ARTICLE | Genetics inMedicine

Genomic characterization of the *RH* locus detects complex and novel structural variation in multi-ethnic cohorts

Marsha M. Wheeler, PHD[1], Kerry W. Lannert, MT (ASCP)[2], Haley Huston, BSc[3], Shelley N. Fletcher, BSc[3], Samantha Harris, BSc[3], Gayle Teramura, BSc[3], Helena J. Maki[2], Chris Frazar, MSc[1], Jason G. Underwood, PHD[1], Tristan Shaffer, BSc[1], Adolfo Correa, MD, MPH, PHD[4], Meghan Delaney, DO, MPH[3,5], Alex P. Reiner, MD, MSc[6], James G. Wilson, MD[7], Deborah A. Nickerson, PHD[1,8] and Jill M. Johnsen, MD[2,9], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Purpose: Rh antigens can provoke severe alloimmune reactions, particularly in high-risk transfusion contexts, such as sickle cell disease. Rh antigens are encoded by the paralogs, *RHD* and *RHCE*, located in one of the most complex genetic loci. Our goal was to characterize *RH* genetic variation in multi-ethnic cohorts, with the focus on detecting *RH* structural variation (SV).

Methods: We customized analytical methods to estimate paralog-specific copy number from next-generation sequencing (NGS) data. We applied these methods to clinically characterized samples, including four World Health Organization (WHO) genotyping references and 1135 Asian and Native American blood donors. Subsequently, we surveyed 1715 African American samples from the Jackson Heart Study.

Results: Most samples in each dataset exhibited SV. SV detection enabled prediction of the immunogenic RhD and RhC antigens in concordance (>99%) with serological phenotyping. RhC antigen expression was associated with exon 2 hybrid alleles (*RHCE*CE-D(2)-CE*). Clinically relevant exon 4–7 hybrid alleles (*RHD*D-CE(4-7)-D*) and exon 9 hybrid alleles (*RHCE*CE-D(9)-CE*) were prevalent in African Americans.

Conclusion: This study shows custom NGS methods can accurately detect *RH* SV, and that SV is important to inform prediction of relevant *RH* alleles. Additionally, this study provides the first large NGS survey of *RH* alleles in African Americans.

Genetics in Medicine (2019) 21:477–486; https://doi.org/10.1038/s41436-018-0074-9

Key Words: *RH*; Structural variation; Hybrid allele; Blood group; Next-generation sequencing

INTRODUCTION

Blood group systems are inherited entities with direct clinical importance in transfusion and transplantation medicine. Blood group antigens are expressed on the surface of red blood cells (RBCs); most are glycoproteins with specificity determined by their oligosaccharide or amino acid sequence.[1] The genes that encode nearly all blood group systems are known[2] and several exhibit substantial genetic complexity and population-specific heterogeneity.

The Rh blood group system contains highly immunogenic antigens and commonly exhibits complex genetic variation including structural variation (SV). It is comprised of >50 different antigens, including the polymorphic RhD (D) and RhCE (C, c, E, and e) antigens. This antigenic diversity stems from genetic variation in two homologous paralogs, *RHD* and *RHCE*, which lie in close proximity at the *RH* locus.[3] At present, *RHD* and *RHCE* encode >280 reported alleles (haplotypes) which include *RHD* gene deletions and *RHD–RHCE* hybrids.[2,4] This level of complexity poses clinical challenges and can provoke significant rates of Rh allosensitization.[5,6] In one study, 45% of chronically transfused African American patients with sickle cell disease (SCD) experienced alloimmunization, primarily due to undetected variation in the Rh blood group system.[5] High rates of Rh alloimmunization persist even when patients receive transfusions from serologically matched African American donors,[5] demonstrating the need for higher-resolution Rh blood group information.

Serology is the mainstay of clinical RBC typing, including Rh. However, serology has known limitations that can be overcome with molecular testing.[7] In clinical laboratories, DNA-based prediction is typically performed using

[1]University of Washington, School of Medicine, Department of Genome Sciences, Seattle, Washington, USA; [2]Bloodworks NW Research Institute, Seattle, Washington, USA; [3]Bloodworks NW Specialty Diagnostics, Red Cell Genomics Laboratory, Seattle, Washington, USA; [4]Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA; [5]Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA; [6]Department of Epidemiology, University of Washington, Seattle, Washington, USA; [7]Department Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA; [8]Brotman Baty Institute for Precision Medicine, Seattle, Washington, USA; [9]Department of Medicine, University of Washington, Seattle, Washington, USA. Correspondence: ,Deborah A. Nickerson debnick@uw.edu) or Jill M. Johnsen jjohnsen@uw.edu)

Submitted 25 January 2018; accepted: 16 May 2018
Published online: 29 June 2018

genotyping platforms (e.g., single-nucleotide polymorphism [SNP] arrays), Sanger sequencing, and variant-specific methods (e.g., polymerase chain reaction with sequence specific primers [PCR-SSP], restriction fragment length polymorphism [RFLP]).[7] These can be used to characterize patients with unexpected alloantibodies, patients at risk for allosensitization, or recently transfused patients. DNA-based methods are also used to identify alleles for which antisera are unavailable and to test for paternal zygosity of the D antigen for pregnancies at risk of hemolytic disease of the fetus and newborn.[7,8] In addition, RBC genotyping methods can aid in discriminating Rh phenotypes, which can produce indeterminate or conflicted serological results.[9] Genotyping methods can discriminate *RH* partial alleles, which lead to missing antigen epitopes and antibody formation when exposed to the conventional antigen.[10] Genetic methods can also discern weak *RH* alleles, which reduce the quantity of antigens on the surface of RBCs but maintain display of the same epitopes as conventional Rh antigens.[11]

Currently, there is growing interest in applying next-generation sequencing (NGS) to Rh antigen prediction.[12–16] NGS can systematically survey for genetic variants, including SV, and is scalable for high-throughput screening. To date, efforts to detect *RH* variation using NGS have shown success in detecting clinically relevant variation but technical challenges have limited the interpretation of *RH* variation and the detection of SV.[12–16] Our primary goal was to develop an *RH* genotyping method that addressed *RH* SV, including *RHD–RHCE* hybrid alleles that alter Rh antigen expression. We customized paralog-specific SV analyses[17] and first applied these methods to four World Health Organization (WHO) RBC genotyping reference samples and to 1135 clinically immunophenotyped and clinically genotyped samples from Asian and Native American blood donors.[18] Subsequently, we applied our methods to survey *RH* variation in 1715 unrelated African American samples from the Jackson Heart Study (JHS). This cohort was whole-genome sequenced (WGS) by the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) program and analyzed in this study to provide the first NGS survey of *RH* alleles for this population.

MATERIALS AND METHODS

Samples

We purchased four WHO reference DNAs (RBC1, RBC4, RBC5, RBC12) from the National Institute for Biological Standards and Control. WHO references were clinically characterized and genotyped by a variety of methods[19] but to our knowledge, not by NGS. These samples represent common European (RBC1, RBC4, RBC5) and African (RBC12) *RH* alleles (Table 1) including alleles encoding D positive (D+), D negative (D−), and combinations of C, c, E, e antigens (Table 1) [19].

Asian and Native American samples ($N = 1168$) were selected from a prior population study of blood donors.[18]

Table 1 Summary of NGS-predicted *RH* alleles, known Rh serology, and DNA variants in WHO reference samples

Sample	Rh serology[a]	ISBT alleles[a]	NGS-predicted antigens[b]	NGS-based alleles[b]	Relevant variation[c]
RBC1	D+	*RHD*01; RHD*01*	D+	*RHD*01; RHD*01*	
	C+c+	*RHCE*C; RHCE*c*	C+c+	*RHCE*C; RHCE*c*	Het. *RHCE*CE-D(2)-CE* hybrid allele
	E+e+	*RHCE*cE; RHCE*e*	E+e+	*RHCE*03; RHCE*01.01*	Het. missense variant (c.676G>C); Het. missense variant (c.48G>C)
RBC4	D+	*RHD*01; RHD*01*	D+	*RHD*01; RHD*01*	
	C+c−	*RHCE*C; RHCE*C*	C+c−	*RHCE*C; RHCE*c*	Hom. *RHCE*CE-D(2)-CE* allele
	E−e+	*RHCE*e; RHCE*e*	E−e+	*RHCE*e; RHCE*01.01*	Het. missense variant (c.48G>C)
RBC5	D−	*RHD*01N; RHD*01N*	D−	*RHD*01N; RHD*01N*	Hom. *RHD* deletion
	C−c+	*RHCE*c; RHCE*c*	C−c+	*RHCE*c; RHCE*c*	
	E−e+	*RHCE*e; RHCE*e*	E−e+	*RHCE*e; RHCE*e*	
RBC12	D−	*RHD*04N.01 (RHDΨ)*	D−	*RHD*04N.01; RHD*01N*	Hemi. variants (37-bp insertion, c.807T>G)[d]; Hemi. *RHD* deletion
	C−c+, V+ VS+	*RHCE*c; RHCE*c*	C−c+	*RHCE*c; (RHCE*01.20.02)*	Het. missense (c.48G>C); Het. missense (c.733C>G)
	E−e+, V+ VS+	*RHCE*e; RHCE*e*	E−e+	*RHCE*e; (RHCE*01.20.02)*	Het. missense (c.48G>C); Het. missense (c.733C>G)

[a]Rh serology and ISBT alleles shown are those previously reported in Boyle et al. (2013) or ISBT v.2.0 110914. ISBT allele names are adapted to avoid assuming phase between C, c and E, e indicative variants
[b]NGS-based predicted antigen expression and NGS-based alleles are predicted based on detected variants shown in the "Relevant variation" column
[c]Complementary DNA (cDNA) positions are relative to NM_016124.3 for *RHD* and NM_020485.4 for *RHCE*
[d]In RBC12, variants under "Relevant variation" co-occurred with 4 other variants that define the *RHD*04N.01*
ISBT International Society of Blood Transfusions, *NGS* next-generation sequencing, *WHO* World Health Organization

Blood samples were collected from consented volunteer donors by Bloodworks Northwest. All samples were previously clinically tested for D and C antigens by serology and for C, c, E, and e genotype using a SNP array, HEA BeadChipTM Kit (Bioarray Solutions Ltd., Immucor).[18] This sample set included 82 samples discrepant between C serology and SNP ($N = 16$) or indeterminate on the SNP array ($N = 66$).

African American samples ($N = 1715$) were selected from JHS samples (phs000964) WGS by the NHLBI TOPMed program. The JHS is a community-based observational study in which individuals were recruited from the tri-county area surrounding Jackson, Mississippi, including a subset who participated in the Atherosclerosis Risk in Communities Study.[20] The samples in this study were randomly selected from the maximum unrelated JHS sample set as identified using KING v1.4.0 (no individuals with a first or second degree relationship).

Library preparation and next-generation sequencing

DNA libraries from WHO and Asian and Native American samples were captured with a targeted panel designed to capture 41 blood group–relevant genes (1473 Kb; Nimblegen, Table S1). For *RH*, this panel captured 269 Kb of continuous sequence including introns, exons, utranslated regions (UTRs), and promoter regions. Library preparation followed a shotgun library construction method[21] and was hybridized in multiplex (22–24 samples per reaction). Sequencing was performed on Illumina HiSeq 2500 machines using paired-end 100 bp reads to a mean coverage of approximately 150×. In total, 1139 samples (1135 Asian and Native American and 4 WHO samples) passed sequencing quality thresholds. No samples were excluded based on performance at the *RH* locus.

JHS African American samples were WGS by the NHLBI TOPMed program. Library preparation for JHS samples similarly followed a shotgun library construction method.[21] Sequencing was performed on Illumina HiSeq X machines using paired-end 150 bp to a mean coverage of approximately 30×. Raw sequence data was aligned to the human reference genome (GRCh37) using BWA-MEM.[22]

Detection of *RH* SV

SV in *RHD* and *RHCE* was identified using an adaptation of methods described previously.[17] SV was identified by leveraging singly unique nucleotides (SUNs) within a repeat masked, pairwise sequence alignment of *RHD* and *RHCE*. SUNs were similarly identified in the Rhesus boxes flanking *RHD*.[3] SUNs were used to anchor DNA sequence *k*-mers ($k = 70$), which were screened for uniqueness against GRCh37 (BLAT v3.5, UCSC). *K*-mers were omitted if they contained >1 perfect match. Read depth was estimated for remaining *k*-mers using a mapping quality ≥40. Copy number was estimated by normalizing using sequencing depth and mean read depth for samples visually confirmed to have no SV. In total, 9189 *k*-mers for *RHD* and *RHCE* and 2054 *k*-mers in the Rhesus boxes informed SV analyses. *K*-mers were distributed across the *RH* locus except for *RHCE* exon 10. *RHD* exon 10 *k*-mers were identified in alignment of the Rhesus boxes. SV breakpoints were identified by change-point analysis using the R changepoint package.[23] SV impacting *RH* exons was prioritized.

Detection of *RH* SNVs and indels and *RH* allele identification

Single-nucleotide variants (SNVs) and small insertions and deletions (indels) were genotyped using GATK HaplotypeCaller and haplotype phased using statistical methods (Beagle v4.1) [24]. Functional annotation was incorporated using SeattleSeq Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). All variants were annotated relative to the RefSeq transcripts, NM_016124.3 (*RHD*) and NM_020485.4 (*RHCE*). To identify *RH* alleles, SNVs, indels, and SVs were cross-referenced with alleles listed by the International Society of Blood Transfusions (ISBT) v2.0 110914, supplemented by information from Rhesusbase.[4] For cross-referencing, complementary DNA (cDNA) coordinates associated with ISBT alleles were converted to GRCh37 coordinates. Chr1:25643553 (NM_016124.3:c.1136) and chr1:25747230 (NM_020485.4:c.48) are variant in GRCh37 relative to ISBT v2.0 110914. Novel variants were selected based on their absence in ISBT v2.0 110914 and prioritized as impactful based on variant function (e.g., predicted loss of function). Genotype quality (GQ) was assessed for novel and annotated ISBT SNVs and indels. Chr1:25643553, which encodes the primary variant of the DAU cluster (the DAU0 allele), had variable GQ because it is present in a multiply-mapping region in exon 8. GQ was low when Chr1:25643553 was variant relative to GRCh37, which contains the DAU0 variant (NM_016124.3:c.1136T). Low GQ resulted from low coverage of *RHD* exon 8 due to the misalignment of reads from this region to its highly homologous region in *RHCE*.

Quantitative multiplex PCR of short fluorescent fragments

To validate NGS-detected *RH* SV, we performed quantitative multiplex PCR of short fluorescent fragments (QMPSF).[25] Fluorescently tagged primers were used to amplify WHO and 18 Asian and Native American samples ($N = 22$) representative of *RHD* gene deletions, *RHD–RHCE* hybrid alleles or deletions/duplications, and to have no SV. QMPSF primers amplified gene-specific *RHD* and *RHCE* exons. *F9* exon 7 and *HFE* exon 2 amplicons served as positive amplification markers and as normalization controls. QMPSF products were separated via capillary gel electrophoresis (ABI 3130xl, Applied Biosystems). Fluorescence peaks were analyzed using the R Fragman package[26] and normalized using the maximum *HFE* peak height.

Combinatorial PCR and Sanger sequencing

To confirm *RHCE*CE-D(2)-CE* alleles (see Results) as hybrid alleles, we designed allele-specific long-range PCRs. Primer pairs were designed to target unique sequences between

intron 1–exon 2 and exon 2–exon 3 (Table S2). PCRs were performed pairing *RHD*- and *RHCE*-specific primers in a combinatorial manner. PCRs consisted of 12.5 µL of Q5 Hot Start High-Fidelity Master Mix (NEB M0494S), 0.5 µM of forward and reverse primers, and 50 ng DNA. Cycling conditions for intron 1–exon 2 were: 98 °C for 30 s followed by 30 cycles of 98 °C for 10 s, 76 °C for 30 s, 72 °C for 6 min, and 72 °C for 2 min. Cycling conditions for exon 2–exon 3 were identical except annealing and extension temperatures were 68 °C for 30 s and 72 °C for 3 min, respectively. PCR was performed on 21 samples (including WHO samples). Two samples with PCR-confirmed *RHCE*CE-D(2)-CE* events were cloned into pMiniT vector (NEB PCR Cloning Kit). Insert-positive clones were Sanger sequenced with vector-specific and gene-agnostic primers (Table S3). Products were aligned against *RHD* and *RHCE* (GRCh37) using Geneious R8 software.

RESULTS

NGS-based characterization of WHO reference samples

We used custom paralog-specific NGS analyses to detect SV at the *RH* locus. These analyses detected SV in all WHO reference samples. In RBC1 and RBC4, NGS analyses detected SV signals (Fig. **1a, c**) indicative of *RHD*-to-*RHCE* hybrid alleles (*RHCE*CE-D(2)-CE*), similar to alleles previously associated with the C+ phenotype.[27,28] Zygosity for this event was consistent with C and c phenotypes (Table **1**)[19]. In RBC5, analyses detected a homozygous *RHD* deletion causal for its reported D- phenotype (Fig. **1b**, Table **1**). In RBC12, analyses detected a hemizygous *RHD* deletion and SV indicative of an exon 9 hybrid allele (*RHCE*CE-D(9)-CE*) (Fig. **1d**). The latter event was not reported previously for RBC12 [19]. Each SV event was validated by QMPSF (Fig. **1**). The one discrepancy between QMPSF and NGS analyses related to whether SV in RBC12 impacts exon 8 in addition to exon 9: QMPSF amplification is suggestive of exon 8 SV, but NGS-based breakpoints predicted exon 8 to be unaffected (Fig. **1d**). The homozygous *RHD* deletion in RBC5 and *RHCE*CE-D(2)-CE* alleles predicted in RBC1 and RBC4 were additionally validated by allele-specific PCR (Fig. S1). PCR confirmed *RHCE*CE-D(2)-CE* events to be hybrid alleles and not separate SV events.

RBC1, RBC4, and RBC12 also harbored SNVs indicative of previously characterized alleles (Table **1**). In RBC1 and RBC4, we detected variants indicative of weak *RHCE* alleles (Table **1**). RBC12 contained hemizygous *RHD* SNVs representative of an *RHD* null allele including a 37-bp insertion and the stop-gained variant casual for its D- phenotype (Table **1**). RBC12 also harbored missense variants associated with the *RHCE*01.20.02* allele and the V+VS+ phenotype, a known finding for RBC12 [19].

NGS-based characterization of clinically characterized Asian and Native American samples

Paralog-specific analyses detected SV in 90% of Asian and Native American samples (Fig. **2a**, genotypes listed in Tables S4–S5). Note, we do not provide representative allele frequencies for these populations because this sample set was selected in a nonrandom manner. The *RHD* deletion was detected in 375 samples (100 homozygotes and 275 hemizygotes, Fig. **2a**). The predicted mean length for this event was 70154 ± 1888 bp and exhibited recombination signals between the flanking Rhesus boxes (similar to Fig. **1b**).[3] *RHCE*CE-D(2)-CE* alleles were detected in 832 samples (388 homozygotes and 444 heterozygotes, Fig. **2a**). The mean length for this event was 4953 ± 238 bp, with the most common variant being 4959 bp in size ($n = 823$) but other differently sized *RHCE*CE-D(2)-CE* were detected and ranged in size from 1038 to 7183 bp.

In 25 samples, we detected SV events impacting other *RHD* and *RHCE* exons, including *RHD* gene duplications and extensive *RHD–RHCE* hybrid alleles (see Figs. **2**a and **3**). Three of these events are annotated in ISBT v2.0 110914: *RHD*D-CE(4-7)-D* (*RHD*01N.07*, Fig. **3b**), *RHD*D-CE(3-9)-D* (*RHD*01N.04*, Fig. **3c**) and *RHD*D-CE(4-8)-D* (*RHD*01N.07*). *RHD*D-CE(4-7)-D* and *RHD*D-CE(4-8)-D* share ISBT allele names because previous genotyping methods could not determine whether exon 8 was affected.[4]

Standard SNV/indel calling methods detected SNVs associated with established serological phenotypes (Table S6, Tables S4–S5). In *RHD*, SNVs indicative of 2 *RHD* null alleles, 7 weak D and Del alleles, and 6 partial D alleles were detected (Table S6). Six samples with weak D and partial D alleles were predicted to inform D phenotype because of compound heterozygosity with *RHD* deletions. For example, one serologically D- sample harbored a splice-site variant (*RHD*DEL1*) and was hemizygous for a *RHD* gene deletion. In *RHCE*, variants were indicative of 10 *RHCE* alleles (Table S6). Predicted loss-of-function variants not reported in ISBT included 1 splice-site variant in *RHD* and 1 splice region variant in *RHCE* (Table S7).

QMPSF and allele-specific PCR for clinically characterized Asian and Native American samples

Using QMPSF, we tested 18 samples that collectively represented a variety of SV events (Fig. **3**, Figs. S2–S4). QMSF validated NGS-predicted events in all samples tested. As above, the discrepancy between QMPSF and NGS analyses related to the size of SV in *RHCE*CE-D(9)-CE* and *RHCE*CE-D(8-9)-CE* alleles (Figs. S3–S4).

Allele-specific PCRs further validated samples encompassing no SV, *RHD* deletions, and *RHCE*CE-D(2)-CE* events ($N = 17$, Figs. S5–S6). Cloning and Sanger sequencing of two samples exhibiting the common *RHCE*CE-D(2)-CE* allele confirmed a *RHCE* intron 1 SNV that was identified by NGS analysis in the larger dataset (chr1:25736299, Fig. S7). This SNV has not been previously reported and is positioned consistent with the *RHCE-RHD* intron 1 breakpoint. The *RHCE*CE-D(2)-CE* intron 2 breakpoint in these two samples was defined by a 109-bp insertion, which has been previously reported.[28]

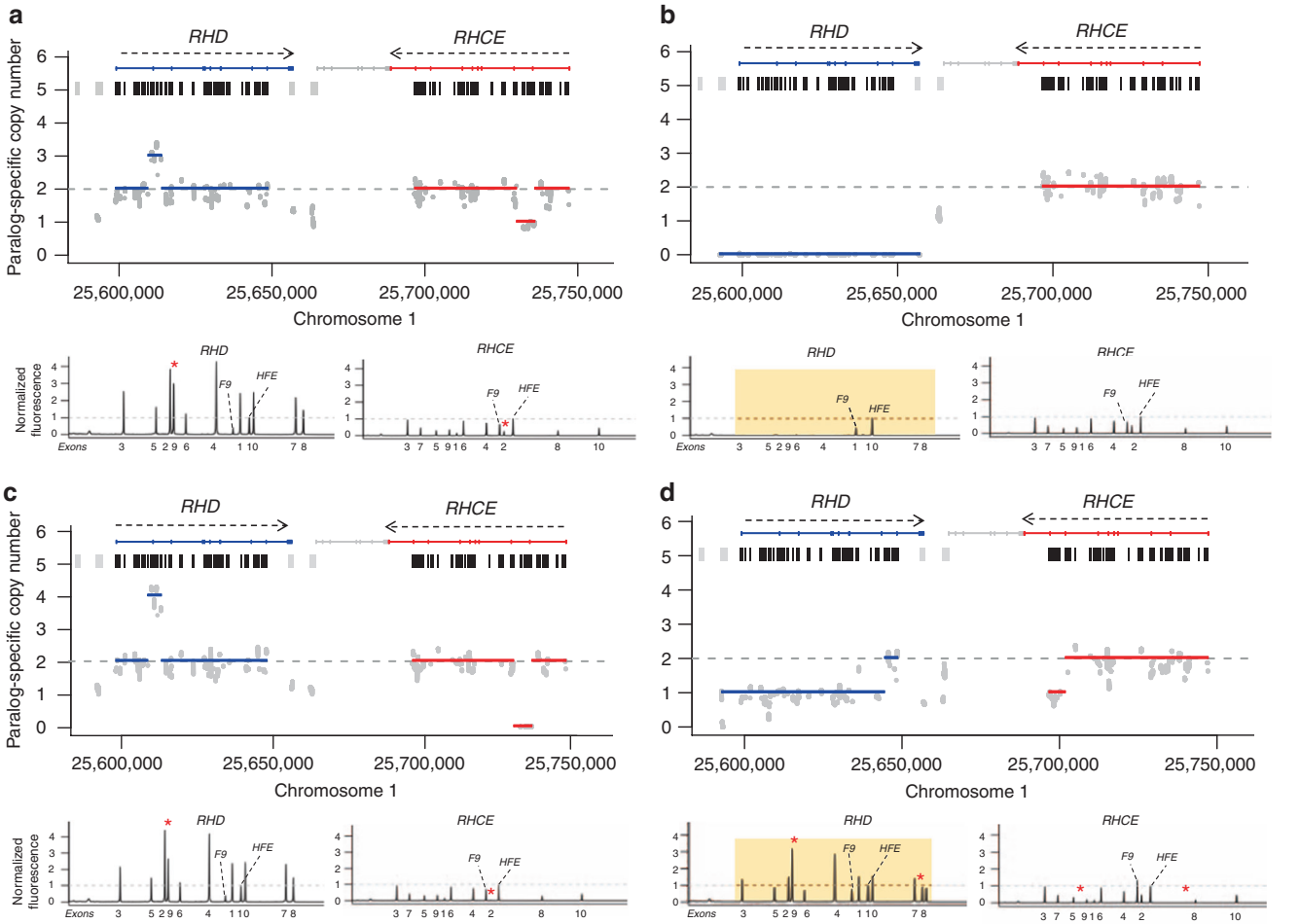

Fig. 1 Structural variation detected in WHO reference samples. a, b, **c**, and **d** Paralog-specific analyses (top) with corresponding quantitative multiplex PCR of short fluorescent fragments (QMPSF) results (below) for RBC1, RBC5, RBC4, and RBC12, respectively. Each paralog-specific panel shows scale *RHD* (blue) and *RHCE* (red) gene schematics (top) and the location of single unique nucleotides within genic regions (black) and in Rhesus boxes (gray). Gray circles within panels represent normalized mean read depth for *k*-mers corresponding to singly unique nucleotides (SUNs). The dashed gray line denotes a copy number of 2; solid blue and red lines indicate inferred copy numbers over the *RHD* and *RHCE* genes, respectively. In QMPSF panels, peak heights are fluorescence measurements normalized to the amplified exon for *HFE*. The *F9* peak serves as an additional positive amplification control. Light yellow panels in QMPSF results for **b** RBC5 and **d** RBC12 highlight *RHD* whole-gene deletions. Red asterisks highlight amplicons with results indicative of structural variation. Note: In **d**, QMPSF amplification of exon 8 is suggestive of structural variation, although exon 8 by next-generation sequencing (NGS) analyses is predicted to be unaffected

Comparisons between NGS-based *RH* alleles with SNP array–based typing and D and C serology

In Asian and Native American samples, NGS-based *RH* alleles were predicted blind to serology and SNP genotyping. NGS-genotype considered SNVs, indels, and SV within each sample. Briefly, D- in this dataset was mainly predicted from homozygous loss of *RHD*. However, one D- sample was predicted to be DEL due to hemizygous loss of *RHD* and the presence of the *RHD*DEL1* allele, a relevant distinction as DEL can provoke anti-D.[29] Another D- sample exhibited hemizygous loss of *RHD* and a deletion of *RHD* exon 3 (see example of exon 3 deletion in Fig. 3f), predicting a partial D phenotype. C and c antigens were predicted based on the presence of *RHCE*CE-D(2)-CE* alleles, while E and e genotypes were assigned using the ISBT annotated *RHCE* missense (NM_020485.4:c.676G>C).

Subsequent comparisons of NGS-genotype with serology showed agreement with the D antigen in 99.8% of samples and with the C+ antigen in 99.2% of samples. Comparisons with clinical SNP-genotype showed 99.9% agreement for prediction of E and e antigens. Direct comparison between all C SNP array predictions and all C NGS-based predictions was not possible due to indeterminate SNP array results in 66 samples (see Methods). In samples that did have SNP-based c and C predictions, our results were 99.5 and 99.7% concordant, respectively. All 66 samples with indeterminate C SNP array calls were predicted by NGS in agreement with serology. Most C SNP indeterminate samples (59/66) were NGS-predicted to be C+; all 66 of these samples were 100% concordant between NGS and serology. Moreover, NGS resolved 9 of 16 samples that were discordant between C SNP array–based genotype and C serology.

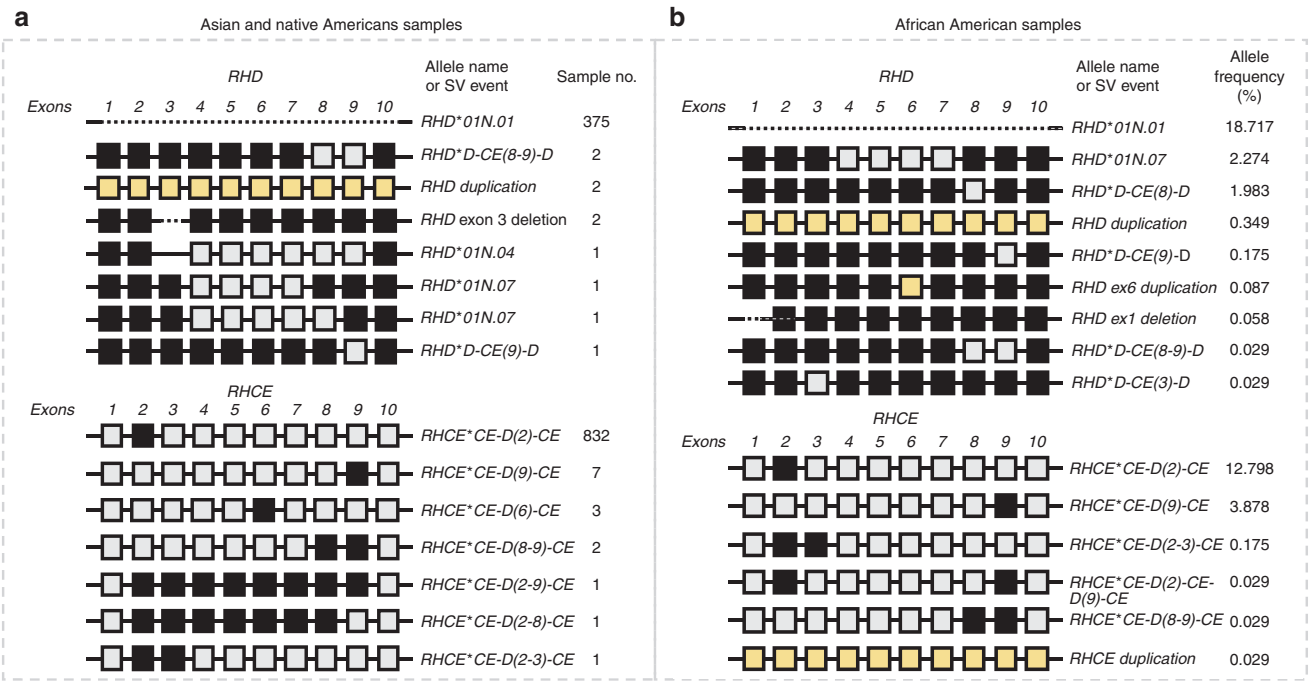

Fig. 2 Schematic summary of structural variation detected in a 1135 Asian and Native American samples and b 1715 Jackson Heart Study, African American samples. For **a**, allele frequencies are not reported because this dataset was selected nonrandomly . In both **a** and **b**, *RHD* and *RHCE* exons are depicted by black and gray boxes (respectively) and oriented 5′ to 3′ for simplification. Yellow boxes correspond to exons exhibiting duplication events. The *RHD* schematic summarizes structural variants detected in *RHD* with its corresponding International Society for Blood Transfusions (ISBT) allele name (if present) and **a** the sample number or **b** its corresponding allele frequency estimate shown as a percent. The *RHCE* schematic summarizes structural variants detected in *RHCE* with its corresponding ISBT allele name (if present) and **a** the sample number or **b** its corresponding allele frequency estimate. For both *RHD* and *RHCE*, structural variation is ordered by **a** sample number or **b** allele frequency estimate shown as a percent. SV structural variation

NGS-based characterization of African American samples

RH SV was detected in 61% of African American samples (Fig. **2b**, genotypes listed in Tables S8–S9). *RHD* gene deletions were present in 586 samples (mean length = 70572 ± 3352 bp) including 56 homozygotes and 530 hemizygotes (Fig. **2b**). *RHCE*CE-D(2)-CE* events were present in 406 samples (mean length = 5216 ± 796 bp) including 33 homozygotes and 373 heterozygotes. We additionally detected hybrid alleles at relatively high prevalence, including *RHD*D-CE(4-7)-D* (*RHD*01N.07*) and *RHCE*CE-D(9)-CE* (Fig. **2b**).

SNVs identified in African American samples were indicative of several *RHD* null alleles, weak D alleles, partial D alleles, and *RHCE* alleles (Table S10–11, Tables S8–S9). SNV-based *RH* alleles with allele frequencies >1% are shown in Table **2**, with previously reported SNP array–based allele frequencies.[30] Note we detected DAU alleles in several samples (Table **2**) but GQ for the primary variant was variable due to sequence homology. In African American samples, we also identified 5 predicted loss-of-function variants not reported in ISBT. In *RHD*, this included 1 splice-site variant and 2 frameshifts. In *RHCE*, this included 1 splice-site variant and 1 stop-gained variant (Table S12).

DISCUSSION

n recent years, there has been growing interest in applying NGS to predict Rh antigens.[12–16] This has been motivated, in part, by the high rates of Rh allosensitization in multiply transfused patients, particularly in African American patients with sickle cell disease.[31,32] In this population, high rates of allosensitization persist even after patients have been matched by serology for D, C, c, E, e antigens and received racially matched RBC transfusions.[5] Evidence suggests this is primarily due to the presence of undetected *RH* variation in patients and donors,[5] emphasizing the need to predict Rh antigens in a systematic and locus-informed manner.

To this end, studies have shown NGS is a viable approach for predicting RBC antigens.[12–16] However, these studies have applied NGS on a limited scale, mostly to a small number of well-characterized individuals, and have been largely insensitive in identifying complex SV, including *RHD–RHCE* hybrid alleles.[12–16] Here, we show customized NGS-based methods can detect known and novel *RH* variation in two large cohorts comprised of individuals of Asian American, Native American, and African American descent.

This customized *RH* method leverages nucleotide differences between *RHD* and *RHCE* to exclude mapping artifacts associated with NGS short read data. This approach enabled SV detection in previously problematic regions including exons 1, 2, and 8 [12,13,15,16] by using information in flanking intronic sequences. Importantly, this approach performs robustly both in targeted capture and whole-genome sequencing, indicating it is generalizable to datasets where NGS spans the *RH* locus. In addition, this approach provides the

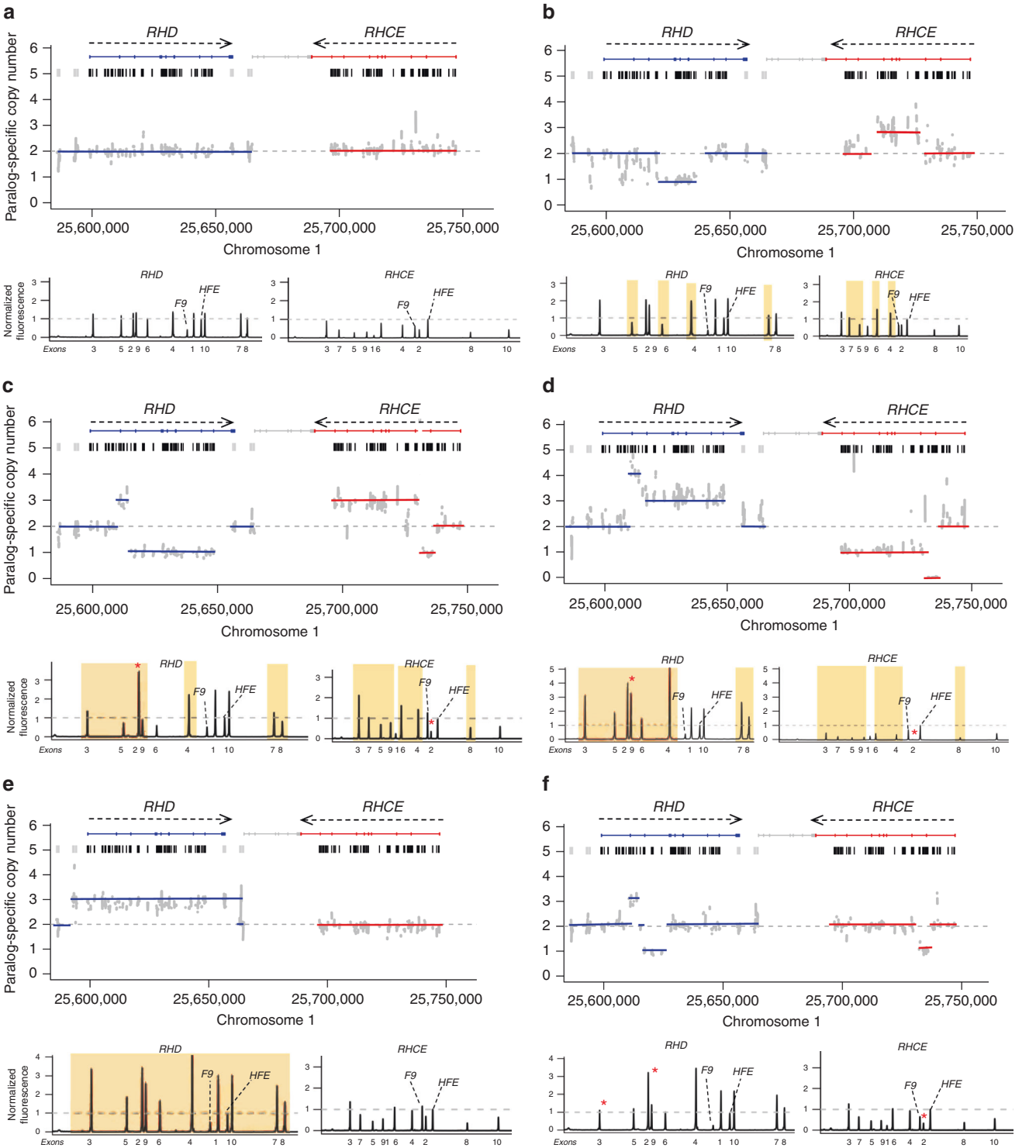

Fig. 3 Selected structural variation detected in Asian and Native American samples. a, **b**, **c**, **d**, **e**, **f** Paralog-specific analyses (top) and corresponding quantitative multiplex PCR of short fluorescent fragments (QMPSF) results (below) for samples exhibiting **a** no structural variation, **b** a *RHD*D-CE(4-7)-D*, **c** a *RHD*D-CE(3-9)-D* and *RHCE*CE-D(2)-CE*, **d** a *RHCE*CE-D(2-9)-CE*, **e** a *RHD* duplication, and **f** a *RHD* exon 3 deletion and *RHCE*CE-D(2)-CE*, respectively. Paralog-specific panels show scale *RHD* (blue) and *RHCE* (red) gene schematics (top) and the location of single unique nucleotides within genic regions (black) and in Rhesus boxes (gray). Gray circles within panels represent normalized mean read depth for 70-mers corresponding to single unique nucleotides. The dashed gray line denotes copy number of 2; solid blue and red lines indicate inferred copy numbers over the *RHD* and *RHCE* genes, respectively. In QMPSF panels, peak heights are fluorescence measurements normalized to the *HFE* control. The *F9* peak is a positive amplification control. Light yellow within QMPSF panels highlight multiple affected exons. Red asterisks highlight individual amplicons indicative of structural variation

Table 2 Prevalent (>1%) single-nucleotide variant (SNV)-based *RHD* and *RHCE* alleles detected in African American samples

Allele name[a]	*Phenotype*[a]	*Allele no.*[b]	*Allele frequency (%)*[b]	*Previously reported frequency (%)*[c]
*RHD*04N.01 (RHDΨ)*	D null	109	3.178	3.4
RHD[186G>T; 410C>T; 455A>C; 602C>G; 667T>G; 819G>A][d]*	DIIIa	40	1.166	1.4
*RHD*03.04*	DIII type 4	77	2.245	0.1
*RHD*09.03*	DAR	49	1.429	1.9
*RHD*10.00*[e]	DAU0	763	22.24	16.1
*RHD*10.03*[e]	DAU3	66	1.924	1.9
*RHCE*01.01*	e weak	1254	36.560	42.8
*RHCE*01.02*	Partial e	67	1.953	1.9
*RHCE*01.06*	Partial eCEAG−	147	4.286	4.5
*RHCE*01.07*	Partial e partial chrS−	41	1.195	1.6
*RHCE*01.20.01*	Partial e partial c, V+ VS+	473	13.790	[f]
*RHCE*01.20.02*	Partial e partial c, V+ VS+	119	3.469	[f]
*RHCE*01.20.03*	Partial e partial c, V −VS+	122	3.557	3.5
*RHCE*cE (RHCE*03)*	E[d]	363	10.583	10.3

[a]Allele names and phenotypes are as designated by ISBT v2.0 110914
[b]The number of alleles present and allele frequency in this dataset
[c]Allele frequencies in African Americans and SCD patients reported in Reid et al.[30]
[d]Novel *RH* allele relative to ISBT v2.0 110914. The "[]" and ";" follow Human Genome Variation Society (HGVS) conventions to denote variants were present on the same chromosome
[e]Genotype quality for the primary variant of the DAU cluster (NM_016124.3:c.1136C>T) was variable due to low coverage in the absence of DAU0 and high sequence homology between *RHD* and *RHCE* exon 8
[f]Alleles were observed in Reid et al.[30] but were reported jointly

ability to detect *RH* SV at scale to measure allele frequencies in large genomic datasets.

We specifically detected *RHCE*CE-D(2)-CE* hybrid alleles as prevalent across all datasets. Similar alleles were reported previously and associated with C+ expression, such as by Carrit et al.[28] However, at present there is a lack of clarity as to whether these alleles are causal for C+ expression. Recent exome studies report exon 2 read depth signals associated with C+, which are indicative of SV;[15,16] however, the majority of modern literature including *RHCE* genotyping references report exon 1 and 2 *RHCE* SNVs as causal for C+[2]. In these large-scale analyses, the most common *RHCE*CE-D(2)-CE* allele spanned ~5 Kb and a subset of samples with *RHCE*CE-D(2)-CE* were validated by multiple orthogonal methods. Sanger sequencing characterized the common *RHCE*CE-D(2)-CE* intronic breakpoints, including an *RHCE* 109 bp "insertion" currently used in C genotyping as well as a previously undetected SNV at the *RHCE*CE-D(2)-CE* breakpoint in *RHCE* intron 1. Our analyses also show *RHCE*CE-D(2)-CE* correctly predicted C serology in 99.2% of clinically characterized samples, strongly supporting that *RHCE*CE-D(2)-CE* is the common cause for C+ antigen expression.

We further identified multiple *RH* hybrid alleles consistent with named ISBT alleles. We identified the clinically known *RHD*01N.07* (*RHD*D-CE(4-7)-D*) in both large cohorts and validated this NGS signature by QMSPF (Fig. **3b**). This allele was prevalent (2.3%) in African Americans, consistent with a recent study reporting this allele to occur in 2.9% of African American individuals and sickle cell disease patients[30] and 10-fold higher than in European populations.[33]

Our methods identified novel *RH* SV alleles that impacted exons 8 and 9. This finding suggests previous genotyping efforts may have been hindered by sequence homology across these exons, a notion supported by our finding of *RHCE*CE-D(9)-CE* allele in the well-characterized WHO reference, RBC12. Notably, *RHCE*CE-D(9)-CE* was common (3.9%) in African American samples. In Asian and Native American samples, QMPSF validated *RHCE*CE-D(9)-CE* alleles but also showed amplification of exon 8 in a subset of samples. QMPSF infers exon 8 copy number through amplification of nearby intronic sequences, leading us to hypothesize intronic variation associated with *RHCE*CE-D(9)-CE* may have impacted this QMPSF result. Alternatively, our NGS-based methods could have excluded exon 8 as part of the SV due to the breakpoint being in a region of high homology.

Although our analyses were focused on SV, we genotyped SNVs indicative of known ISBT alleles. Notably, in an Asian American sample we detected hemizygous loss of *RHD* and an *RHD* splice-site variant causal for the DEL phenotype (*RHD*DEL1). This correlated with the D- phenotype reported in this blood donor, but this is a relevant finding as DEL is not null for D protein expression and can provoke D

alloimmunization. This DEL allele has been reported as a common cause of D- in Asian populations;[29] although, in this study of Asian Americans homozygous loss of *RHD* was the primary cause of D-. We further found weak and partial *RH* alleles known to be prevalent and clinically consequential in African populations (Table 2). Consistent with previous NGS work,[15] we detected common *RHD* SNVs in African Americans indicative of DAU alleles. The primary DAU0 SNV had variable genotype quality leading us (and others)[15] to provide caution when interpreting DAU allele frequencies derived from NGS. The limitation we observed was low coverage in the absence of the DAU0 SNV due to increased sequence homology with *RHCE*. Additional customization of NGS analyses, such as the use of an alternative mapping locus, should resolve this limitation. Separately, in RBC12, we detected SNVs indicative of *RHD*04N.01* and *RHCE*01.20.02*. In African Americans, we detected *RHD*04N.01* at a frequency of ~3% (Table 2), consistent with allele frequencies reported by other studies in individuals of African descent.[30] *RHD*04N.01* co-occurred with hemizygous *RHD* gene deletions predicting D- in 1.4% of African Americans, while 3.2% of African Americans were D- due to homozygous *RHD* gene deletions.

In summary, our results show the ability of NGS-based methods to systematically identify *RH* SV and detect known, complex, and novel *RH* SV. This represents the first scale study of *RH* variation in Asian and Native Americans and the largest population survey of *RH* SV in African Americans to date. We found complex SV to be common suggesting additional clinically relevant *RH* variation remains undiscovered. Altogether, this study shows locus-informed genomic approaches can detect *RH* alleles and characterize complex genetic variation in large and diverse datasets.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0074-9) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank our colleagues at Bloodworks NW and the Nickerson Lab for their advice and assistance, particularly Danielle Drury-Stewart, Thomas Walsh, Colleen Lammers, Ken Setran, Yanyun Wu, James Zimring, Karen Nelson, Barbara Konkle, Colleen Davis, Stephanie Krauter, Josh Smith, Peggy Robertson, Steven Lee, and Qian Yi. This study was supported by an NHLBI RS&G Pilot Project (HHSN268201100037C) and a Cardiovascular Research Training Grant. Whole-genome sequencing (WGS) for the TOPMed program was supported by the NHLBI. WGS for the Jackson Heart Study (JHS) (phs000964.v1.p1) was performed at the University of Washington (UW) Northwest Genomics Center (HHSN268201100037C). Centralized data harmonization was provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples. The JHS is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from the NHLBI and the National Institute for Minority Health and Health Disparities (NIMHD).

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Reid, ME, Lomas-Francis, C & Olsson, ML. *The Blood Group Antigen FactsBook*. London, UK. (Academic Press, 2012).
2. Storry JR, et al. International society of blood transfusion working party on red cell immunogenetics and terminology: report of the Seoul and London meetings. *ISBT Sci Ser*. 2016;11:118–22.
3. Wagner FF, Flegel WA. RHD gene deletion occurred in the Rhesus box. *Blood*. 2000;95:3662–8.
4. Wagner FF, Flegel WA. The rhesus site. *Transfus Med Hemother*. 2014;41:357–63.
5. Chou ST, et al. High prevalence of red blood cell alloimmunization in sickle cell disease despite transfusion from Rh-matched minority donors. *Blood*. 2013;122:1062–71.
6. Sippert E, et al. Variant RH alleles and Rh immunisation in patients with sickle cell disease. *Blood Transfus*. 2015;13:72–7.
7. Hillyer CD, Shaz BH, Winkler AM, Reid M. Integrating molecular technologies for red blood cell typing and compatibility testing into blood centers and transfusion services. *Transfus Med Rev*. 2008;22:117–32.
8. Westhoff CM. Molecular DNA-based testing for blood group antigens: recipient-donor focus. *ISBT Sci Ser*. 2013;8:1–5.
9. Denomme GA, Dake LR, Vilensky D, Ramyar L, Judd WJ. Rh discrepancies caused by variable reactivity of partial and weak D types with different serologic techniques. *Transfusion*. 2008;48:473–8.
10. Castilho L, et al. High frequency of partial DIIIa and DAR alleles found in sickle cell disease patients suggests increased risk of alloimmunization to RhD. *Transfus Med*. 2005;15:49–55.
11. Wagner FF, et al. Molecular basis of weak D phenotypes. *Blood*. 1999;93:385–93.
12. Stabentheiner S, et al. Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. *Vox Sang*. 2011;100:381–8.
13. Fichou Y, Audrézet M-P, Guéguen P, Le Maréchal C, Férec C. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol*. 2014;167:554–62.
14. Lane WJ, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion*. 2016;56:743–54.
15. Chou ST, et al. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv*. 2017;1:1414–22.
16. Schoeman EM, 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.
17. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. *Science*. 2010;330:641–6.
18. Delaney M, et al. Red blood cell antigen genotype analysis for 9087 Asian, Asian American, and Native American blood donors. *Transfusion*. 2015;55:2369–75.
19. Boyle J, et al. International reference reagents to standardise blood group genotyping: evaluation of candidate preparations in an international collaborative study. *Vox Sang*. 2013;104:144–52.
20. Taylor HA. The Jackson Heart Study: an overview. *Ethn Dis*. 2005;15: S6–1–3.
21. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009;461:272–6.
22. Li, H Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013;1303:3997.
23. Killick, R, and Eckley IA. changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software,*58:1–19.

24. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007;81:1084–97.
25. Fichou Y, et al. A convenient qualitative and quantitative method to investigate RHD-RHCE hybrid genes. *Transfusion*. 2013;53:2974–82.
26. Covarrubias-Pazaran G, Diaz-Garcia L, Schlautman B, Salazar W, Zalapa J. Fragman: an R package for fragment analysis. *BMC Genet*. 2016;17:62.
27. Poulter M, Kemp TJ, Carritt B. DNA-based rhesus typing: simultaneous determination of RHC and RHD status using the polymerase chain reaction. *Vox Sang*. 1996;70:164–8.
28. Carritt B, Kemp TJ, Poulter M. Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. *Hum Mol Genet*. 1997;6:843–50.
29. Kwon DH, Sandler SG, Flegel WA. DEL phenotype. *Immunohematology*. 2017;33:125–32.
30. Reid ME, Halter-Hipsky C, Hue-Roye K, Hoppe C. Genomic analyses of RH alleles to improve transfusion therapy in patients with sickle cell disease. *Blood Cells Mol Dis*. 2014;52:195–202.
31. Aygun B, Padmanabhan S, Paley C, Chandrasekaran V. Clinical significance of RBC alloantibodies and autoantibodies in sickle cell patients who received transfusions. *Transfusion*. 2002;42: 37–43.
32. Lasalle-Williams M, et al. Extended red blood cell antigen matching for transfusions in sickle cell disease: a review of a 14-year experience from a single center (CME). *Transfusion*. 2011;51:1732–9.
33. Wagner FF, Frohmajer A, Flegel WA. RHD positive haplotypes in D negative Europeans. *BMC Genet.* 2001;2:10.