Supplementary Materials for

Resistance to malaria through structural variation of red blood cell invasion receptors

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

correspondence to: spencer@well.ox.ac.uk; dominic.kwiatkowski@sanger.ac.uk

This PDF file includes:

Supplementary Text

I. Additional details of CNV calling and evaluation

A. Refinement of CNV calls

We manually curated the HMM-based set of CNV calls as follows. We considered sets of copy number variable segments that consistently occurred in the same individuals as forming the same CNV. We inspected these and the remaining segments that were only found in combination with other CNVs and made the following modifications to the calls:

(1) Four individuals were identified as heterozygous for overlapping deletions (three heterozygous for DEL1 and DEL2, and one heterozygous for DEL1 and DEL5) and we updated their genotype calls. Three of the individuals were in trios, and segregation of the variants was consistent with compound heterozygous genotypes.

(2) A short deletion around the 3' end of *GYPB* was found in five of the nine carriers of the multi-segment variant DUP4. Inspection of coverage profiles indicated that this deletion was present in all the individuals in which the rest of the copy number segments matched, and we include the deleted segment as part of DUP4.

(3) In one trio carrying a duplication with disjoint segments (DUP5), the parent was called to have a triplication within the first segment while the child was only called as carrying a duplication; the third carrier also had a similar triplication and so we report the variant with a triplicated segment but note that there is uncertainty about the location of this change.

(4) A deletion downstream of *GYPE* was found in one of the two DUP6 carriers; inspection of the coverage profile indicated that this deletion was also carried by the second DUP6 carrier and so both copy number changes are included as part of DUP6.

B. Singleton CNVs

We note several caveats to the interpretation of the singleton variants. First, some singletons may represent a slightly different call of a more common variant, such as those that largely overlap with DUP1 (e.g., DUP9-DUP13), DEL1 (e.g., DEL9) or DEL2 (e.g., DEL11 and DEL12). Consistent with this, several of these were subsequently phased onto haplotypes that cluster with the corresponding common variant. Second, the multisegment singletons could represent a heterozygous genotype of two overlapping variants that we did not disentangle (i.e., DUP21, DUP26). Third, short variants found in a single individual may be due to sampling noise in a small number of windows. We therefore conservatively count a minimum of 11 singleton variants (DEL10, DEL13, DUP14, DUP17, DUP19, DUP22, DUP23, DUP24, DUP25, DUP27, DUP28).

C. Inheritance of CNVs

To assess the inheritance of the CNVs, we identified complete trios in which variants were segregating. Of the 207 sequenced MalariaGEN trios, 13 CNVs were found to be segregating in a total of 73 trios (**Table S4**). Transmission of the CNVs from parent to child was consistent with Mendelian inheritance except for DUP13, which was carried by a child but neither parent. Inspection of the coverage profiles suggested that one parent also showed increased coverage and likely carried the variant as well. We also observe significant undertransmission of DUP1 from heterozygous parents. There are some indications we may be missing DUP1 genotype calls; for example, a handful of singletons overlap with DUP1 and may be the same variant (**Fig. S3**), and several 1000 Genomes individuals who do not carry DUP1 in our analysis carry a duplication similar to DUP1 in the 1000 Genomes structural variant calls (**Fig. S4**). However, we do not observe increased coverage in this region for any of the children of untransmitting heterozygous parents.

D. Comparison with 1000 Genomes Phase 3 structural variant calls

We compared the CNV calls from the HMM with those released in the 1000 Genomes Phase 3 paper on structural variants (*23*) (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/). This data set consists of multiple classes of structural variants called by nine different algorithms that were then merged. In the glycophorin region, it contains 15 variants of at least 3,200 bp including six deletions, three duplications, and six multi-allelic CNVs (here all with one duplication and one deletion alternative allele; **Fig. S4A**). Like the HMM-called CNVs, these variants span the glycophorin region, although the set of variants is largely different. Of the 2504 Phase 3 individuals, 2317 do not carry a CNV allele in either call set, 164 carry a CNV allele in both call sets, and 23 individuals carry a CNV allele in one call set but not the other (13 in Phase 3 and not the HMM and 10 in the HMM and not Phase 3). In all cases where a copy number variant is called in both data sets, the copy number called is the same (although for seven individuals the 1000 Genomes call also includes a variant with a different copy number genotype; **Fig. S4B**). Conservatively then, 2474/2504 (98.8%) carry the same overall copy number call in both data sets. We note that most of the individuals with a CNV in the 1000 Genomes call set carry multiple overlapping variants with up to eight different variants carried by a single individual (**Fig. S4B**). While a few of the 1000 Genomes variants match closely to the HMM variants (**Fig. S4A**), these are often assigned to individuals along with variants that overlap them (**Fig. S4B**). Notably, the one 1000 Genomes individual carrying DUP4 is assigned duplication alleles for four variants, all of which are found in other individuals.

II. Relation to known blood groups

CNVs that delete or alter the extracellular sequences of expressed glycophorins may affect the MNS blood group antigens, encoded by *GYPA* and *GYPB*, present at the red cell surface. Among the large CNVs we identify, several have a predicted functional impact that corresponds to a known MNS blood group antigen or class of variant. In turn, other CNVs affect the genes in ways not predicted to affect the expression of blood group antigens, and there are structural variants reported to underlie blood group phenotypes that we do not observe. We describe the relationship between the CNVs and known MNS blood group phenotypes briefly here, and summarize them in **Table S6**.

A. Deletion of *GYPB*

The absence of GYPB antigens, which results from homozygous loss of function mutations affecting *GYPB*, corresponds to the S−s−U− blood group phenotype. We identify four non-singleton CNVs that fully delete *GYPB* (DEL1, DEL2, DEL4 and DEL6) and one that involves the deletion of the majority of *GYPB* protein-coding sequence (DEL8; deleting all but the 12 amino acids of the signal peptide encoded in the first exon). Singleton variant DEL10 also deletes the full *GYPB* coding sequence; DEL9 and DEL12 delete *GYPB* as well but likely correspond to DEL1 and DEL2, respectively. The expected frequency of *GYPB* deletion homozygotes in our dataset, based on the frequencies of these variants, is commensurate with estimates of S−s−U− frequencies from the literature (**Fig. S23**).

However, none of these deletion variants corresponds to the primary mutational event reported to underlie the S−s−U− phenotype. The reported breakpoint, which was found by restriction mapping, lies within intron 1 of *GYPE* and *GYPB* connecting exon 1 of *GYPB* with the rest of the *GYPE* gene (*8, 31, 70-72*). In contrast, the deletion breakpoints observed here lie outside the genes, with the exception of DEL8. While DEL8 does retain exon 1 of *GYPB*, it also retains exon 1 of *GYPE*. It may be that the variant reported in the literature was not sampled here, but this survey shows that other *GYPB* deletions, in particular DEL1 and DEL2, which are present across many different populations, are likely the most common.

To validate the breakpoint of the most prevalent *GYPB* deletion, we designed primers to amplify across the DEL1 breakpoint predicted from the HMM (see below). This localizes the DEL1 breakpoint to an identical 120 bp at chr4:144835160-144835280 in the *GYPE* unit of the segmental duplication and chr4:144945398-144945518 in the *GYPB* unit of the segmental duplication (**Fig. S6**).

B. *GYPB-A* hybrids

Two blood group phenotypes are known to result from *GYPB-A* hybrids and differ based on where the junction between *GYPB* and *GYPA* is (*33, 72*). The Dantu antigen

is encoded by a junction within intron 4, connecting GYPB exon 4 to GYPA exon 5 as discussed in the main text for DUP4. The other such blood group phenotype is GP.Sch, which corresponds to a junction within intron 3 that connects GYPB exon 2 (*GYPB* exon 3 is not expressed) with GYPA exon 4, thereby encoding the St^a antigen. DUP2 has a breakpoint within intron 3 and thus is predicted to encode the St^a blood group antigen, although we do not map the exact breakpoint. St^a has been primarily identified in east Asian populations including Chinese, Taiwanese, and Japanese, where it ranges in frequency from 0-6% across populations, and multiple mutational origins have been reported (*33, 73, 74*). We observe DUP2 mainly in the Chinese populations from 1000 Genomes, at comparable frequency (**Fig. 2**). One singleton CNV, DUP27, may also encode a *GYPB-A* hybrid (**Fig. S3**)

C. *GYPA-B* hybrids

GYPA-B hybrids can encode different blood group antigens depending on where the A-B junction occurs (intron 3 or intron 4) and whether the encoded protein carries the S or s determining amino acid from GYPB exon 4 (in the case of hybrids with an intron 3 junction) (*8, 72*). We identified one singleton variant predicted to encode a *GYPA-B* hybrid (DEL13, carried by NA20867 from the Gujarati Indian from Houston, Texas 1000 Genomes population) but have not determined which of these it corresponds to.

D. *GYPE-A* hybrids

GYPE-A hybrids are not known to underlie any of the MNS blood group phenotypes, and to our knowledge, no such molecular variant has been reported (*16, 72*). Here, we find three rare variants, one doubleton (DUP8) and two singletons (DUP23 and DUP24), predicted to encode *GYPE-A* hybrid genes, although they may not be expressed at the protein level. All four carriers of these variants are from South Asian populations from the 1000 Genomes.

E. Whole gene duplications

Whole gene duplications (without creation of any hybrid) are not predicted to alter the specificity of blood group antigens, and to our knowledge no such molecular variant has been reported. We observe two non-singleton CNVs (DUP3, and DUP7) and five singleton CNVs (DUP14, DUP17, DUP19, DUP25 and DUP26) predicted to duplicate *GYPE*, *GYPB*, *GYPA*, or both *GYPE* and *GYPB.* We additionally note that DUP6 is predicted to duplicate the majority of *GYPA*, including the entire protein-coding sequence.

F. Deletion of *GYPA* or of both *GYPA* and *GYPB*

Homozygous loss of function of GYPA, or of both GYPA and GYPB, are very rare but reported in the literature as blood group phenotypes En(a-) and M^k, respectively,

with no known deleterious effects (*33*). We do not observe any CNVs that involve deletion of *GYPA*.

We have not directly characterized the many blood group variants that are caused by SNPs or gene conversion events and this remains a challenging task for future work.

III. Validation of variant breakpoints by Sanger sequencing

A. DEL1

To design assays for the DEL1 breakpoint, we focused on a reference alignment of the 11 kb upstream from the transcription start sites of *GYPA*, *GYPB* and *GYPE* where the putative DEL1 breakpoint lies (**Fig. S5**). In between the transcription start site and the putative breakpoint there is a ~3 kb sequence that is unique to the *GYPE* unit of the segmental duplication (see **Fig. 1A**), with the putative breakpoint situated approximately 3 kb beyond in the homologous sequences. A PCR was designed to amplify a 5 kb segment of DNA beginning in the *GYPE* specific sequence (with primer GYP_DEL1_F6) and running for ~4 kb into the homologous region (to primer GYP_DEL1_R4A; **Fig. S5** and **Table S5**). This was followed by a nested PCR using primers situated inside or overlapping with the first round primers (GYP_DEL1_F1 and GYP_DEL1_R4C) to generate a final 4.8 kb product. The reverse primers are common to all three glycophorin sequences, allowing amplification of both wild-type product (sequence fully from the *GYPE* unit of the segmental duplication) and DEL1 product (*GYPE* unit – *GYPB* unit hybrid sequence). For Sanger sequencing, three different reverse primers were designed (GYP_DEL1_R1, GYP_DEL1_R2, and GYP_DEL1_R3), which are spaced ~1 kb apart and are found in all three units of the segmental duplication.

The initial PCR product was generated in a 20μL PCR using 2μL of 20ng/μL gDNA, 1μL of 10μM specific *GYPE* forward primer (GYP_DEL1_F6), 1μL of 10μM common reverse primer (GYP_DEL1_R4A), 10μL of Phusion Taq master Mix (Phusion® High-Fidelity PCR Kit, NEB, Hitchin, UK) and 6μL of water. PCR cycling conditions were: 98°C for 30 seconds; then 35 cycles of 98°C for 10 seconds, 68.5°C for 30 seconds, 72°C for 5 minutes; followed by a final extension of 72°C for 5 minutes. The PCR product was diluted 1:10 in water and used in a 50μL nested PCR with 2.5μL DNA, 2.5μL of 10uM forward primer (GYP_DEL1_F1), 2.5μL of 10μM reverse primer (GYP_DEL1_R4C), 25μL Phusion Master Mix and 17.5μL of water. PCR cycling conditions were: 98°C for 30 seconds; then 35 cycles of 98°C for 10 seconds, 68.5°C for 30 seconds, 72°C for 5 minutes; followed by a final extension of 72°C for 5 minutes. Five microliters of the nested PCR product was run on a 0.7% agarose gel (90V for 45 minutes) to check the reaction band size (4.8 kb) and purity of the product. Either the remaining 45 μL of PCR product was run on a new gel and the 4.8 kb band excised for purification, or the PCR reaction was directly purified. Purification was undertaken with the Qiagen Qiaex II DNA extraction Kit (Qiagen, Crawley, UK). Products were diluted and sent for Sanger sequencing at GATC

Biotech (Constance, Germany) using the sequencing primers (GYP_DEL1_R1, GYP_DEL1_R2 and GYP_DEL1_R3; **Fig. S5** and **Table S5**). Sanger sequences from two individuals homozygous for DEL1 and one individual not carrying any CNVs using sequencing primer GYP_DEL1_R2 are shown in **Fig. S6**. This localizes the breakpoint to chr4:144835160-144835280 in the *GYPE* unit of the segmental duplication and chr4:144945398-144945518 in the *GYPB* unit of the segmental duplication.

B. DUP4

A 4.1 kb fragment around the predicted *GYPB-A* hybrid breakpoint located between exons 4 and 5 was amplified by PCR. One primer (GYPA_Exon6_Fwd) was designed to the *GYPA* reference sequence between exons 5 and 6 and lies in a unique sequence not present in the other two genes (**Table S10**). The second primer (GYPB_753_Rev) was designed to the *GYPB* reference sequence just upstream of exon 3, but has only three differences from the homologous location in the *GYPA* reference sequence (**Table S10**). It is therefore possible that the PCR would amplify a *GYPB-A* hybrid product as well as a fully *GYPA* product (**Fig. S17**). A 50µL PCR was performed using 10µL of 1ng/µL gDNA, 2ul of 10µM forward primer, 2µl of 10µM reverse primer, 25µl of Phusion Taq master Mix (Phusion® High-Fidelity PCR Kit, NEB, Hitchin, UK) and 11µl of water. PCR cycling conditions were: 98°C for 30 seconds; then 35 cycles of 98°C for 10 seconds, 68.5°C for 30 seconds, 72°C for 4 minutes; followed by a final extension of 72 °C for 5 minutes.

In practice, we find that this reaction always result in a PCR product, as expected if it is also amplifying full *GYPA* sequences. In order to remove the wholly *GYPA* product, a restriction enzyme site (BlpI [5'…GC/TNAGC…3']) was identified between exon 4 and the putative breakpoint that cleaves *GYPA* sequence but not the hybrid sequence, predicted to be *GYPB* sequence by this point (**Fig. S17**). Ten microliters of PCR product was mixed with 2µl of NEB CutSmart® buffer, 0.5µl (5U) of BlpI enzyme (NEB) and 7.5µl of water. The reaction mixture was incubated at 25 °C for 1 hour. PCR products and restriction digests were separated on a 0.7% agarose gel (90V for 45 minutes) and visualised using ethidium bromide staining. The uncut 4.1 kb band indicates the presence of the *GYPB-A* hybrid product while a pair of bands at 2.7 kb and 1.4 kb indicates the cut *GYPA* products (**Fig. S18**); hybrid carriers have all three bands. The 4.1 kb band was excised from the gel and purified using the Qiagen Qiaex II gel purification kit (Qiagen, Crawley,UK). Fragments were diluted and prepared as required, and sent for Sanger sequencing by GATC Biotech (Constance, Germany) using primers located either side of the putative breakpoint (**Table S10**). Samples identified as non-hybrid carriers (all PCR product cut by the enzyme) were used for sequencing without enzyme digestion, as a negative control. A sample of sequence results is shown in **Fig. S19.**

IV. Formation of complex variants

A. Simulation of unequal crossing over

We implemented a computer program in C++ to iteratively simulate unequal crossing over, allowing breakpoints to occur at any of the six locations observed for DUP4 with no constraint based on homology. To do this, we encoded the reference haplotype as a series of seven segments ending in coverage breakpoints (i.e., as the string 0123456; **Fig. 6A**). We first computed all haplotypes formed by unequal crossover of this haplotype with itself, recording the resulting haplotypes and the positions of the breakpoints, which we refer to as 'generation 1' haplotypes. For example, one possible event leading to deletion of *GYPE* represented in this notation is

```
Left ancestor:              0|1 2 3 4 5 6
Right ancestor:       0 1 2|3 4 5 6
Resulting haplotype:        0 3 4 5 6
```

where the vertical bars denote the breakpoints and grey text denotes the part of the haplotype not inherited by the descendant.

We then iteratively generated all possible haplotypes in generations 2 and 3, at each stage allowing unequal crossing over between any two haplotypes from the previous generation(s), again assuming all recombination occurs between the numbered segments. In total, we found 30 possible haplotypes in generation 1, 5,500 in generation 2, and 189,738,750 in generation 3. Of these, a total of 852, all in generation 3, match the DUP4 copy number profile (1 copy of segment 0, 2 of segment 1, 1 of segment 2, 0 of segment 3, 2 of segment 4, 3 of segment 5, and 1 of segment 6).

To assess the total number of distinct crossover events required to produce each generation 3 haplotype, we considered the recorded histories of crossover events leading to the haplotypes. Each such history lists the seven potentially distinct haplotypes ancestral to the final haplotype: two immediate ancestor haplotypes, two direct ancestors of each of the direct ancestor haplotypes, and the reference haplotype. The number of distinct haplotypes among these ancestors is equal to the number of unequal crossover events in the path leading to the final haplotype. Among the 852 generation 3 haplotypes matching the DUP4 copy number profile, the minimum number of distinct ancestral haplotypes was four. Of these, the only segment order that included all three of the observed breakpoint connections was the one predicted (0121545456; **Fig. 6B**), which was achieved with four unequal crossover events by a total of 39 possible histories. These 39 histories involved a fourth event that either independently created another hybrid (a second 5-4 event) or duplicated an existing hybrid via unequal crossing over between two hybrid-carrying haplotypes (a 4-5 event between two chromosomes each already having experienced the same 5-4 event). We illustrate one possible sequence of ancestral

events in **Fig. S20**, presented over four generations, allowing the first three events to occur with a reference haplotype.

B. Relationships between CNVs

If DUP4 arose by a series of unequal crossover events, we might expect to find intermediate copy number variants ancestrally related to DUP4. In this population survey, two deletions have one end that falls in the same 1600 bp bin as one of the DUP4 breakpoints (**Fig. 1C**; DEL2 and DEL5), but closer inspection of coverage suggests that these breakpoints are likely to be different, and we do not observe these or other CNV haplotypes clustering with DUP4 haplotypes (**Fig. S9**). Any intermediates may be at too low frequency to have been sampled in the populations sequenced here. In addition to Dantu NE, two other Dantu+ variants (Ph type and MD type) have been reported in the literature, and these could be ancestrally related, although as yet each has only been found in a single, unrelated individual (*72, 75, 76*). Alternatively, DUP4 formation may have involved more complex mutational steps with fewer or no intermediates.

There is a relationship between haplotypes that carry DUP1 and DEL4 alleles (**Fig. S9**). The 1600 bp bin in which DUP1 ends is adjacent to the bin in which DEL4 begins, and a single bin to the right of DEL4 also shows high coverage in DUP1 carriers (**Fig. 1B**), along with several neighboring bins excluded due to low mappability. This is consistent with a scenario where DUP1 arose by unequal crossing over between a DEL4 haplotype (where the two disjoint duplicated sequences were adjacent) and a reference haplotype (**Fig. S24**).

V. Calling glycophorin CNVs from Illumina assay intensity data

Modelling CNV genotypes We denote by X_i the CNV genotype of individual i. All individuals are diploid, so X_i consists of an unordered pair of CNV alleles,

$$X_i = (X_i^1, X_i^2)$$

Here, we consider the three overlapping variants DEL1, DEL2 and DUP4, as well as non-CNV haplotypes (denoted 'WT'), so that

$$X_i^k \in \{\text{WT}, \text{DEL1}, \text{DEL2}, \text{DUP4}\}$$

For robustness we also include an additional diploid state, termed 'OTHER', described below.

We number the SNPs $1, \cdots, N$ and write I_{ij} for the intensity values of individual i at SNP j, and $I_i = (I_{i1}, I_{i2}, \cdots)$. Intensity values can be linked to CNV genotype using Bayes' theorem,

$$P(X_i|I_i) \propto P(I_i|X_i)P(X_i)$$

This requires specifying prior probabilities on the CNV genotype, $P(X_i)$. We specify prior probabilities by assuming each CNV haplotype has a 1% probability of occuring, and genotypes occur in Hardy-Weinberg proportions. We additionally modify this by assigning a 0.1% probability to the 'OTHER' genotype, downweighting other genotypes accordingly so that the total prior probability is 1.

We treat intensities at seperate assays as independent given CNV status, so that

$$P(I_i|X_i) = \prod_{j=1}^{N} P(I_{ij}|X_i)$$

Intensities at a given SNP are assumed to depend on the CNV genotype through the underlying SNP genotype, as

$$P(I_{ij}|X_i) = \sum_Z P(I_{ij}|Z)P(Z|X_i)$$

where Z sums over the possible genotypes for sample i at the SNP. As with the CNV genotype X, we consider Z as an unordered pair $Z = (Z^1, Z^2)$, where Z^1 and Z^2 are the allelic types at the SNP carried by the first and second haplotypes. We assume a maximum copy number of 3 for each assay on each haplotype. Thus, letting r denote reference allele and n denote non-reference allele, we have

$$Z^k \in \{\emptyset, r, n, rr, rn, nn, rrr, rrn, rnn, nnn\}$$

where $\emptyset$ denotes 0 copies, r denotes one reference allele, rn one reference and one non-reference allele, etc. The allelic type of each haplotype depends on the corresponding CNV carried by that haplotype,

$$P(Z|X_i) = P(Z^1|X^1)P(Z^2|X^2)$$

Specifying SNP genotype based on CNV genotype We now describe how we specify $P(Z^i|X^i)$. Let f_j denote the frequency of the non-reference allele at SNP j among non-CNV carriers in the reference panel, and let $c_j(X)$ denote the copy number at SNP j of haplotype X, as learned from the HMM path. For SNPs with $c_j(X) = 1$ (normal copy number) we set

$$P(Z|X) = \begin{cases} f_j & \text{if } Z = n \\ 1 - f_j & \text{if } Z = r \\ 0 & \text{otherwise} \end{cases}$$

For SNPs with $c_j(X) = 0$ (i.e. SNPs inside DEL1 or DEL2, or within the deleted segment of DUP4) we set

$$P(Z|X) = \begin{cases} 1 & \text{if } Z = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

These two rules account for wild type and deletion alleles as well as regions outside increased DUP4 copy number. For SNPs in duplicated or triplicated segments of DUP4, we infer the duplicated allele from the position of the additional cluster containing DUP4 carriers in the Omni 2.5M data for reference panel individuals (c.f. Fig. 3a), with copy number taken from the HMM path for DUP4. Additionally, we allow a 1% probability that one of the duplicated/triplicated alleles is different. E.g. for triplicated segments for which we infer the additional alleles as n in reference panel individuals, we set

$$P(Z|X) = \begin{cases} 0.99 & \text{if } Z = nnn \\ 0.01 & \text{if } Z = rnn \\ 0 & \text{otherwise} \end{cases}$$

Finally, for the CNV genotype 'OTHER' we assume a uniform weighting across all possible SNP genotypes for copy numbers $0 - 3$.

Model for intensities given SNP genotype We model the intensity values at each SNP using a mixture of bivariate normal distributions, with a seperate mixture component for each possible SNP genotype. Specifically, given a SNP genotype $Z = z$, we assume

$$P(I_{ij}|Z = z) = MVN(m_z, \Sigma_z)$$

Here, the mean m_z and variance Σ_z of these clusters is fit using a parameterised model with 13 parameters per SNP, an extension of the basic model for

the three diploid clusters, as follows. Let $r(g)$ and $n(g)$ denote the number of copies of the ref and non-ref allele in SNP genotype g, and $t(g) = r(g) + n(g)$ be the total number of alleles in SNP genotype g. We set

$$m_z = \begin{pmatrix} m_{z1} \\ m_{z2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} h_1 \cdot \frac{r(g)^{\alpha_1}}{\left(\frac{t(g)}{2}\right)^{\gamma_1}} \\ h_2 \cdot \frac{n(g)^{\alpha_2}}{\left(\frac{t(g)}{2}\right)^{\gamma_2}} \end{pmatrix}$$

For diploid genotypes this reduces to

$$m_z = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \begin{pmatrix} h_x \cdot r(g)^{\alpha_0} \\ h_y \cdot n(g)^{\alpha_1} \end{pmatrix}$$

Here,

- μ_1 and μ_2 reflect the intensities of the X and Y channel given no copies of the relevant allele. (These are typically small and positive, and can be estimated from the homozygote clusters.)
- h_1 and h_2 reflect the X and Y intensities of the diploid heterozygote cluster, relative to μ_x and μ_y.
- α_1 and α_2 are attentuation parameters that reflect nonlinearity in the channel intensities for diploid samples. Values < 1 are most plausible and imply that intensity values increase sublinearly with the number of alleles.
- γ_1 and γ_2 are included to model interference between r and n alleles. If $\gamma > 0$ then increasing the number of n alleles (respectively r alleles) reduces X (respectively Y channel intensity), even though the number of alleles targetted by that channel remains unchanged.

The μ, h and α parameters together specify a model of intensities for diploid genotypes, and they can be estimated from the positions of diploid homozygous and heterozygous clusters. The γ parameters affect the relationship between these clusters and non-diploid copy number clusters. The γ_i can also be interpreted as reflecting the growth in intensity for genotypes with only one type of allele (reference or non-reference); for example, for genotypes containing only the reference allele the model for x becomes

$$\mu_1 + h_1 \cdot 2^{\gamma_1} \cdot r(g)^{(\alpha_1 - \gamma_1)}$$

so that the x position of genotype r is at $\mu_1 + h_1 \cdot 2^{\gamma_1}$ while that for rr is at $\mu_1 + h_1 \cdot 2^{\alpha_1}$.

To avoid overfitting we specify a mild beta(1,2) prior on γ_i/α_i. This choice of prior ensures that $0 \leq \gamma_i \leq \alpha_i$ which reflects the expected behaviour that additional copies of a given allele will lead to higher intensity on the corresponding axis.

Modelling cluster covariances We adopt a simple form of cluster covariance Σ_z as follows. Suppose $\Sigma_{het} = \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$ is a variance-covariance matrix for the heterozygote cluster. We assume the variance-covariance matrix of other clusters are determined from this by scaling in the x and y direction. Specifically we assume the matrix is

$$\begin{pmatrix} (m_{z1}-\mu_1)^2\,\sigma_x^2 & (m_{z1}-\mu_1)\,(m_{z2}-\mu_2)\,\rho_{xy}\sigma_x\sigma_y \\ (m_{z1}-\mu_1)\,(m_{z2}-\mu_2)\,\rho_{xy}\sigma_x\sigma_y & (m_{z2}-\mu_2)^2\,\sigma_y^2 \end{pmatrix}$$

To model the observed spread in intensities even when the number of alleles is zero, we additionally add a diagonal matrix with parameters σ_1, σ_2 to the variance-covariance matrix.

Fitting the model

We chose SNPs likely to be informative for this analysis based on the presence of well-defined extra clusters in visual inspection of cluster plots (c.f. in Figure 3a). In total we based inference on assays for SNPs rs1822842, rs1808991, rs9997931, rs9799404, kgp22831194, kgp11638798, kgp8150242, kgp21216198, kgp20708880, kgp20743622, kgp21249626, rs6844670, kgp2941248, rs3936169, rs11728240, rs4374581, and rs13103731, as listed in the Omni 2.5M chip manifest (version '4v1_D').

We fit the model iteratively, alternating between fitting cluster parameters at each SNP and calling CNV genotypes. We first initialise parameters using cluster positions estimated from individuals imputed to carry no CNV in the data (using these to estimate the μ, h, and α parameters as described above). and by setting $\gamma = 0$ reflecting no interference between the two alleles. At each iteration we form a set of CNV genotype calls by taking CNV genotypes with posterior probability of at least 0.75. We then refit cluster parameters based on these CNV genotypes using the optim() function in R. We run eleven iterations in total, each with 200 iterations of Nelder-Mead simplex optimisation. Finally, we take CNV genotypes called with 75% posterior probability as the output of the method.

We inspected model fit by plotting ellipses delineating 95% probability distribution of each cluster on cluster plots of each SNP, with individuals coloured by inferred genotype. Figure S14 shows the final model fit in Kenya.

Comparison of intensity-based and imputed CNV genotype calls

The CNV genotype counts for each dataset are listed in Table S8. A comparison of intensity-based genotype calls for DUP4 with imputed genotype calls in the Gambia, Kenya and Malawi is presented in Table S9. In Gambia, two samples are called by the intensity method as carrying DUP4 in heterozygous form. However, we find no evidence from imputation that these are DUP4 carriers (posterior probabilty from imputation = 0). Inspection of these samples on

cluster plots suggests intensities may also be consistent with non-DUP4 copy number and we interpret these as probable mis-calls.

Supplementary Figures

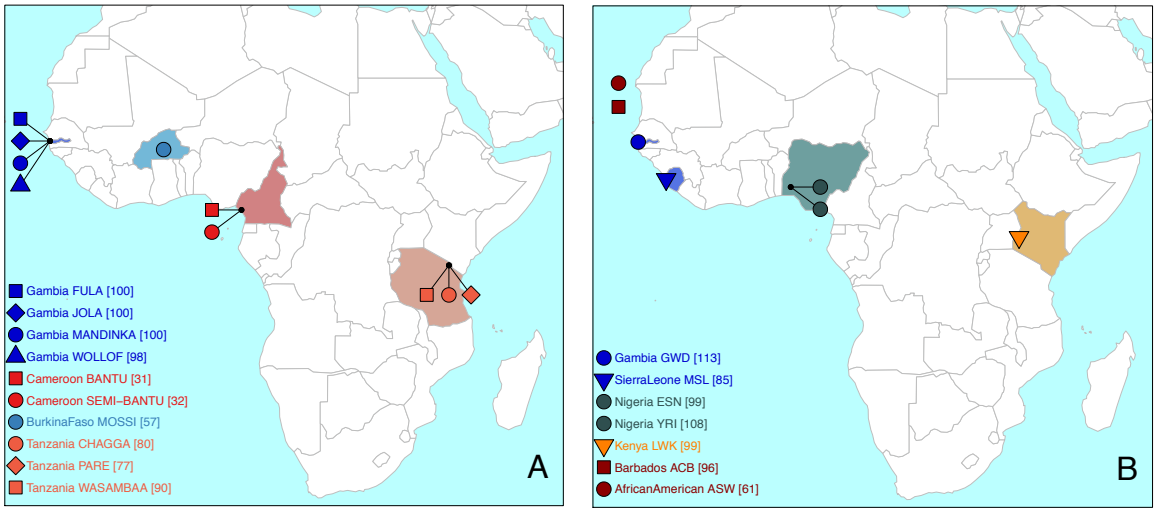


Figure S1. Geographic origin of the individuals collected for sequencing by MalariaGEN partners (A) and African individuals sequenced in the 1000 Genomes Project Phase 3 (B). The approximate sampling location is indicated on the map. The legend gives the country, population and number of individuals after QC. Admixed African American groups are arbitrarily indicated on the left of the map.

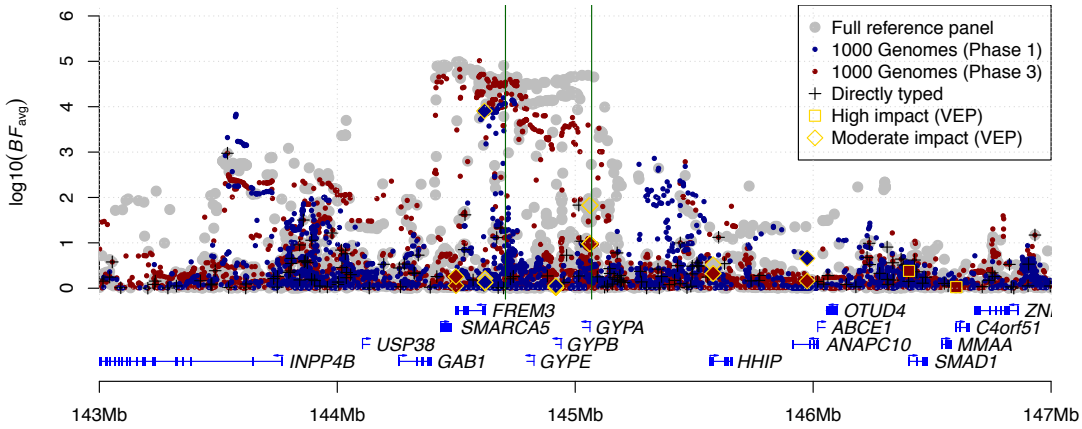


Figure S2. Impact of the reference panel on the evidence for association. Evidence for association (BF_{avg}, computed as in (*14*); y axis) across a 4 Mb region on chromosome 4 around the glycophorin genes (x axis). Colored circles represent SNPs and indels imputed from the 1000 Genomes Phase 1 reference panel (blue), the 1000 Genomes Phase 3 reference panel (red), or the full reference panel (grey). Genotyped SNPs are denoted with black plusses. Yellow-outlined diamonds and squares denote variants annotated by Variant Effect Predictor to have a functional effect classified as moderate or high impact according to the Ensembl IMPACT rating. Vertical green lines demarcate the glycophorin segmental duplication. Protein-coding genes are shown below.

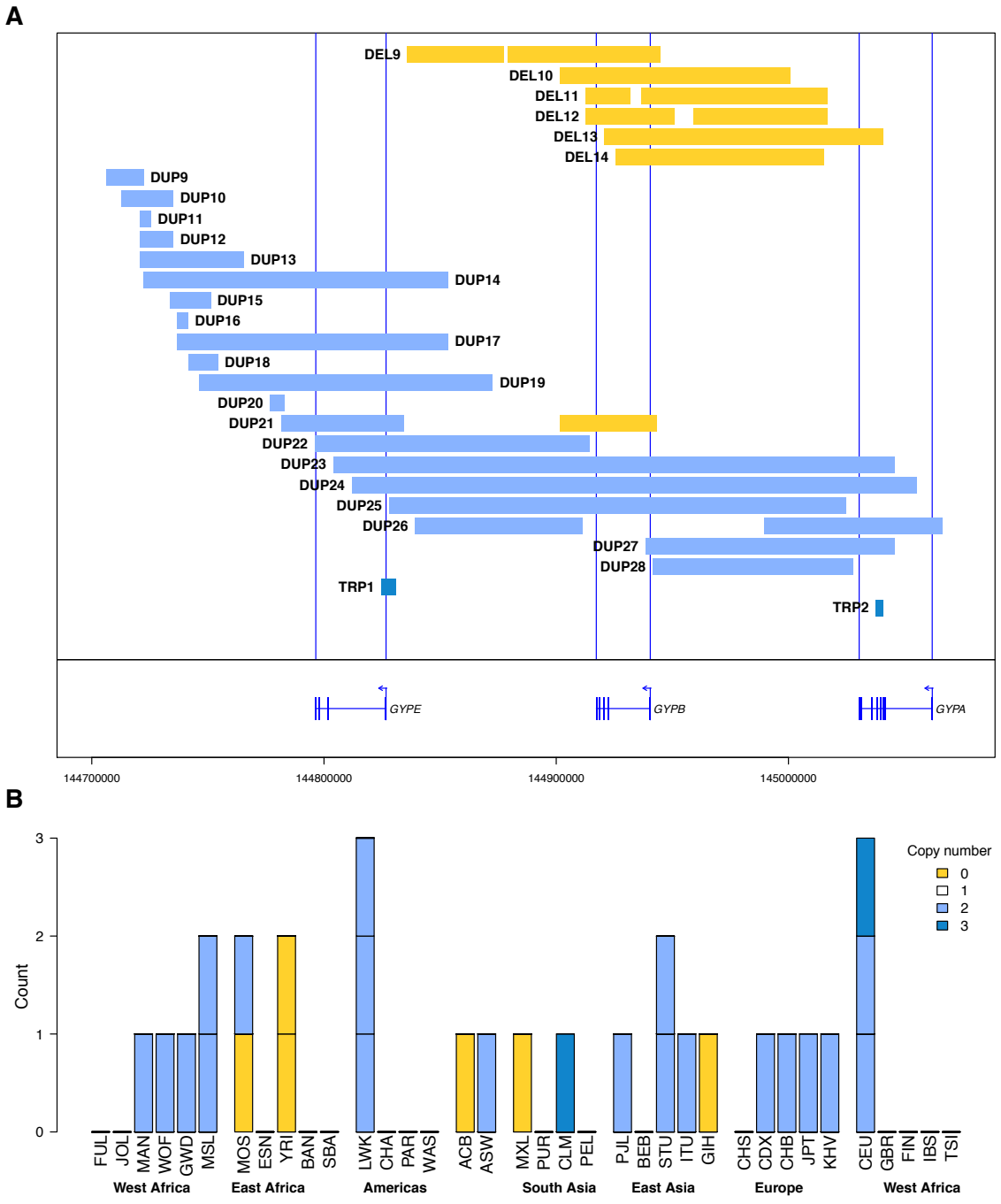


Figure S3. Copy number variants identified in a single unrelated individual (A) and their distribution across populations (B). As in Fig. 1B, variants are indicated with deletion in yellow, duplication in light blue, and triplication in dark blue. Variants are numbered within category by position from left to right. Blue vertical lines mark the locations of the three glycophorin genes, which are shown below the variants. We note several caveats about interpreting these singleton variants (*22*).

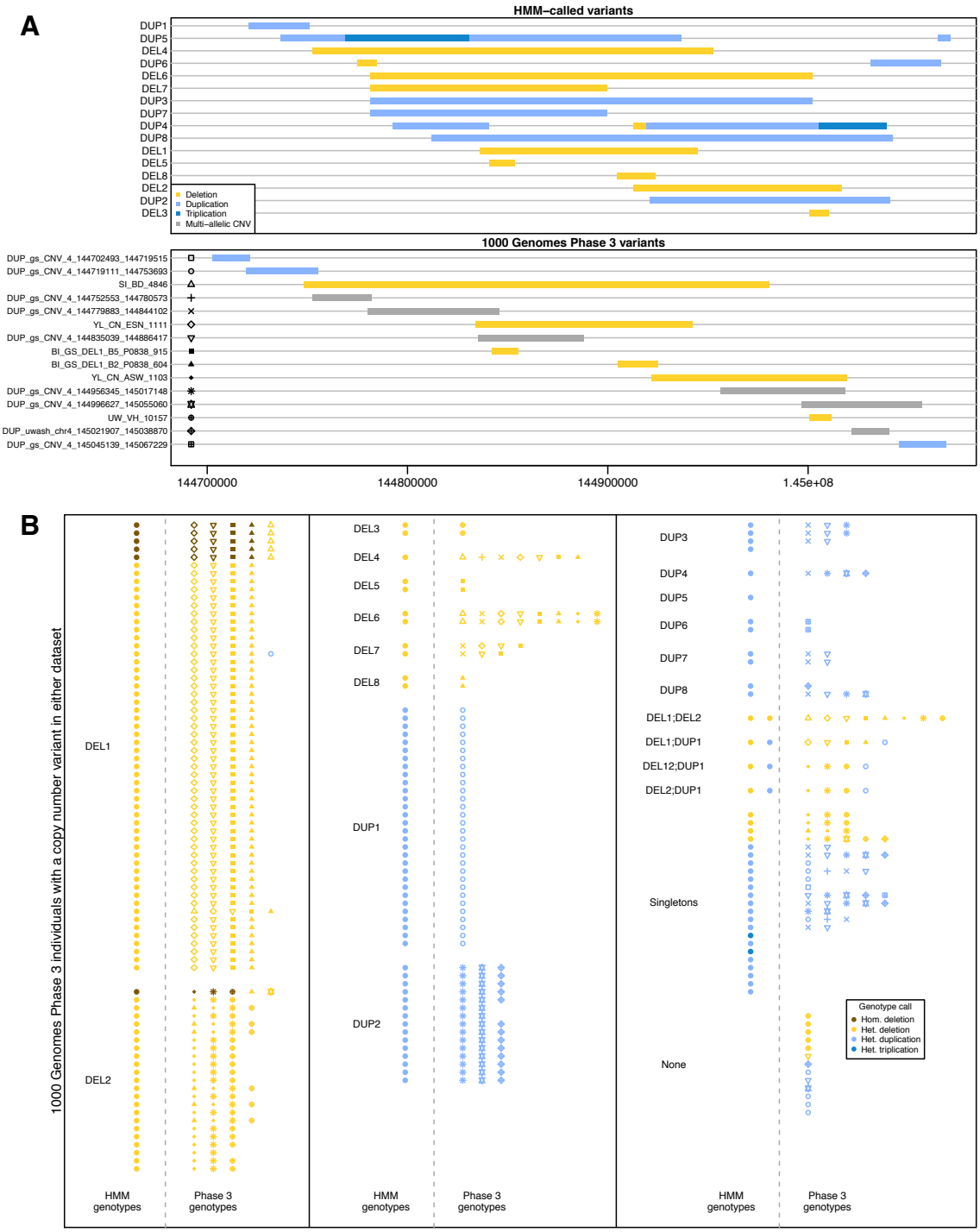


Figure S4. Comparison of HMM and 1000 Genomes CNV genotype calls for 1000 Genomes Phase 3 individuals. **(A)** Non-singleton CNVs in the glycophorin region called by our HMM method (top) and CNVs called by the 1000 Genomes Phase 3 structural variant analysis (bottom), ordered by position. Variants are colored by the copy number of the non-reference allele, with deletion in yellow, duplication in blue, triplication in dark blue, and variants from the 1000 Genomes

call set that have multiple non-reference alleles (here all have one deletion allele and one duplication allele) in gray. **(B)** CNV genotypes for the 187 individuals with a non-reference allele at a CNV in either of the call sets. In each panel of the plot, a row represents an individual, with the HMM genotype calls on the left of the dashed line labeled with the variant(s) carried, and the 1000 Genomes genotype calls on the right, with variants indicated by the shapes designated in **(A)**. When an individual carries non-reference alleles at more than one variant, they are positioned next to each other on the same row, colored by each genotype call.

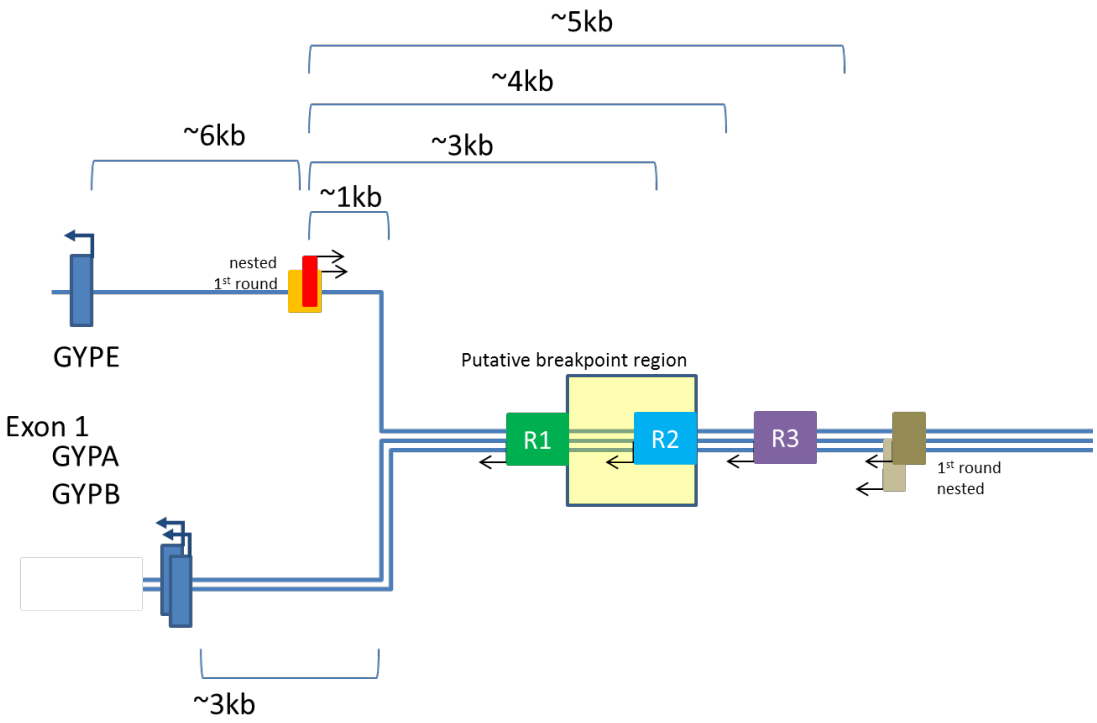


Figure S5. Schematic of the homology upstream of the glycophorin genes and PCR design to identify the DEL1 breakpoint. DNA is represented by the blue lines and shows the separation in homology ~3 kb upstream of the transcription start sites, where the *GYPE* unit of the segmental duplication contains ~3 kb of sequence not present in the *GYPB* or *GYPA* units. The putative breakpoint for DEL1 is located within the fully homologous region approximately 3 kb from this split in homology (indicated by the yellow box). The approximate locations of the first round, nested, and three sequencing primers (R1, R2, and R3) are marked.

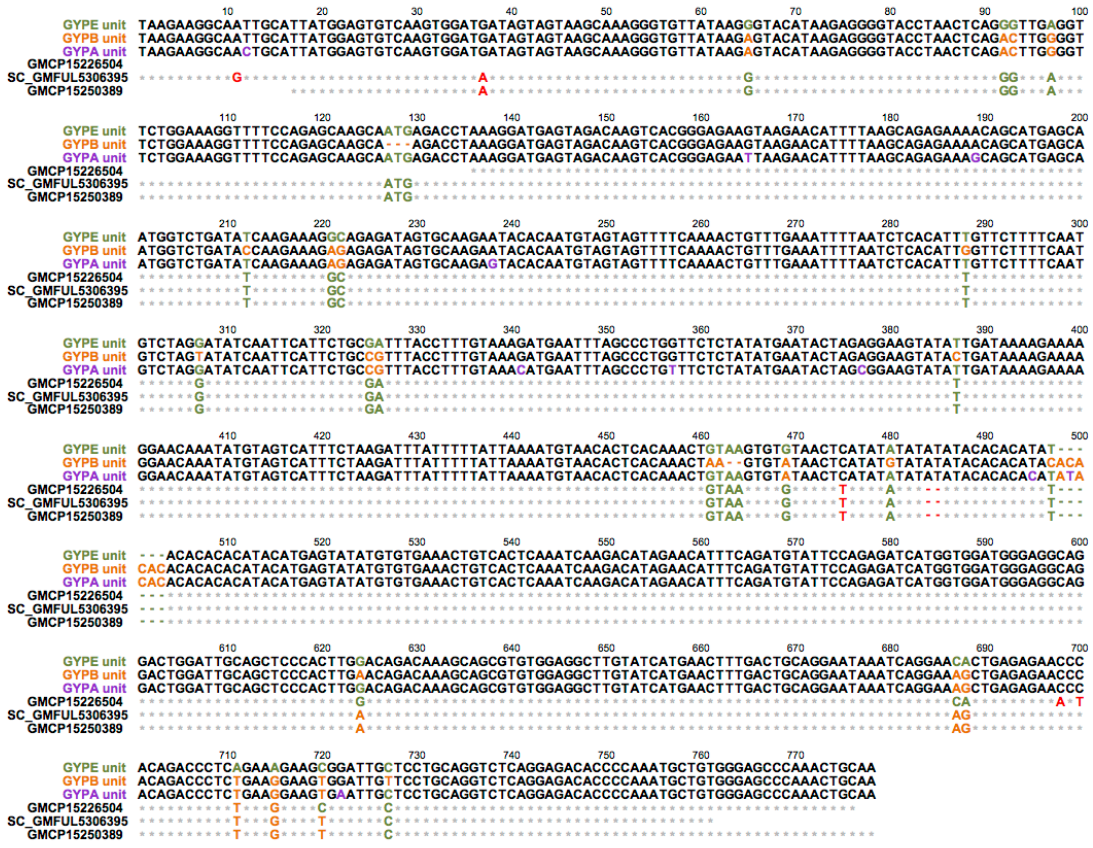


Figure S6. Sanger sequence across the DEL1 breakpoint. The top three lines show a multiple sequence alignment of homologous sequences upstream of the transcription start site of the three glycophorin genes from chr4:144834663-144835434, chr4:144944900-144945672, and chr4:145066241-145067018 in the *GYPE*, *GYPB*, and *GYPA* units of the glycophorin segmental duplication, respectively. Sanger sequences from the sequencing primer GYP_DEL1_R2 (**Table S5**) are shown for a control individual with no CNVs (GMCP15226504), one sequenced individual genotyped as a DEL1 homozygote by the HMM (SC_GMFUL5306395) and one GWAS individual imputed as a DEL1 homozygote (GMCP15250389). Because we are looking for a switch from *GYPE* to *GYPB*, differences between the reference sequence for the *GYPE* and *GYPB* units are colored green and orange, respectively. The Sanger sequences and the *GYPA* sequence are colored orange if they match *GYPB* at that site and green if they match *GYPE*. Grey stars indicate positions in the Sanger sequences that match both *GYPE* and *GYPB*. Where the Sanger sequences match none of the reference sequences, they are colored red. Where the *GYPA* sequence differs from both *GYPE* and *GYPB* it is colored purple. The DEL1 breakpoint is between positions 503 and 624 in this alignment.

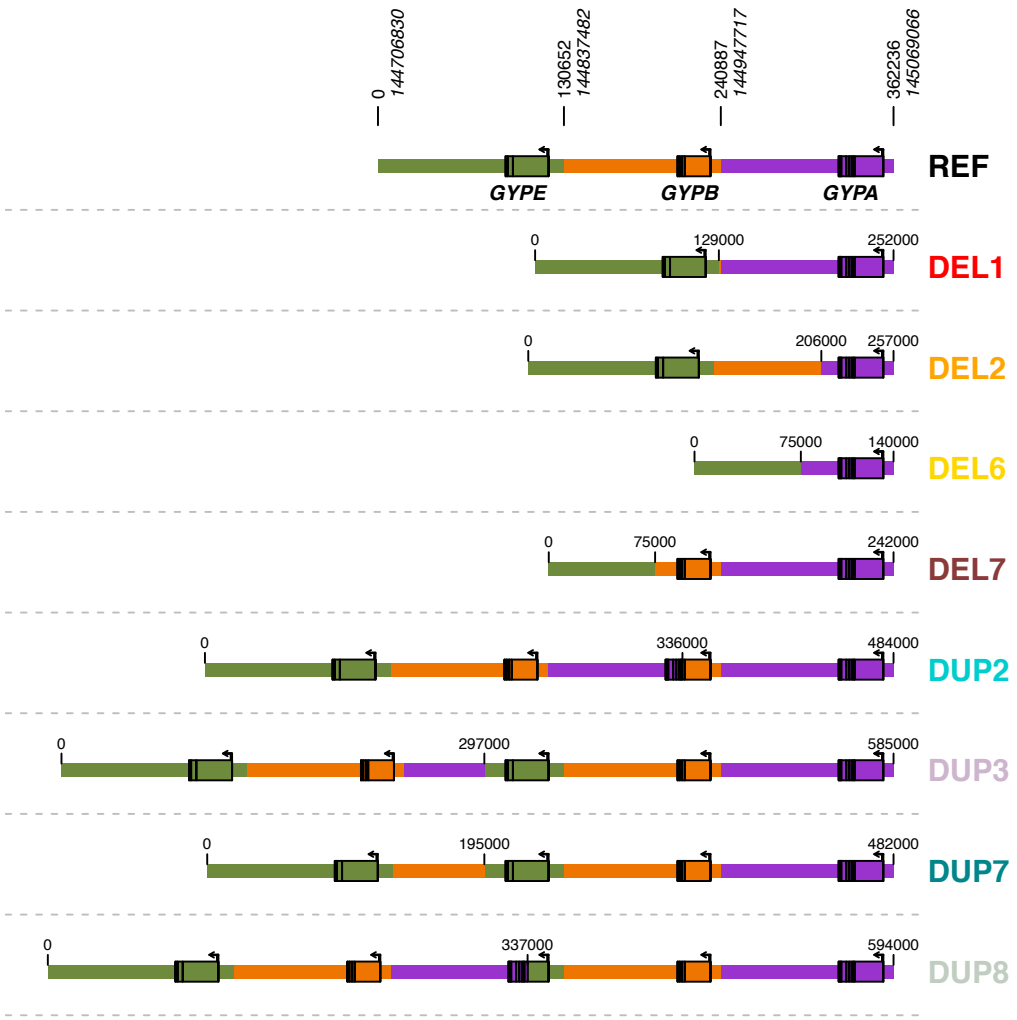


Figure S7. Predicted chromosomal structure of CNVs with a single pair of homologous breakpoints. A reference chromosome is shown at the top, with relative and absolute positions (GRCh37) indicated above followed by the eight CNVs with a single pair of homologous breakpoints. The three segmentally duplicated sequences are indicated by the three colors, with the gene in each shown in the same color and exons in black. Numbers above each variant indicate the total length and the approximate position of the breakpoint. DUP2 and DUP8 are predicted to create hybrid glycophorin genes (*GYPB-A* and *GYPE-A*, respectively).

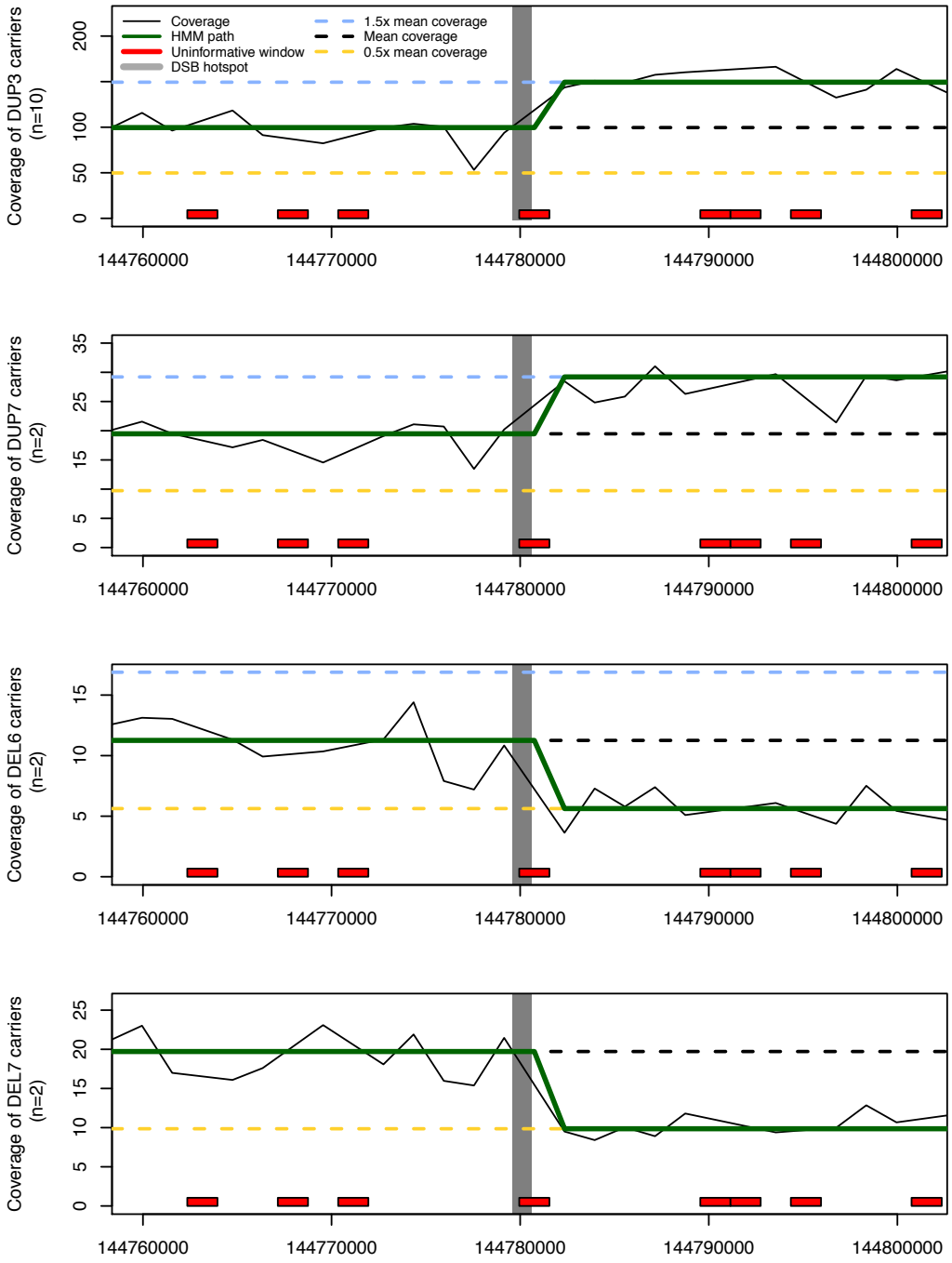


Figure S8. Coverage around the CNV breakpoint shared by DUP3, DUP7, DEL6, and DEL7. Coverage, shown as a black line connecting the midpoint of each informative 1600 bp window, is pooled across heterozygous carriers for each variant and averaged over sites with mappability > 0.9. The mean and expected coverage for other copy numbers is estimated outside the glycophorin region and shown with horizontal dashed lines. The HMM path is shown in green. Uninformative windows (<400 mappable sites) are marked in red along the bottom. The DSB hotspot from (*24*) is shaded in gray.

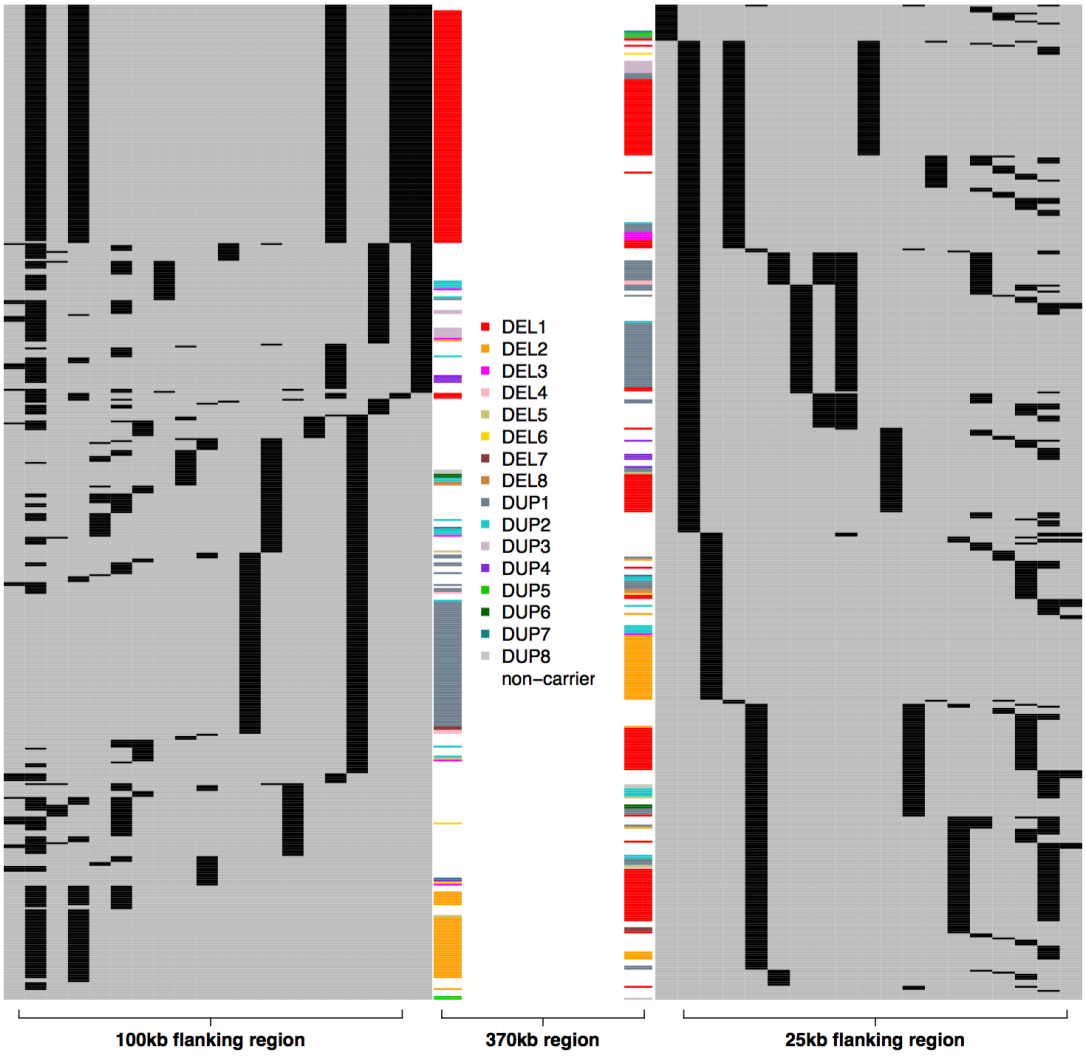


Figure S9. Phased haplotypes carrying CNVs in the reference panel. Reference panel haplotypes over the 100 kb (left) and 25 kb (right) flanking the glycophorin region, with minor allele calls in black and major allele calls in gray. Variants were thinned outwards in each direction to have a minimum 2% minor allele frequency and pairwise r^2 of at most 0.5. Haplotypes were then clustered by ordering lexicographically outward, separately to the left and right. For visualisation purposes, non-CNV-carrying haplotypes are condensed so that identical non-carrier haplotypes fill at most ten rows. Colors indicate CNV-carrying haplotypes according to the legend. HG02554 contributes a haplotype visible as the single purple segment clustering away from other DUP4 haplotypes at the bottom left of the plot; the corresponding right segment clusters with other DUP4 haplotypes indicating a probable switch error.

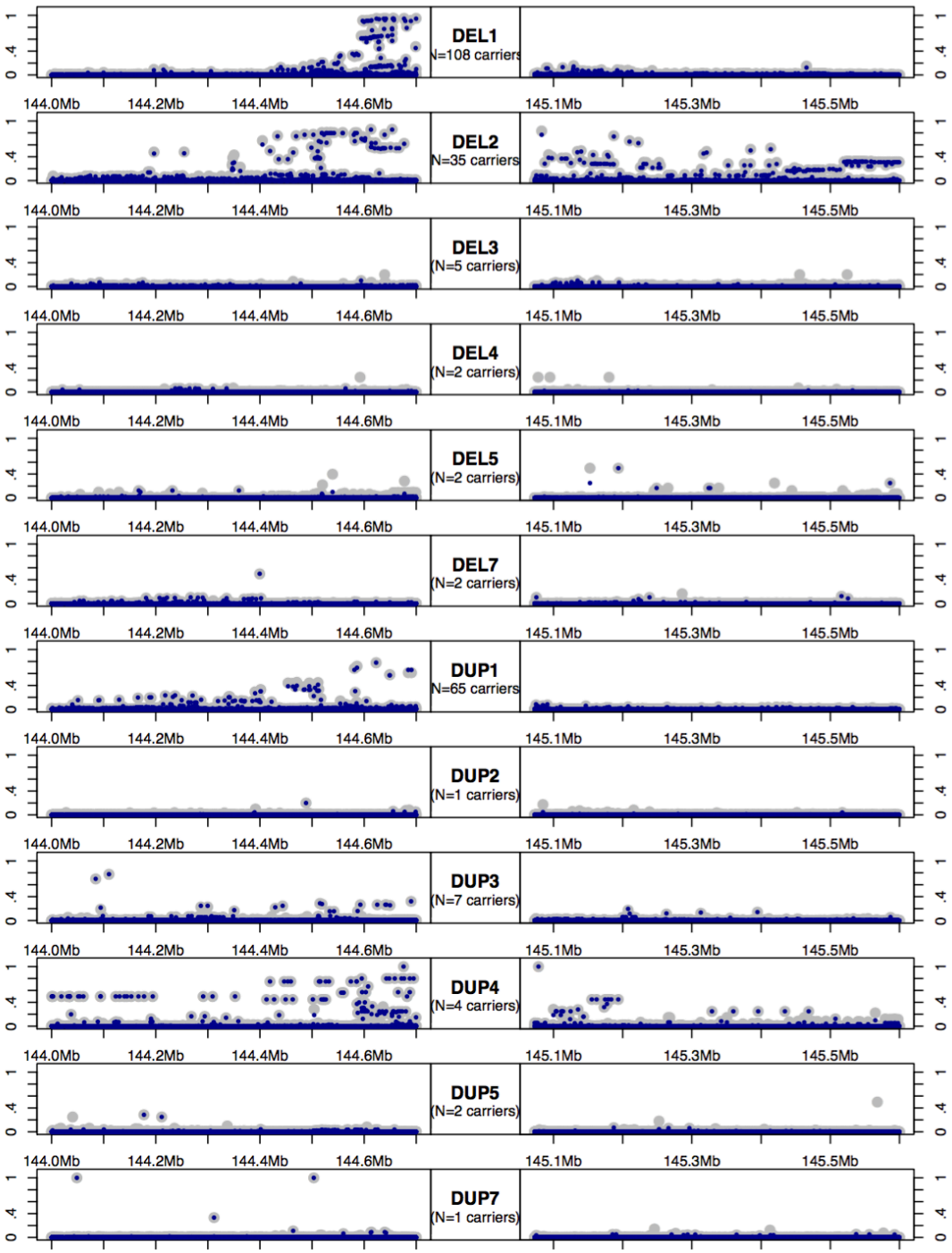


Figure S10. Linkage disequilibrium (LD) between CNVs and SNPs surrounding the glycophorin region. Grey dots show the genotypic r^2 (correlation between allele dosage) between each variant in the regions immediately flanking the glycophorin region and the labelled CNVs, while blue dots show LD computed from SHAPEIT-phased haplotypes. LD is computed using the 1046 African reference panel individuals without parents in the combined reference panel, and only CNVs observed in African individuals are shown.

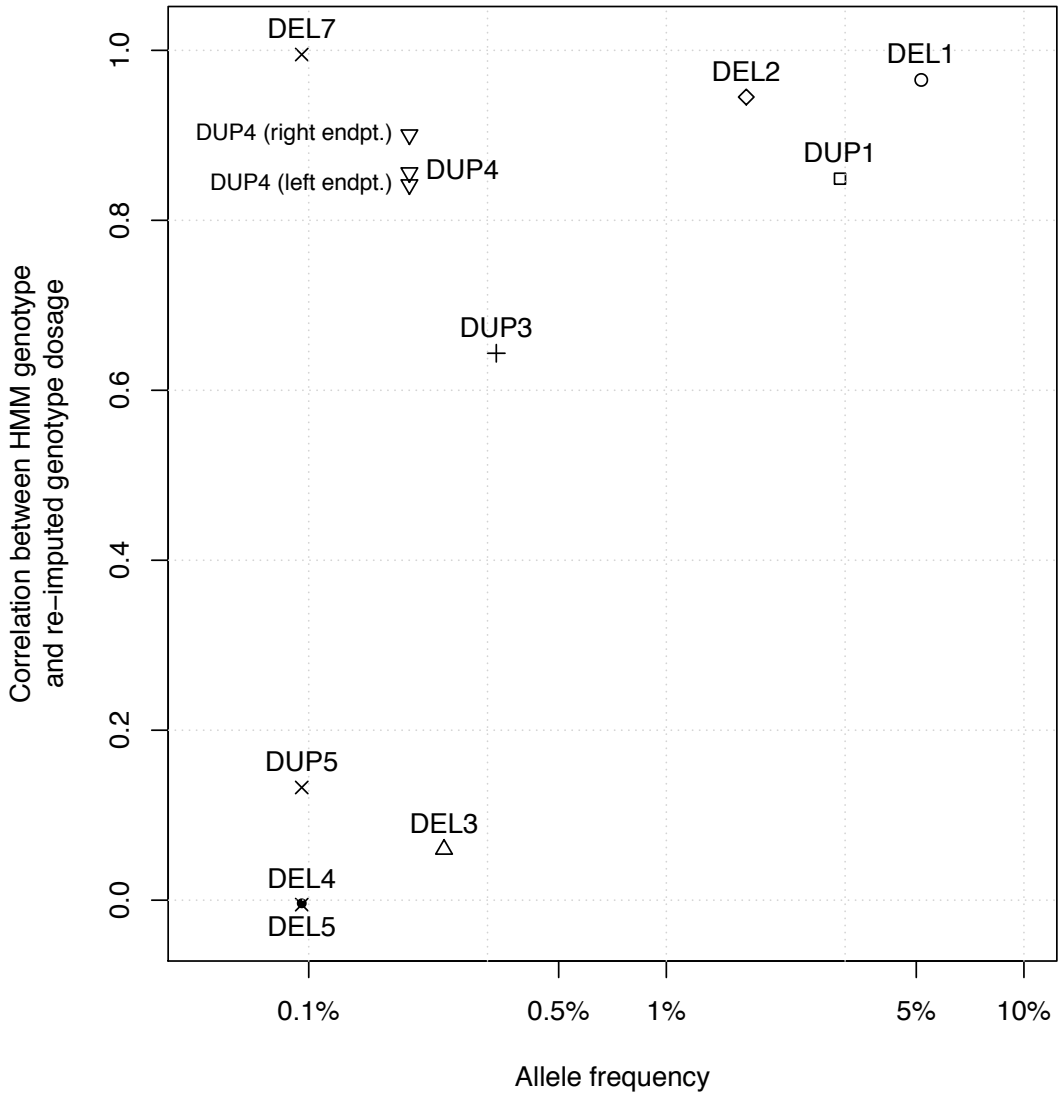


Figure S11. Assessment of imputation performance by cross-validation. The correlation between HMM-based genotype calls and re-imputed genotype dosage (y axis) in cross-validation, plotted against allele frequency (x axis), for the 10 CNVs that were observed in at least two unrelated African reference panel samples. Points shown reflect imputation at the CNV midpoint except where labelled.

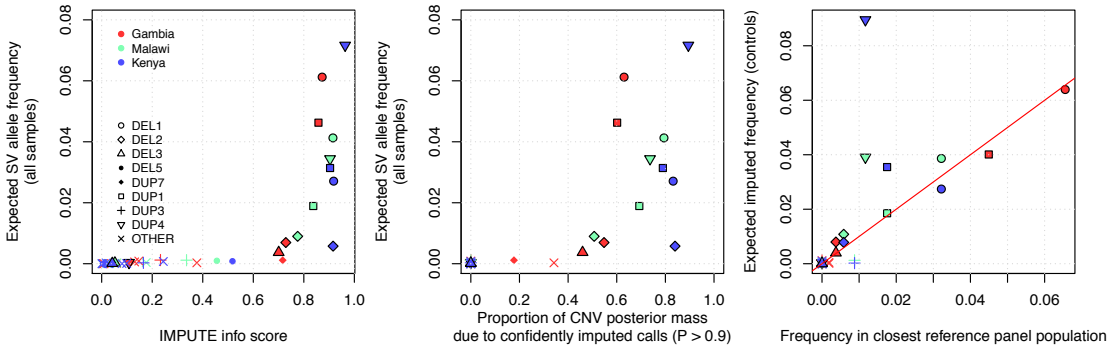


Figure S12. Imputation performance. Left and middle panels: IMPUTE info score (x axis, left panel) and proportion of posterior mass on non-reference CNV calls that is due to confidently imputed genotypes (x axis, middle panel), plotted against expected imputed allele frequency (y axis) for all CNVs imputed into the Gambia, Malawi and Kenya GWAS data. Points are colored to denote population, and shape denotes the CNV. Right panel: expected imputed frequency of each CNV in control samples (y axis) plotted against the estimated frequency in the geographically closest reference panel population (the Gambia - Gambian reference panel groups; Malawi, Kenya - Tanzanian reference panel groups).

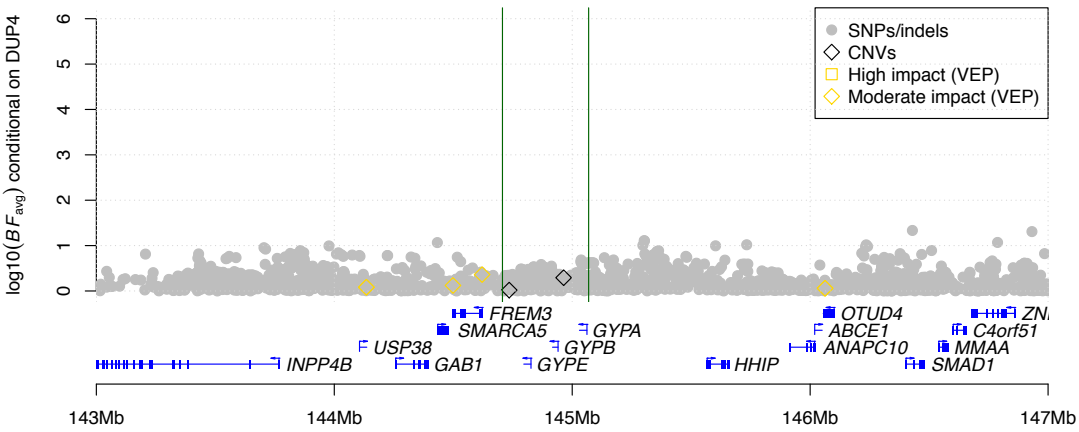


Figure S13. Association signal conditional on DUP4. Evidence for association (BF_{avg}, computed as in (*14*); y axis) after conditioning on imputed genotypes at DUP4. Both SNPs and indels (grey circles) and CNVs (black diamonds) are shown. For other details, see legend of **Fig. S2**.

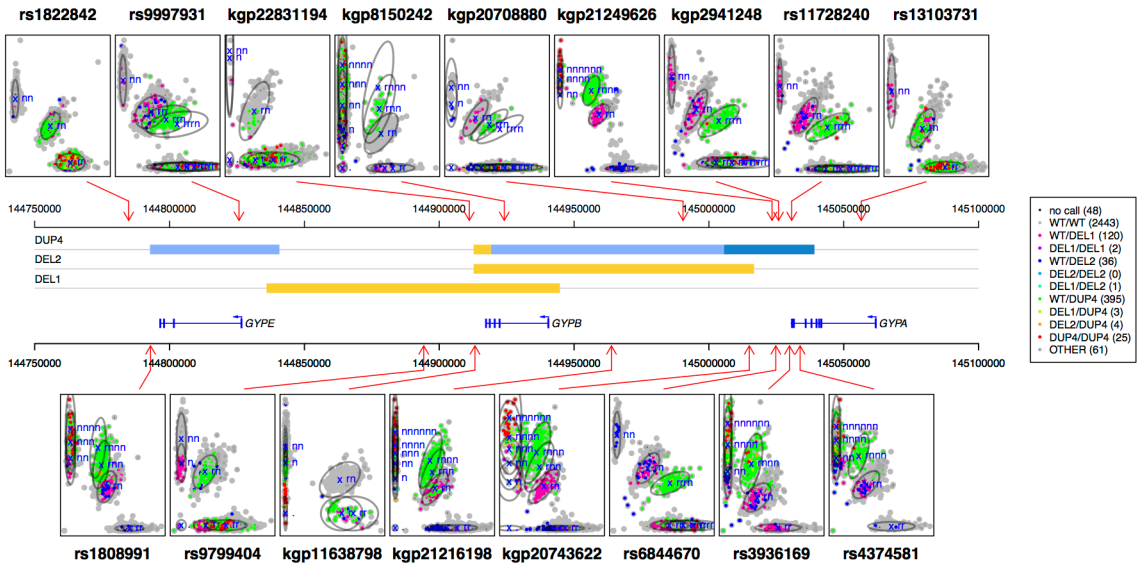
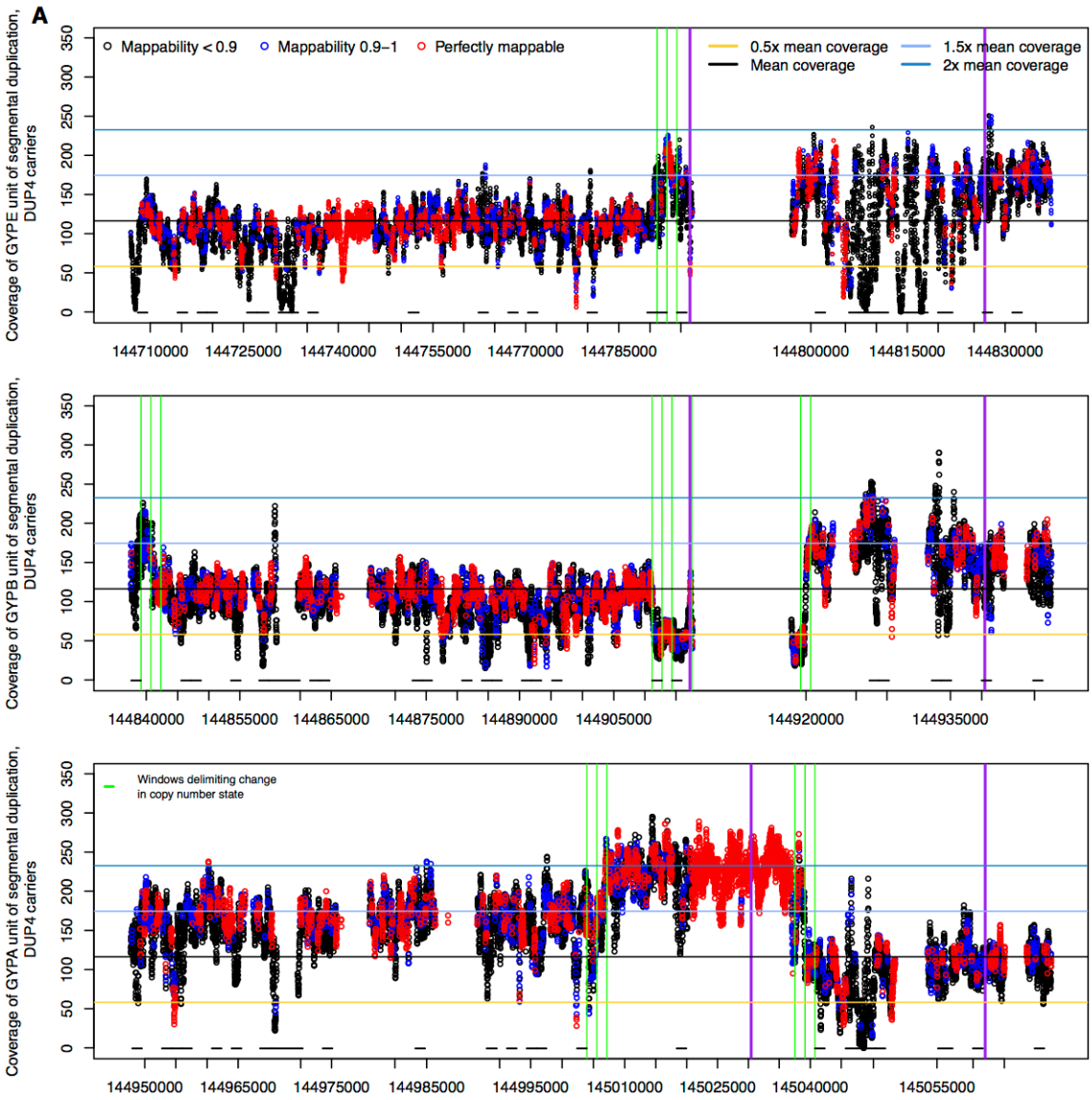


Figure S14. Detail of intensity-based CNV calling in Kenya. Top and bottom: normalized microarray intensities for the reference (x axis) and non-reference (y axis) allele probes, for a subset of Illumina microarray SNP assays mapping to the region (plot labels) as described in (*22*). Individuals are colored by inferred CNV genotype according to the legend on the right, with 'no call' referring to individuals having less than 75% posterior probability of any CNV genotype. The number of individuals with each inferred type is also given in the legend. Blue crosses and grey ellipses denote the modeled mean and covariance of each possible genotype cluster given the set of CNVs considered, with the corresponding genotype given in blue text; the clustering model is fit based on intensity values as described in (*22*). Middle: the mapping position of probes and copy number profile of CNVs included in the analysis, and the position of the glycophorin genes.

Figure S15. Pooled coverage across DUP4 carriers along a multiple sequence alignment of (A) the glycophorin segmental duplication and (B) the subset of the segmental duplication encoding the glycophorin genes. Coverage is shown per site, pooled across the nine heterozygous DUP4 carriers. The three rows correspond to the three units of the segmental duplication, with sites directly above each other in homologous positions from the multiple sequence alignment. Gaps in each panel therefore represent sites not present in a repeat unit. Sites are colored by mappability and the expected coverage for different copy numbers, estimated from coverage outside the glycophorin region, is indicated with horizontal lines. Black bars along the bottom of the plot indicate the 1600 bp windows that were excluded from the copy number inference due to too few mappable sites (<400). Vertical green lines demarcate pairs of windows where copy number state transitions occur. The positions of the three glycophorin genes are indicated with vertical purple lines in **(A)**, and this region is shown in **(B)** with the exons numbered and shaded in purple. As is customary for these glycophorin genes, exon numbering corresponds to the exons of *GYPA* (pseudoexons in *GYPB* and *GYPE* not shown). For ease of display, only every 10^{th} site is plotted.


A
Mappability < 0.9
Mappability 0.9–1
Perfectly mappable
0.5x mean coverage
Mean coverage
1.5x mean coverage
2x mean coverage
Coverage of GYPE unit of segmental duplication, DUP4 carriers
Coverage of GYPB unit of segmental duplication, DUP4 carriers
Coverage of GYPA unit of segmental duplication, DUP4 carriers
Windows delimiting change in copy number state
0
50
100
150
200
250
300
350
144710000
144725000
144740000
144755000
144770000
144785000
144800000
144815000
144830000
144840000
144855000
144865000
144875000
144890000
144905000
144920000
144935000
144950000
144965000
144975000
144985000
144995000
145010000
145025000
145040000
145055000

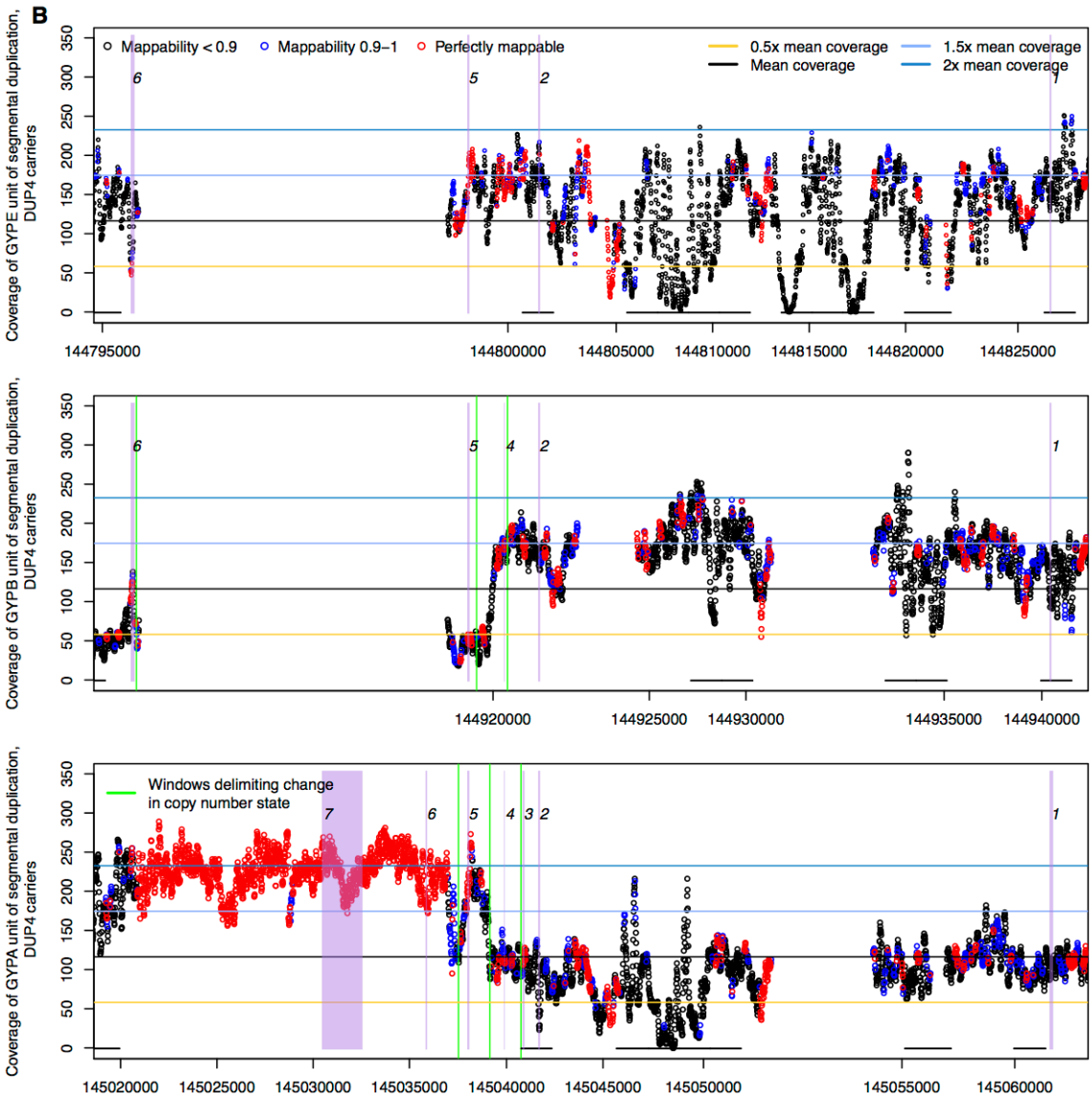
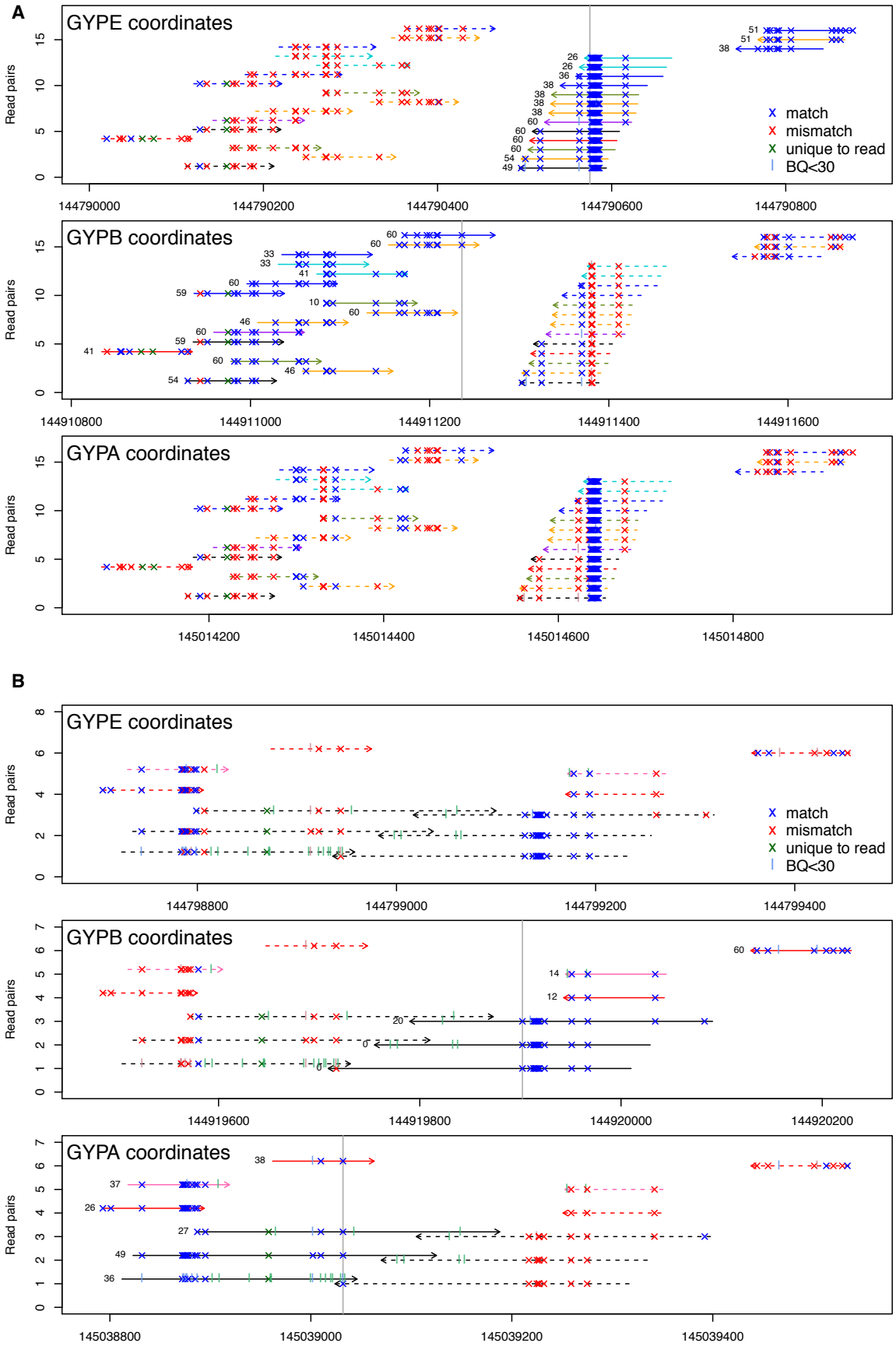
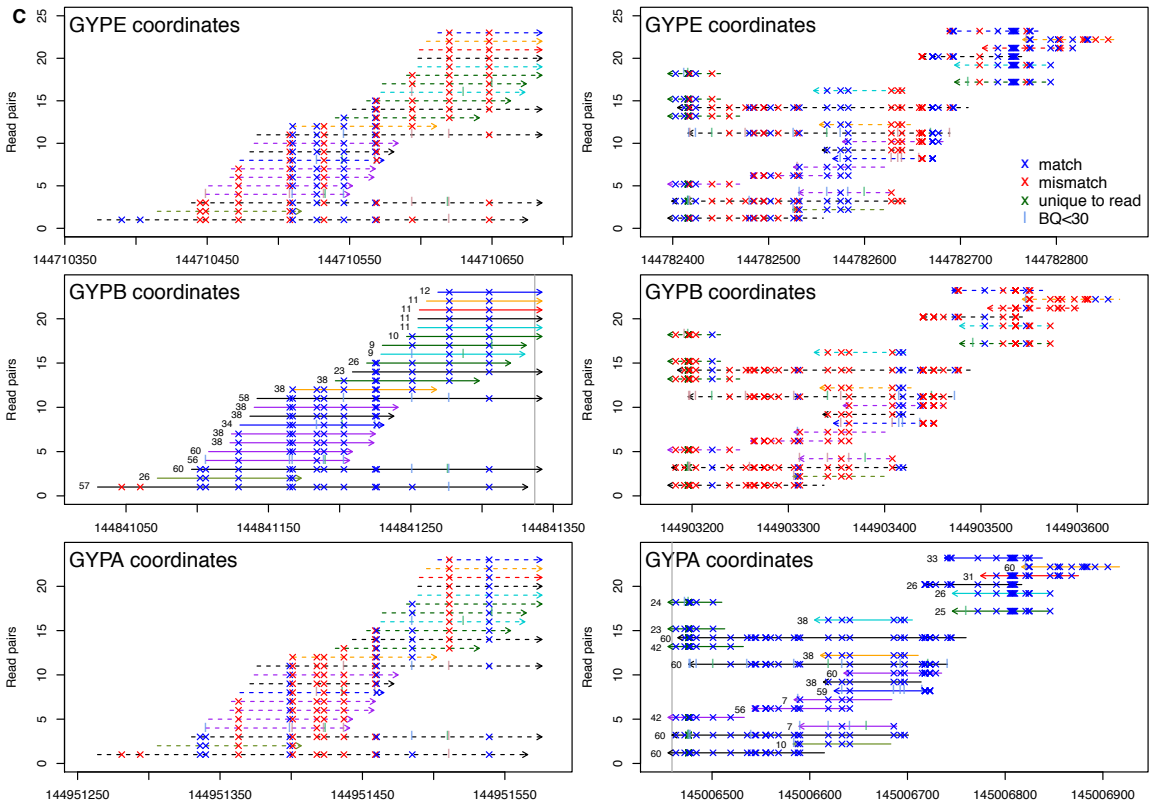

B
Mappability < 0.9
Mappability 0.9–1
Perfectly mappable
0.5x mean coverage
Mean coverage
1.5x mean coverage
2x mean coverage
Coverage of GYPE unit of segmental duplication, DUP4 carriers
Coverage of GYPB unit of segmental duplication, DUP4 carriers
Coverage of GYPA unit of segmental duplication, DUP4 carriers
Windows delimiting change in copy number state
144795000
144800000
144805000
144810000
144815000
144820000
144825000
144920000
144925000
144930000
144935000
144940000
145020000
145025000
145030000
145035000
145040000
145045000
145050000
145055000
145060000

Figure S16. Discordant read pairs supporting the connections between copy number breakpoints. Read pairs with discordant mappings near the DUP4 breakpoints are shown at the mapped position (indicated by a solid line) as well as the two homologous positions in the segmental duplication (dashed lines). The mapping quality assigned by BWA mem is indicated next to the mapped position. Matches and mismatches to variable sites in the multiple sequence alignment are indicated in blue or red respectively, with an 'x' for sites with high base quality (BQ≥30) or a '|' for sites with low base quality (BQ<30); positions where the read differs from all three reference positions are shown in green. Positions that are not indicated are identical across all three reference sequences and the read. The colors of the read lines indicate which individual carrier they are from. **(A)** First and third copy number change points, showing a connection from segment 2 to segment 1 (as numbered in **Fig. 6**). This breakpoint is delimited to 144,790,429-144,790,575 in the *GYPE* unit and 144,911,236-144,911,381 in the *GYPB* unit. **(B)** Fourth and sixth copy number change points, showing a connection from segment 5 to segment 4. This breakpoint is delimited to 144,919,717-144,919,902 in the *GYPB* unit and 145,039,032-145,039,217 in the *GYPA* unit, in agreement with Sanger sequence across this region (**Fig. S19**). **(C)** Second and fifth copy number change points, showing a connection from segment 1 to segment 5. This breakpoint is delimited to 3 bp of microhomology from 144,841,338-144,841,340 in the *GYPB* unit and 145,006,456-145,006,458 in the *GYPA* unit. Vertical gray lines mark the informative positions between which the breakpoints can be localized.


A
GYPE coordinates
GYPB coordinates
GYPA coordinates
Read pairs
match
mismatch
unique to read
BQ<30
144790000
144790200
144790400
144790600
144790800
144910800
144911000
144911200
144911400
144911600
145014200
145014400
145014600
145014800
B
GYPE coordinates
GYPB coordinates
GYPA coordinates
Read pairs
match
mismatch
unique to read
BQ<30
144798800
144799000
144799200
144799400
144919600
144919800
144920000
144920200
145038800
145039000
145039200
145039400


c
GYPE coordinates
Read pairs
144710350
144710450
144710550
144710650
GYPE coordinates
Read pairs
144782400
144782500
144782600
144782700
144782800
match
mismatch
unique to read
BQ<30
GYPB coordinates
Read pairs
144841050
144841150
144841250
144841350
GYPB coordinates
Read pairs
144903200
144903300
144903400
144903500
144903600
GYPA coordinates
Read pairs
144951250
144951350
144951450
144951550
GYPA coordinates
Read pairs
145006500
145006600
145006700
145006800
145006900

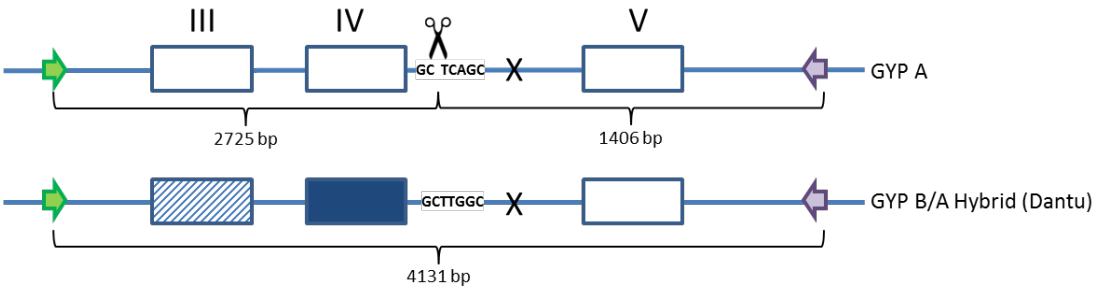


Figure S17. Schematic of the *GYPA* and *GYPB-A* hybrid PCR products. Exons are indicated by boxes and labelled with respect to *GYPA*. PCR primers (**Table S10**) are indicated by arrows and span a 4.1 kb region including exons 3 (III) to 5 (V). *GYPA* exons are colored white and *GYPB* exons are colored blue. The pseudo-exon of GYPB (III) is indicated by the blue hatching. The putative location of the breakpoint between *GYPB* and *GYPA* giving rise to the hybrid is identified with 'X', and the BlpI restriction site sequence (located at 145039618-145039624 in the *GYPA* reference and at 144799538-144799544 in the *GYPB* reference) is highlighted with an indication that the *GYPA* sequence is cleaved. The resulting fragment sizes are given (schema not to scale).

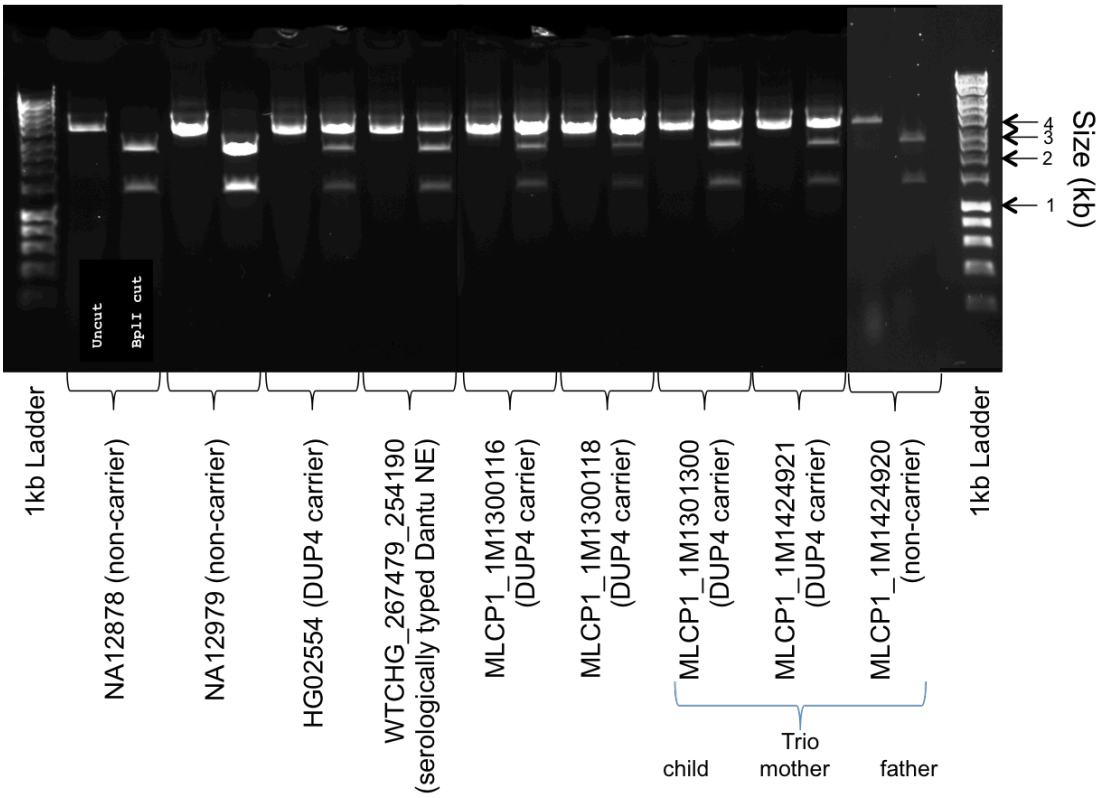


Figure S18. Agarose gel image of restriction digests of PCR fragments from *GYPA* and *GYPB-A* hybrid products. Samples are displayed in pairs showing the uncut PCR product (from PCR primers in **Table S10** and shown in **Fig. S17**) on the left and the BlpI digested material on the right of each well pair. The enzyme cuts the *GYPA* sequence into two fragments (2.7 kb and 1.4 kb), while leaving the Dantu *GYPB-A* hybrid intact at 4.1 kb. The prediction of hybrid carrier status is indicated. The first three samples were all obtained from the Coriell Biorepository and are part of the HapMap/1000 Genomes projects. Sample WTCHG_267479_254190 is from a serologically defined Dantu carrier obtained from the International Blood Group Reference Laboratory in Bristol, UK. The other five samples are from the collection forming the Kenyan GWAS dataset, and include a trio as indicated.

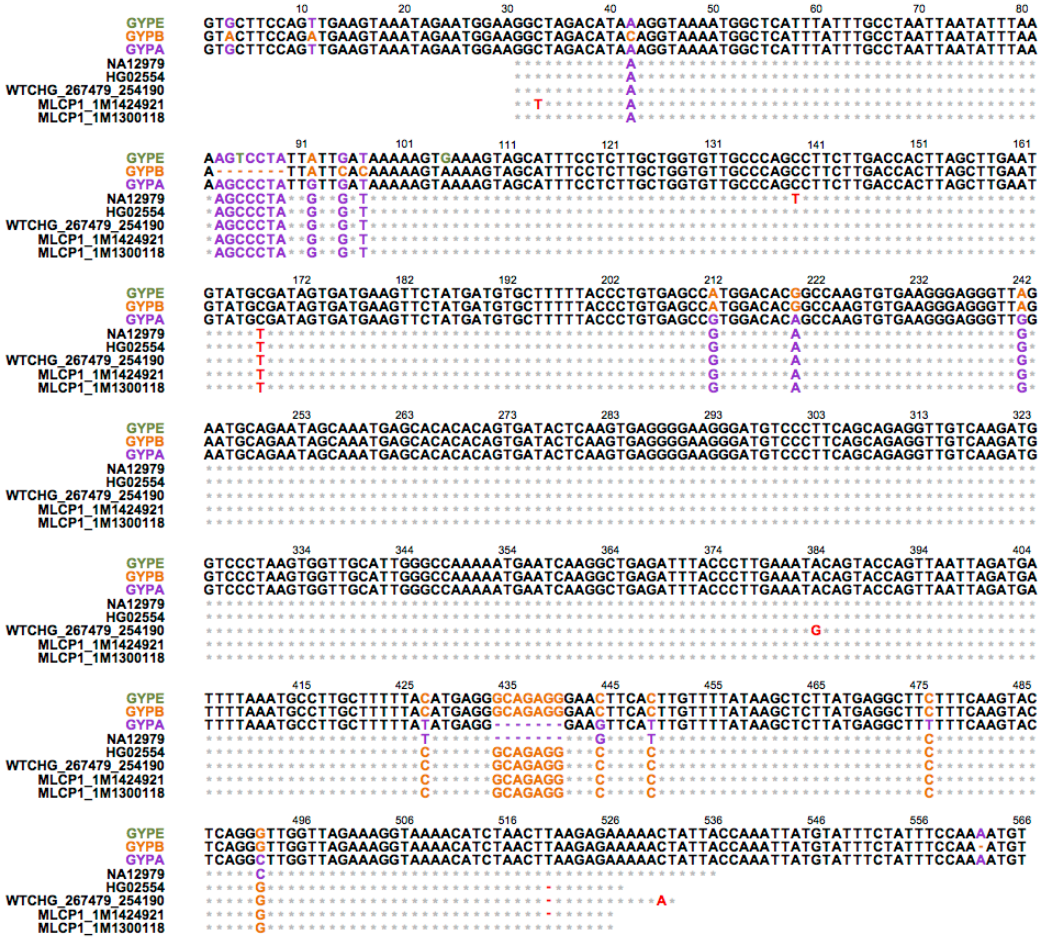


Figure S19. Sanger sequence across the *GYPB-A* hybrid gene breakpoint. The top three lines show a multiple sequence alignment of homologous sequence from intron 4 of the three glycophorin genes from chr4:144798703-144799268, chr4:144919483-144920037, and chr4:145038791-145039349. The sequences below are consensus Sanger sequences from a 1000 Genomes non-DUP4 carrier (NA12979), the 1000 Genomes DUP4 carrier (HG02554), the individual serologically typed as a Dantu carrier (WTCHG_257579_254190), and two Kenyan individuals imputed to carry DUP4. The bases highlighted in purple indicate a match to *GYPA* and orange a match to *GYPB* at sites that differentiate *GYPA* from *GYPB*. Sequenced bases represented by a grey star match both *GYPA* and *GYPB* reference sequence, and those colored in red are different from all reference locations. Sites in the alignment unique to *GYPE* are shown in green in the *GYPE* reference sequence. The hybrid breakpoint is between positions 242 and 427 in this alignment. The sequencing primers are located at positions 5-30 and 537-562 (**Table S10**).

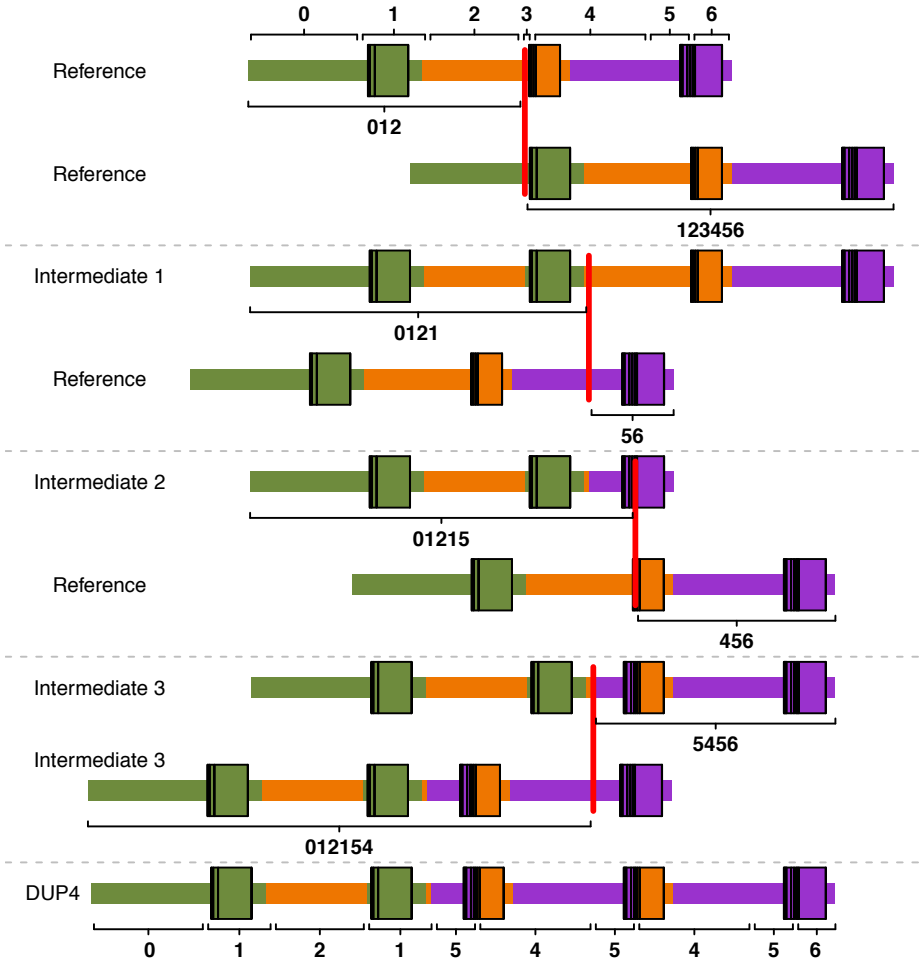


Figure S20. Model of a possible series of unequal crossover events leading to DUP4. This model shows three unequal crossover events, marked by red vertical lines, that correspond to each of the three breakpoint pairs. This leads to a chromosome with a single *GYPB-A* hybrid, and an additional event between two such chromosomes then occurs to duplicate the hybrid. Alternative orderings of these steps as well as more complex mutational events or additional involvement of two non-reference chromosomes are also possible (*22*). Segments of the reference sequence are numbered as in **Fig. 6**.

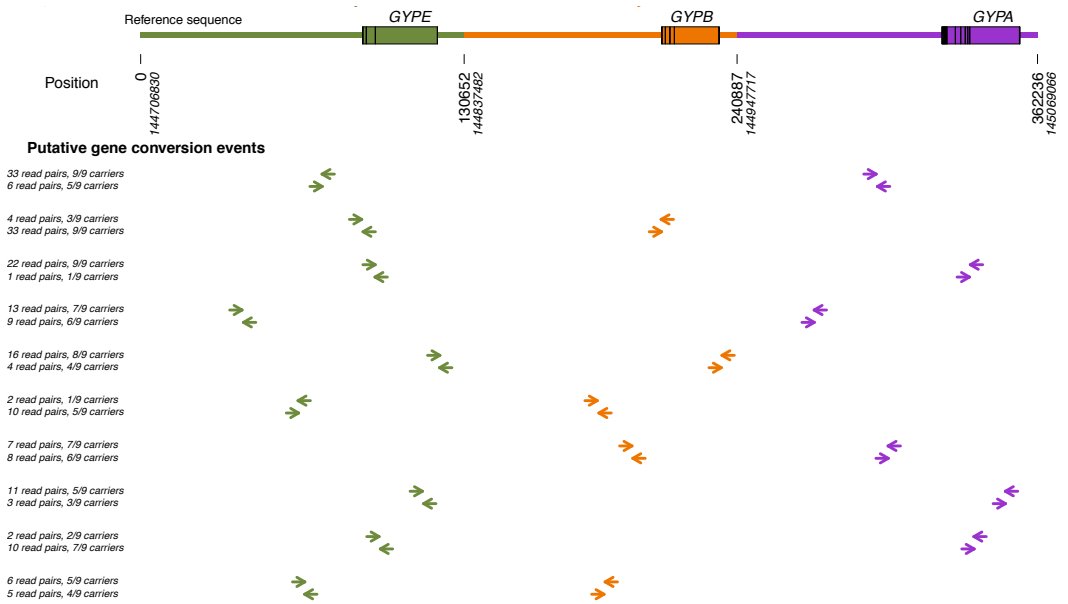
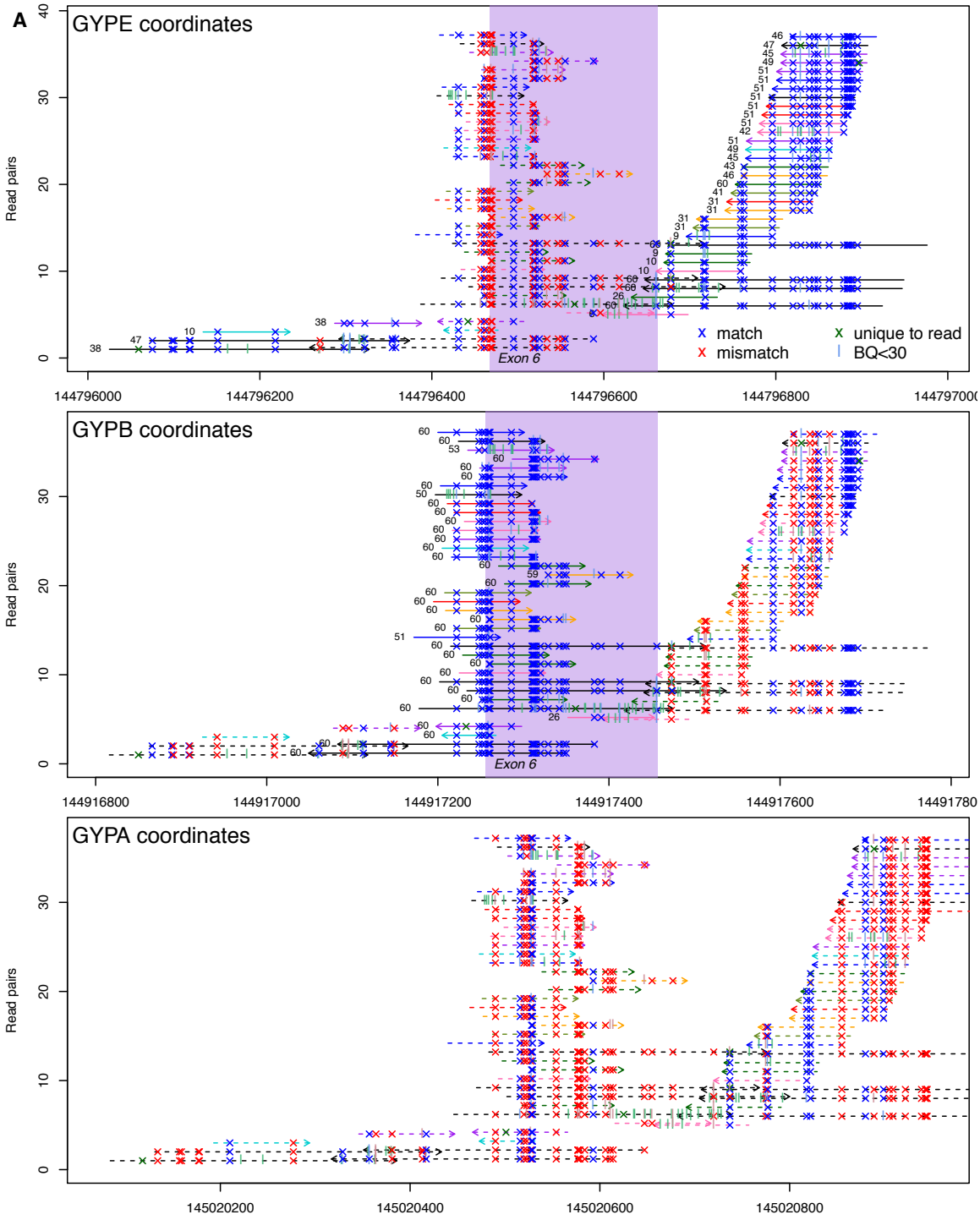
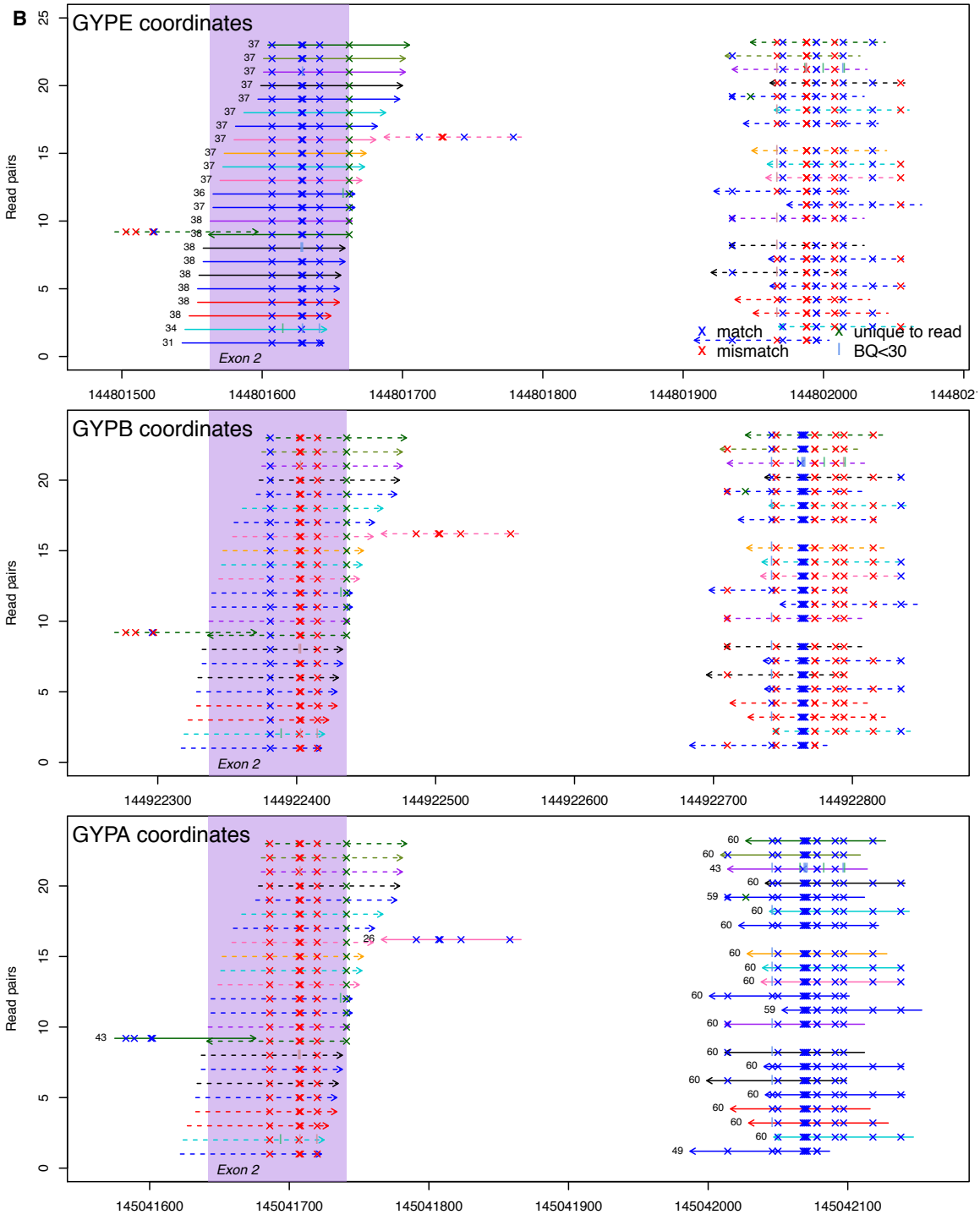


Figure S21. Discordant read pairs supporting putative gene conversion events in DUP4 carriers. Each set of arrows represents a group of read pairs mapped to the same location (both ends of both reads within 1 kb of each other), where the distance between the paired reads is >1000 bp. Arrows indicate the strand and position mapped to the human reference, shown above, with colors indicating the segmentally duplicated sequence and coordinates both relative to the segmental duplication and the GRCh37 assembly. These mapping patterns are consistent with gene conversion, where clustered read pairs indicate a connection to another of the homologous locations and then back again. The number of read pairs and how many of the nine DUP4 carriers they were observed in is shown on the left, for each cluster of mapping locations. Clusters with more than 10 read pairs are shown.

Figure S22. Read pairs supporting a putative gene conversion event of *GYPB* exon 6 into *GYPE* (A) and of *GYPE* exon 2 into *GYPA* (B) in DUP4 carriers. Exons are marked in purple. Matches and mismatches to the multiple sequence alignment of the segmental duplication are marked as in **Fig. S16,** and the colors of the read lines indicate distinct individual carriers. These are the second and third events shown in **Fig. S21**, respectively; no others overlap exons. The event in **(B)** appears to reflect the M/N blood group polymorphism, which is determined by three SNPs in *GYPA* exon 2. The human reference sequence at both *GYPA* and *GYPB* encodes the N allele at these 3 SNPs, but the M allele is present in the reference at *GYPE* exon 2.


A
GYPE coordinates
Read pairs
Exon 6
match
mismatch
unique to read
BQ<30
144796000
144796200
144796400
144796600
144796800
GYPB coordinates
Exon 6
144916800
144917000
144917200
144917400
144917600
GYPA coordinates
145020200
145020400
145020600
145020800


B
GYPE coordinates
Read pairs
Exon 2
match
unique to read
mismatch
BQ<30
144801500
144801600
144801700
144801800
144801900
144802000
GYPB coordinates
Exon 2
144922300
144922400
144922500
144922600
144922700
144922800
GYPA coordinates
Exon 2
145041600
145041700
145041800
145041900
145042000
145042100

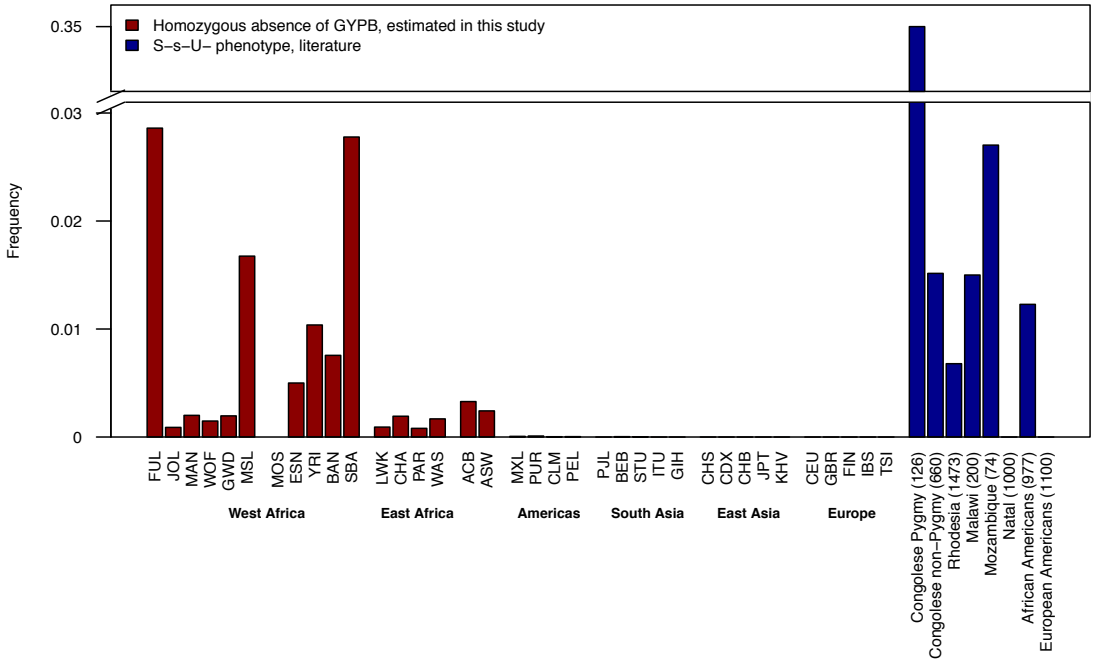


Figure S23. Frequency of S−s−U− phenotype inferred from *GYPB* deletion allele frequencies in this study and from serological reports in the literature. The estimates for Congolese are from (*77*), estimates for European and African Americans are from (*78*), and estimates for other populations are from (*32*). The number in parentheses is the number of individuals tested. For some studies, S−s−U− is inferred from a subset of the anti-S, anti-s, and anti-U assays. Data for anti-S and anti-s tests across many additional non-African populations support 0% frequency of S−s− phenotypes (*79*).

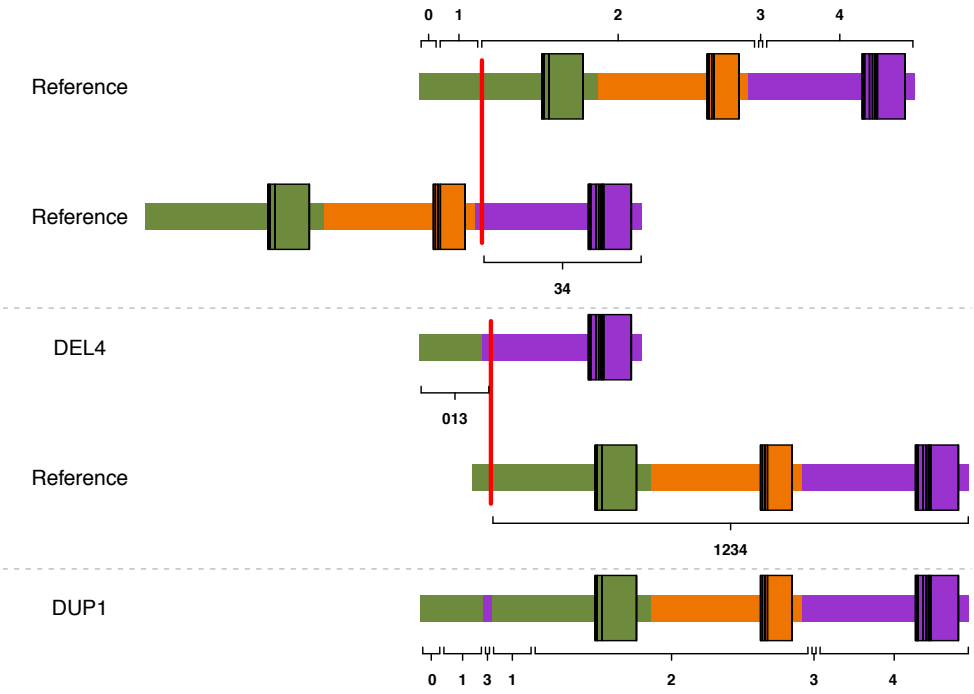


Figure S24. Model of unequal crossover events where DUP1 results from NAHR on a DEL4 background. As in **Fig. S20**, unequal crossover events are marked by red vertical lines across the two chromosomes involved. Here, the numbered segments of the reference are delineated by the breakpoints of both DEL4 and DUP1, including an additional duplicated segment in DUP1 carriers (segment 3). The first event creates DEL4, and the second event creates DUP1 by NAHR between DEL4 and a reference chromosome, leading to segment order 0131234 in DUP1; 0123134 is also plausible with a different misalignment in the second event. The second event, but not the first, occurs between homologous positions in the segmental duplication.

Supplementary Tables

Population	Population (ab.)	Country	Number	Mean coverage	Number of trios +(duos)
Fula	FUL	The Gambia	100	8.7	31 +(1)
Jola	JOL	The Gambia	100	9.5	32 +(1)
Mandinka	MAN	The Gambia	100	10.0	33
Wollof	WOF	The Gambia	98	9.9	32 +(1)
Bantu	BAN	Cameroon	31	10.1	5 +(3)
Semi Bantu	SBA	Cameroon	32	10.1	8
Mossi	MOS	Burkina Faso	57	10.2	0
Chagga	CHA	Tanzania	80	10.5	21 +(2)
Pare	PAR	Tanzania	77	10.0	22 +(2)
Wasambaa	WAS	Tanzania	90	10.9	23 +(6)
Total			765	9.9	207 + (16)

Table S1. Individuals sequenced from MalariaGEN. The number of trios confirmed by genetic data is given in the last column as well as the number of duos (where only one parent-child relationship was confirmed).

Country	Institution	Ethics Approving Committee	Ethics Committee	Local IDs
The Gambia	Medical Research Council Unit, The Gambia	MRC Gambia and Gambia Government	MRC/Gambia Government Ethics Committee	SCC1156
Cameroon	University of Buea	Institutional Research Board, University of Buea Government of Cameroon	Institutional Research Board Provincial Delegate for Public Health	University of Buea ethical clearance 07-12-2005 D7.1.A/MPH/SWP/PDPH/PS.CH/2340/811
Burkina Faso	Centre National de Recherche et de Formation sur le Paludisme	Ministry of Health & Ministry of Science and Education	Health Research Ethics Committee	No. 2007-048
Tanzania	Joint Malaria Programme, Kilimanjaro Christian Medical Centre	London School of Hygiene and Tropical Medicine National Institute for Medical Research (NIMR), Tanzania	London School of Hygiene and Tropical Medicine Ethics Research Committee NIMR Research Coordinating Committee	4093 NIMR/HQ/R.8a/Vol.IX/611

Table S2. Study sites and ethics approving institutions for sequenced samples included in this study.

Super-population	Population (ab.)	Population name	Number	Mean coverage
African	ACB	African Caribbean in Barbados	96	7.1
African	ASW	African-American in Southwest US	61	6.0
African	ESN	Esan in Nigeria	99	6.6
African	GWD	Gambian in Western Division, The Gambia	113	7.7
African	LWK	Luhya in Webuye, Kenya	99	5.9
African	MSL	Mende in Sierra Leone	85	6.7
African	YRI	Yoruba in Ibadan, Nigeria	108	6.0
American	CLM	Colombian in Medellin, Colombia	94	6.2
American	MXL	Mexican Ancestry in Los Angeles, California	64	5.7
American	PEL	Peruvian in Lima, Peru	85	6.9
American	PUR	Puerto Rican in Puerto Rico	104	6.2
East Asian	CDX	Chinese Dai in Xishuangbanna, China	93	5.5
East Asian	CHB	Han Chinese in Beijing, China	103	6.3
East Asian	CHS	Han Chinese South	105	5.6
East Asian	JPT	Japanese in Tokyo, Japan	104	6.5
East Asian	KHV	Kinh in Ho Chi Minh City, Vietnam	99	7.2
European	CEU	Utah residents (CEPH) with Northern and Western European ancestry	99	7.2
European	FIN	Finnish in Finland	99	5.0
European	GBR	British in England and Scotland	91	6.4
European	IBS	Iberian populations in Spain	107	5.9
European	TSI	Toscani in Italia	107	6.2
South Asian	BEB	Bengali in Bangladesh	86	6.2
South Asian	GIH	Gujarati Indian in Houston, TX	103	5.8
South Asian	ITU	Indian Telugu in the UK	102	6.5
South Asian	PJL	Punjabi in Lahore, Pakistan	96	6.7
South Asian	STU	Sri Lankan Tamil in the UK	102	6.8
Total			2504	6.4

Table S3. Individuals sequenced in the 1000 Genomes Project Phase 3.

Variant	# trios	# het parents	# transmitted	# hom parents	# transmitted	*P*-value for het transmissions
DEL1	36	35	17	5	5	1
DUP1	25	26	5	1	1	0.0025*
DEL2	7	7	2	0	NA	0.45
DUP3	2	2	2	0	NA	0.5
DUP4	3	3	2	0	NA	1
DEL3	2	2	1	0	NA	1
DEL4	2	2	1	0	NA	1
DEL5	1	1	1	0	NA	1
DUP5	1	1	1	0	NA	1
DUP13	1	0**	1	0	NA	1
DUP16	1	1	0	0	NA	1

*This may be due to missing DUP1 calls, see (*22*).

**This Mendelian error is likely due to a missed call of DUP13 in one parent.

Table S4. Observations and transmissions of copy number variants in MalariaGEN trios. For each variant observed in the trios, the number of trios in which it is segregating is shown, followed by the number of heterozygous (het) and homogyzous (hom) parents and the number of each who transmitted the variant to the child in the trio, respectively. The *P*-value for heterozygous transmissions is calculated by a binomial test, with *P*=0.5 and the number transmitted as the number of successes.

Primer name	Primer Sequence (5'-3')	Dir	Matching Location(s) (GRCh37)
First Round PCR			
GYP_DEL1_F6	TTTCGCTAGTAGTATTTGTCCGTGTC	F	**144832560-144832585 (*GYPE*)**
GYP_DEL1_R4A	GAGGGAGCAGATAGTTGGTTTATGA	R	**144837351-144837375 (*GYPE*)** *144947586-144947610 (GYPB)* *145068936-145068960 (GYPA)*
Nested PCR			
GYP_DEL1_F1	TATTTGTCCGTGTCCCAAGA	F	**144832572-144832591 (*GYPE*)**
GYP_DEL1_R4C	CAGATAGTTGGTTTATGAATTCTTATCC	R	**144837341-144837368 (*GYPE*)** *144947576-144947603 (GYPB)* *145068926-145068953 (GYPA)*
Sequencing primers			
GYP_DEL1_R1	CGATGGACTTAGAGGCAACTG	R	**144834442-144834462 (*GYPE*)** *144944679-144944699 (GYPB)* *145066024-145066044 (GYPA)*
GYP_DEL1_R2	GGATGTGTGTTCAGGAGCTG	R	**144835462-144835481 (*GYPE*)** **144945700-144945719 (*GYPB*)** **145067046-145067065 (*GYPA*)**
GYP_DEL1_R3	TTTTCCTGAAGTTTTGATTGTTTG	R	**144836417-144836440 (*GYPE*)** **144946656-144946679 (*GYPB*)** *145068006-145068029 (GYPA)*

Table S5. PCR and sequencing primers used to amplify and sequence across the DEL1 breakpoint. The positions in bold show an exact match to the primer sequence while those in italics represent homologous binding sites with few mismatches. All primers were sourced from IDT (Leuvan, Belgium).

Type of variant	Observed variants	Predicted MNS blood group phenotype
Absence of GYPB	DEL1, DEL2, DEL4, DEL6, DEL8, *DEL10*	S−s−U− (when homozygous)
GYPB-A hybrid	DUP2 DUP4 *DUP27*	GP.Sch GP.Dantu * (GP.Sch or GP.Dantu)
GYPE-A hybrid	DUP8, *DUP23, DUP24*	None
GYPA-B hybrid	*DEL13*	* (GP.Hil[+], GP.JL[++] or GP.Sat)
Whole gene duplications (only)	DUP3, DUP7, *DUP14, DUP17, DUP19, DUP25, DUP26*	None
Absence of GYPA	None	En(a-) (when homozygous)
Absence of both GYPA and GYPB	None	M^k (when homozygous)

*Depends on precise breakpoint (not determined)
[+]Previously known as Miltenberger V
[++]Previously known as Miltenberger XI

Table S6. Consequences of structural variants on the glycophorin genes and their predicted effect on MNS blood group phenotype. Singleton variants are shown in italics.

Population	Cerebral malaria OR (95% CI)	Severe malarial anaemia OR (95% CI)	Other severe malaria OR (95%CI)
Malawi	**0.66 (0.44-1.00)**	**0.72 (0.23-2.24)**	**0.88 (0.56-1.38)**
Kenya	**0.58 (0.43-0.77)**	**0.68 (0.42-1.10)**	**0.61 (0.45-0.82)**
Meta-analysis	**0.60 (0.47-0.76)**	**0.69 (0.45-1.08)**	**0.68 (0.53-0.87)**

Table S7. Effect of DUP4 on subphenotypes. Odds ratios and 95% confidence intervals for estimates of the effect of DUP4 on severe malaria subphenotypes, computed by multinomial logistic regression with outcome levels control, cerebral malaria (CM) case, severe malarial anaemia (SMA) case, or other severe malaria case. Individuals recorded as having both CM and SMA were excluded.

Genotype	The Gambia	Malawi	Kenya
WT/WT	4093	2091	2443
WT/DEL1	536	157	120
WT/DEL2	71	41	36
WT/DUP4	1	141	395
DEL1/DEL1	30	2	2
DEL1/DEL2	3	0	1
DEL1/DUP4	0	3	3
DEL2/DEL2	0	1	0
DEL2/DUP4	1	4	4
DUP4/DUP4	0	4	25
Other	111	37	61
Not called (<75% posterior probability)	74	35	48

Table S8. Genotype calls inferred from intensities in the Gambia, Kenya and Malawi.

The Gambia	Imputed genotypes				
Intensity-based genotypes		0	1	2	(no call)
	0	4912			3
	1	2			
	2				
	(no call)	3			

Malawi	Imputed genotypes				
Intensity-based genotypes		0	1	2	(no call)
	0	2327	6		25
	1	1	131		18
	2		1	3	
	(no call)	1	2		1

Kenya	Imputed genotypes				
Intensity-based genotypes		0	1	2	(no call)
	0	2692	6		27
	1	2	380		1
	2		3	21	
	(no call)	1	3		

Table S9. Comparison of imputed genotypes at DUP4 with intensity-based calls in the Gambia, Kenya and Malawi. Cells show the number of individuals in each population with the given DUP4 genotype, as called based on intensity data (rows) and by imputation (columns). Calls are made based on having at least 75% posterior probability on the given genotype; 'no call' reflects individuals with less than 75% posterior probability of any call. Blank cells represent zeroes.

Primer ID	Primer Sequence 5'-3'	Matching location(s) (GRCh37)
	PCR primers	
GYPA_Exon6_Fwd	CTTCGATAAGCTGTGTTGTATGGATGT	**145036894-145036920 (*GYPA*)**
GYPB_753_Rev	GGGATGTGGGAGAATTTGTCTTTCATGATACGCTG	*145040991-145041025 (GYPA)* **144921685-144921719 (*GYPB*)** *144800912-144800945 (GYPE)*
	Sequencing primers	
GYP_ABE_3000_REV	TCCAGATGAAGTAAATAGAATGGAA	*145038796-145038820 (GYPA)* **144919488-144919512 (*GYPB*)** *144798708-144798732 (GYPE)*
GYP_ABE_2450_FWD	TTTGGAAATAGAAATACATAATTTGG	**145039320-145039345 (*GYPA*)** **144920012-144920037 (*GYPB*)** **144799239-144799264 (*GYPE*)**

Table S10. PCR and sequencing primers used to amplify exons 3 to 5 of *GYPA* and *GYPB-A* hybrids. The positions in bold show an exact match to the primer sequence while those in italics represent homologous binding sites with few mismatches. All primers were sourced from IDT (Leuvan, Belgium).

Population sampled	# Dantu positive	# tested	Frequency (%)	Reference
North London Blood Transfusion Centre	1*	44,112	0	(*46*)
African American donors in Chicago, USA	5	1,000	0.5	(*48*)
Predominantly African American donors in Dayton, USA	5	2,200	0.23	(*48*)
German donors in Cologne, Germany	0	1,000	0	(*48*)
Admixed in South African Cape region	‡	‡	1.1	(*47*)

*An admixed individual from Mauritius
‡Not provided

Table S11. Estimates of the frequency of the Dantu blood group phenotype in the literature. To our knowledge, these are the only published population frequency estimates.