & Molecular Biology, University of Buea, PO Box 63, Buea, South West Region, Cameroon.
[10] KEMRI-Wellcome Trust Research Programme, PO Box 230-80108, Kilifi, Kenya.
[11] Nuffield Department of Medicine, NDM Research Building, Roosevelt Drive, Headington, Oxford OX3 7FZ, UK.
[12] Faculty of Medicine, Department of Medicine, Imperial College, Exhibition Road, London SW7 2AZ, UK.
[13] Joint Malaria Programme, Kilimanjaro Christian Medical Centre, PO box 2228, Moshi, Tanzania.
[14] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.
[15] Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Queen Elizabeth Central Hospital, College of Medicine, PO Box 30096, Chichiri, Blantyre 3, Malawi.
[16]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK.
[17] Blantyre Malaria Project, Queen Elizabeth Central Hospital, College of Medicine, PO Box 30096, Chichiri, Blantyre 3, Malawi.
[18] College of Osteopathic Medicine, Michigan State University, East Lansing, MI 48824, USA.
[19] International Blood Group Reference Laboratory, NHS Blood and Transplant, 500 North Bristol Park, Filton, Bristol BS34 7QH, UK.
[+]spencer@well.ox.ac.uk; dominic.kwiatkowski@sanger.ac.uk

Human genetic variation in the region between *FREM3* and *GYPE* was recently reported as associated with resistance to severe malaria in sub-Saharan Africa. By constructing a new imputation reference panel from sequence data with an increased representation of African populations, we show that the signal of association extends across the nearby genes *GYPA* and *GYPB*, which encode receptors for erythrocyte invasion by *Plasmodium falciparum*. Detailed analysis of large copy number variation reveals that the association can be explained by a complex structural variant that has lost *GYPB* but gained two hybrid genes, each with a GYPB extracellular domain and GYPA intracellular domain. This rearrangement, which is globally rare but carried by one in six children in parts of Kenya, encodes the Dantu blood group antigen, and reduces the risk of severe malaria by up to 40%. These findings provide the first direct evidence of the role malaria has played in shaping the landscape of variation at this polymorphic locus.

Malaria parasites cause human disease by invading and replicating inside red blood cells, which can lead to severe, life-threatening complications. Genome-wide association studies (GWAS) of malaria have the potential to discover natural mechanisms of resistance, pointing to genetic variants that influence the outcome of infection and may be under selection. In a recent GWAS, we identified alleles associated with protection from severe malaria on chromosome 4, lying between *FREM3* and a cluster of three paralagous genes encoding the glycophorins GYPE, GYPB and GYPA [1]. Although the association signal did not extend across this gene cluster, GYPA and GYPB are strong candidates for the underlying biological basis of the genetic association, as they are abundant erythrocyte surface proteins that are receptors for host cell invasion by *Plasmodium falciparum* [2, 3], the main malaria parasite in Africa [4].

Further analysis of this association signal is complicated by two main factors. First, the GWAS samples were collected at multiple locations in Africa, where levels of human genetic diversity are higher than in other parts of the world, and African populations remain relatively underrepresented in genome variation reference panels. Second, the glycophorin genes are in a region of segmental duplication known to harbour multiple forms of structural variation that contribute to the highly polymorphic MNS blood group system, but have not been characterized by next generation sequence data [5, 6]. Moreover, it has been proposed that genetic diversity at the glycophorins is particularly high in African populations, due at least in

part to selection driven by malaria [2, 7-9]. Here, we aim to capture additional relevant variation, including structural variation, to gain a better understanding of the association signal in this region.

Results

An African-enriched reference panel in the glycophorin region

We began by constructing a reference panel with improved representation of sub-Saharan African populations from countries where malaria is endemic. We performed genome sequencing of 765 individuals from 10 ethnic groups in the Gambia, Burkina Faso, Cameroon and Tanzania, including 207 family trios (100 bp paired end (PE) reads, mean coverage 10x; **Methods** and **Supplementary Tables 1** and **2**). We focused on a region surrounding the observed association signal (chr4:140Mb-150Mb; GRCh37 coordinates). Genotypes at SNPs and short indels in the region were called and computationally phased using published methods [10-12] and combined with Phase 3 of the 1000 Genomes Project [13] to obtain a reference panel of 3,269 individuals, including 1,269 Africans and a further 157 individuals with African ancestry (**Methods, Supplementary Fig. 1** and **Supplementary Tables 1** and **3**). We imputed variants from this panel into the published severe malaria GWAS dataset comprising 4,579 cases of severe malaria and 5,310 population controls from the Gambia, Kenya and Malawi and tested for association as described previously (**Methods**; [1]). The signal of association, previously identified and replicated at SNPs lying between *FREM3* and *GYPE*, extends over a region of at least 700 kb, and includes linked variants within *GYPA* and *GYPB*, where association is only apparent with the additional African reference data (**Supplementary Fig. 2**).

Identification of copy number variants

We next assessed copy number variation in the glycophorin region (defined here as the segmental duplication within which the three genes lie; chr4:144,706,830-145,069,066; **Methods**) for the sequenced reference panel individuals. The high level of sequence identity between the duplicate units presents a considerable challenge for short read sequence analysis due to ambiguous mapping (**Fig. 1b**) [14]. We therefore focused on changes in read depth at sites of high mappability. Specifically, we developed a hidden Markov model (HMM) to infer the underlying copy number state for each individual in 1600 bp windows using read depth averaged over sites above a set mappability threshold, and excluded windows with too few (<400) such sites. We adjusted for systematic variation along the genome

with a window-specific factor computed across individuals. We then grouped individuals carrying similar copy number paths to assign CNV genotypes (**Methods**).

Across the 3,269 samples, we identified eight deletions and eight duplications that were found in ≥2 unrelated individuals (referred to below as non-singleton CNVs), as well as at least 11 singleton variants (**Fig. 1a**, **Supplementary Fig. 3** and **Supplementary Note**). For reference, we label these variants by copy number type (DEL for deletion, DUP for duplication), and number them in order of frequency. To validate the CNV calls we analysed transmission in family trios and observed segregation as expected with few exceptions (**Supplementary Table 4** and **Supplementary Note**). We also compared the CNV calls with the 1000 Genomes Project structural variant analysis [15], and found highly consistent copy number inference (98.8% of individuals have the same copy number call) but improved resolution of overlapping variants in our analysis (**Supplementary Fig. 4** and **Supplementary Note**). Validation of the breakpoint of the most common variant (DEL1) by Sanger sequencing further confirmed the accuracy of the calls made by our HMM (**Supplementary Note**).

The variants ranged in length from 3.2 kb (the minimum possible with our method) to >200 kb and included deletions and duplications of entire genes. Loss of *GYPB* was a common feature, with five different forms of *GYPB* deletion among the non-singleton CNVs (**Fig. 1a**). Hybrid gene structures were another common feature, with two non-singleton CNVs predicted to generate *GYPB-A* or *GYPE-A* hybrids (**Supplementary Fig. 5**). Some variants are predicted to correspond to known MNS blood group antigens while others have not previously been reported (**Supplementary Note**). Of the non-singleton CNVs, half (8/16) had a single pair of breakpoints in homologous parts of the segmental duplication, consistent with formation via non-allelic homologous recombination (NAHR; **Fig. 1c**). Of these, four share a breakpoint position, which coincides with a double-strand break (DSB) hotspot active in a PRDM9 C allele carrier ([16]; **Fig. 1c** and **Supplementary Fig. 6**).

CNVs in the glycophorin region were observed more frequently in Africa than other parts of the world (**Fig. 1d**). In this dataset, the combined frequency of glycophorin CNVs in African populations was 11% compared to 1.1% in non-African populations, and most of the non-singleton CNVs (13/16) were found in individuals of African ancestry. Among fourteen different ethnic groups sampled in Africa, the estimated

frequency ranged from 4.7-21% with the highest frequencies in west and central African populations.

Association with severe malaria

We sought to incorporate CNVs into the phased reference panel with the aim of imputing into our GWAS dataset. Computational phasing of CNVs is challenging, as published methods do not model CNV mutational mechanisms or non-diploid copy number at smaller variants within CNVs. To work around this, we excluded SNPs and short indels within the glycophorin region and relied on the trio structure of sequenced individuals to resolve haplotype phase between CNVs and flanking SNPs (**Methods**, **Supplementary Figs. 7 and 8**) [11]. We evaluated how well CNVs can be imputed by comparing CNV genotype calls from the HMM with re-imputed genotype probabilities in the 1046 unrelated African individuals in our reference panel, in a cross-validation framework (**Methods**). This analysis predicts reasonable imputation performance for the three highest frequency CNVs, DEL1, DEL2, and DUP1, as well as for DUP4 and the rarer DEL7 (correlation between true and re-imputed genotypes > 0.8; **Supplementary Fig. 9**). The combined reference panel, which may be of use for imputation in other studies, is available online (**Online Resources**).

We used this panel to impute CNVs into the severe malaria GWAS samples and tested for association as before (**Methods**). One of the imputed CNVs, DUP4, is associated with decreased risk of severe malaria (odds ratio, OR=0.59; 95% CI 0.48-0.71, $P = 7.4\text{x}10^{-8}$ using an additive model with fixed-effect meta-analysis across populations; **Fig. 2a**). Across populations, evidence for association at DUP4 is among the strongest of any variant in our data. Moreover, conditioning on the imputed genotypes at DUP4 in the statistical association model removes signal at all other strongly associated variants including the previously reported markers of association (e.g., $P_{conditional}$=0.54 at rs186873296; **Fig. 2f, Supplementary Fig. 10**). DUP4 has an estimated heterozygous relative risk of 0.61 (95% CI 0.50-0.75) and its genetic effect appears to be consistent with an additive model, although the low frequency of homozygotes makes it difficult to distinguish the extent of dominance (homozygous relative risk 0.31; 95% CI 0.09-1.06; n=23 homozygotes). Analysis of different clinical forms of severe malaria showed that DUP4 reduced the risk of both cerebral malaria and severe malarial anaemia to a similar degree (**Supplementary Table 5**). While we noted some evidence of additional associations in the region, including a possible protective effect of DEL2 (OR=0.61; 95%CI=0.41-0.92, P=0.02),

these results are compatible with a primary signal of association that is well explained by an additive effect of DUP4.

DUP4 is imputed with high confidence in both east African populations (**Fig. 2d**), where it is at substantially higher frequency than in the reference panel (**Supplementary Fig. 11**). To independently confirm the imputed DUP4 genotypes, we analysed SNP microarray data for intensity patterns indicative of copy number variation (**Fig. 3a**) using a Bayesian clustering model (**Supplementary Note**). This method correctly classifies the sequenced DUP4 carriers, and classification of GWAS samples was highly concordant with the imputed DUP4 genotypes in the east African populations (r^2=0.97 in Kenya; r^2=0.88 in Malawi). Surprisingly, both imputation and the microarray intensity analysis suggest there may be no copies of DUP4 present among the 4791 Gambian individuals in the GWAS. This large frequency difference places DUP4 as an outlier compared with imputed variants at a similar frequency in the Gambia or in Kenya genome-wide (empirical P=1.7x10^{-3} and P=5x10^{-3}, respectively; **Fig. 4a; Methods**). Computation of haplotype homozygosity (**Fig. 4c**) provides evidence that DUP4 is carried on an extended haplotype (empirical P=9.6x10^{-3} for iHS [17, 18] compared with variants of similar frequency genome-wide; **Fig. 4d**) that may have risen to its current frequency in Kenya relatively recently. We note that DUP4 is also absent from all but two of the reference panel populations (**Fig. 1d**).

The physical structure of DUP4

The copy number profile of DUP4 is complex, with a total of six copy number changes that cannot have arisen by a single unequal crossover event from reference-like sequences (**Fig. 1a** and **Fig. 3b**). At the gene level, this copy number profile corresponds to duplication of *GYPE*, deletion of the 3' end of *GYPB*, duplication of the 5' end of *GYPB* and triplication of the 3' end of *GYPA*. To begin to understand the functional consequences of DUP4, we sought to reconstruct the physical arrangement of this variant by pooling data across the nine carriers in the sequenced reference panel (eight Wasambaa individuals from Tanzania including three parent-child pairs, and a single African Caribbean individual from Barbados). First, analysis of coverage along a multiple sequence alignment of the segmental duplication corroborated the location of the six copy number changes from the HMM, with two pairs of breakpoints at homologous locations in the alignment (**Supplementary Fig. 12**).

Next, we looked for sequenced read pairs spanning CNV breakpoints, which provide direct evidence of the structure of the underlying DNA. We identified read pairs that were mapped near breakpoints but with discordant positions (MQ>=1, absolute insert size > 1000 bp), including longer read data we generated for the 1000 Genomes individual who carries DUP4 (HG02554; 300 bp PE reads on Illumina MiSeq to 13x coverage; **Methods** and **Online Resources**). Discordant read pairs supported the connection between each pair of homologous breakpoints as well as between the remaining two breakpoints, which lie in non-homologous sequence (**Fig. 5a** and **Supplementary Fig. 13**). Based on the combined evidence from copy number changes, discordant read pairs, and homology between inferred breakpoints, we generated a model of the DUP4 chromosome that contains five glycophorin genes (**Fig. 5b**).

A prominent functional change on this structure is the presence of two *GYPB-A* hybrid genes, supported by several read pairs within intron 4 of *GYPA* and *GYPB* and the copy number profile. We confirmed the hybrid sequence by PCR-based Sanger sequencing of a 4.1 kb segment spanning the breakpoint (**Fig. 5b**, **Methods** and **Supplementary Note**). These data localize the breakpoint to a 184 bp section of *GYPA* and *GYPB* where the two genes have identical sequence (**Supplementary Figs. 13** and **14)**. If translated, the encoded protein would join the extracellular domain of GYPB to the transmembrane and intracellular domains of GYPA, creating a peptide sequence at their junction that is characteristic of the Dantu antigen in the MNS blood group system (**Fig. 5d**) [19, 20]. Moreover, like DUP4, the most common Dantu variant (termed NE type, here referred to as Dantu NE) is reported to have two such hybrid genes and lack a full *GYPB* gene [21]. We sequenced genomic DNA from an individual serologically determined to be Dantu positive, and of NE type (150 bp PE reads on Illumina HiSeq to 18x coverage; **Methods** and **Online Resources**) and analysed it using our HMM. The coverage profile and HMM-inferred copy number path, indistinguishable from those of DUP4 carriers, confirm identification of DUP4 as the molecular basis of Dantu NE (**Fig. 5c**).

To our knowledge, the full structure of this Dantu variant has not been previously reported [6]. In addition to duplicate *GYPB-A* hybrid genes, these data reveal a duplicated copy of *GYPE* and the precise location of six breakpoints. Either complex mutational events or a series of at least four unequal crossover events are needed to account for the formation of this variant (confirmed by simulation; **Methods** and **Supplementary Fig. 15**). However, we find no potential intermediates and no

obvious relationship between DUP4 and other structural variant haplotypes in the present dataset (**Supplementary Fig. 7** and **Supplementary Note**). Further analysis of discordant read pairs identifies a number of shorter discrepancies relative to the reference sequence that are consistent with gene conversion events (**Supplementary Fig. 16**) and could be functionally relevant (e.g., **Supplementary Fig. 17**).

Discussion

Here we use whole-genome sequence data in combination with novel statistical methods to identify at least 27 CNVs in the glycophorin region that segregate in global populations. In this study, 14% of sub-Saharan African individuals carry a variant that affects the genic copy number relative to the reference assembly. Our description of these variants from genome sequence data complements the existing literature on antigenic variation associated with the MNS blood group system and offers novel insights. We find, for example, that the frequency of *GYPB* deletion is broadly commensurate with previous surveys of the S−s−U− blood group phenotype linked to absence of the GYPB protein, but the *GYPB* deletions in our data differ from the reported molecular variant (**Supplementary Note**; [6, 22-24]).

Of the array of glycophorin CNVs identified, one (DUP4) is associated with resistance to severe malaria and fully explains the previously reported signal of association [1]. While there may be other functional mutations on this haplotype, we propose that the direct consequences of this rearrangement are likely to drive the underlying causal mechanism for resistance to severe malaria. DUP4 was not present in the 1000 Genomes Phase 1 reference panel (used in [1]), and exists as a singleton in the 1000 Genomes Phase 3 reference panel. Thus, as previously observed at the sickle cell locus [25], mapping of the association signal by imputation was only possible with the inclusion of additional individuals in the reference panel.

Through additional sequencing, we have shown that DUP4 corresponds to the variant encoding the Dantu+ (NE type) blood group phenotype, thus linking the hybrid genes predicted in the reconstructed sequence to a serologically distinct hybrid protein that is expressed on the red blood cell [20, 26]. The few existing studies of Dantu+ (NE type) erythrocytes indicate high levels of the hybrid GYPB-A protein and lower levels of GYPA than wild type cells [26, 27] and a single study reports parasite growth to be impaired [28]. Dantu NE is one of the many glycophorin

variants that have been hypothesized to influence malaria susceptibility, but the results here are the first evidence of a protective effect in natural populations.

These findings then raise the question of how DUP4 protects against malaria. GYPA and GYPB are exclusively expressed on the erythrocyte surface and are targeted by parasites during invasion [3, 29]. *P. falciparum* EBA175 binds to the extracellular portion of GYPA [30], which is preserved in DUP4. *P. falciparum* EBL1 binds to the extracellular portion of GYPB [31] which is duplicated in DUP4 but joined to intracellular GYPA. The significance of the extra copy of *GYPE* or the absence of full *GYPB* in DUP4 is uncertain, since *GYPE* is not known to be expressed at the protein level [2, 32], and there is no evidence that absence of *GYPB* alone confers protection (**Fig. 2a**). GYPA and GYPB are known to form homodimers as well as heterodimers in the red cell membrane [24], so these copy number changes could have complex functional effects. There are physical interactions between GYPA and band 3 (encoded by *SLC4A1*) at the red cell surface [33] and parasite binding to GYPA appears to initiate a signal leading to increased membrane rigidity [34]. Thus the GYPB-A hybrid proteins seen in DUP4 could potentially affect both receptor-ligand interactions and the physical properties of the red cell membrane during the process of parasite host cell invasion.

Previous surveys of the Dantu blood group antigen have indicated that it is rare ([24, 35-37]; **Supplementary Table 6**). We find that DUP4 is absent or at very low frequency outside parts of East Africa, with a frequency difference and extended haplotype consistent with a recent rise in frequency in Kenya. In contrast, the malaria-protective variant causing sickle-cell anaemia (rs334 in *HBB*), which is thought to be under balancing selection, has a similar frequency in both the Gambia and Kenya (**Fig. 4a**). One possibility for why DUP4 is not more widespread, given its strong protective effect against malaria, is that it has arisen only very recently without time for gene flow to facilitate its dispersion. Alternatively, this frequency distribution may also be consistent with balancing selection, for example if it protects only against certain strains of *P. falciparum* that are specific to East Africa. The glycophorin region is near a signal of long-term balancing selection, and measures of polymorphism in both the human glycophorins and *P. falciparum* EBA175 have been suggestive of diversifying selection [7-9, 38, 39]. Although apparently not directly related to these signals, current selection on DUP4 may represent a snapshot of the long-term evolutionary processes acting at this locus. Mapping the allele frequency of DUP4 across additional populations may help clarify the nature of selection.

Recent GWAS have confirmed three other loci associated with severe malaria (*HBB*, *ABO*, *ATP2B4*), all of which are also related to red blood cell function [1, 40]. However, this association with *GYPA* and *GYPB* stands out by directly involving variation in invasion receptors. In the case of *P. vivax,* it is well known that removal of the Duffy receptor from the red blood cell prevents an essential interaction with the parasite to confer strong protection against infection [41]. For *P. falciparum*, although one essential interaction is known [42], GYPA and GYPB are among several alternative red cell surface receptors that *P. falciparum* can use to gain entry into the cell and so are not thought to be required [43]. Intriguingly, there is marked variation among *P. falciparum* strains in preference for different invasion pathways in vitro [29]; field studies that account for parasite heterogeneity and tests for genetic interactions may be important in determining whether DUP4 has a direct effect on parasite invasion. While the precise functional mechanism remains to be determined, the findings presented here show that evolution has hit upon a way to alter these invasion receptors for *P. falciparum* to resist severe infection. Further work to understand this mechanism may lead us toward new parasite vulnerabilities that could be utilized in future interventions against this deadly disease.

Figure 1. Copy number variants in the glycophorin region. (a) Sequence coverage in 1600 bp windows and copy number for CNVs identified in ≥2 unrelated individuals. Black dashes show the mean sequence coverage across heterozygous carriers, normalized by the mean coverage of those individuals outside this region and the bin-specific factor for each window (**Methods**). Individuals carrying multiple variants were excluded from the computation and only bins with at least 25% mappable sites are displayed, as were input to the HMM. A horizontal gray line indicates the expected coverage without copy number variation The inferred CNVs are indicated with deletion in yellow, duplication in light blue, and triplication in dark blue. Blue vertical lines mark the locations of the three genes. **(b)** Mappability in the glycophorin region. Rows show: (i) the number of mappable sites in the 1600 bp windows used for copy number inference (windows with less than 25% mappable sites not shown); (ii) the maximum identity with the homologous locations in the segmental duplication in the same 1600 bp windows, as inferred from a multiple sequence alignment, with mean of 0.96 indicated with a red dashed line; (iii) sequences of at least 100 bp that are unique to one (black) or two (grey) out of the three segmentally duplicated units; (iv) protein-coding genes; (v) location of the segmentally duplicated units. **(c)** Positions of breakpoints of variants shown in **(a)**, coloured as the variant names in **(a)** and shaded by whether the variant has a single pair of homologous breakpoints, a single pair of non-homologous breakpoints, or is a multi-segment CNV. The recombination rate from LD-based recombination maps [44, 45] and locations of DSB hotspots [16] are annotated below. **(d)** Frequency of each CNV in the sampled populations, estimated from unrelated individuals. Populations are grouped based on geographical proximity; population abbreviations can be found in **Supplementary Tables 1** and **3**.

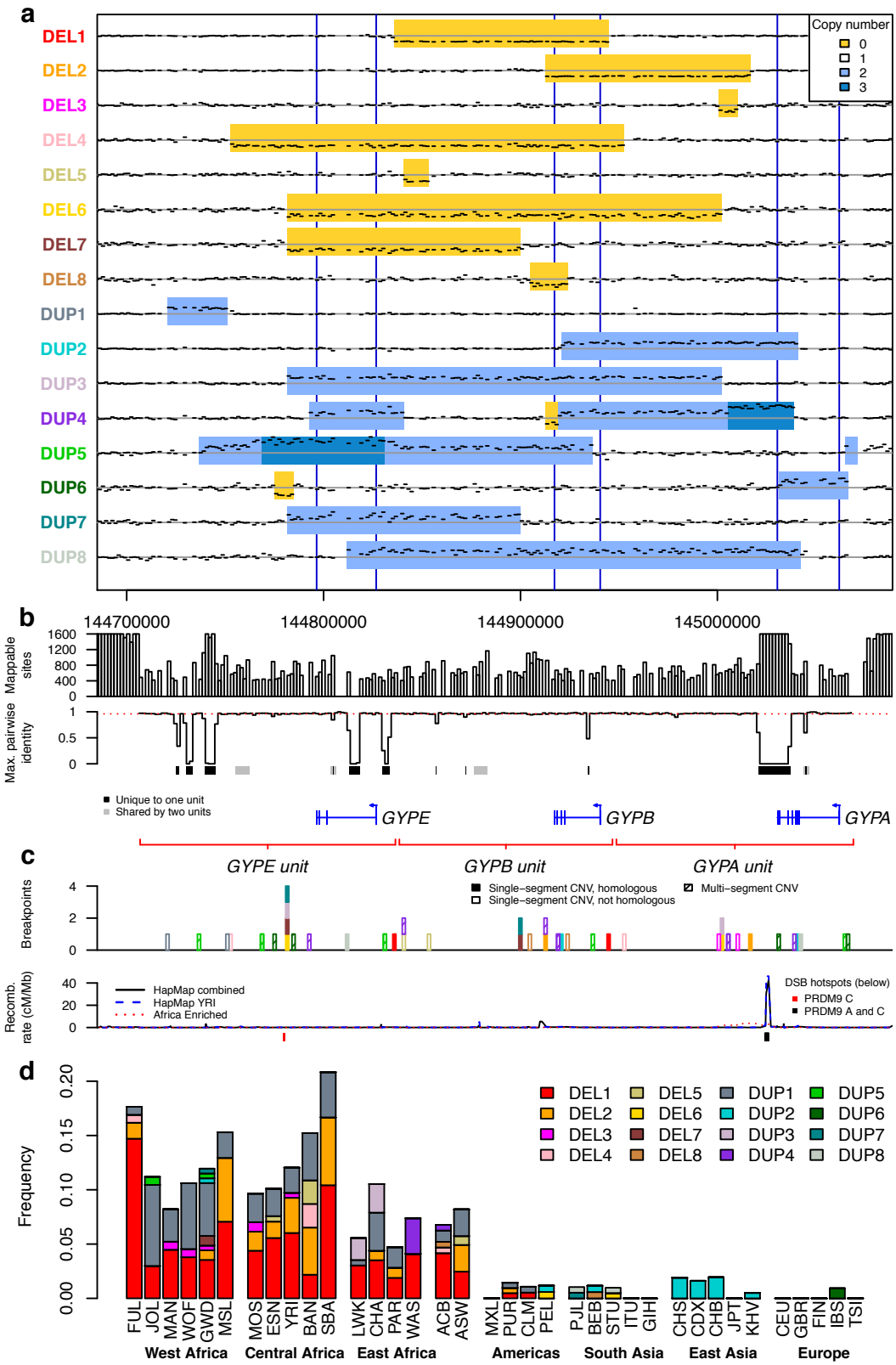

a
DEL1
DEL2
DEL3
DEL4
DEL5
DEL6
DEL7
DEL8
DUP1
DUP2
DUP3
DUP4
DUP5
DUP6
DUP7
DUP8
Copy number
0
1
2
3
144700000
144800000
144900000
145000000
b
Mappable sites
Max. pairwise identity
Unique to one unit
Shared by two units
GYPE
GYPB
GYPA
c
GYPE unit
GYPB unit
GYPA unit
Breakpoints
Single−segment CNV, homologous
Single−segment CNV, not homologous
Multi−segment CNV
Recomb. rate (cM/Mb)
HapMap combined
HapMap YRI
Africa Enriched
DSB hotspots (below)
PRDM9 C
PRDM9 A and C
d
Frequency
DEL1
DEL2
DEL3
DEL4
DEL5
DEL6
DEL7
DEL8
DUP1
DUP2
DUP3
DUP4
DUP5
DUP6
DUP7
DUP8
FUL JOL MAN WOF GWD MSL
MOS ESN YRI BAN SBA
LWK CHA PAR WAS
ACB ASW
MXL PUR CLM PEL
PJL BEB STU ITU GIH
CHS CDX CHB JPT KHV
CEU GBR FIN IBS TSI
West Africa
Central Africa
East Africa
Americas
South Asia
East Asia
Europe

Figure 2. Evidence of association. (a) The evidence for association (-log10 *P*-value for association test, y axis) for SNPs, short indels, and CNVs across the glycophorin region (x axis). *P*-values are computed by meta-analysis across three African populations under an additive model of association. Points are coloured by LD with DUP4, as computed in east African reference panel populations. Directly typed SNPs are denoted with black plusses, and DEL1, DEL2, DUP1 and DUP4 are denoted with diamonds. Black triangles represent SNPs where the association signal was previously reported and replicated in further samples. **(b)** The copy number of DEL1, DEL2, DUP1 and DUP4, as in **Fig. 1a. (c)** Protein-coding genes in the region. **(d)** Comparison of expected imputed allele frequency (x axis) and IMPUTE info score (y axis), for the four annotated CNVs. **(e)** Detail of the evidence for association at DUP4. Coloured circles and text show the estimated allele frequency of DUP4 in population controls and severe malaria cases in each population. To the right is the odds ratio and 95% confidence interval for DUP4 heterozygotes (diamonds) and homozygotes (circles) relative to non-carriers. The bottom two rows represent effect sizes in a fixed-effect meta-analysis. Sample sizes (number of controls/number of cases) are denoted to the left. **(f)** Comparison of association test *P*-values conditioning on five principal components (y axis, as in panel **(a)**), and additionally conditioning on genotypes at DUP4 (x axis) for all variants in panel **(a)**.

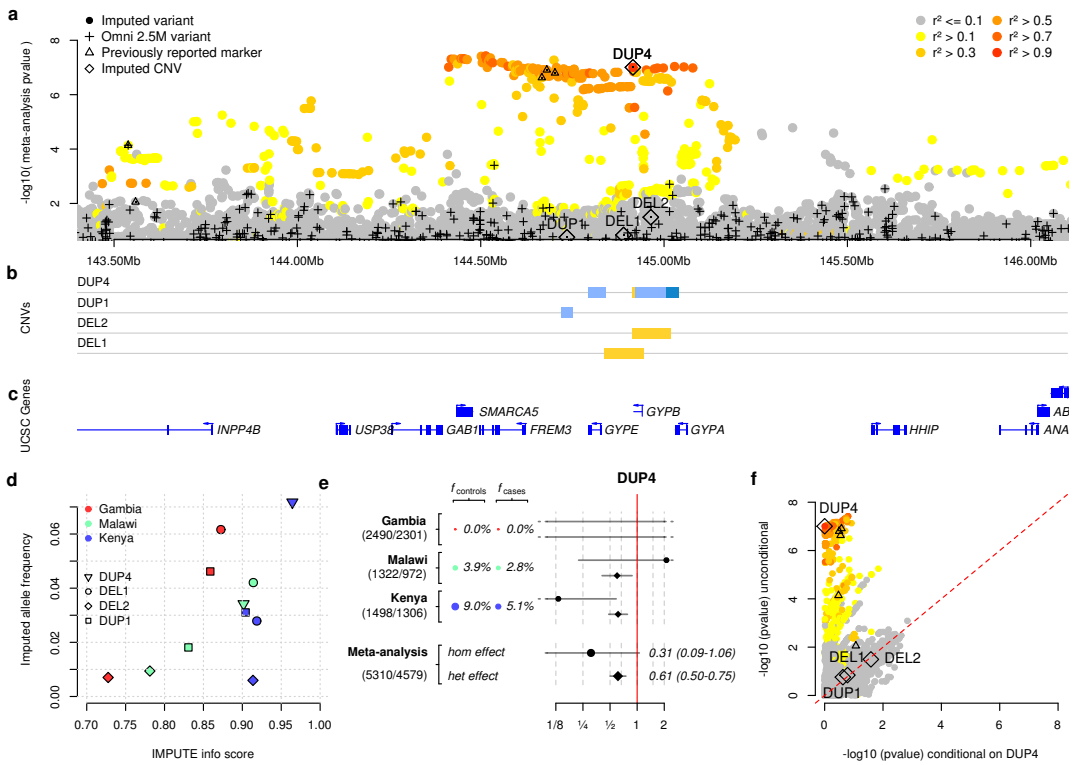


Figure 3. The effect of DUP4 on SNP array intensities. (a) Normalised Illumina Omni 2.5M intensity values at selected SNP assays across the glycophorin region for reference panel individuals (top row; N=367 individuals from Burkina Faso, Cameroon, and Tanzania) and Kenyan GWAS individuals (second row, N = 3,142). Blue and yellow points represent individuals heterozygous or homozygous for DUP4 respectively, as determined by the HMM in reference panel individuals and by imputation in Kenya (genotypes with posterior probability at least 0.75). Arrows denote the mapping location of these SNPs. **(b)** HMM path for a single DUP4 individual. **(c)** Position of the glycophorin genes and exons.

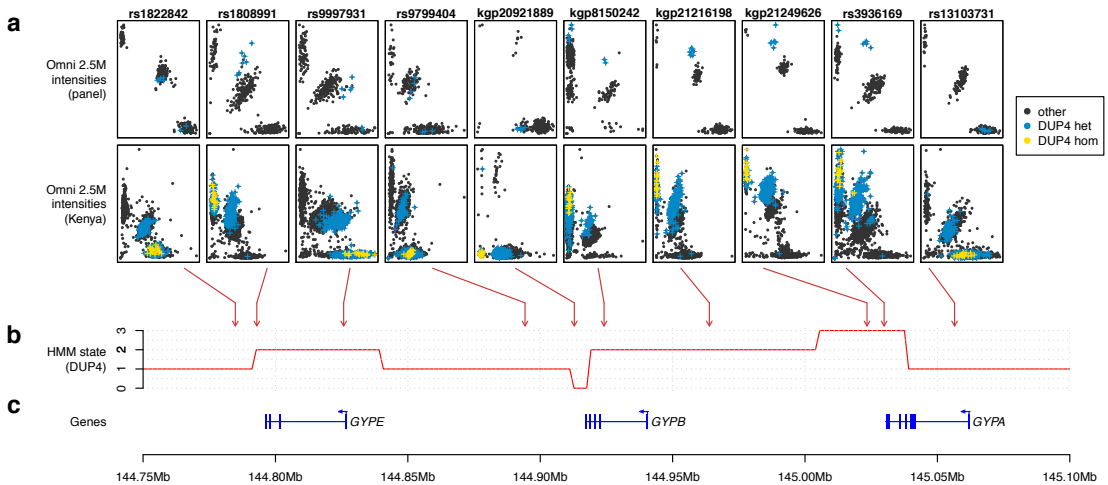


Figure 4. DUP4 frequency and haplotype homogyzosity. (a) The empirical joint allele-frequency spectrum for population controls in the Gambia (x axis) and Kenya (y axis) in 0.5% frequency bins between 0 and 20% based on genome-wide imputation reported previously [1]. Bins are coloured according to the number of variants, and the frequencies of DUP4 and rs334 are highlighted. The histograms show the marginal distribution of all SNPs in the Gambia in the same frequency bin as DUP4 in Kenya (8.5-9%, top) and the marginal distribution of all SNPs in Kenya in the same frequency bin as DUP4 in the Gambia (0-0.5%, right). **(b)** The estimated frequency of DUP4 in East African populations, shown as point estimates with 95% confidence intervals. The number of haplotypes sampled is given in brackets. Estimates are from population controls in the GWAS or, indicated by a dagger, from the HMM genotype calls in the reference panel. The overall frequency of DUP4 in the Kenyan controls (f_{Kenya} = 0.09) is shown as a dotted vertical line. **(c)** Extended haplotype homozygosity (EHH) computed outward from the glycophorin region for DUP4 haplotypes (purple) and non-DUP4 haplotypes (red) in Kenya, after excluding other variants within the glycophorin region. The distribution of unstandardised iHS for all typed SNPs within 1% frequency of DUP4 in Kenyan controls. The position of DUP4 in this distribution is denoted by a red line, with empirical *P*-value annotated. **(d)** Recombination maps [44, 45] across the 142.5-147.5Mb region of chromosome 4. **(e)** The 272 haplotypes imputed to carry DUP4 in Kenya, clustered on 1 Mb extending in either direction from the glycophorin region, which is shaded in blue and was excluded from this analysis. The bar on the left depicts the population for each haplotype with colours as in panel **(b)**. **(f)** A random sample of 272 non-DUP4 haplotypes clustered on the same region. **(g)** Protein-coding genes.

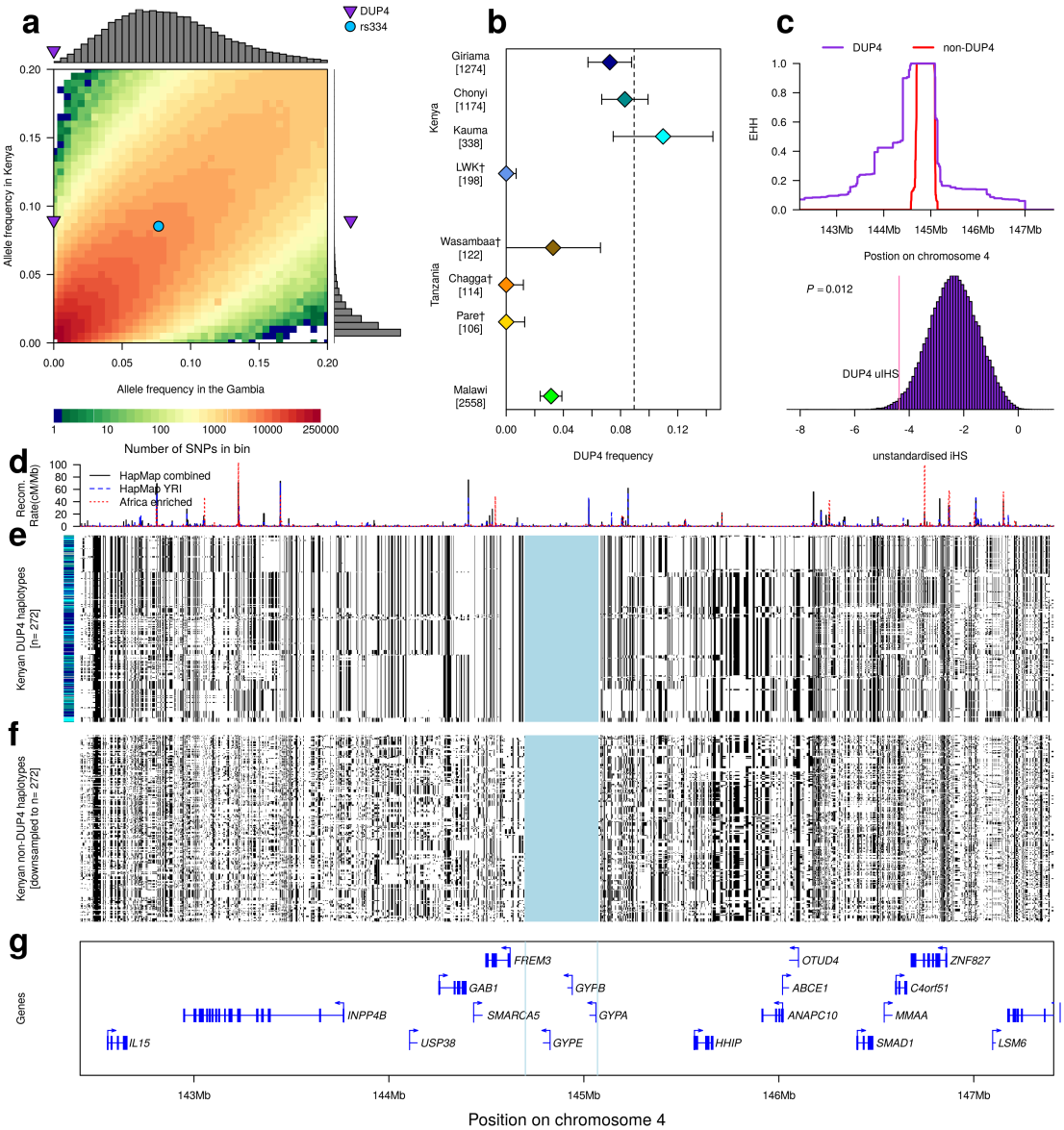
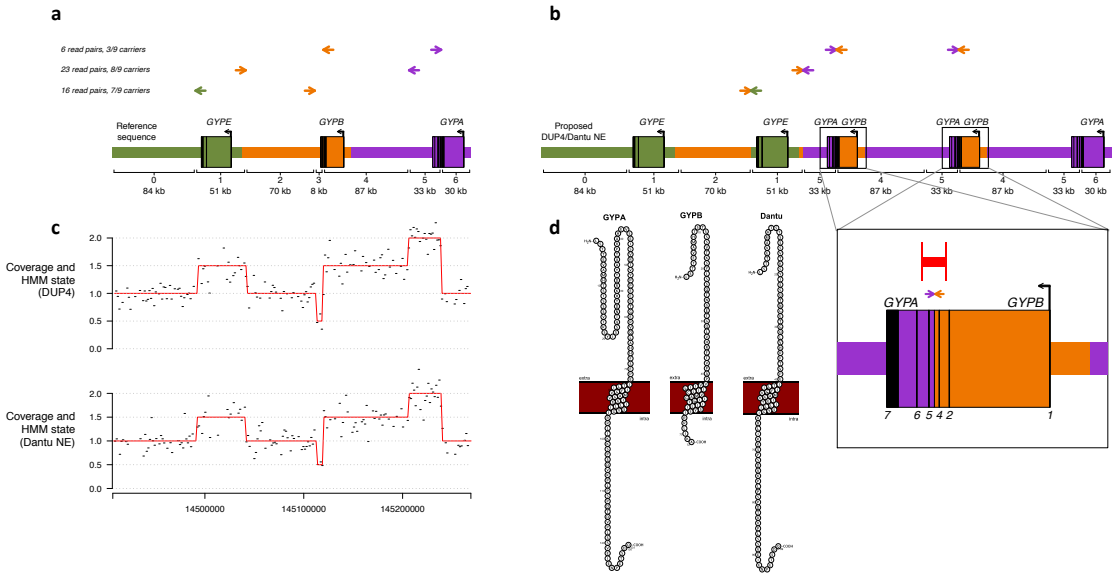

a
DUP4
rs334
Allele frequency in Kenya
Allele frequency in the Gambia
Number of SNPs in bin
b
Kenya
Tanzania
Giriama [1274]
Chonyi [1174]
Kauma [338]
LWK† [198]
Wasambaa† [122]
Chagga† [114]
Pare† [106]
Malawi [2558]
DUP4 frequency
c
DUP4
non-DUP4
EHH
Postion on chromosome 4
P = 0.012
DUP4 uIHS
unstandardised iHS
d
Recom. Rate(cM/Mb)
HapMap combined
HapMap YRI
Africa enriched
e
Kenyan DUP4 haplotypes [n= 272]
f
Kenyan non-DUP4 haplotypes [downsampled to n= 272]
g
Genes
FREM3
GAB1
GYPB
INPP4B
SMARCA5
GYPA
IL15
USP38
GYPE
HHIP
OTUD4
ZNF827
ABCE1
C4orf51
ANAPC10
MMAA
SMAD1
LSM6
Position on chromosome 4

Figure 5. The structure of DUP4. (a) Discordant read pairs mapped near DUP4 copy number changes. Coloured arrows represent read pairs from heterozygous DUP4 carriers, with paired reads shown on the same horizontal line and the direction of the arrows depicting the strand and position as mapped to the human reference sequence. The number of such read pairs and distinct carriers in which they are found is given to the left. A schematic of the reference sequence is below with colors indicating the segmentally duplicated units. Brackets delineate the segments with different copy number in DUP4, which are numbered and labeled with their length to the nearest kb. **(b)** The structure of DUP4, inferred by connecting sequence at breakpoints based on sequence homology and discordant read pairs. Arrows depict the concordant positions of the read pairs in **(a)** on this structure, and the order of reference segments is shown below. Inset: detail of the inferred *GYPB-A* hybrid genes, indicating the positions of discordant read pairs (arrows), PCR primers (vertical red lines) and the resulting product (horizontal red line). **(c)** Normalized coverage in 1600 bp windows (black) and HMM path (red) for a DUP4 carrier (top) and for an individual serotyped as Dantu+ (NE type; bottom), on the same x axis as **(a)**. **(d)** Protein sequences of GYPA (NP_002090), GYPB (NP_002091), and the Dantu hybrid within the cell membrane depicting the extracellular, transmembrane, and intracellular domains as visualised with protter [46].



References

1. Malaria Genomic Epidemiology Network, et al., *A novel locus of resistance to severe malaria in a region of ancient balancing selection.* Nature, 2015. **526**(7572): p. 253-7.
2. Cartron, J.-P. and P. Rouger, *Molecular basis of human blood group antigens*. Blood cell biochemistry. 1995, New York: Plenum Press. xx, 492 p.
3. Satchwell, T.J., *Erythrocyte invasion receptors for Plasmodium falciparum: new and old.* Transfus Med, 2016. **26**(2): p. 77-88.
4. Bhatt, S., et al., *The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015.* Nature, 2015. **526**(7572): p. 207-11.
5. Patnaik, S.K., W. Helmberg, and O.O. Blumenfeld, *BGMUT Database of Allelic Variants of Genes Encoding Human Blood Group Antigens.* Transfus Med Hemother, 2014. **41**(5): p. 346-51.
6. Blumenfeld, O.O. and C.H. Huang, *Molecular genetics of glycophorin MNS variants.* Transfus Clin Biol, 1997. **4**(4): p. 357-65.
7. Baum, J., R.H. Ward, and D.J. Conway, *Natural selection on the erythrocyte surface.* Mol Biol Evol, 2002. **19**(3): p. 223-9.
8. 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, 2011. **88**(6): p. 741-54.
9. Wang, H.Y., et al., *Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites.* Mol Biol Evol, 2003. **20**(11): p. 1795-804.
10. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* Am J Hum Genet, 2007. **81**(5): p. 1084-97.
11. Delaneau, O., J.F. Zagury, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* Nat Methods, 2013. **10**(1): p. 5-6.
12. Garrison, E. and G. Marth *Haplotype-based variant detection from short-read sequencing*. ArXiv e-prints, 2012. **1207**.
13. The 1000 Genomes Project Consortium, et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
14. Derrien, T., et al., *Fast computation and applications of genome mappability.* PLoS One, 2012. **7**(1): p. e30377.
15. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.
16. Pratto, F., et al., *DNA recombination. Recombination initiation maps of individual human genomes.* Science, 2014. **346**(6211): p. 1256442.
17. Voight, B.F., et al., *A map of recent positive selection in the human genome.* PLoS Biol, 2006. **4**(3): p. e72.
18. Sabeti, P.C., et al., *Detecting recent positive selection in the human genome from haplotype structure.* Nature, 2002. **419**(6909): p. 832-7.
19. Dahr, W., et al., *Hybrid glycophorins from human erythrocyte membranes. I. Isolation and complete structural analysis of the hybrid sialoglycoprotein*

from Dantu-positive red cells of the N.E. variety. Eur J Biochem, 1987. **166**(1): p. 31-6.
20. Blumenfeld, O.O., A.J. Smith, and J.J. Moulds, *Membrane glycophorins of Dantu blood group erythrocytes.* J Biol Chem, 1987. **262**(24): p. 11864-70.
21. Huang, C.H. and O.O. Blumenfeld, *Characterization of a genomic hybrid specifying the human erythrocyte antigen Dantu: Dantu gene is duplicated and linked to a delta glycophorin gene deletion.* Proc Natl Acad Sci U S A, 1988. **85**(24): p. 9640-4.
22. Rahuel, C., et al., *Erythrocyte glycophorin B deficiency may occur by two distinct gene alterations.* Am J Hematol, 1991. **37**(1): p. 57-8.
23. Lowe, R.F. and P.P. Moores, *S-s-U-red cell factor in Africans of Rhodesia, Malawi, Mozambique and Natal.* Hum Hered, 1972. **22**(4): p. 344-50.
24. Daniels, G., *Human blood groups : Geoff Daniels ; foreword to first edition by Ruth Sanger*. 3rd ed. 2013, Chichester, West Sussex: John Wiley & Sons. ix, 544 p.
25. Jallow, M., et al., *Genome-wide and fine-resolution association analysis of malaria in West Africa.* Nat Genet, 2009. **41**(6): p. 657-65.
26. Dahr, W., et al., *The Dantu erythrocyte phenotype of the NE variety. I. Dodecylsulfate polyacrylamide gel electrophoretic studies.* Blut, 1987. **55**(1): p. 19-31.
27. Merry, A.H., et al., *The use of monoclonal antibodies to quantify the levels of sialoglycoproteins alpha and delta and variant sialoglycoproteins in human erythrocyte membranes.* Biochem J, 1986. **233**(1): p. 93-8.
28. Field, S.P., et al., *Glycophorin variants and Plasmodium falciparum: protective effect of the Dantu phenotype in vitro.* Hum Genet, 1994. **93**(2): p. 148-50.
29. Wright, G.J. and J.C. Rayner, *Plasmodium falciparum erythrocyte invasion: combining function with immune evasion.* PLoS Pathog, 2014. **10**(3): p. e1003943.
30. Orlandi, P.A., F.W. Klotz, and J.D. Haynes, *A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of Plasmodium falciparum recognizes the terminal Neu5Ac(alpha 2-3)Gal- sequences of glycophorin A.* J Cell Biol, 1992. **116**(4): p. 901-9.
31. Mayer, D.C., et al., *Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1.* Proc Natl Acad Sci U S A, 2009. **106**(13): p. 5348-52.
32. Rahuel, C., J.F. Elouet, and J.P. Cartron, *Post-transcriptional regulation of the cell surface expression of glycophorins A, B, and E.* J Biol Chem, 1994. **269**(52): p. 32752-8.
33. Huang, C.H., et al., *Human red blood cell Wright antigens: a genetic and evolutionary perspective on glycophorin A-band 3 interaction.* Blood, 1996. **87**(9): p. 3942-7.
34. Chasis, J.A., et al., *Signal transduction by glycophorin A: role of extracellular and cytoplasmic domains in a modulatable process.* J Cell Biol, 1988. **107**(4): p. 1351-7.
35. Contreras, M., et al., *Serology and genetics of an MNSs-associated antigen Dantu.* Vox Sang, 1984. **46**(6): p. 377-86.
36. Moores, P., E. Smart, and I. Marais, *The Dantu Phenotype in Southern Africa.* Transfus Med, 1992. **2**(Suppl. 1): p. 68.

37. Unger, P., et al., *The Dantu erythrocyte phenotype of the NE variety. II. Serology, immunochemistry, genetics, and frequency.* Blut, 1987. **55**(1): p. 33-43.
38. Leffler, E.M., et al., *Multiple instances of ancient balancing selection shared between humans and chimpanzees.* Science, 2013. **339**(6127): p. 1578-82.
39. Verra, F., et al., *Contrasting signatures of selection on the Plasmodium falciparum erythrocyte binding antigen gene family.* Mol Biochem Parasitol, 2006. **149**(2): p. 182-90.
40. Timmann, C., et al., *Genome-wide association study indicates two novel resistance loci for severe malaria.* Nature, 2012. **489**(7416): p. 443-6.
41. Langhi, D.M., Jr. and J.O. Bordin, *Duffy blood group and malaria.* Hematology, 2006. **11**(5): p. 389-98.
42. Crosnier, C., et al., *Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum.* Nature, 2011. **480**(7378): p. 534-7.
43. Hadley, T.J., et al., *Falciparum malaria parasites invade erythrocytes that lack glycophorin A and B (MkMk). Strain differences indicate receptor heterogeneity and two pathways for invasion.* J Clin Invest, 1987. **80**(4): p. 1190-3.
44. International HapMap Consortium, et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
45. Hinch, A.G., et al., *The landscape of recombination in African Americans.* Nature, 2011. **476**(7359): p. 170-5.
46. Omasits, U., et al., *Protter: interactive protein feature visualization and integration with experimental proteomic data.* Bioinformatics, 2014. **30**(6): p. 884-6.

Acknowledgements

We thank all the study participants and the members of the MalariaGEN consortium. A list of researchers involved at each study site can be found at https://www.malariagen.net/projects/consortial-project-1/malariagen-consortium-members.

The MalariaGEN Project is supported by the Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundations of the National Institutes of Health (566) as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by the Wellcome Trust (090770/Z/09/Z). This research was supported by the Medical Research Council (G0600718; G0600230; MR/M006212/1). Chris C.A. Spencer was supported by a Wellcome Trust Career Development Fellowship (097364/Z/11/Z). The Wellcome Trust also provides core awards to The Wellcome Trust Centre for Human Genetics (090532/Z/09/Z) and the Wellcome Trust Sanger Institute (098051).

Eric Achidi received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 242095 – EVIMalaR and the Central African Network for Tuberculosis, HIV/AIDS and Malaria (CANTAM) funded by the European and Developing Countries Clinical Trials Partnership (EDCTP). Thomas N. Williams is funded by Senior Fellowships from the Wellcome Trust (076934/Z/05/Z and091758/Z/10/Z) and through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 242095 – EVIMalaR. The KEMRI-Wellcome Trust Programme is funded through core support from the Wellcome Trust. Carolyne Ndila is supported through a strategic award to the KEMRI-Wellcome Trust Programme by the Wellcome Trust (084538). Tanzania/KCMC/JMP received funding from MRC grant number (G9901439). The Malawi-Liverpool-Wellcome Trust Clinical Research Programme (MLW) is a Major Overseas Programme of the Wellcome Trust. Malcolm Molyneux was funded by a Wellcome Trust Research Leave Fellowship. Vysaul Nyirongo was supported on the MLW core grant.

We thank the staff of the WTSI Sample Logistics, Genotyping, Sequencing and Informatics facilities and the WTCHG High-Throughput Genomics core for their contributions to sample handling and generation and processing of sequence data.

Methods

1) Sample collection and sequencing 23
A) Collection and sequencing of African individuals 23
B) Sequence data curation 23
C) 1000 Genomes sequence data 23

2) Overview of the glycophorin region 24

3) Construction of a regional reference panel 24
A) Identification of polymorphic sites and genotype likelihood computation 24
B) Genotype calling and phasing 25

4) Imputation and association testing at SNPs and short indels 25
A) Imputation and association testing 25
B) Functional annotation of variants 26

5) Identification of copy number variation 26
A) Method to call copy number variation 26
B) CNV calling in the 3,269 sequenced individuals 27

6) Phasing and imputation of CNVs 27
A) Initial phasing of CNVs 27
B) Haplotype-based curation and re-phasing of CNVs 28
C) Cross-validation of CNV imputation in the reference panel 29
D) Imputation of CNVs 29

7) Analysis of association at CNVs 30
A) Association with severe malaria 30
B) Conditional tests of association 30
C) Association with clinical subtypes 30

8) Population genetic analysis 30
A) Frequency differentiation 30
B) Haplotype homozygosity 31

9) Resolution of the structure of DUP4 31
A) Discordant read pair analysis 31
B) Molecular assays and Sanger sequencing 32
C) Sequencing and copy number analysis of a serologically-typed Dantu+ individual 32
D) Simulations of DUP4 formation 32

10) Online Resources 33

11) Methods references 34

1) Sample collection and sequencing

A) Collection and sequencing of African individuals

Blood samples from a total of 773 healthy individuals from 10 ethnic groups in four countries in sub-Saharan Africa were collected by partners in the MalariaGEN consortium (**Online Resources**) as part of ongoing projects (**Supplementary Fig. 1a** and **Supplementary Tables 1** and **2**). Individuals were from the general population, with most collected in family trios except in Burkina Faso where individuals are unrelated. Genomic DNA was extracted and sequencing was performed on Illumina HiSeq 2000 at the Wellcome Trust Sanger Institute with 100 bp paired end reads to an average of 10x coverage. Reads were mapped to the GRCh37 human reference genome with additional sequences as modified by the 1000 Genomes Project (hs37d5.fa; [13]), using BWA [47] with base quality score recalibration (BQSR) and local realignment around known indels as implemented in GATK [48, 49].

B) Sequence data curation

We used GATK HaplotypeCaller to compute an initial set of genotype likelihoods across samples at a genome-wide set of variants, including polymorphic sites from 1000 Genomes Phase 3. We computed average coverage across the genome for each individual using BEDTools genomecov [50] and excluded seven Bantu individuals from Cameroon with less than 2x coverage across the genome, and one Wollof individual from the Gambia with less than 6x coverage and greater than 10% missing call rate in the GATK analysis. All further analyses described here are based on the 765 non-excluded individuals.

We inferred the sex of sequenced samples based on the ratio of X chromosome coverage to autosomal coverage. To infer family relationships, we used lcMLkin [51] to compute maximum likelihood pairwise kinship estimates from the GATK-estimated genotype likelihoods at a thinned set of ~26,000 SNPs genome-wide. We then ran PRIMUS [52] to infer pedigrees from the kinship estimates and compared the inferred and reported relationships. Based on this we manually curated the family structure of sequenced samples by removing relationships incompatible with trio structure (IBD1 < 0.9 for parent-child relationship), swapping three individuals between trios with clear sample mixups, and exchanging parental labels in two trios to be consistent with the genetic sex of the parents. The curated dataset contains 207 trios, 16 duos, and 115 individuals without nominal close relationships. All trios and duos are unrelated to each other except for one extended family in the Wollof from the Gambia, which consists of a quad (two parents and two children, here encoded as two trios), where one of the children is a parent in an additional trio.

C) 1000 Genomes sequence data

The 2,504 individuals from 26 populations in the 1000 Genomes Phase 3 release [13] were analysed. Bam files containing reads mapped to GRCh37 were

downloaded from the 1000 Genomes FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/; **Supplementary Fig. 1b** and **Supplementary Table 3**).

2) Overview of the glycophorin region

The glycophorin gene cluster on chromosome 4 results from segmental duplication events in the ancestor to African great apes [53] and is not related in sequence to *GYPC* on chromosome 2. We identified the region of segmental duplication, here referred to as the glycophorin region, as chr4:144,706,830-145,069,066, and the three paralogous units of the segmental duplication as: *GYPE* unit, chr4:144,706,830-144,837,481; *GYPB* unit, chr4:144,837,482-144,947,716; *GYPA* unit, chr4:144,947,717-145,069,066. Each gene occupies ~30 kb toward the end of its ~120 kb repeat unit. We generated a multiple sequence alignment of the reference sequence for the three units by running kalign [54] with default parameters, and calculated pairwise identity in 1600 bp windows along this alignment (**Fig. 1b**). This multiple sequence alignment is available in **Online Resources**. The glycophorin genes are transcribed on the negative strand, and we adopt the convention of numbering exons and introns as they occur in the *GYPA* transcript (exon 3 is a pseudoexon in *GYPB* and exons 3 and 4 are pseudoexons in *GYPE*). We focus on the three protein-coding genes, but note that a long noncoding RNA is annotated between *GYPE* and *GYPB* (LOC101927636). The coordinates here, and throughout the paper, are given with respect to GRCh37.

3) Construction of a regional reference panel

A) Identification of polymorphic sites and genotype likelihood computation

To construct a regional reference panel we focused on the 10 Mb region chr4:140-150Mb, including a 500 kb margin at either end. We first assembled a list of previously identified SNPs and short indels from the 1000 Genomes Phase 3 [13], the Illumina Omni 2.5M array, the ExAC project [55] and the European Variation Archive (downloaded on 24th February 2016; www.ebi.ac.uk/eva/), totaling 421,670 variants. We then used freebayes v1.0.2 [12] to calculate genotype likelihoods at these sites as well as at newly identified putatively polymorphic sites, across the 3,269 sequenced individuals. Freebayes was run in 10 kb chunks across the region. We filtered freebayes output to include all previously identified variants and high quality novel variants, i.e., novel variants with quality (QUAL) > 1, presence of supporting reads on both strands (SAF > 0 and SAR > 0) and both sides of the variant (RPL>1 and RPR>1), and high quality per alternate observation (QUAL/AO>10). In total, the filtered output contained 424,909 variants, of which 412,795 were among the previously identified variants.

B) Genotype calling and phasing

We next aimed to generate a high-quality set of genotype calls in the 765 individuals collected by MalariaGEN partners. We focused on sites with combined mean coverage between 7.5x and 13.5x, excluding approximately 4% and 1% of variants with depth below or above this range respectively, and restricted to variants with allele count at least 2, leaving a total of 117,531 variants. We produced an initial set of genotype calls at these variants using BEAGLE 4.0 (beagle.r1399; [10]) without specifying family information. Based on the initial calls, we removed variants with more than five mendelian errors in the 207 trios, strong evidence for deviation from Hardy-Weinberg equilibrium (P_{HWE} < 1x10^{-3} in the 542 individuals without parents in the dataset) or more than one alternate allele. We then re-ran BEAGLE on the remaining set of 111,167 variants, including family information, to produce genotypes calls. We phased these genotypes using SHAPEIT, specifying 400 selected and 200 random conditioning states, an effective population size of 17,469, and 10 burn-in, 8 pruning and 50 main iterations, and including trio information.

Finally, to form a joint reference panel across all individuals, we merged the phased haplotypes with the 1000 Genomes Phase 3 haplotypes [13] at the overlapping set of variants. The merged reference panel, which contains 96,676 variants in the region chr4:139.5-150.5Mb, is available in **Online Resources**.

4) Imputation and association testing at SNPs and short indels

A) Imputation and association testing

We used both the 1000 Genomes Phase 3 reference panel [13] and the merged reference panel described above to impute genotypes into the previously published sets of severe malaria cases and population controls from the Gambia, Kenya and Malawi [1]. In brief, we ran IMPUTE2 [56, 57] in 2 Mb chunks, using Illumina Omni 2.5M genotype calls as previously described. A total of 5,621 (the Gambia), 5,203 (Malawi), and 5,583 (Kenya) genotyped variants were included in the 11 Mb region considered. We specified 1,000 copying states (-k_hap 1000), an effective population size of 20,000 (-Ne 20000) as recommended for African populations, and a buffer region of 500 kb. We removed reference panel individuals with parents in the panel before imputation.

We tested for association in each population under additive, dominant, recessive and heterozygote models using SNPTEST. We restricted analysis here to severe malaria cases with nonzero parasitaemia, as measured by blood slide at the time of admission, as this is likely to be the most accurate phenotype. We meta-analysed association results across populations as previously described [1], computing both fixed-effect meta-analysis and a model-average Bayes factor (BF_{avg}) [1]. Meta-analysis results under the two reference panels for imputation, as well as under the previously published 1000 Genomes Phase 1 reference panel imputation, are shown in **Supplementary Fig. 2**.

B) Functional annotation of variants

We used Variant Effect Predictor (VEP; [58]) to annotate the functional consequences of all variants with evidence of association (BF_{avg}>1). We noted one variant with VEP IMPACT rating 'moderate' with some evidence of association (rs147343123; missense in exon 1 of *GYPA*, predicted deleterious; association test BF_{avg}=17.34, P=1.4x10^{-3} for 1000 Genomes imputation; BF_{avg}=121.3, P = 1.5x10^{-4} for full panel imputation). We also noted that the nonsynonymous variant in *FREM3* exon 1 previously found to have evidence of association (rs181620317; see [1]), is imputed at much lower frequency using both the 1000 Genomes Phase 3 panel and the full reference panel (e.g. frequency = 0.2% in Kenya using the full reference panel), and shows no evidence for association in these imputations (P > 0.1).

5) Identification of copy number variation

A) Method to call copy number variation

To identify large copy number variants (CNVs) across the glycophorin region, we implemented a hidden Markov model (HMM) to infer the underlying copy number state from the observed read counts. The input to the HMM is read depth, averaged over sites in windows of fixed length (here, 1600 bp) for each individual. To reduce the problems with mapping in the region, we included only:

- reads with at least a mapping quality (MQ) of 20
- mappable sites, defined as sites with mappability >0.9, where mappability of a site is the mean value of the CRG mappability track for all 100-mers overlapping that site [14]
- windows with ≥25% of sites fulfilling this criterion; windows with fewer mappable sites were considered uninformative

We modeled the mean depth of coverage for individual i at window j, $d_{i,j,}$ as normally distributed with mean and variance dependent on the assumed underlying copy number k (k=2 is the normal diploid state):

$$d_{i,j} \sim Norm\left(\frac{k}{2}w_j\mu_i, \left(\frac{k}{2}\sigma_i\right)^2\right) \quad k \in \{0,1,2,3,4\}$$

For copy number k=0 (homozygous deletion), we used a truncated normal distribution (truncated at 0) and assigned a variance of 0.04 to allow for spurious mapping. To account for systematic variation in coverage along the genome, we estimated a window-specific factor (w_j) proportional to how much individuals with no copy number variation (k=2) are above or below their mean in that window. These define the emission probabilities for the HMM. We used a fixed transition matrix that assumes a low rate of switching with approximately 0.999 probability of remaining in the current copy number state; 1x10^{-4} for leaving the diploid state; 0.001 probability of returning to diploid (k=2) from a non-diploid state; and 1x10^{-5} probability of switching among non-diploid states.

We estimated μ_i, σ_i and w_j by starting with an initial guess assuming everyone to be diploid across the region, and then running the Viterbi algorithm separately for each individual. We then recalculated μ_i and σ_i for each individual, only including windows in which the copy number is inferred to be 2, and w_j for each window only including individuals in which the copy number is inferred to be 2. We iterated this algorithm until no further changes in the inferred underlying copy number paths were observed for any individuals.

B) CNV calling in the 3,269 sequenced individuals

We applied the HMM method described above to the full set of 3,269 individuals with sequence data (see Sample collection and sequencing) in an 850 kb region including the glycophorin genes (chr4:144.35-145.20Mb). After inferring the copy number state paths for each individual, we then considered variants to be the same across individuals if the direction of copy number change was the same and both end points were within three bins of each other. Heterozygous triplications and homozygous duplications were differentiated by looking across individuals; variants that were always found in copy number state 5 were attributed as triplications. We excluded copy number variable segments that covered only a single bin or were outside the segmental duplication. We then considered copy number variable segments that were always found together to be a single CNV, and manually refined the few other copy number variable segments that were not found separately from other CNVs (**Supplementary Note**). This process identified 16 non-singleton variants and 28 singleton variants (**Fig. 1a** and Supplementary **Fig. 3**), although we note several caveats about the singleton CNVs including some that likely correspond to more common variants (**Supplementary Note**).

To validate the CNVs and genotype calls, we assessed inheritance in the MalariaGEN trios and compared genotype calls for the 1000 Genomes individuals with those released in the 1000 Genomes Phase 3 paper on structural variants ([15]; **Supplementary Table 4, Supplementary Fig. 4** and **Supplementary Note**). We also designed PCR primers on either side of the DEL1 variant and generated Sanger sequence that confirmed and localized this breakpoint (**Supplementary Note**).

6) Phasing and imputation of CNVs

A) Initial phasing of CNVs

We investigated whether CNVs could be accurately phased relative to surrounding SNP variation in the regional reference panel (see Construction of a regional reference panel). Collectively, non-singleton CNVs in our dataset cover a total of 350 kb (**Fig. 1a**), which extends over most of the region of segmental duplication surrounding the glycophorin genes. In principle, inference of CNV haplotype phase might benefit from copy-number-aware genotype calls at smaller variants within CNVs. However, implementing this is challenging as the state-of-the art phasing methods assume samples are diploid. A further possible

issue is that read mapping, and hence the quality of SNP genotype calls, is likely to be impaired in such regions of segmental duplication.

Motivated by these observations, we took the following approach to phasing CNVs which leverages the family trio structure of sequenced individuals to infer accurate phase, using both SHAPEIT2 [11] and MVNCALL [59]. First, we focused on the 765 sequenced individuals collected by MalariaGEN partners. We removed variants within the region of segmental duplication (here taken as chr4:144.7-145.07Mb) from the reference panel, and replaced these with CNV genotype calls, to form a single file with genotypes at the CNVs and flanking SNPs and indels. SHAPEIT2 requires each variant to be assigned a single genomic position. For each variant longer than 10 kb we included the genotypes for that variant at the start-, mid- and endpoints of the CNV; for variants less than 10 kb we used the midpoint. We then ran SHAPEIT2 with parameters as for SNP/indel phasing to produce phased genotype calls, treating each CNV as a separate, biallelic variant. Next, to phase CNVs into the 1000 Genomes Project individuals, we extracted the Omni 2.5M sites with allele frequency > 1% from the 1000 Genomes Project phased haplotypes and removed the region of segmental duplication. We used MVNCALL to phase each non-singleton CNV into this scaffold, again placing each CNV greater than 10 kb in length at its start-, mid-, and end positions.

We assessed accuracy of phasing by considering patterns of LD between CNVs and variants in the left and right flanking regions (**Supplementary Figs. 7 and 8**). We noted high LD between some CNVs and variants to the left of the region, including for DEL1, DEL2, DUP1 and DUP4. LD estimated from phased haplotypes captured most of the correlation between genotypes across the region, as expected in an outbred population if haplotype phase is accurate. The small deviations observed may be due to the presence of a small number of switch errors, or potentially to population substructure.

B) Haplotype-based curation and re-phasing of CNVs

Some singleton CNVs have similar copy number profiles to more common CNVs in the HMM-based calls (**Supplementary Note**) and after phasing we observed that haplotypes carrying several of these clustered with the corresponding common variant (DEL9 with DEL1; DEL11, DEL12 and DEL14 with DEL2; DUP15 and DUP18 with DUP1). We reasoned that these are likely to represent the same variant, with differences in calling potentially due to noise in coverage profiles or variation in the other chromosome. We merged these singletons with the corresponding non-singleton CNV for subsequent phasing. We also noted that haplotypes carrying DEL4 cluster with DUP1, which shares a similar breakpoint (**Fig. 1**). A plausible explanation for this is that DUP1 arose by NAHR on a DEL4 background (**Supplementary Note**).

Three individuals in the 1000 Genomes data carry CNVs that are not singletons in the overall dataset but are private to that individual within the 1000 Genomes data (HG01986, carrying DEL4; HG02554, carrying DUP4, and HG02585, carrying DUP5). Of these, we noted in particular that HG02554 appears to have a

switch error in the 1000 Genomes (explaining clustering of opposite haplotypes with other DUP4 haplotypes on either side of the region; **Supplementary Fig. 7**). Because our approach phases variants in the 1000 Genomes separately and these are singletons in that dataset, we excluded these three individuals for re-phasing.

With these modifications, we repeated the procedure described above, using SHAPEIT2 and MVNCALL to re-phase CNVs and flanking short variants, and merged the two reference panels for subsequent imputation.

C) Cross-validation of CNV imputation in the reference panel

To evaluate how well CNVs are likely to be imputed in our GWAS dataset, we performed a cross-validation experiment using the African reference panel individuals as follows. For each individual, we removed the individual and his/her family members (if present) from the phased reference panel haplotypes to form a subsetted panel. We also extracted genotype calls for that individual at variants on the Illumina Omni 2.5M array from the reference panel genotypes, excluding variants within the glycophorin region. We used these genotypes and the subsetted panel to re-impute CNVs for that individual.

We evaluated CNV re-imputation by computing the correlation between HMM-based genotype calls and genotype dosages from the re-imputation (**Supplementary Fig. 9**) and by direct comparison of HMM and re-imputed calls. We note that DEL4 carriers were imputed with some confidence as carrying DUP1, consistent with the shared haplotype for these variants and the higher frequency of DUP1. Given the functionally distinct nature of these variants, this affects interpretation of imputed DUP1 genotypes.

Among CNVs > 10kb in length, we noted little variation between the three imputation locations (leftmost breakpoint, midpoint, and rightmost breakpoint of the CNV), except for DUP4, where the right endpoint had slightly higher imputation performance (**Supplementary Fig. 9**). For all analyses presented in this paper we refer to the midpoint imputation of each CNV.

D) Imputation of CNVs

We used the combined panel to impute CNVs into the three GWAS datasets. Imputation settings were as described above (see Imputation and association testing). To ensure imputation was based on flanking SNPs, we removed SNPs within the glycophorin region from the genotype data before imputation. We evaluated imputation performance in the GWAS data by comparing the overall expected allele frequency against the IMPUTE info metric and another metric of confidence in imputed CNV call probabilities, the proportion of expected frequency of CNV heterozygote or homozygote that is due to genotypes with at least 90% probability (**Fig. 2d** and **Supplementary Fig. 11a,b**). We also compared the expected frequency in control samples with the frequency in the geographically nearest reference panel population (**Supplementary Fig. 11c**).

7) Analysis of association at CNVs

A) Association with severe malaria

We tested for association with each CNV in each population and computed both fixed-effect and bayesian meta-analyses as for other variants (see Imputation and association testing, above). To directly estimate the effect of heterozygote and homozygote genotypes, we modified SNPTEST to fit the logistic regression model with a separate parameter for heterozygote and homozygote genotypes in a missing data likelihood framework that integrates over imputation uncertainty.

B) Conditional tests of association

We used QCTOOL v2 (http://www.well.ox.ac.uk/~gav/qctool_v2/) to extract the additive and heterozygote imputed dosages of DUP4 for each individual, and repeated the association test and meta-analysis conditioning on these dosages (**Fig. 2f** and **Supplementary Fig. 10**).

C) Association with clinical subtypes

Severe malaria-affected children in our data are recorded as either having cerebral malaria (CM), severe malarial anaemia (SMA), or other nonspecific severe malaria phenotype (OTHER). To assess the association of DUP4 with these subphenotypes, we fit a multinomial logistic regression model with these outcome levels using population controls as a baseline (**Supplementary Table 5**). A small number of individuals are annotated as both CM and SMA and were excluded from this analysis.

8) Population genetic analysis

A) Frequency differentiation

We used estimated minor allele frequencies (MAF) at both typed and imputed variants from [1] for the 2,490 population controls from the Gambia and 1,498 population controls from Kenya to investigate the extent to which the observed frequency difference of DUP4 is extreme relative to other variants genome-wide. For this comparison, we included all autosomal variants having IMPUTE info >0.75 and estimated frequency ≥0.5% in at least one of the populations (14,973,426 variants in total). We binned variants into 0.5% MAF bins (**Fig. 4a**) and noted the frequency estimates for DUP4 (f_{Kenya}=0.0895 and f_{Gambia}=0.0003) and, for comparison, the sickle cell anaemia-causing allele (rs334:T; MAF=0.0853 in Kenya and 0.0766 in the Gambia based on direct Sequenom typing of this SNP in these samples [60]).

To quantify the extent to which DUP4 is an outlier, we computed two empirical *P*-values based on the marginal distributions. Specifically, we computed $P_{Gambia|Kenya}$ as the proportion of variants with MAF less than or equal to f_{Gambia} in the Gambia, among all variants with MAF within 1% of f_{Kenya} in Kenya (empirical $P_{Gambia|Kenya}<1.2\times10^{-6}$; 0 of 831,956 variants in this bin). Similarly, we computed $P_{Kenya|Gambia}$ as the proportion of variants with MAF equal to or greater than f_{Kenya}

in Kenya, among all variants with MAF within 1% of f_{Gambia} in the Gambia (empirical $P_{Kenya|Gambia}$=1.7x10^{-3}; 2,851 of 1,710,922 variants in this bin).

We note that the computation of $P_{Gambia|Kenya}$ is sensitive to the frequency of DUP4 in the Gambia, which may be underestimated due to poor imputation performance. To account for this we recomputed $P_{Gambia|Kenya}$ assuming 1% frequency in the Gambia (adjusted empirical $P_{Gambia|Kenya}$=5x10^{-3}) and refer to this value in the main text.

B) Haplotype homozygosity

To assess haplotype homozygosity, we first used SHAPEIT2 to phase imputed CNV genotypes onto haplotypes defined by directly-typed SNPs in the region chr4:139.5-150.5Mb, excluding SNPs in the glycophorin region. We specified 400 selected and 200 random copying states (--states 400, --states-random 200), an effective population size of 17,469 as recommended for African populations, and included reference panel haplotypes to inform phasing. EHH [18] was then computed around DUP4 using the *rehh* R package. We used a custom script to compute an unstandardized integrated haplotype score (uiHS) [17], using recombination rate estimates from the HapMap combined recombination map [44].

To generate a distribution of uiHS at SNPs with a similar frequency to DUP4, we computed uiHS at those genotyped SNPs [1] where the human ancestral allele could be inferred using the six primate EPO alignments (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/), and the derived allele frequency was in the 2% frequency bin centred around f_{Kenya} (0.0795-0.0995). We computed an empirical *P*-value as the proportion of SNPs with uiHS less than or equal to that observed at DUP4 (*P*=0.0119; 996 of the 83,419 variants in this bin). We note that the exclusion of the glycophorin region from the estimate at DUP4 is likely to be conservative, since in effect it adds a constant term equal to the recombination length of the removed interval (approximately 0.15 cM) to both the numerator and denominator of the statistic.

9) Resolution of the structure of DUP4

A) Discordant read pair analysis

We began by remapping reads from each of the nine heterozygous DUP4 carriers genome-wide using bwa mem, which has better performance for longer reads [61]. Because even longer reads should facilitate more confident mapping around the breakpoints, we also obtained DNA from Coriell for HG02554, the 1000 Genomes sample carrying DUP4, and generated 300 bp PE sequence data on Illumina MiSeq at the High Throughput Genomics core at the Wellcome Trust Centre for Human Genetics at the University of Oxford (**Online Resources**). We mapped these reads to the same GRCh37 human reference (hs37d5.fa) with bwa mem, yielding 13x genomic coverage. For all re-mapped bam files, we marked

duplicates with Picard MarkDuplicates and excluded duplicate reads from further analysis.

We then used samtools to pull out read pairs where both reads had a primary alignment to the glycophorin region with MQ≥1 and an absolute insert size ≥1 kb. For the 300 bp data, we allowed one of the reads in the pair to be mapped to the glycophorin region with MQ=0. Across samples, there were 434 such read pairs. We grouped read pairs where both ends were mapped within 1 kb of each other and identified those near the HMM-inferred breakpoints.

To view how uniquely the discordant read pairs matched each of the three possible homologous positions in the segmental duplication, we used the mapped position and the cigar string to place each read into the multiple sequence alignment and then identified positions in the read with a match or mismatch to each of the three aligned reference sequences using custom scripts in R (e.g., **Supplementary Fig. 13)**.

B) Molecular assays and Sanger sequencing

Briefly, a 4.1 kb fragment spanning the *GYPB-A* hybrid breakpoint (located between exons 4 and 5) was amplified by PCR using primers designed against *GYPB* and *GYPA* sequences. In practice, it is difficult to design specific primers due to the high homology and the assay amplifies both *GYPA* and *GYPB-A* hybrid sequences. To separate these, we identified a restriction enzyme site (BlpI [5'…GC/TNAGC…3']) that cleaves the *GYPA* sequence but not that of the hybrid. PCR products were digested and then separated on an agarose gel. We excised the 4.1 kb band and obtained Sanger sequence of the region around the putative breakpoint. A full description of the design, primers and protocols is given in the **Supplementary Note**.

C) Sequencing and copy number analysis of a serologically-typed Dantu+ individual

We obtained DNA from the International Blood Group Reference Laboratory in Bristol, UK from an archived reference sample that had been serologically typed as Dantu+ (NE type). This sample was collected in 1992 and the individual was originally from Natal, South Africa. The DNA was sequenced with 150 bp PE reads on Illumina HiSeq 4000 by the High Throughput Genomics core at the Wellcome Trust Centre for Human Genetics at the University of Oxford (**Online Resources**). Reads were mapped to the same GRCh37 human reference genome (hs37d5.fa) with bwa mem, yielding 18x coverage. To generate CNV calls, we ran our HMM method on this individual alone, without using the window-specific factors.

D) Simulations of DUP4 formation

To determine possible routes to formation of DUP4, we implemented a computer program in C++ to iteratively generate structural rearrangements via unequal crossing over, allowing breakpoints to occur at any of the six locations observed

for DUP4 with no constraint based on homology. In brief, we encode the reference haplotype as a string of seven segments delineated by the coverage breakpoints (i.e., as the string 0123456; **Fig. 5a**). We ran the program through three generations of events, where the first generation produced all possible events between two reference haplotypes and the second and third generations allowed unequal crossing over between any two haplotypes from previous generations. This brute force search allows us to place a lower limit of four on the number of events required to form DUP4. For additional details on the program and search, see **Supplementary Note**.

10) Online Resources

MalariaGEN website: http://www.malariagen.net
Multiple sequence alignment of three glycophorin segmental duplication units: *[URL]*
Combined reference panel: *[URL]*
Additional sequence data for HG02554: *[URL]*
Sequence data for Dantu NE carrier: *[URL]*

11) Methods references

47. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
48. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
49. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
50. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
51. Lipatov, M., et al., *Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data.* 2015.
52. Staples, J., et al., *PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent.* Am J Hum Genet, 2014. **95**(5): p. 553-64.
53. Rearden, A., et al., *Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution.* J Biol Chem, 1993. **268**(3): p. 2260-7.
54. Lassmann, T., O. Frings, and E.L. Sonnhammer, *Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features.* Nucleic Acids Res, 2009. **37**(3): p. 858-65.
55. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
56. Howie, B., J. Marchini, and M. Stephens, *Genotype imputation with thousands of genomes.* G3 (Bethesda), 2011. **1**(6): p. 457-70.
57. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
58. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
59. Menelaou, A. and J. Marchini, *Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold.* Bioinformatics, 2013. **29**(1): p. 84-91.
60. Malaria Genomic Epidemiology Network, *Reappraisal of known malaria resistance loci in a large multicenter study.* Nat Genet, 2014. **46**(11): p. 1197-204.
61. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. ArXiv e-prints, 2013. **1303**.