






Multiple *GYPB* gene deletions associated with the U– phenotype in those of African ancestry

William J. Lane ^{1,2} Nicholas S. Gleadall,^{3,4} Judith Aeschlimann,⁵ Sunitha Vege,⁵ Alba Sanchis-Juan,^{3,4,6} Jonathan Stephens,^{3,4,6} Jensyn Cone Sullivan ¹ Helen H. Mah,¹ Maria Aguad,¹ Robin Smeland-Wagman,¹ Matthew S. Lebo,^{1,2,7,8} Prathik K. Vijay Kumar,⁸ Richard M. Kaufman ^{1,2} Robert C. Green ^{2,8,9,10} Willem H. Ouwehand,^{3,4,11} and Connie M. Westhoff ⁵

BACKGROUND: The MNS blood group system is defined by three homologous genes: *GYPB*, *GYPE*, and *GYPB*. *GYPB* encodes for glycophorin B (GPB) carrying S/s and the “universal” antigen U. RBCs of approximately 1% of individuals of African ancestry are U– due to absence of GPB. The U– phenotype has long been attributed to a deletion encompassing *GYPB* exons 2 to 5 and *GYPE* exon 1 (*GYPB*01N*).

STUDY DESIGN AND METHODS: Samples from two U– individuals underwent Illumina short read whole genome sequencing (WGS) and Nanopore long read WGS. In addition, two existing WGS datasets, MedSeq (n = 110) and 1000 Genomes (1000G, n = 2535), were analyzed for *GYPB* deletions. Deletions were confirmed by Sanger sequencing. Twenty known U– donor samples were tested by a PCR assay to determine the specific deletion alleles present in African Americans.

RESULTS: Two large *GYPB* deletions in U– samples of African ancestry were identified: a 110 kb deletion extending left of *GYPB* (DEL_B_LEFT) and a 103 kb deletion extending right (DEL_B_RIGHT). DEL_B_LEFT and DEL_B_RIGHT were the most common *GYPB* deletions in the 1000 Genomes Project 669 African genomes (allele frequencies 0.04 and 0.02). Seven additional deletions involving *GYPB* were seen in African, Admixed American, and South Asian samples. No samples analyzed had *GYPB*01N*.

CONCLUSIONS: The U– phenotype in those of African ancestry is primarily associated with two different complete deletions of *GYPB* (with intact *GYPE*). Seven additional less common *GYPB* deletion backgrounds were found. *GYPB*01N*, long assumed to be the allele commonly encoding U– phenotypes, appears to be rare.

The MNS blood group system is one of the most complex, consisting of three homologous genes, *GYPB*, *GYPE*, and *GYPB*, which encode 49 known antigens.¹ The three genes arose from the duplication of an ancestral gene via homologous recombination at Alu repeat sequences.² In the human reference genome all three genes are in a reverse direction relative to their expression and tandem to each other (ordered *GYPE*, *GYPB*, *GYPB*).³ *GYPB* and *GYPB* encode the red blood cell (RBC) membrane proteins, glycophorin A (GPA) and glycophorin B (GPB), with M/N and S/s antigens being those commonly considered when performing antibody identification and compatibility testing.¹ *GYPE* does not encode any known antigens. Recombination between *GYPB* and *GYPB* form hybrid genes responsible for low incidence antigens (Hil, MINY, TSEN, Dantu, SAT, etc.). Deletion events result in null phenotypes designated En(a–) for *GYPB*, U– for *GYPB*, and M^kM^k for absence of both GPA and GPB.

RBCs of approximately 1% of individuals of African ancestry are U– in which *GYPB* encoded antigens are absent including

From the ¹Department of Pathology and ¹⁰Division of Genetics, Department of Medicine, Brigham and Women’s Hospital; the ²Harvard Medical School; the ⁷Laboratory for Molecular Medicine; the ⁸Partners Personalized Medicine; the ⁹Broad Institute of MIT and Harvard, Boston, Massachusetts; the ³Department of Haematology, University of Cambridge; the ⁴NHS Blood and Transplant; the ⁶NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust; the ¹¹Wellcome Sanger Institute, Cambridge, UK; and the ⁵New York Blood Center, New York, New York.

Address reprint requests to: William J. Lane, MD, PhD, Pathology Department, Brigham and Women’s Hospital, Hale Building for Transformative Medicine, Rm 8002L, 60 Fenwood Rd, Boston, MA 02115; e-mail: wlane@bwh.harvard.edu

Received for publication January 10, 2020; revision received February 25, 2020, and accepted April 2, 2020.

doi:10.1111/trf.15839

© 2020 AABB

TRANSFUSION 2020;99:999;1–14

S/s and the high frequency U antigen.⁴ The U⁻ phenotype appears to protect from malaria infection, since red blood cells deficient for GPB are partially resistant to *Plasmodium falciparum* invasion.⁵ U⁻ individuals are at risk of developing anti-U following exposure to U⁺ RBCs with the potential for severe hemolytic transfusion reactions and hemolytic disease of the fetus and newborn.⁵ In 1990, Southern blotting of a sample from a S-s-U⁻ individual revealed a large deletion encompassing *GYPB* exons 2 to 5 and *GYPE* exon 1 (designated *GYPB*01N*).⁶ *GYPB*01N* was assumed to be the most common genetic basis of the U⁻ phenotype. The M^kM^k null phenotype (M-N-S-s-U⁻) involves a large deletion of *GYPB* exons 2-7 and *GYPB* exons 1-5 (designated *GYP*01N* to convey loss of expression of both GPA and GPB).¹ *GYP*He(GL)* and *GYP*SAT* represent hybrid genes containing regions from both *GYPB* and *GYPE*, which lead to a U⁻ phenotype along with the expression of the new antigens He and SAT.¹

Next generation sequencing (NGS) has revolutionized DNA sequencing.⁷ In addition to the detection of single nucleotide variants, NGS data—especially that derived from whole genome sequencing (WGS)—can be analyzed for structural variations.⁷ We and others have used NGS for blood group genotyping and for detection of structural variations with a primary focus on the Rh blood group system.^{3,8-13} Here we revisit the genetics of the U⁻ phenotype using both Illumina short read and Nanopore long read WGS data of known U⁻ samples, analyze two public WGS datasets, and develop a sequence specific PCR assay to test African American U⁻ blood donors for the *GYPB* deletions found by our WGS analysis.

METHODS

Study samples

We analyzed *GYPB* deletions in samples from two existing public WGS datasets and performed WGS on two serologically known S-s-U⁻ samples (UNEG-001, UNEG-002). An additional 20 serologically typed S-s-U⁻ African-American donor samples underwent *GYPB* deletion sequence-specific PCR using primers designed from the WGS data analysis.

The existing WGS datasets consisted of GRCh37/hg19 aligned BAM files: high coverage 30x genomes from the MedSeq Project (n = 110)^{3,14} and low coverage 7x genomes from 1000 Genomes Project (n = 2535).¹⁵ The MedSeq Project genomes are available through dbGaP under study accession phs000958. The 1000 Genomes Project phase 3 genomes were downloaded from <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3>. The data sets and samples are summarized in Table 1. Overall, in total, the data reflect analysis of genomes of 704 individuals of African ancestry.

Amplification and Sanger sequencing of potential breakpoints was performed on archived DNA from MedSeq Project sample MEDSEQ-110 with approval from the Partners HealthCare Human Research Committee (IRB). DNA from 1000 Genomes Project samples (n = 12) was obtained from Coriell Cell Repositories.

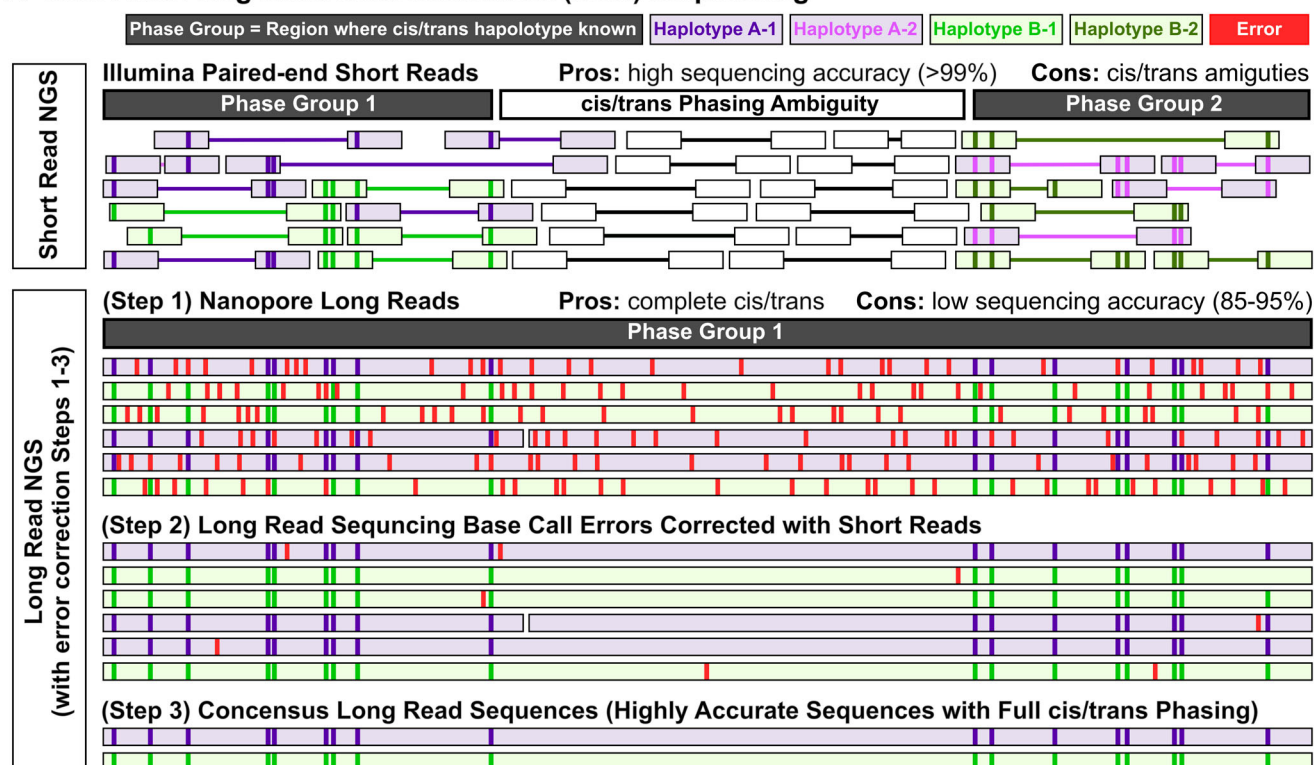
Copy number analysis of the glycoporphin (GYP) locus

Sequencing coverage was extracted from WGS alignment files using BEDTools v2.17.0¹⁶ and the average depth of coverage was calculated. The 110 MedSeq Project genomes were initially analyzed using read depth copy number for *GYPB*, *GYPE*, and *GYPE* introns and exons as previously detailed.³ For this study, samples with evidence of *GYPB* deletions were studied in greater detail to determine the full extent of the deletions. Briefly, the average depth of coverage over 100 and 2500 base pair (bp) bins was calculated over a large region (referred to as the *GYP* locus) including *GYPB*, *GYPE*, and the surrounding intergenic regions (chr4:144,600,000-145,225,000, GRCh37/hg19). The intragenic region from chr4:144,600,000-144,700,000 was used to determine average background depth of coverage. Copy number was calculated using: $\text{CopyNumber} = \frac{\text{AverageCoverage}_{\text{bin}}}{\text{AverageCoverage}_{\text{background}}} \times 2$ with the allele zygosity assigned using the following ranges: 1x (copy number < 0.5), 2x (copy number >= 0.5 and < 1.5), and 3x (copy number >= 1.5). The Integrative Genomics Viewer (IGV)¹⁷ was used to verify sequence identity, depth of coverage (i.e., the number of

TABLE 1. Datasets and samples

Source	Data type	Purpose	Ethnic breakdown
MedSeq project	Short read WGS (30x high coverage)	Study of genomic sequencing in randomized patients	110 genomes: 89 European ancestry, 13 African ancestry, 4 Asian, and 4 Hispanic
1000 genomes project	Short read WGS (7x low coverage)	Genomic sequencing large populations	2,535 genomes: 669 African (AFR), 515 East Asian (EAS), 505 European (EUR), 494 South Asian (SAS), 352 Admixed American (AMR)
This study	Short read WGS (30x high coverage) + long read WGS for one	WGS of two known U ⁻	2 African American ancestry
This study	Sanger sequencing	Confirmation of <i>GYPB</i> deletions in 20 known U ⁻	20 African American ancestry

A Short and Long Read Next Generation (NGS) Sequencing



B Detecting Deletions in Short and Long Read NGS Data

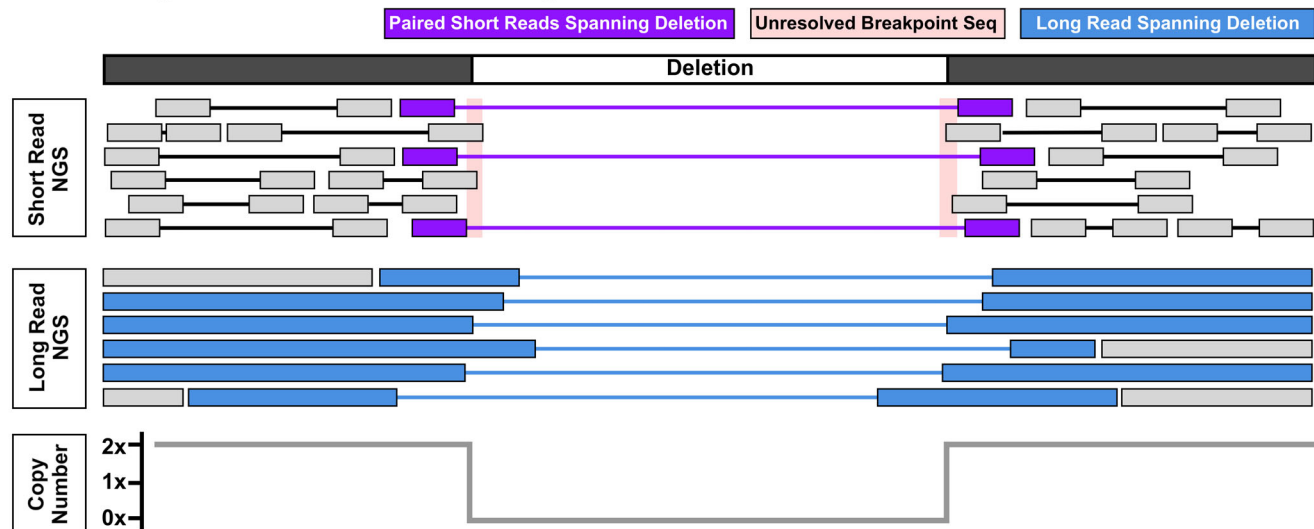
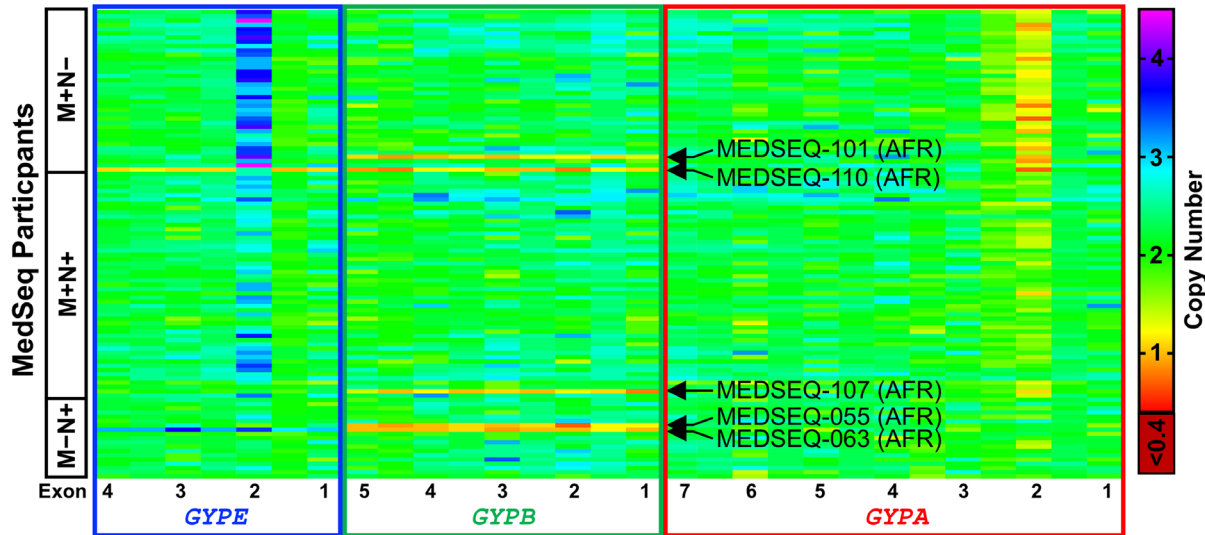


Fig. 1. Illustration of Paired-end Short Read and Long Read NGS. Examples of short and long read NGS over a heterozygous region. Each fragment read is represented as a colored box and the heterozygous nucleotides indicated by colored lines. (A) The phase groups (dark gray boxes) represent regions over which the genetic changes could be determined to be on the same or separate chromosomes (i.e., cis/trans haplotype relationship). Between phase group 1 and 2 the white boxes represent regions of high sequence similarity and changes in phase group 1 (A1, B1) cannot be linked to phase group 2 (A2, B2) in short read NGS. 1) Long read NGS have a higher sequencing base calling error rate (red) but can be corrected using short read NGS; 2) and self-corrected using multiple overlapping reads; 3) to determine a final cis/trans phased consensus sequence. (B) Short and long read sequence depiction for detection of large structural deletions. For short read NGS, the purple boxes illustrate paired short reads with one part of the fragment sequence located left of the deletion and second part to the right. For long read NGS, the blue boxes illustrate fragment reads spanning the deletion. The sequence copy number (0x, 1x, 2x) is shown. [Color figure can be viewed at wileyonlinelibrary.com]

A MedSeq Project: *GYPA*, *GYPB*, *GYPE* Copy Number Changes



B MedSeq Project: *GYPB* Deletions

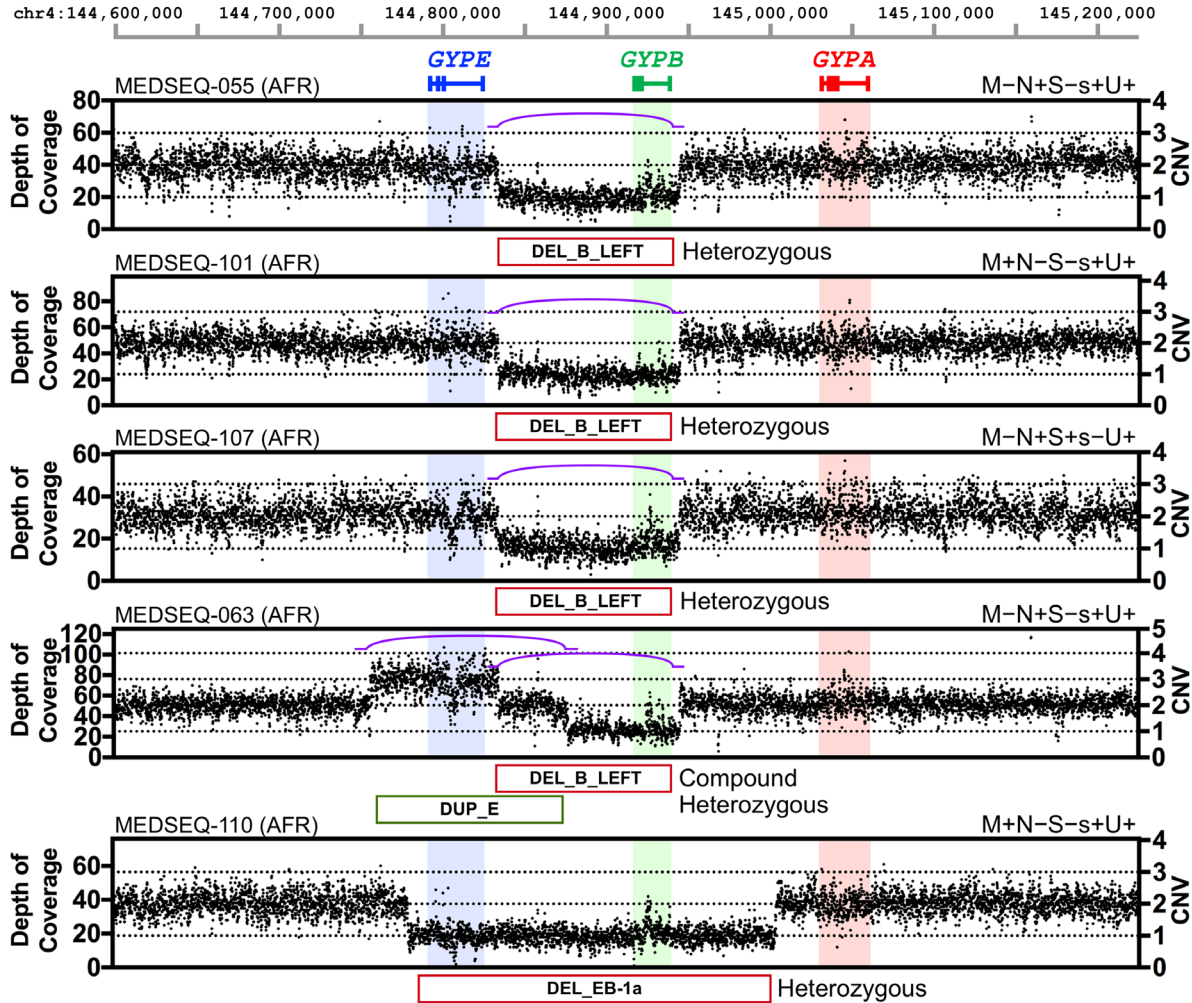


Fig. 2. Legend on next page.

times a specific site was sequenced), and paired reads (i.e., two reads per DNA fragment with the first in forward orientation and the second in reverse-complement orientation). The copy number data was analyzed for patterns consistent with the previously reported *GYP* locus deletions as defined by the International Society of Blood Transfusion (ISBT) blood group allele table for the MNS blood group system (v4.1 170119).¹

Illumina short read and nanopore long read WGS of U– samples

See Supplemental Information for details about the WGS workflow.

Deletion breakpoint identification

Figure 1A and B illustrate the process of determining deletion breakpoints from short and long read WGS. Known repeat regions located over the *GYP* locus were downloaded from UCSC Table Browser¹⁸ and used to annotate the reference sequence in IGV. To locate the potential start and end of the *GYPB* deletions, the WGS data was visualized in IGV to identify where the read depth decreased and if reads spanning the deletion were present (Fig. 1B). For short read WGS, although there are reads near the deletion, it can be difficult to determine the breakpoint with high confidence especially in large regions of high homology. In long read WGS the reads span the deleted region, which makes it possible to define the exact breakpoint sequence. However, short read Illumina sequencing has a higher base calling accuracy than long read Nanopore sequencing.¹⁹ FMLRC v1.0.0²⁰ was run to use the short reads to error correct the long reads over the *GYP* locus (chr4:144,600,350-145,225,690, GRCh37/hg19). The *GYPB* deletions were identified by the long read structural variant caller Sniffles v1.0.1²¹ and viewed in IGV¹⁷ (with link supplementary alignments turned on) and the long sequence reads that spanned the deletions were identified and improved by determining a consensus sequence. The consensus sequence was determined at positions covered by two or more reads with concordant bases (other positions called ambiguous). Reference genome sequences for the region implicated in the

breakpoint were downloaded using TogoWS²² and aligned to the error corrected long read consensus sequence using Clustal Omega v1.2.3.²³ The exact deletion breakpoints were identified by finding the positions at which the WGS sequence diverged from the reference sequence located before and after the deletion.

See Supplemental Information for details about Sequence-specific PCR and Sanger Sequencing of Deletion Breakpoints.

RESULTS

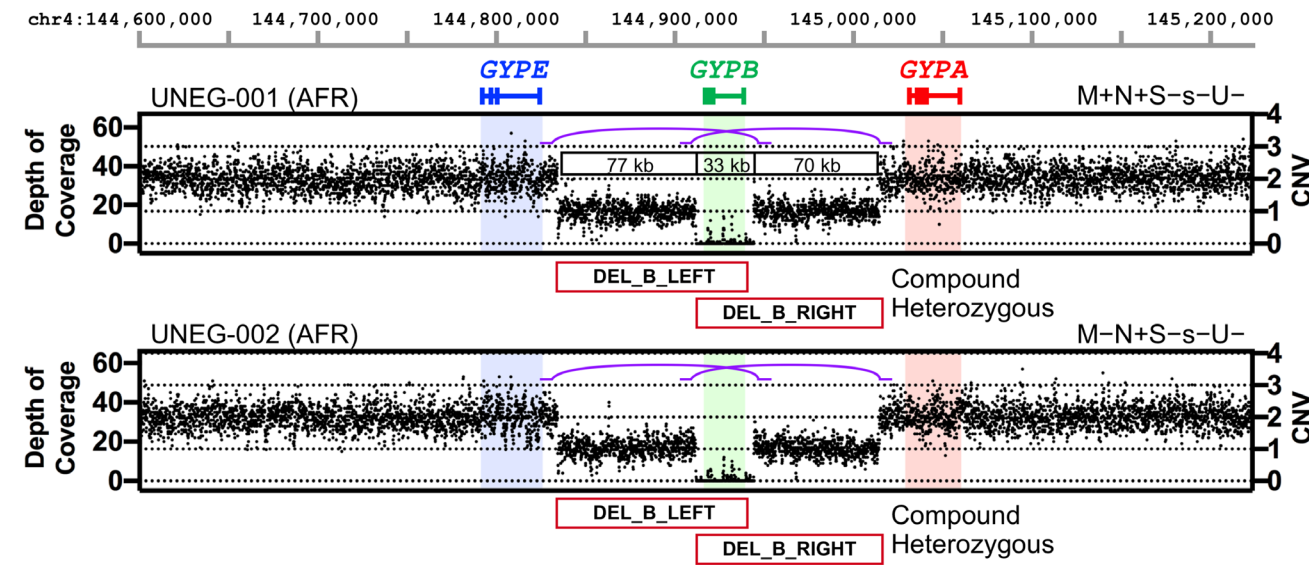
Novel *GYP* deletions in the MedSeq project genomes

We analyzed the 110 MedSeq Project genomes^{3,14} for structural variations in the *GYP* locus with a specific focus on *GYPB* deletions. The sequence read depth of coverage (i.e., the number of times a specific sequence position was seen) was used to calculate the copy number over each exon and intron for *GYPB*, *GYPB*, and *GYPE* (Fig. 2A). *GYPB* deletion results in a copy number decrease over the deleted region, as does misalignment to a homologous gene.¹³ As we previously reported, *GYPB**M exon 2 sequences (Fig. 2A, yellow) partially misalign to *GYPE* exon 2 revealed as an increase in *GYPE* exon 2 (Fig. 2A, blue). *GYPB* exon 2 was therefore excluded when screening samples for *GYP* locus deletions.

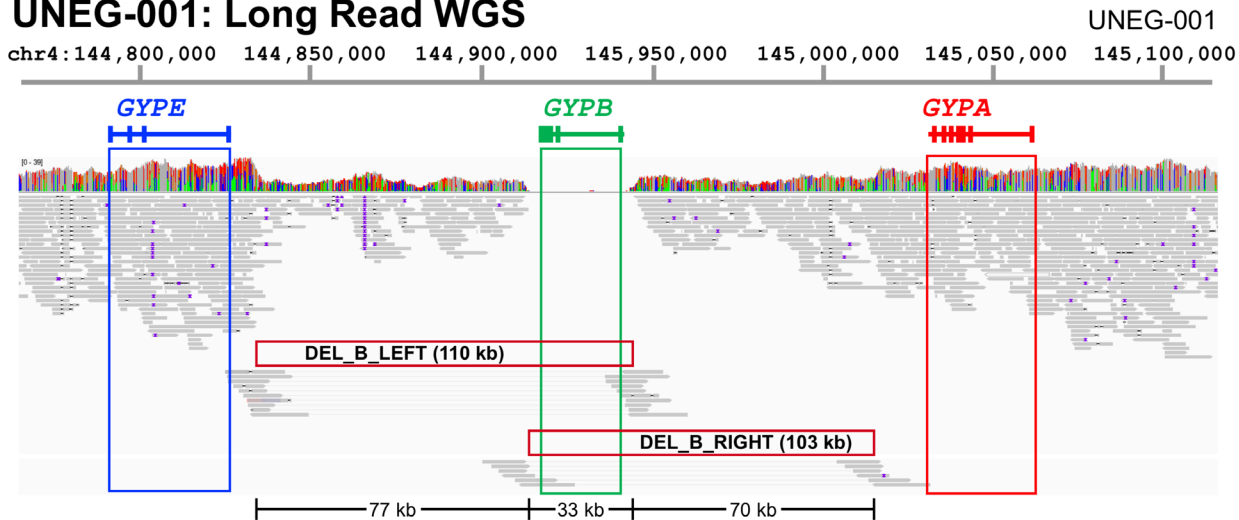
Five of thirteen MedSeq African ancestry genomes showed *GYPB* deletions (Fig. 2A and B). Three (MEDSEQ-055, -101, and -107) showed only one copy (1x) over the entire *GYPB* gene, one (MEDSEQ-063) was 1x over *GYPB* and increased 3x over *GYPE*, and one (MEDSEQ-110) was 1x over both *GYPB* and *GYPE* (Fig. 2A). Figure 2B shows details of the gene alignments. All three samples with *GYPB* 1x copy (MEDSEQ-055, -101, and -107) were heterozygous for a large 110 kb deletion spanning the entirety of *GYPB* and extending upstream (left) (designated DEL_B_LEFT). MEDSEQ-063 was heterozygous for the same left extending *GYPB* deletion (DEL_B_LEFT) with and a *GYPE* duplication, all supported by paired reads spanning the deleted and

Fig. 2. MNS *GYP* locus Copy Number Changes in MedSeq Project WGS Data. (A) Heat map shows the copy number changes for *GYPE*, *GYPB*, and *GYPB* exons and introns for all 110 MedSeq Project genomes (ordered by M and N antigen status for clarity). Each horizontal row represents results from one sample and the columns demark the exons and introns. Exon numbers are indicated on the bottom axis in their genomic ordering (i.e., genes are *GYPE*, *GYPB*, and *GYPB* all in the reverse direction). Five genomes show decreases in coverage of *GYPB* (yellow/orange). Note also the decrease in exon 2 of *GYPB**M due to misalignment to exon 2 of *GYPE* (blue), which we have previously reported.¹³ (B) Depth of coverage plot of short read WGS (left axis) and copy number variation (CNV, right axis) of *GYPE*, *GYPB*, and *GYPB* regions (100 bp bins) for the five African (AFR) MedSeq Project genomes in A above. Each gene region is highlighted in color (*GYPE*: blue, *GYPB*: green, *GYPB*: red) with the genes in a reverse direction (i.e., exon 1 is on the right). The location of relevant paired reads are shown with purple curved lines. The evident structural variations are shown boxed at the bottom of each plot. Three samples (MEDSEQ-055, -101, and -107), are heterozygous for a left extending *GYPB* deletion (DEL_B_LEFT), one sample is compound heterozygous for DEL_B_LEFT deletion and a *GYPE* (DUP_E) duplication, and one sample is heterozygous for *GYPE* and *GYPB* (DEL_EB-1a) deletion. [Color figure can be viewed at wileyonlinelibrary.com]

A UNEG-001 and UNEG-002: Short Read WGS



B UNEG-001: Long Read WGS



C UNEG-001: Deletion Breakpoints

DEL_B_LEFT Breakpoint Sequence

Before Breakpoint After

5' - GAAAAGGAACAAATATGTAGTCATTTCTAAGATTTATTTTATTAATAATGTAACTACAAACTGTAAGTGTGTAACCTT
ATATATATATACACACATATACACACATACATGAGTATATGTGTGAACTGTCACTCAAATCAAGACATAGAACATTT
AT (n) / AC (n)
CAGATGTATTCCAGAGATCATGGTGGATGGGAGGCAGGACTGGATTGCAGCTCCCACTTGAACAGACAAAGCAGCGTGTG
L1PBa
GAGGCTTGATCATGAACCTTGACTGCAGGAATAAATCAGGAAAGCTGAGAGAACCCACAGACCCTCTGAAGGAAGTGA-3'

DEL_B_RIGHT Breakpoint Sequence

Before Breakpoint After

5' - TTCTCATGCTGCTATAAAAACTGCGCAAGACTGTGTAATTTATAAAGGAAAGAGGTTTGAAGTGTATCTACAGTTTGCAT
THE1C
GGCTTGAAGGTCTTGAATACTTACAATCATGACCAAGGGGAAACAAACACATCTTTCTTACATAGTGGCAGGAAGGA
THE1C
GAAGAATGAGAGCTGAGTGAAGGGGAAGCTCCTTTATAAACTATCAGATTATGTGAGAATTTATCTACTCTCATGAGA
THE1C
ATAGCATAGGGGAAACCACCGCAATGATTCAAGTACCTCCCATGGGTTCTCCCATGACAAGTGGGGATTATTGGAAC-3'

THE1C

Fig. 3. Legend on next page.

duplicated regions. In MEDSEQ-063, the duplicated and deleted regions overlap, reflected as a copy number 2 region (Fig. 2B). MEDSEQ-110 was heterozygous for a large 224 kb deletion (no paired reads) designated DEL_EB-1 to denote the deletion spans both *GYPB* and *GYPE*. None of the Med-Seq Project genomes carried a structural variation consistent with the originally described⁶ genetic basis of the U– phenotype with deletion of *GYPB* exons 2 to 5 and deletion of *GYPE* exon 1 (*GYPB*01N*).¹

GYP deletions in U– African Americans

To explore this unexpected finding, short read WGS (Illumina) was performed on samples from two African American individuals with a S–s–U– serologic RBC phenotype (UNEG-001 and UNEG-002). Copy number calculation showed that the genomes in both were 0x over a 33 kb region covering all *GYPB* with 1x copy number regions extending 77 kb to the left and 70 kb to the right of *GYPB*. *GYPA* and *GYPE* were 2x (Fig. 3A). Although the decreases in alignment could be the result of misalignment, no appreciable increase in other regions of the *GYP* locus was evident in the data. The deletion pattern could also be explained by compound heterozygosity of two different overlapping deletions. One deletion appeared to be the previously detected left extending intragenic *GYPB* deletion (above, Fig. 2B) (DEL_B_LEFT) and the other a *GYPB* deletion with a right extending intragenic region deletion (designated DEL_B_RIGHT). Paired reads were found spanning the proposed breakpoints, in agreement with a compound heterozygous deletion of *GYPB* as the basis of the U– phenotype in both samples.

To unambiguously define the compound heterozygous deletions, long read WGS (Nanopore) was performed on UNEG-001. The long read WGS showed clonal reads spanning two different deletion breakpoint regions (Fig. 3B), confirming the presence of two independent *GYPB* deletions; one 110 kb deletion extending left (DEL_B_LEFT) and a second 103 kb deletion extending right (DEL_B_RIGHT). Automated deletion detection using Sniffles confirmed the presence of these same two deletions: DEL_B_LEFT deleted from chr4:144,835,510-144,945,751 (STD_quant_start = 560.224509; STD_quant_stop =

563.441656) and DEL_B_RIGHT deleted from chr4:144,913,336-145,016,591 (STD_quant_start = 320.061452; STD_quant_stop = 315.031480).

Although Nanopore WGS produces long reads, the base calling accuracy is only 85-95%.¹⁹ To identify the exact breakpoint sequences, the base calls for the long reads were error corrected using the short read WGS data which has a >99% base calling accuracy (described in Fig. 1A).¹⁹ The error corrected long reads were combined to generate a consensus sequence which was aligned to the human reference genome sequences. The occasional nucleotide differences before and after the breakpoint were used to identify the exact breakpoint region in the long read (Fig. 3C). Each of the deletions occurred in regions of highly similar DNA sequence with the DEL_B_LEFT breakpoint occurring in a 120 bp region of identity containing AT(n)/AC(n) and LIPBa repeats (*GYPE*-*GYPB* intergenic region chr4:144,835,160-144,835,279 and *GYPB*-*GYPA* intergenic region chr4:144,945,398-144,945,517) and DEL_B_RIGHT breakpoint occurring in a 130 bp region of identity containing a THE1C repeat (*GYPE*-*GYPB* intergenic region chr4:144,912,872-144,913,001 and *GYPB*-*GYPA* intergenic region chr4:145,016,127-145,016,256) (Fig. 3C, italics). Sequence-specific PCR primers were designed spanning DEL_B_LEFT and DEL_B_RIGHT and used to Sanger sequence the breakpoint regions in both U– individuals (UNEG-001 and UNEG-002), confirming the breakpoint sequences identified by WGS. (Figs. S1 and S2, available as supporting information in the online version of this paper, show the full breakpoint region nucleotide sequence alignment).

GYP deletions in 1000 genomes project dataset

Analysis of 2535 low coverage whole genomes from the 1000 Genomes Project revealed seven different large deletions involving *GYPB* in 89 individuals (Fig. 4): 81 African (AFR), 5 Admixed American (AMR), 3 South Asian (SAS). Table S1, available as supporting information in the online version of this paper, shows the allele frequencies for the specific *GYP* deletions based on analysis of the 1000 Genomes Project data.

Fig. 3. MNS *GYP* locus Copy Number Changes in Short and Long Read WGS from Two U– Individuals. (A) Depth of coverage plot of short read WGS (left axis) and copy number variation (CNV, right axis) of *GYPE*, *GYPB*, and *GYPA* regions (100 bp bins) for U– samples (UNEG0-001, –002) of African American (AFR) ancestry. Each gene region is highlighted (*GYPE*: blue, *GYPB*: green, *GYPA*: red) with the genes in a reverse direction (i.e., exon 1 is on the right). The location of relevant paired reads are shown with purple curved lines. The evident structural variations are shown boxed at the bottom of each plot. Both are compound heterozygous with DEL_B_LEFT and DEL_B_RIGHT deletions of *GYPB*. (B) Long read WGS from sample UNEG-001. The *GYP* genes and the size of the deletions (kb) identified to be *in trans* by the clonal nature of long-range sequencing reads are designated and confirm compound heterozygosity for DEL_B_LEFT and DEL_B_RIGHT. Note: *GYPE*, *GYPB*, and *GYPA* are in reverse direction (i.e., exon 1 is on the right). (C) Sanger sequence of deletion breakpoint in UNEG-001. The breakpoint region is indicated in italics. Nucleotide changes before and after the breakpoint are underlined. Repeat regions located in the breakpoint region are shown [AT(n), AC(n), LIPBa and THE1C]. [Color figure can be viewed at wileyonlinelibrary.com]

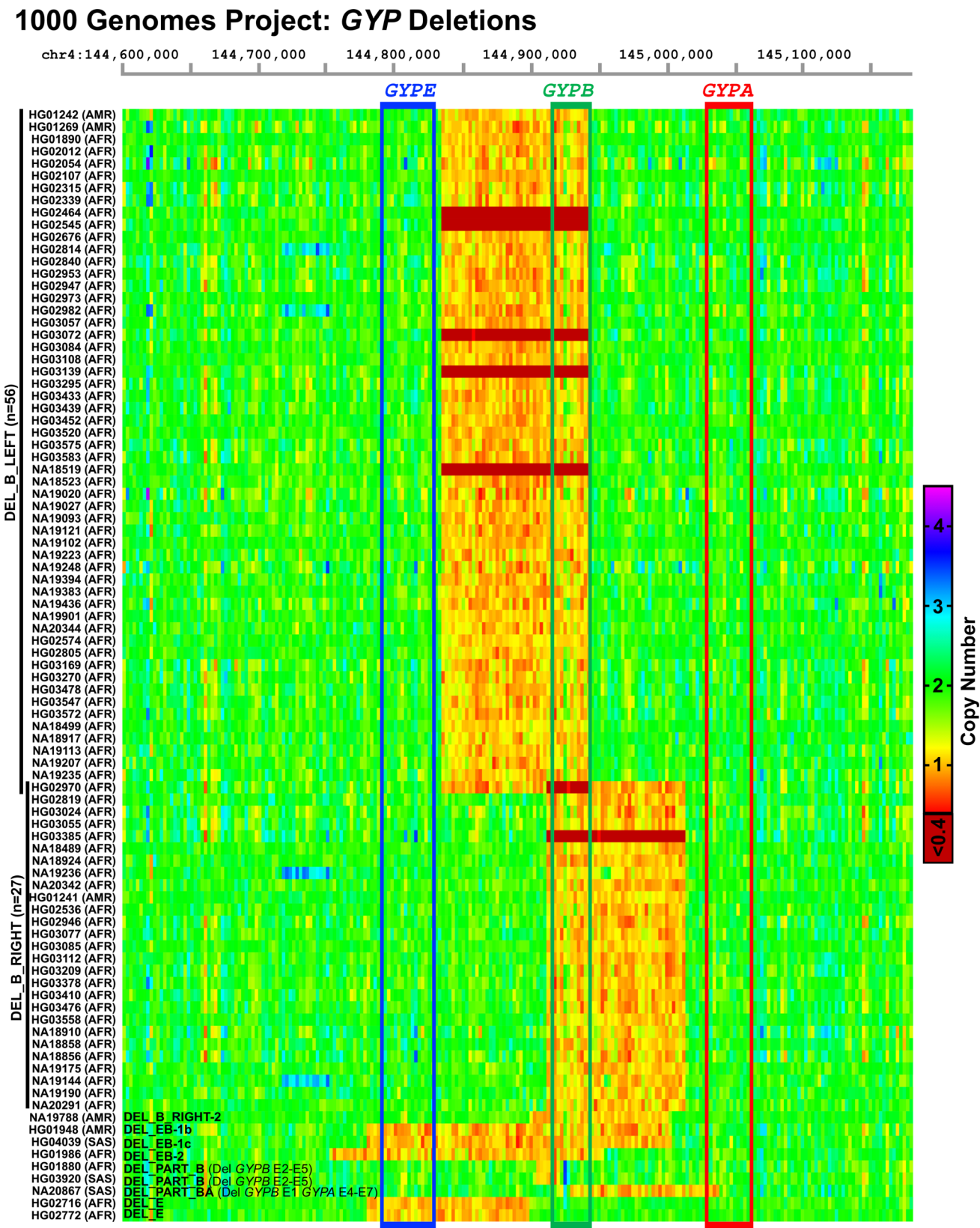


Fig. 4. Legend on next page.

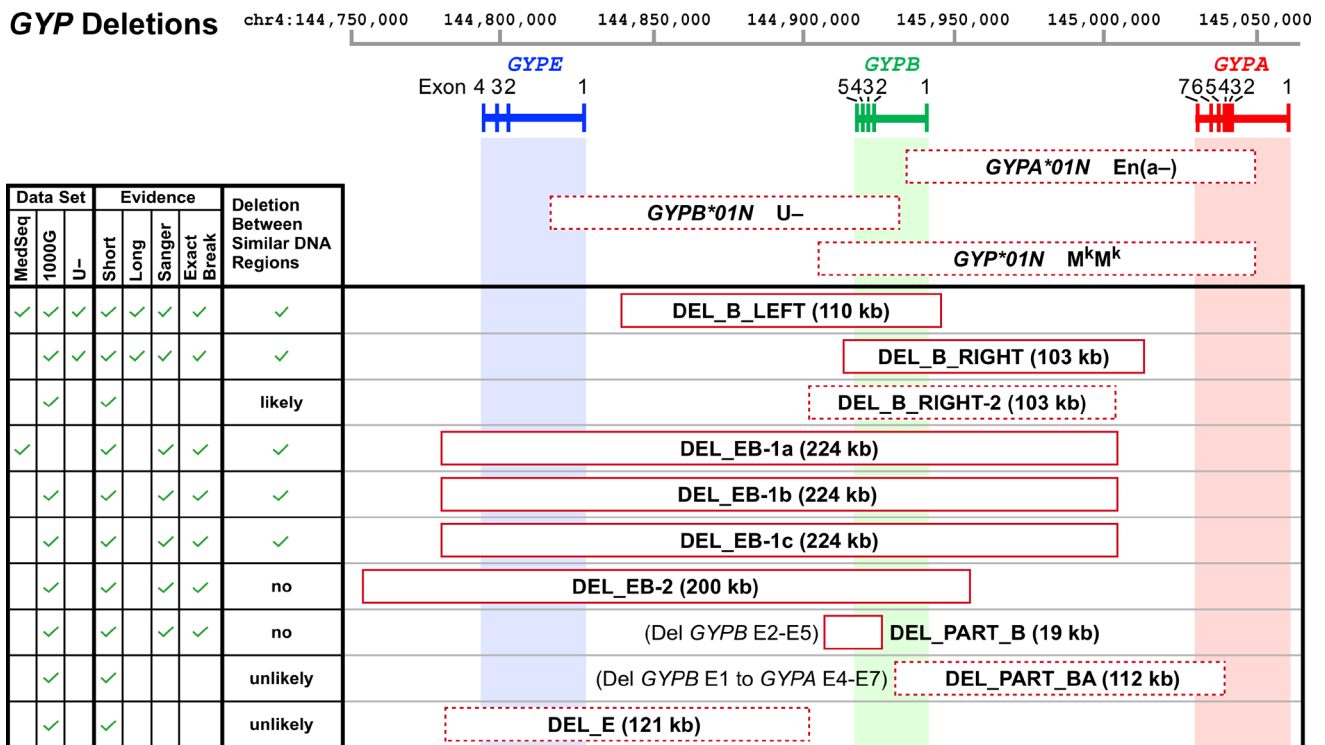


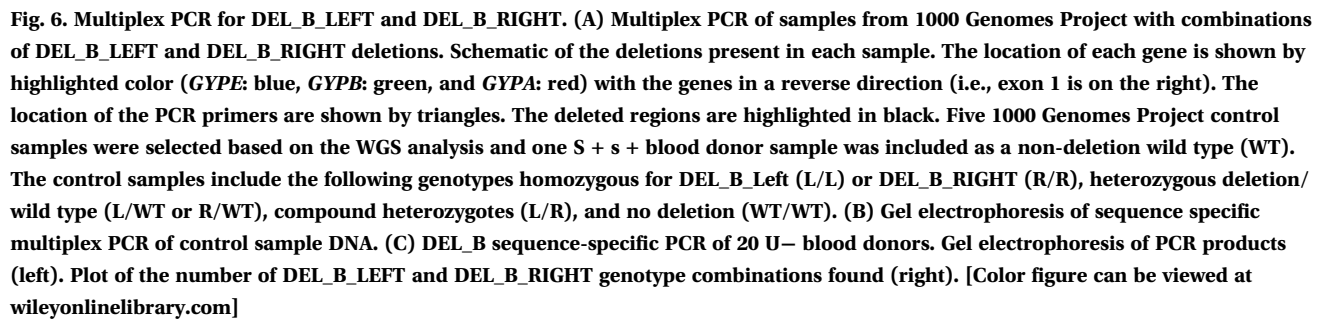
Fig. 5. Diagrammatic Summary of GYP Locus Deletions. The location of *GYPE*, *GYPB*, and *GYPA* are highlighted with the genes in reverse direction (i.e., exon 1 is on the right). The deleted regions are shown as boxes identified by the specific deletion and size in kb. The three previously published *GYP* locus deletions,^{1,6} associate with the En(a-) phenotype (*GYPA*01N*), the U- phenotype (*GYPB*01N*), and the M^kM^k phenotype (*GYP*01N*), are included (top) depicted as dotted rectangles to indicate the exact breakpoints have not been reported or studied. The location of the 10 *GYP* locus deletions reported in this study are shown as rectangles along with the data set in which the deletion was found, evidence to support (short and/or long read WGS, Sanger breakpoint sequencing), and if the deletion is in a region of DNA similarity. The three novel breakpoints not confirmed by Sanger are shown as dotted rectangles. See supplemental data for the breakpoint sequences and the genomic locations. [Color figure can be viewed at [wileyonlinelibrary.com](#)]

The two most common deletions were DEL_B_LEFT ($n = 61$) and DEL_B_RIGHT ($n = 28$). There were five AFR individuals homozygous for DEL_B_LEFT, one AFR individual homozygous for DEL_B_RIGHT and one compound heterozygous AFR individual with both DEL_B_LEFT and DEL_B_RIGHT. As such, the predicted prevalence of U- would be 1% (7 of 669) among those of African ancestry in the 1000 Genomes Project. Heterozygous DEL_B_LEFT and DEL_B_RIGHT samples were also evident (Fig. S3, available as supporting information in the online version of this paper.).

Several other unique deletions involving *GYPB* were observed in AFR, AMR, and SAS individuals (Fig. 4 and Fig. S4, available as supporting information in the online

version of this paper). Three samples, HG01948 [AMR-Peruvian], HG04039 [SAS-Sri Lankan Tamil], and HG01986 [AFR-Barbadian], were heterozygous for deletion of both *GYPB* and *GYPE*, with the first two corresponding to the heterozygous DEL_EB-1 deletion found in MEDSEQ-110 (Fig. 2B) and the latter a more left-shifted 200 kb deletion, designated DEL_EB-2. Four other samples had three other novel *GYPB* deletions (Fig. 4 and Fig. S4, available as supporting information in the online version of this paper). NA19788 [AMR-Mexican American] had a 103 kb deletion spanning the entirety of *GYPB* and extending downstream (right), designated DEL_B_RIGHT-2 since it is the same size and general location as the previous DEL_B_RIGHT but has a small 12 kb shift. HG01880 [AFR-Barbadian] and HG03920 [SAS-Bengali]

Fig. 4. MNS GYP locus Copy Number Changes in 1000 Genomes Project WGS. Heat map showing the copy number changes for *GYPE*, *GYPB*, and *GYPA* for 1000 Genomes Project samples with *GYPB* deletions ordered by location. Each horizontal row represents results from one sample and the column colors reflect the sequence read depth-based copy number (2,500 bp bins). Each gene region is highlighted (*GYPE*: blue, *GYPB*: green, *GYPA*: red) with the genes in a reverse direction (i.e., exon 1 is on the right). Ninety-one samples having apparent deletions are shown; 55 DEL_B_LEFT, 26 DEL_B_RIGHT, one DEL_B_LEFT/DEL_B_RIGHT compound heterozygote, and nine samples with seven other novel deletions. [Color figure can be viewed at [wileyonlinelibrary.com](#)]



Sanger sequencing was used to confirm the breakpoints of the *GYPB* deletions identified by WGS for MEDSEQ-110 and 1000 Genomes Project samples, and summarized in Table S2, available as supporting information in the online

version of this paper, by ethnic group, population location and the specific *GYP* deletion. Sequencing of DEL_B_LEFT (homozygous HG02464 [AFR-Gambian] and heterozygous HG01890 [AFR-Barbadian]) and DEL_B_RIGHT (homozygous HG03385 [AFR-Mende in Sierra Leone] and heterozygous HG02819 [AFR-Gambian]) were consistent with the individual breakpoints found in the compound heterozygous African American UNEG-001 and UNEG-002 (Fig. S1, available as supporting information in the online version of this paper, LEFT Breakpoint Sequence and Fig. S2, available as supporting information in the online version of this paper, RIGHT Breakpoint Sequence). Amplification and Sanger sequencing of the deletion breakpoints was also attempted on the eight other samples with potential novel deletions involving *GYPB* (Table S2, available as supporting information in the online version of this paper). Short read WGS data was used to locate the approximate location and

primers (Table S3, available as supporting information in the online version of this paper) located just outside of the deletion were used for amplification. All three DEL_EB-1 samples, characterized by deletion of both *GYPE* and *GYPB*, had slightly different breakpoint locations within a few hundred bp of each other (Fig. S5, available as supporting information in the online version of this paper), designated for MEDSEQ-110 (African American) as DEL_EB-1a, for HG01948 (Peruvian) as DEL_EB-1b, and for HG04039 (Sri Lankan Tamil) as DEL_EB-1c. Figure S6, available as supporting information in the online version of this paper, shows the breakpoint sequence for HG01986 (Barbados African), designated DEL_EB-2, with deletion of *GYPE* and *GYPB* but left-shifted by 200 kb. Sequence products from HG01880 (Barbados African) and HG03920 (Bengali) confirmed the location of the deletion region for DEL_PART_B characterized by partial deletion of *GYPB* extending from exon 2 through 5 (Fig. S7, available as supporting information in the online version of this paper). We were not successful in amplifying products from NA19788 [AMR-Mexican American] spanning the proposed DEL_B_RIGHT-2, or from NA20867 [SAS-Gujarati Indian] with partial deletion of *GYPB* exon 1 and *GYPB* exons 4-7, designated DEL_PART_BA.

Figure 5 summarizes the 10 novel *GYP* locus deletions identified in this study, 9 of which include all or part of *GYPB* (Table S4, available as supporting information in the online version of this paper, indicates the GRCh37/hg19 chromosomal coordinates). Interestingly, although most of the deletions occurred between regions of high similarity, at least two occurred in dissimilar regions (Fig. 5). DEL_EB-1a/b/c all occurred in a large stretch of highly similar DNA sequence (at least a 10,000 bp stretch is sequence with 95% identity). The DEL_EB-1a/b/c samples did not have paired reads spanning the deletion copy number changes, which can be explained by the large homology in this region, since paired read structural variant detection requires that the paired reads span a breakpoint with a sequence on either side sufficiently different as to cause alignment to the other region.

GYPB deletions in U– African American blood donors

WGS data analysis revealed nine novel deletions involving *GYPB* that would predict a U– phenotype in homozygotes. To determine the prevalence of these specific deletions in African American individuals, DNA from 20 known U– blood donors were tested using sequence-specific primers (Table S3, available as supporting information in the online version of this paper) and multiplex PCR. Figure 6A shows the *GYP* locus diagram and location of the deletion and non-deletion (wild type) control primers and five 1000 Genomes Project DNA samples selected as control samples based on our WGS analysis and one non-deletion wildtype S + s + blood donor. Figure 6B shows the multiplex PCR results of the control samples

illustrating the gel electrophoresis patterns associated with different combinations of DEL_B_LEFT, DEL_B_RIGHT, and the non-deletion wildtype. Figure 6C shows that among the 20 U– donors, 10 were homozygous for DEL_B_LEFT, 9 were compound heterozygous for DEL_B_LEFT and DEL_B_RIGHT, and 1 was homozygous for DEL_B_RIGHT. The results of each sample are summarized in Table S5, available as supporting information in the online version of this paper, along with the M, N, S, s, and U RBC phenotypes. The U– phenotype is reported to be primarily associated with the M–N+ form of *GYPB*.⁵ Similarly, here 12 of the 20 U– samples were M–N+. However, both DEL_B_LEFT and DEL_B_RIGHT were found in M–N+ and M+N– samples, indicating that neither deletion is exclusively associated with or linked in cis to a particular form of *GYPB*.

DISCUSSION

We report that the U– phenotype in those of African ancestry is primarily defined by two different genetic backgrounds consisting of complete deletion of *GYPB* with intact *GYPE*. Twenty-two serologically U– African American donors tested were homozygous for either a 110 kb deletion, DEL_B_LEFT, encompassing and extending left of *GYPB*, or homozygous for a 103 kb deletion, DEL_B_RIGHT, encompassing and extending right of *GYPB*, or were compound heterozygous DEL_B_LEFT/DEL_B_RIGHT. The two most common *GYPB* deletions present in the MedSeq Project sequence data and the 1000 Genomes Project data were also DEL_B_LEFT and DEL_B_RIGHT, with five AFR samples homozygous for DEL_B_LEFT, one homozygous for DEL_B_RIGHT, and one compound heterozygous with a predicted prevalence of 1% (7 of 669) U– individuals among those of African ancestry in the 1000 Genomes Project data.

We also identified seven other less common novel deletions involving *GYPB* in WGS data from individuals of African, Admixed American, and South Asian ancestry. Among the total of nine different deletions of all or part of *GYPB*, four backgrounds also included deletion of *GYPE* (DEL_EB-1a/b/c and – 2). Two were partial deletions of *GYPB* (exons 2-5) (DEL_PART_B), or deletion of *GYPB* exon 1 extending through exons 4-7 of *GYPB* (DEL_PART_BA). It is probable that these two partial *GYPB* deletions lead to a U– phenotype since DEL_PART_B involves the same region of *GYPB* as *GYPB*01N* and DEL_PART_BA likely removes the *GYPB* promoter. A tenth *GYP* locus deletion included *GYPE* only, and the impact on expression of glycophorins, if any, is unknown. The previously reported U– deletion of *GYPB* exon 2-5 with deletion of *GYPE* exon 1 (*GYPB*01N*)^{1,6} was not found in any samples from the MedSeq Project and 1000 Genomes Project datasets. Lastly, we also identified a potential heterozygous *GYPE* duplication (DUP_E) of unknown phenotypic importance in MEDSEQ-063 (AFR).

The molecular background of the U– phenotype was first investigated by Huang et al.²⁴ in 1987 using Southern

blot analysis which suggested that the absence of GPB was the result of the deletion of the entire *GYPB* gene, but the size of the deletion could not be confirmed. In 1989, Tate et al.²⁵ found similar results by Southern blot analysis, but reported that the 5' non-coding region of *GYPB* appeared to be present. In 1990, Vignal et al.⁶ performed Southern blot analysis of DNA from a French S–s–U– blood donor and identified a deletion of *GYPB* exons 2 to 5 and *GYPE* exon 1 as the molecular basis of the U– phenotype and taken together with the previous reports was assumed to be the primary genetic change underlying the U– phenotype and given the ISBT allele designation *GYPB*01N*.¹ Although the ethnicity of the proband was not included in that report,⁶ the U– sample was identified as “Fav” provided by M. Girard. A literature search found a previous report²⁶ of a Caucasian family identified as “Fav” with four rare S–s–U– family members. The data here indicate that the genetic background *GYPB*01N* is a rare Caucasian allele, as it was not found in our study of diverse individuals from 1000 Genomes Project data and individuals of African ancestry where the S–s–U– phenotype is most often found.

For prediction of the U– phenotype for blood transfusion, SNP-based assays target *GYPB* exons 4 and 5,^{27–29} which would also result in an accurate phenotype for samples with eight of the nine novel *GYPB* deletions reported here and likely accounts for why these alternative molecular backgrounds have gone undetected in the transfusion field. In the malaria literature, a recent report from the MalariaGen study investigated 1269 African genomes for *GYP* locus structural variations and found that *GYP*Dantu* offered resistance to malaria infection.³⁰ Among other structural variations noted were two common deletions they designated DEL1 and DEL2,³⁰ which are consistent with DEL_B_LEFT and DEL_B_RIGHT reported here. The MalariaGen study also found large deletions they designated DEL4 and DEL6³⁰ that included loss of *GYPB* and *GYPE*, consistent with DEL_EB-2 and DEL_EB-1, respectively, found here and noted a *GYPE* duplication DUP7,³⁰ consistent with DUP_E reported here. Indeed, the MalariaGen DEL1 Sanger sequence breakpoint sequence was consistent with that found here for DEL_B_LEFT (the MalariaGen study did not include Sanger sequencing of any other deletions). While this manuscript was under review, Gassner et al.³¹ also reported deletions consistent with DEL_B_LEFT, DEL_B_RIGHT, and DEL_PART_B.

Limitations to our study include that it is not possible to serologically confirm the 1000 Genomes Project genomic findings since only DNA and not red blood cell samples are available as archived material and participants are not available for follow up. Although the 1000 Genomes Project data are low coverage genomes with variable coverage, such limitations appear to be mitigated by the large size of these structural variations. Sanger sequencing confirmed seven out of the nine deletions involving *GYPB*. The unconfirmed deletions (DEL_B_RIGHT-2 and DEL_PART_BA) will require

further investigation as we were not successful in designing PCR primers that specifically amplified the proposed breakpoint regions.

Our results highlight the power of combining short and long read WGS to identify deletion breakpoint regions. By combining short and long read data we were able to identify the exact breakpoint locations for DEL_B_LEFT and DEL_B_RIGHT at level of fidelity equal to Sanger sequencing. In addition, for samples with only short read WGS this information alone narrowed down the location of the breakpoint to within 100s of bp, allowing for Sanger sequencing of the breakpoints. Given the vast size of the introns (1000s of bp) and the high homology between the *GYP* locus genes, this WGS-based guidance greatly simplified PCR primer selection for Sanger sequencing.

We propose that DEL_B_LEFT and DEL_B_RIGHT be named *GYPB*05N.01* and *GYPB*05N.02*, respectively consistent with Gassner et al.³¹ If the particular assay only detects the presence or absence of *GYP* locus exons and not the exact deletion breakpoints, the allele could be called at a lower resolution (e.g., *GYPB*05N*). Based on the 1000 Genomes Project data, the largest of the datasets, the allele frequencies for those for African ancestry were DEL_B_LEFT 0.04 (60 of 1338) and DEL_B_RIGHT 0.02 (27 of 1338), with a predicted 1% prevalence of U– phenotype in individuals of African ancestry. This predicted prevalence agrees with the published 1% serologic U– prevalence,⁴ suggesting that these two deletions represent the majority of U– phenotypes in those of African ancestry. For the other seven novel *GYPB* deletions, homozygous or compound heterozygous samples with serologic U antigen typing will be needed to assign alleles. Such an effort could be undertaken by screening serologically typed U– samples using the *GYPB* deletions sequence-specific PCR reactions reported here.

Our results indicate that there are many distinct genetic mechanisms underlying the U– phenotype and that there is still much to learn with regards to *GYP* locus structural variations. The information reported here is key for bioinformatic interpretative algorithms to accurately determine genotype/phenotype correlations from NGS data. The forthcoming WGS datasets from large genome sequencing projects will provide an unprecedented opportunity for future validation of these results along with continued discovery.

SEQUENCES

The deletion breakpoint sequences have been deposited to GenBank for DEL_B_LEFT (MN946505), DEL_B_RIGHT (MN946506), DEL_PART_B (MN958889), DEL_EB-1a (MN969924), DEL_EB-1b (MN969925), DEL_EB-1c (MT084352), and DEL_EB-2 (MT084353).

ACKNOWLEDGMENTS

The MedSeq Project was supported by the NHGRI (U01-HG006500). WJL is supported by the Brigham and Women's Hospital Pathology


Department Stanley L. Robbins MD Memorial Research Fund Award. CMW is supported by the Doris Duke Charitable Foundation (2011097 and 2015133). RCG is supported by the National Institutes of Health (U19-HD077671, U01-HG008685, R01-HG009922 and R01-HL143295), as well as funding from the Department of Defense and the Franca Sozzani Fund. The authors thank the staff and participants of the MedSeq Project. The authors also thank the staff of the New York Blood Center and Brigham and Women's Hospital Blood Bank. The research in the Ouwehand laboratory is supported by grants from the National Institute for Health Research and NHS Blood and Transplant.

CONFLICT OF INTEREST

RCG receives compensation for advising the following companies: AIA, Applied Therapeutics, Genome Medical, Glenn Biggs Institute, Helix, Humanity, and Verily. The other authors have disclosed no conflicts of interest.

REFERENCES

1. International Society of Blood Transfusion. Red cell immunogenetics and blood group terminology [Internet]. [cited 2017 Sep 1]. Available from: <https://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/>
2. Onda M, Kudo S, Fukuda M. Genomic organization of glycophorin A gene family revealed by yeast artificial chromosomes containing human genomic DNA. *J Biol Chem* 1994;269:13013-20.
3. Lane WJ, Westhoff CM, Gleadall NS, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol* 2018;5:e241-51.
4. Issitt PD, Anstee DJ. *Applied Blood Group Serology*. Durham, NC: Montgomery Scientific Publications; 1998. p. 1208.
5. Daniels G. *Human Blood Groups*. Hoboken, NJ: John Wiley & Sons; 2013. p. 544.
6. Vignal A, Rahuel C, London J, et al. A novel gene member of the human glycophorin A and B gene family. *Molecular cloning and expression*. *Eur J Biochem* 1990;191:619-25.
7. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. *Front Genet* 2015; 6:138.
8. Baronas J, Westhoff CM, Vege S, et al. RHD zygosity determination from whole genome sequencing data. *J Blood Disorder Transfus* 2016;7:1-5.
9. Möller M, Jöud M, Storry JR, et al. ErythroGene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Adv* 2016;1:240-9.
10. Schoeman EM, Lopez GH, McGowan EC, et al. Evaluation of targeted exome sequencing for 28 protein-based blood group systems, including the homologous gene systems, for blood group genotyping. *Transfusion* 2017;57:1078-88.
11. Chou ST, Flanagan JM, Vege S, et al. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv* 2017;1:1414-22.
12. Wheeler MM, Lannert KW, Huston H, et al. Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. *Genet Med* 2019;21:477-86.
13. Lane WJ, Vege S, Mah HH, et al. Automated typing of red blood cell and platelet antigens from whole exome sequences. *Transfusion* 2019;59:3253-63.
14. Vassy JL, Lautenbach DM, McLaughlin HM, et al. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* 2014;15:85.
15. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015; 526:68-74.
16. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841-2.
17. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178-92.
18. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res* 2004;32(Database issue):D493-6.
19. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* 2018;19:90.
20. Wang JR, Holt J, McMillan L, et al. FMLRC: hybrid long read error correction using an FM-index. *BMC Bioinformatics* 2018;19:50.
21. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;15:461-8.
22. TogoWS [Internet]. [cited 2019 Oct 26]. Available from: <http://togows.org/>
23. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* 2018;27:135-45.
24. Huang CH, Johe K, Moulds JJ, et al. Delta glycophorin (glycophorin B) gene deletion in two individuals homozygous for the S—U— blood group phenotype. *Blood* 1987;70:1830-5.
25. Tate CG, Tanner MJ, Judson PA, et al. Studies on human red-cell membrane glycophorin A and glycophorin B genes in glycophorin-deficient individuals. *Biochem J* 1989;263:993-6.
26. Sondag-Thull D, Girard M, Blanchard D, et al. S-s-U-phenotype in a Caucasian family. *Exp Clin Immunogenet* 1986;3:181-6.
27. Hashmi G, Shariff T, Seul M, et al. A flexible array format for large-scale, rapid blood group DNA typing. *Transfusion* 2005;45:680-8.
28. Hashmi G, Shariff T, Zhang Y, et al. Determination of 24 minor red blood cell antigens for more than 2000 blood donors by high-throughput DNA analysis. *Transfusion* 2007;47:736-47.
29. Finning K, Bhandari R, Sellers F, et al. Evaluation of red blood cell and platelet antigen genotyping platforms (ID CORE XT/ID HPA XT) in routine clinical practice. *Blood Transfus* 2016; 14:160-7.
30. Leffler EM, Band G, Busby GBJ, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* 2017;356(6343):1-12.

31. Gassner C, Denomme GA, Portmann C, et al. Two Prevalent ~100-kb **GYPB** Deletions Causative of the GPB-Deficient Blood Group MNS Phenotype S-s-U- in Black Africans. *Transfus Med Hemother* 2020. Available from: <https://www.karger.com/https://doi.org/10.1159/000504946>. 

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1: Supporting Information..