ARTICLE

Accurate long-read sequencing allows assembly of the duplicated *RHD* and *RHCE* genes harboring variants relevant to blood transfusion

Zhe Zhang,[1,8] Hyun Hyung An,[2,8] Sunitha Vege,[3] Taishan Hu,[4] Shiping Zhang,[1] Timothy Mosbruger,[4] Pushkala Jayaraman,[4] Dimitri Monos,[4,5] Connie M. Westhoff,[3] and Stella T. Chou[2,6,7,*]

Summary

Next-generation sequencing (NGS) technologies have transformed medical genetics. However, short-read lengths pose a limitation on identification of structural variants, sequencing repetitive regions, phasing of distant nucleotide changes, and distinguishing highly homologous genomic regions. Long-read sequencing technologies may offer improvements in the characterization of genes that are currently difficult to assess. We used a combination of targeted DNA capture, long-read sequencing, and a customized bioinformatics pipeline to fully assemble the *RH* region, which harbors variation relevant to red cell donor-recipient mismatch, particularly among patients with sickle cell disease. *RHD* and *RHCE* are a pair of duplicated genes located within an ~175 kb region on human chromosome 1 that have high sequence similarity and frequent structural variations. To achieve the assembly, we utilized palindrome repeats in PacBio SMRT reads to obtain consensus sequences of 2.1 to 2.9 kb average length with over 99% accuracy. We used these long consensus sequences to identify 771 assembly markers and to phase the *RHD-RHCE* region with high confidence. The dataset enabled direct linkage between coding and intronic variants, phasing of distant SNPs to determine *RHD-RHCE* haplotypes, and identification of known and novel structural variations along with the breakpoints. A limiting factor in phasing is the frequency of heterozygous assembly markers and therefore was most successful in samples from African Black individuals with increased heterogeneity at the *RH* locus. Overall, this approach allows *RH* genotyping and *de novo* assembly in an unbiased and comprehensive manner that is necessary to expand application of NGS technology to high-resolution *RH* typing.

Introduction

Next-generation sequencing (NGS) technology has enabled expansion of genetic testing for numerous diagnostic applications, transforming many aspects of clinical medicine. Despite these advances, NGS with short-read lengths of ~150–300 base pairs (bp) hinders the ability to accurately map or assemble reads from regions with structural variation, repetitive sequence, high guanine-cytosine (GC) content, or duplicated and highly homologous genes.[1,2] Short-read NGS does not support direct variant phasing, leaving analysis highly dependent on reference genomes, which are known to be imperfect and problematic for genes with a high degree of heterogeneity among different populations.[1–3] Long-read sequencing (LRS) platforms may be able to overcome specific limitations of short-read NGS-based gene assembly for duplicated or homologous genes with high degrees of variation.[2,4] Third generation sequencing platforms are capable of sequencing through traditionally difficult sequence templates with high GC content[5] and enable direct phasing of variants located multiple kilobases apart and assembly of complex genomic regions such as the major histocompatibility complex.[6,7]

In the field of transfusion medicine, genetic characterization of blood group systems has identified the associated polymorphisms responsible for most of the clinically relevant red cell antigens that contribute to incompatibility in blood transfusions.[8] The Rh system is one of the most diverse and complex and comprised of two linked homologous genes, *RHD* (MIM: 111680) and *RHCE* (MIM: 111700), that encode the common red-cell-specific Rh antigens designated D, C, c, E, and e.[9] The two genes are proximally located in reverse orientation and span ~175 kb (chr1: 25,258,883–25,425,327) with ~30 kb between them. *RHD* is flanked by a pair of upstream and downstream 9.2 kb sequences that share >97% similarity and are in identical orientation, termed Rhesus boxes.[10] The most common RhD-negative phenotype is due to deletion of the *RHD* gene that occurred by an unequal cross-over between the Rhesus boxes leaving a single hybrid box.[10,11]

RH genetic heterogeneity is not uncommon and is more often found in individuals of African descent. For patients with sickle cell disease (SCD [MIM: 603903]) who frequently

[1]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [2]Division of Hematology, Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [3]Immunohematology and Genomics, New York Blood Center, New York, NY 11101, USA; [4]Immunogenetics Laboratory, Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [5]Department of Pathology and Laboratory Medicine, Perelman Schools of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [6]Division of Transfusion Medicine, Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[7]Present address: Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Abramson Research Center Room 316D, Philadelphia, PA 19104, USA
[8]These authors contributed equally
*Correspondence: chous@chop.edu
https://doi.org/10.1016/j.ajhg.2021.12.003.

require blood transfusion, *RH* genetic polymorphisms that change the conformation of Rh proteins in the red cells of both patients and healthy blood donors contribute to incompatibility.[12,13] This is demonstrated as production of antibodies to Rh proteins following transfusion (termed alloimmunization) despite transfusions that are Rh matched by traditional immunologic methodologies.[12–14]

RH genotyping currently uses a combination of PCR-based approaches and commercial DNA array-based testing to target multiple regions of the *RH* locus. NGS-based mapping of complementary DNA (cDNA)-annotated red cell and platelet antigens can provide accurate blood group antigen typing for most blood groups via whole-genome sequence (WGS) data.[15,16] For *RH*, alignment of NGS sequence reads is more difficult because of the high homology between *RHD* and *RHCE*,[17,18] and short-read sequences do not allow for full-gene phasing for *RHD-RHCE* haplotype assembly or allele assignment in samples with multiple polymorphisms and distant nucleotide variations. This has made it difficult to establish a standard reference sequence without artifact. Longer reads covering a larger number of assembly markers would have the advantage of reducing such ambiguity.

In this study, we combined our targeted DNA enrichment method, region-specific enrichment (RSE), which captures long DNA segments of up to 20 kb in length,[19] with long-read sequencing on the Pacific Biosciences (PacBio) platform. Although PacBio single-molecule real-time (SMRT) reads may allow for full assembly of the *RHD-RHCE* region, they have much lower sequencing accuracy (10% to 15% error rate) compared to Illumina short reads (<1% error rate).[5] Thus, we hypothesized that palindrome sequences found within the same SMRT full reads could be used for generation of a consensus sequence with higher accuracy based on two assumptions. First, the palindrome sequences in the same read originate from an identical DNA fragment, which is duplicated during whole-genome amplification. Second, sequencing errors happen randomly; therefore, the consensus of multiple palindrome sequences would have higher accuracy on the basis of the "wisdom of the crowd" principal. Using a custom PAClindrome bioinformatics pipeline, we successfully assembled the *RHD-RHCE* region with accurate consensus sequences at high depth among Black subjects who have a high frequency of *RH* genetic variation.

Material and methods

Sample selection and DNA preparation

We obtained samples from two White and nine Black individuals to represent varied *RH* genotypes (Table 1) under a protocol approved by the Children's Hospital of Philadelphia Institutional Review Board. *RH* genotyping had been performed as described previously.[12,13] Genomic DNA was isolated from fresh whole blood on the EZ1 Advanced XL instrument with EZ1 DNA Blood Kits and EZ1 Advanced XL DNA blood cards (QIAGEN, Maryland) and stored at 4°C.

Targeted capture oligonucleotide primer design

RSE primers were designed for the capture of targeted regions with a custom-designed software program RSE Antholigo as previously described.[19,20] AnthOligo is available as a public web application (see web resources). Forty-nine capture oligonucleotides were designed to target the *RHD-RHCE* region including the 30 kb between the two genes with an average distance of ~3.3 kb between primers (Table S1). We targeted additional genes simultaneously to include a minimum amount of genomic material for capture efficiency (data not shown).[19]

Targeted DNA capture by RSE

We performed each RSE reaction by using 2 μg fresh genomic DNA with 4.5 μM primers and the RSE Kit (Generation Biotech), as previously described.[19] In brief, we denatured DNA at 95°C for 5 min and incubated it at 64°C for 20 min to anneal primers and incorporate biotinylated dNTPs via the polymerase. Targeted DNA fragments were purified with streptavidin-coated magnetic beads (Generation Biotech) at 64°C for 30 min, followed by a wash with dH_2O. We eluted captured DNA from beads by incubating the samples at 80°C for 20 min followed by magnetic separation.

Whole-genome amplification and sequencing

We amplified captured DNA with the REPLI-g Midi Kit (QIAGEN) without denaturation by incubating the RSE sample with REPLI-g master mix at 30°C for 4 h followed by polymerase inactivation 65°C for 5 min. DNA capture and amplification efficiencies were assessed by quantitative PCR (qPCR) of targeted loci on the RSE captured material and the amplified sample prior to PacBio sequencing. See Table S2 for primers. Sequencing was performed on the PacBio Sequel System with two SMRT cells per sample.

SNP arrays

To validate variants, we performed whole-genome SNP genotyping on Infinium Omni2.5 SNP arrays (Illumina) by using standard protocol. Raw data were processed by GenomeStudio V2011.1. Genotyping calls with quality score less than 0.2 were ignored. We used theta and R values to determine zygosity and copy number, respectively. All heterozygous calls had theta values between 0.25 and 0.75.

PAClindrome bioinformatics pipeline

We developed a custom bioinformatics pipeline to identify palindrome repeats from long PacBio SMRT reads and draw consensus sequences from the repeats for higher sequence accuracy. The pipeline was deployed on a computer cluster for parallel processing of millions of SMRT reads through the following steps. For each SMRT read 20 kb or longer, the full read was trimmed into successive 400-base segments with 300-bases overlapping between the two closest segments. All segments were aligned back to the full reads to search for matches other than themselves. One sub-region within the full read matched by an unbroken set of segments and having the highest total alignment score was selected as the seed. The seed sequence was aligned again to the full read to identify its matches as palindrome repeats. Palindromes were then aligned to each other via multiple sequence alignment (MSA). We obtained the consensus sequences from an MSA matrix by using the simple majority rule at each base. The weighted majority rule can be used as an alternative, but it did not significantly improve the accuracy of consensus sequences in this dataset.

Table 1. Subject, race, and *RHD* and *RHCE* genotypes with presumed haplotypes (noted as allele 1 and allele 2) performed by conventional methods and the red cell Rh phenotypes and presumed haplotypes determined by serologic typing

Subject	Race	*RHD* allele 1	*RHCE* allele 1	*RHD* allele 2	*RHCE* allele 2	Serologic Rh phenotype/ haplotypes
UPID 10	Black	*RHD Ψ (RHD*08N.01)*	**ce48C (RHCE*01.01)*	*deleted D (RHD*01N.01)*	**ce254G (RHCE*01.06.01)*	D-C-c+E-e+ ce/ce
UPID 70	Black	**DIIIa-CEVS (4-7)-D (RHD*03N.01)*	**ce^{S} (RHCE*01.20.03)*	*deleted D (RHD*01N.01)*	**ce48C (RHCE*01.01)*	D-C+c+E-e+ Ce/ce
UPID 83	Black	**weak partial D 4.0 (RHD*09.03)*	**ce48C (RHCE*01.01)*	**DAU0 (RHD*10.00)*	**ce48C (RHCE*01.01)*	D+C-c+E-e+ Dce/ce or Dce/Dce
UPID 164	Black	**DAU0 (RHD*10.00)*	**ce (RHCE*01)*	*DAU3 (RHD*10.03)*	**ce48C (RHCE*01.01)*	D+C-c+E-e+ Dce/ce or Dce/Dce
UPID 19	Black	*RHD (RHD*01)*	**ce (RHCE*01)*	*RHD (RHD*01)*	**ce (RHCE*01)*	D+C-c+E-e+ Dce/ce or Dce/Dce
UPID 30	Black	**DAU0 (RHD*10.00)*	**ce48C (RHCE*01.01)*	*RHD (RHD*01)*	**Ce (RHCE*02)*	D+C+c.E-e+ DCe/Dce
UPID 333	Black	**DAU0 (RHD*10.00)*	**ce48C (RHCE*01.01)*	*RHD (RHD*01)*	**ce733G (RHCE*01.20.01)*	D+C-c+E-e+ Dce/ce or Dce/Dce
UPID 18	Black	*RHD (RHD*01)*	**cE (RHCE*03)*	*RHD (RHD*01)*	**ce (RHCE*01)*	D+C-c+E+e+ Dce/DcE
UPID 3	Black	*deleted D (RHD*01N.01)*	**ce (RHCE*01)*	*deleted D (RHD*01N.01)*	**ce48C (RHCE*01.01)*	D-C-c+E-e+ ce/ce
UID 1	White	*RHD (RHD*01)*	**Ce (RHCE*02)*	*RHD (RHD*01)*	**Ce (RHCE*02)*	D+C+c-E-e+ DCe/DCe
UID 2	White	*RHD (RHD*01)*	**cE (RHCE*03)*	*deleted D (RHD*01N.01)*	**ce (RHCE*01)*	D+C-c+E+e+ DcE/ce

UID, unique identifier; UPID, unique patient identifier. (**ce^{s}* = ce48C,733G, 1006T).

Linkage analysis and phasing

We performed phasing of the *RHD-RHCE* region by sequentially establishing linkage between two adjacent assembly markers. In a specific sample, both markers had to be heterozygous with the frequency of either allele higher than 15% and located within the same 20 or more consensus sequences. We also used SMRT reads in the raw data not containing any palindrome repeats to confirm linkage between markers, especially those distant from one another. All alleles were named according to the International Society of Blood Transfusion (ISBT) working group that develops and maintains guidelines for blood group antigen and allele nomenclature for use in transfusion medicine and related sciences.[21]

Statistical methods

Statistical analysis was performed with R packages. We estimated sequencing error rate by comparing palindromes of the same SMRT read to each other on the basis of the assumption that they were clones of the same original DNA fragment. Phasing of two neighboring heterozygous markers used consensus sequences including both markers and significance of their linkage was tested by Fisher's exact test ($p < 10e{-}5$). The phasing continued until no more markers with significant linkage were available.

Data and code availability, bioinformatics software

We used BLASR (The PacBio long-read aligner, see web resources) to align SMRT reads to the reference genome (GRCh38) or to themselves through the "PAClindrome" pipeline. MSA of palindrome repeats was done with the MUSCLE algorithm.[22] Alignment of raw SMRT reads and consensus sequences was visualized by Integrative Genomics Viewer (IGV). All other data processing and analysis steps were performed with custom R code and existing R/Bioconductor packages, including Biostrings and GenomicAlignment. Raw sequencing data are archived at the SRA database (Bioproject ID PRJNA775954). Documentation and custom code of bioinformatics analysis are available at GitHub (see web resources).

Results

Study samples

The 11 samples chosen were from two White individuals and nine Black individuals with SCD selected for diverse *RH* alleles previously determined by *RHD* and *RHCE* genotyping (Table 1). We selected individuals with structural variations including *RHD* deletion and exon translocation as well as *RH* variants characterized by one or more SNPs in the coding region and samples for which phasing of changes was uncertain. The two White individuals were included as test samples with less genetic variation expected compared to Black individuals.

Examination of the RH loci

Reference sequences (GRCh38) demonstrate the high sequence similarity between *RHD* (GenBank: NM_016124) and *RHCE* (GenBank: NM_020485) genes (Figure 1). There are long stretches of exact identity (Table S3A) that confound sequence assignment and contig assembly. Notably, intron 9 of *RHCE* starts with a *RHCE*-specific sequence ~1 kb long and ends with a 4,289-base region that is identical between the genes, which causes short NGS reads to map and be assigned to either gene. Additional sequence complexities within the *RH* locus include duplicated transposable DNA elements and a pair of identical 128-base sequences within *RHD* intron 2. Long stretches of identical sequences complicate the assembly of the *RHD-RHCE* region because they not only increase the ambiguity of read mapping but also facilitate structural rearrangement. In contrast, gene-specific insertion-deletions (INDELs) (Table S3B) aid assignment. Both genes

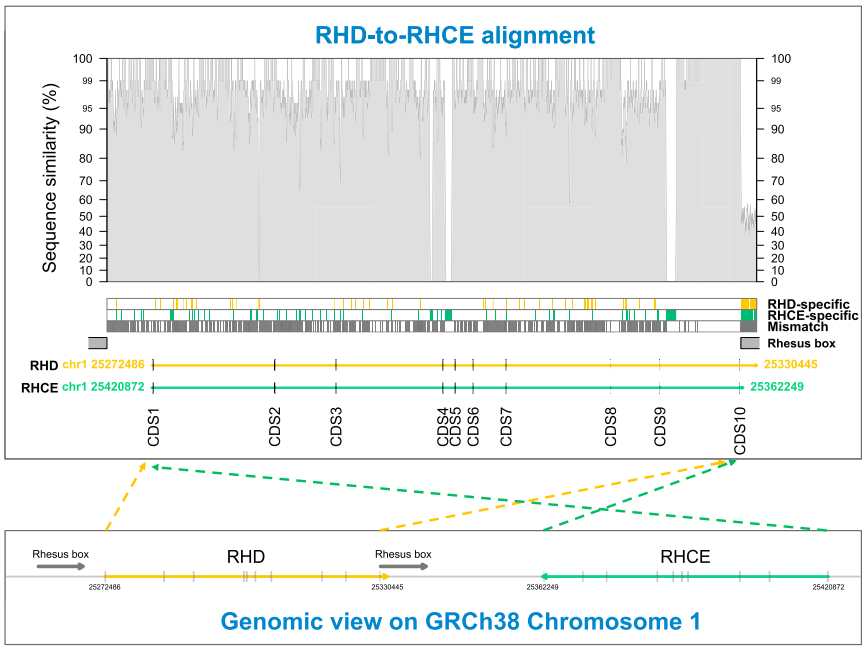

Figure 1. Chromosome location and structure of *RHD* and *RHCE* and their alignment to one other based on the GRCh38 reference genome

Duplication of RHCE inserted *RHD* between two Rhesus boxes (bottom), which includes 62,420 total bases: 60,341 bases are matches between two genes, 1,630 bases are mismatches, and 449 bases are unique to *RHD* (top). There are also 2,351 bases unique to *RHCE*, including long *RHCE*-specific regions in introns 3, 4, and 9 (green blocks). Bold vertical lines along the gene represent coding sequences (CDSs). Sequencing similarity on y axis was calculated on the basis of the percentage of matched bases within each 100-base window.

contain a variety of low-complexity sequences, such as mononucleotides up to 35 bases long and tandem repeats up to 44 bases long (Table S3C). Low-complexity sequences are hotspots for sequencing errors but are potentially useful as assembly markers because of their higher frequency of INDELs.

Determination of sequence accuracy

Sequencing libraries from the 11 samples each generated 0.92 to 1.38 million PacBio SMRT reads, 40.7% to 69.4% of which contained at least one pair of palindromic subreads (Table S4). The palindromes did not completely overlap, and some were mismatched with each other by greater than 10% because of sequencing errors. There were often gaps of irregular sizes within and between the palindromes, and the repeats sometimes had a nested structure caused by rounds of multiple displacement amplification (MDA), which necessitated the development of a custom analysis pipeline (Figure S1).[23–26] The number of palindromes found in each SMRT read varied from two to over 200. We developed the "PAClindrome" bioinformatics pipeline to identify palindrome repeats in full SMRT reads and then draw consensus sequences from the repeats. The pipeline determined the boundaries of palindrome repeats in full reads and drew their consensus sequence after aligning the repeats to each other. We tested and confirmed that sequencing errors were mostly random in palindromes except at mononucleotide sites.

Visual inspection of read-to-genome alignment suggested that palindrome consensus sequences had much higher accuracy than the original SMRT reads (Figure S3). We performed *in silico* analysis to determine the accuracy of palindrome consensus sequences compared to the raw reads. We started with a random subset of SMRT reads containing at least 60 palindromes. Alignment of these raw reads to each other showed an average error rate > 10% in all subjects. Next, a permutation procedure randomly selected two subsets of non-overlapping palindromes from the same SMRT read. As the number of palindromic repeats used for the consensus increased from three to 30 the average error rate dropped (Figure 2). When consensus sequences were drawn from six palindromes, the error rate was ~1% and plateaued at ~0.25% when over 20 palindromes were used.

The 11 samples each obtained 124,724 to 260,861 high-quality consensus sequences with 2.1 to 2.9kb average length and 98.74% to 99.13% average accuracy (Table S4). The consensus sequences were enriched at the *RHD-RHCE* region by our RSE protocol and had average sequencing depth from 86 to 248 (Figure S4). UPID 3 had much lower coverage across the whole *RHD* locus because of homozygous *RHD* deletion. Nevertheless, some consensus sequences of UPID 3 were mistakenly aligned to *RHD*, most commonly near the end of intron 9 where *RHD* and *RHCE* have a long stretch of identical sequences. Such misalignment reflected the complexity of assembling this region and was potentially caused by a combination of several factors including translocation between the two genes. A commonly observed example of misalignment was the low sequencing coverage of UPID 70 at the middle of *RHD* because of translocation (exon 4–7) from *RHCE* into *RHD in trans* to deleted *RHD*.

Assembly of RHD and RHCE

Heterozygous SNPs were identified from the high-quality long consensus sequences as assembly markers. Small INDELs were not used as markers because they had higher error rates in both the original PacBio SMRT reads and the palindrome consensus sequences. A total of 771 unique markers were found in all 11 samples, and 366 (47.5%) markers were heterozygous in more than one subject. The two samples from White individuals (UIDs) had only seven and 15 markers, in contrast to the nine samples from Black individuals (UPIDs), which had between 55 and 295 assembly markers (mean = 163.4). The 15 most

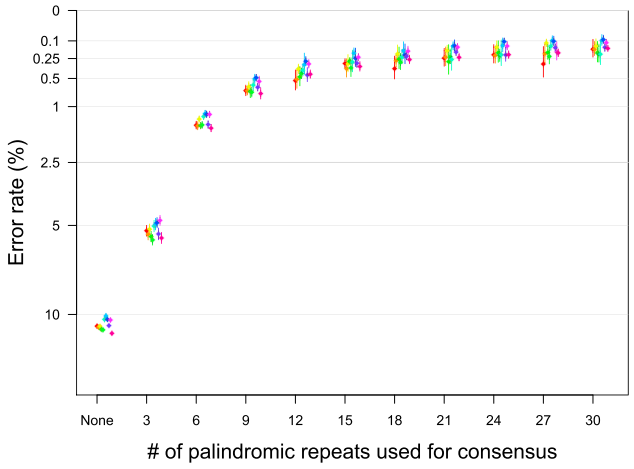

Figure 2. Improving sequencing accuracy as more palindromes were used to draw consensus sequences
A permutation procedure randomly selected two subsets of non-overlapping palindromes from the same SMRT read. Two consensus sequences were drawn independently from both subsets and aligned to each other to obtain their error rate as the percentage of unmatched. Colors represent 11 subjects in this study, which has almost identical trajectory of error rate changes. The error rate in the original palindrome sequences had an average greater than 10%, consistently declined as the number of palindromes increased, and plateaued between 0.1 and 0.25%.

common markers were heterozygous in six samples and were either intronic (*RHCE* intron 1–3) or intergenic SNPs located near the *RHCE* 5′-UTR (Table S5).

The distance between adjacent markers ranged from 1 to 1,918 bases (mean = 220.3) and 13 pairs of adjacent markers were greater than 1 kb away from each other (Figure S5A). Marker frequency was relatively lower in *RHD* introns 1, 2, and 8 and *RHCE* intron 6. The sequencing depth at heterozygous assembly markers ranged from 16 to 646 (mean = 157.1) (Figure S5B). There was no significant correlation between average sequencing depth and the sensitivity of calling heterozygous variants across the 11 subjects of this study, suggesting sequencing depth was not a limiting factor to assembly in this study.

Among all assembly markers, 759 (98.4%) are known SNPs recorded in the dbSNP database. Their global frequencies of alternative alleles ranged from 0.004% to 99.46% (mean = 14.6%) according to the 1000 Genomes database and gnomAD. The vast majority of the assembly markers were intronic (64.5%) or intergenic (33.6%). Among the 15 exonic markers, 11 were missense, including known variants *RHD* c.1136C>T (p.Thr379Met), *RHCE* c.48G>C (p.Trp16Cys), *RHCE* c.676G>C (p.Ala226Pro), and *RHCE* c.733C>G (p.Leu245Val).

To validate these assembly markers, we characterized seven of the 11 subjects with Infinium Omni2.5 SNP arrays (Illumina). Among approximately 2.3 million SNPs on the microarray, 46 are located within the *RHD-RHCE* region and 22 are within the 771 assembly markers (Table S6). PacBio and SNP array had 100% agreement on the 154 genotyping calls (Figure S6).

Adjacent heterozygous markers were linked to construct assembly contigs (Figure 3). The linkage between two markers was established by multiple palindrome consensuses consistently containing high-quality allele calls at both markers. The assembly contigs were sequentially phased by adding one linked marker at a time until there was a gap too large to be bridged by any two markers. The longest distance between two adjacent markers that could be directly linked was about 9.5 kb. Assembly gap length ranged from 5.7 to 40.1 kb (mean = 14.5 kb). The UID 1 sample, homozygous for *RHD* and *RHCE* (DCe/DCe), was highly homogeneous across the *RHD-RHCE* region and had the two largest assembly gaps. When the distances were 5 kb or greater, the likelihood of directly linking two adjacent markers was 22 of 36 (61.1%).

The whole *RHD-RHCE* region was successfully assembled for four samples from Black individuals (UPIDs 10, 70, 83, 164, Table 2, Figure 3). Additionally, the *RHD* alleles of one sample without *RHD* deletion (UPID 19) and the *RHCE* alleles of three samples (UPIDs 19, 30, 333) were fully assembled. All samples for which full assembly was possible had a high frequency of heterozygous markers. Successful assembly relied on both the total number of markers and their distribution. Although UPID 333 had the highest numbers of markers, the 5′ region of *RHD* could not be fully assembled because of a lack of heterozygous markers. However, it was possible to phase the 3′ end of *RHD* to the *RHCE* loci, showing the c.1136T SNP marker to be in *cis* with c.48C and the conventional c.1136C to be in *cis* with c.773G SNP. Thus, we confirmed the presumed haplotypes for UPID 333 as *RHD*RHD* in *cis* with *RHCE*ce733G* and *RHD*DAU0* in *cis* with *RHCE*ce48C*.

Comparison to current RH genotyping methods

We identified all clinically relevant SNPs as determined previously by current *RH* genotype assays for all samples (Figure 3). Of the four samples that achieved full assembly of the *RHD* to *RHCE* region, UPIDs 10 and 164 were determined to have a different *RH* haplotype linkage than presumed on the basis of frequency. *RH* genotyping for UPID 10 presumed *RHD*deleted D* in *cis* with *RHCE*ce254G* and *RHD*RHDΨ* in *cis* with *RHCE*ce48C*. PacBio sequencing and *RH* assembly identified all the SNPs associated with these alleles: the 37 base pair (bp) insertion in intron 3 and the five SNPs associated with *RHD*RHDΨ*,[27] as well as 254C>G and 48G>C in *RHCE* (Figure 3). In addition, we identified a 5.6 kb translocation of *RHD* into *RHCE* (from intron 8 to intron 9 and encompassing exon 9) that was not detected when using current *RH* exon-specific sequencing. Full assembly of both *RHD* to *RHCE* regions determined that this allele, *RHCE*ce254G-D(9)-ce*, was novel and found in *cis* with *RHD*Ψ* (Figure 3).

The *RH* genotype for UPID 164 presumed *RHD*DAU0* in *cis* with *RHCE*ce* and *RHD*DAU3* in *cis* with *RHCE*ce48C*. PacBio sequencing followed by *RH* assembly instead demonstrated conventional *RHD* in *cis* with *RHCE*ce* (and *RHD*DAU3* in *cis* with *RHCE*ce48C*). By phasing

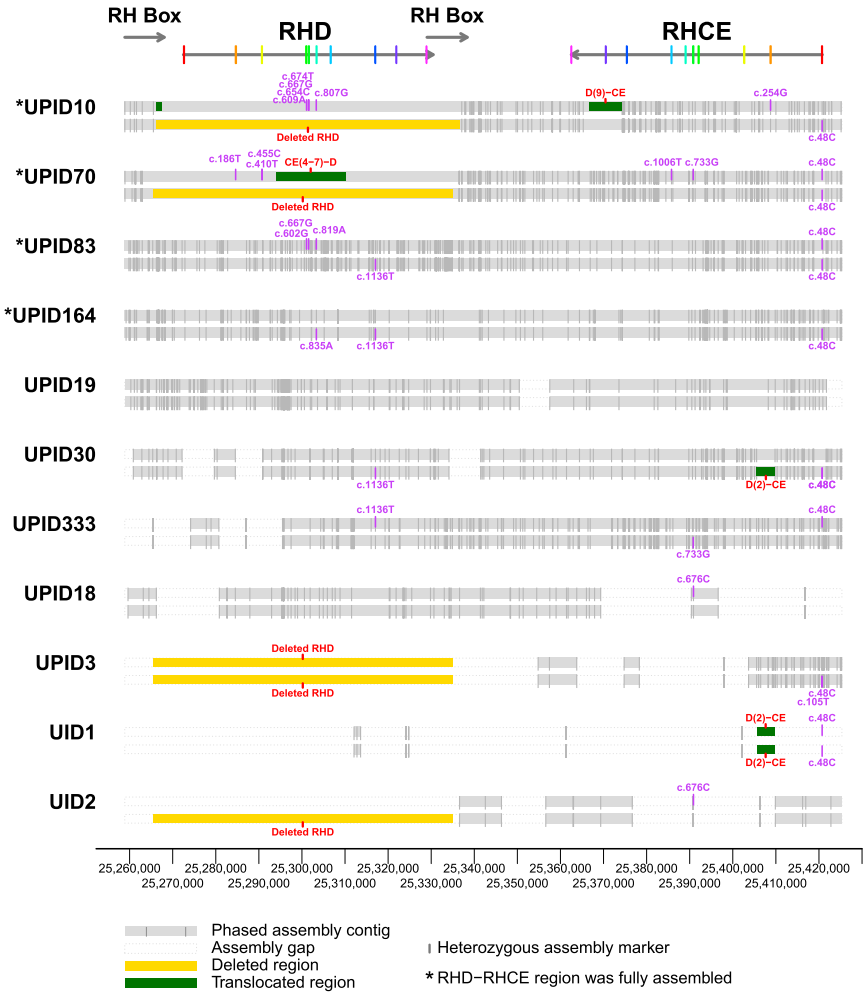

Figure 3. Assembly of the *RHD-RHCE* region in 11 individuals
Gray boxes represent assembly contigs and vertical gray lines represent heterozygous assembly markers. White boxes represent assembly gaps. Markers within the same contig were linked, while two distal markers with any assembly gap between them are not linked. Exonic markers are labeled in purple; all result in amino acid except c.105T in UPID 3. RHD deletion and translocation events are illustrated as yellow and green boxes, respectively.

the entire *RHD* to *RHCE* regions, we found two SNPs in *RHD* intron 7 (genomic position 25316671 and 25316673) ~400 bp upstream of the c.1136C>T change (genomic position 25317062) interrogated for DAU alleles (Figure S7). This region corresponds to the location of *RHD*-specific primers used for PCR-RFLP to detect the 1136C/T SNP; we determined that the presence of these two SNPs that resemble *RHCE* rather than *RHD* caused failure of the *RHD*-specific 5′ PCR primer to anneal and amplify, resulting in allele drop-out of the conventional *RHD* in this sample.

Linkage between variants

We identified 14 sets of assembly markers that are linked to each other in multiple subjects (Table S7). These sets included six to 15 heterozygous markers that were 6.7–35.7 kb in total length and phased together in two to six subjects. Markers in these sets were distributed across the whole *RHD-RHCE* region, relatively enriched at the first half of *RHCE*, and identified within *RHD* or *RHCE* or in nearby intergenic region. Those in overlapping sets had no linkage to each other. Exonic variants were only found in three sets: *RHD* c.1136C>CT (genomic position 25317062), *RHCE* c.48G>C (genomic position 25420739), and *RHCE* c.733C>G (genomic position 25390817).

To validate linkage, we compared minor allele frequencies (MAFs) of variants in different populations, assuming that alleles linked in a population should have similar frequencies. African, European, and global MAFs obtained from whole-genome sequencing data were retrieved from the 1000 Genomes database and gnomAD (Figure 4 and Figure S8). In Africans, variants within the same marker set have much more similar MAFs to each other than variants not in the same set, regardless of the distance between variants. In Europeans, while MAFs are much lower in general, the alternative allele frequencies are consistent in 13 of the 14 sets. The only exception is the set of markers that includes *RHD* c.1136T, which has much lower frequency in Europeans than the other variants in the same set. However, the frequency of c.1136T in Europeans may be an alignment artifact in the databases and has not been reproduced by Sanger sequencing. The actual frequency of c.1136T in Europeans may be similar to the other linked MAFs. On the other hand, *RHCE* c.733C>G and its nearby intronic variants maintained their linkage in Europeans, although the MAFs are very different between the two populations.

RHCE c.48G>C (rs586178) has been associated to four nearby intronic variants in almost all genome-wide association study (GWAS) populations (Table 3 and Figure S9). The SNP c.48G>C itself and the four variants globally associated to it are all members of a seven-marker set, which included another two markers that were not measured by GWASs. Similarly, rs3091242, an intergenic SNP located between *RHD* and *RHCE*, has been associated to five nearby variants in all GWAS populations and they are the exact same variants in a six-marker set (Table S8).

Structural variations

Major structural variations including *RHD* deletion and exon translocation were identified in multiple subjects

Table 2. Summary of RHD-RHCE assembly

ID	Contig	Marker	Gap length (min–max bp)	Completed (kb)	Completed % (whole region)	Completed % (*RHD*)	Completed % (*RHCE*)
UPID 10	1	152	0	166.4	100.0	100.0[a]	100.0
UPID 70	1	178	0	166.4	100.0	100.0[a]	100.0
UPID 83	1	295	0	166.4	100.0	100.0	100.0
UPID 164	1	196	0	166.4	100.0	100.0	100.0
UPID 19	2	171	7076	155.8	93.6	100.0	100.0
UPID 30	4	154	6,363–7,492	143.2	86.1	76.6	100.0
UPID 333	4	202	6,399–8,793	136.4	82.0	71.6	100.0
UPID 18	4	68	14,624–21,037	101.3	60.9	85.6	22.7
UPID 3	4	55	5,653–19,910	34.2	62.2	100.0[a]	38.0
UID 2	5	15	7,655–19,003	45.3	68.9	100.0[a]	43.4
UID 1	6	7	11,152–40,954	1.6	0.9	2.7	0.0

The numbers of assembly contigs and markers, the range of assembly gaps, the total length of assembled region, and the percentage of the whole region and individual genes that was successfully assembled per sample.
[a]Hemizygous or homozygous *RHD* deletion.

through the assembly of the *RHD-RHCE* region. Long and accurate consensus sequences of palindromes allowed for successful phasing of these translocated regions and identification of their breakpoints down to each nucleotide base pair. One such structural variation is the deletion of the *RHD* gene, which has been shown to occur between upstream and downstream Rhesus boxes that share ~97% sequence similarity.[10,28] With *RHD* deletion, the two Rhesus boxes fuse to generate a hybrid box. Using long reads to fully phase the hybrid boxes, we confirmed the conventional fusion site of the two Rhesus boxes to an 889 bp region within the 1,463 bp region of identical sequence located in the second half of the hybrid box in three samples (Figure 5A). One *RHD* deletion event found in UPID 10 identified a novel alternative fusion site ~100 bases downstream of the conventional site, located within a 34 bp window that is identical between the two boxes and starts with a 12-base palindrome sequence (chr1: 25,266,232- 25,266,266).

Translocation of exon 2 from *RHD* to *RHCE* along with a 109 bp insertion in intron 2 contributes to the genetic basis for RhC expression and was found in two samples as expected (one homozygous in UID 1 and 1 heterozygous in UPID 30).[29] All three *RHCE*Ce* alleles had the same 5′ and 3′ translocation breakpoints, encompassing a 4.1 kb translocation (Figure 5B). The 5′ translocation breakpoint was narrowed down to a 109 bp region of high sequence similarity between *RHD* and *RHCE*, located ~1 kb upstream of exon 2. We identified a SNP within the 5′ translocation breakpoint (g.25409808) of all three *RHCE*Ce* alleles that did not resemble conventional *RHD* or *RHCE* (Figure 5B, g.25409808, highlighted green) but one that is commonly found in both *RHD* (g.25283634C>A, rs114582484, MAF = 0.26) and *RHCE* (g.25409808A>T, rs28594470, MAF = 0.37), suggesting that this SNP may be associated with this translocation event. The 3′ breakpoint in intron 2 was located between g.25405592 and g.25405593, where the 109 base insertion occurred in all samples. We confirmed that the inserted 109 bases were composed of three parts: bases 1–20 of no known origin; bases 21–85, which were reverse-complementary duplicates of an earlier sequence in *RHD* intron 2 (chr1: 25,287,190–25,287,254); and bases 85–109, which were exact duplicates of bases immediately before the breakpoint in *RHD*.

In UPID 70, full assembly confirmed the translocation of *RHCE* exons 4 to 7 to the *RHD* locus resulting in the hybrid gene *RHD*DIIIa-CEVS(4-7)-D* found in *cis* to *RHCE*ceS* (c.48G>C, c.733C>G, and c.1006G>T) (Figure 3).[30] Upstream of the translocation start site, the *RHD*-coding variants associated with DIIIa exons 2 and 3 (c.186G>T, c.410C>T, and c.455A>C) were identified. Within the translocated region, we identified the *RHCE*ceS* exon 5 (c.733C>G) and exon 7 (c.1006G>T) SNPs, as well as intronic SNPs associated with the *RHCE*ceS* allele (Figure 5C). The 5′ breakpoint for the CE(4-7)-D translocation was located within a 72 bp window of identical sequence between *RHD* and *RHCE* ~3 kb downstream of exon 3 (g.25293908 ~g.25293979). The 3′ breakpoint was located in a 32 bp window that is identical between *RHD* and *RHCE* (g. 25310218 ~g. 25310250), approximately 3.5 kb downstream of exon 7. Several smaller translocation events were nested around each translocation breakpoint, ranging in size from 20 bp to 4,393 bp (Figure S10).

For the 5.6 kb translocation of *RHD* exon 9 into *RHCE*ce* found in UPID 10, the 5′ breakpoint was located within a small 7 bp window (g. 25374191 ~g. 25374197) ~1 kb upstream of exon 9. The 3′ translocation breakpoint was narrowed down to a 1,303 bp region of intron 9, located at the beginning of a 4.3 kb region that is identical between *RHD*

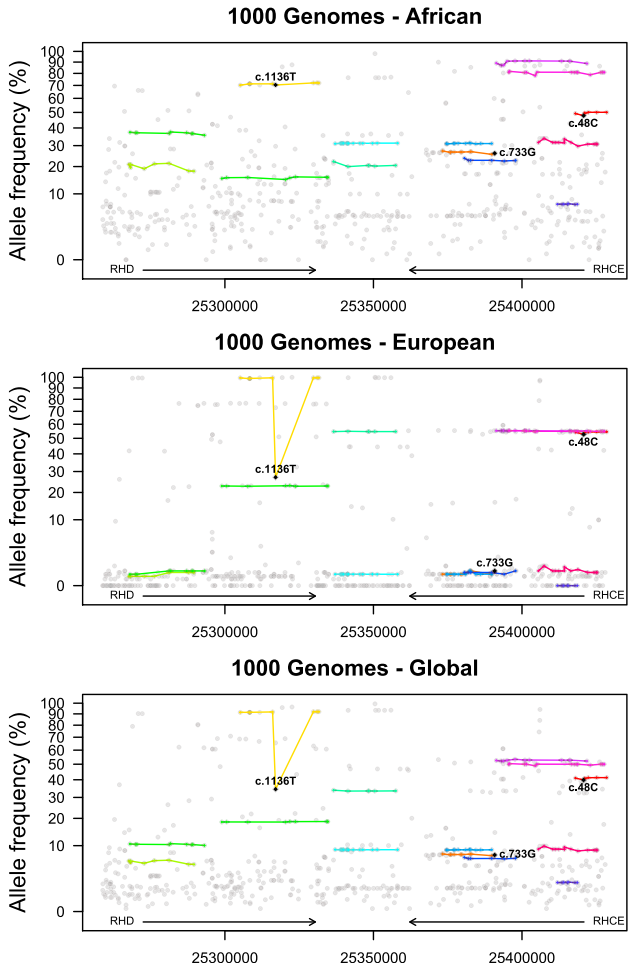

Figure 4. Frequency of alternative alleles according to the 1000 Genomes database

Each colored line represents a set of markers linked to each other in 11 subjects of this study. All other assembly markers within the region were colored in gray. y axis is the known frequency of alternative alleles in one or global population according to whole-genome sequencing data.

and *RHCE* (g. 25369890 ~g. 25368587, Figure 5D). UPID 10 harbored multiple SNPs in this region that resembled neither *RHD* or *RHCE* reference sequences and did not have a clear origin.

Discussion

The success of assembling *RHD-RHCE* and similar genomic regions with high-throughput sequencing technologies has several limiting factors: sequencing depth at the targeted region, accuracy of variant calling, length of sequencing reads, and availability of assembly markers in individual subjects. Using a targeted capture protocol, high-throughput long-read sequencing, and custom PAClindrome bioinformatics pipeline, we successfully assembled the *RHD-RHCE* region with accurate consensus sequences at high depth primarily among Black subjects who have increased *RH* genetic variation. High accuracy reads > 5 kb in length were extremely valuable to assemble regions with repetitive sequences and high frequency of structural variants. The strategy described here addresses the problem of ambiguities in *RH* genotypes and similar genes that exist with conventional genotyping methods, including allele dropout and phasing haplotypes. In samples with sufficient heterozygous SNPs, the phase resolution of *RHD* and *RHCE* alleles is a major advantage over current assays.

Palindrome repeats commonly exist in high-throughput sequencing datasets generated by chimera formation during whole-genome amplification (i.e., multiple displacement amplification [MDA]) (Figure S1).[23,31,32] Of 13.4 million zero-mode waveguide (ZMW) reads among 11 samples, 65.5% had at least two continuous long read (CLR) subreads generated by multiple passes of circular sequencing with an average of 4.4 subreads. In addition, PAClindrome identified palindromes in 55.0% of the reads with an average of 9.7 palindromes per read. In 47.0% of the full reads, the number of palindromes identified by PAClindrome was higher than the number of CLR subreads. Our pipeline used both types of subreads, originated from circular sequencing and MDA, to draw consensus sequences. Using this strategy, we identified 64.9% more full reads that included at least six total palindromes (CLR and MDA), which allowed for high accuracy consensus (Figure 2).

Our analysis of public datasets in Sequence Read Archive identified palindromes in many other PacBio datasets. The palindrome repeats introduce bias into procedures such as variant calling and gene assembly and need to be computationally detected and corrected.[7,32,33] The existing tool Pacasus corrects PacBio sequencing errors by splitting reads into palindromes via recursively aligning each read to itself. Our PAClindrome pipeline not only identifies palindromes through self-alignment but also uses palindromes to generate their own consensus sequences via multiple sequence alignment. These consensus sequences had much improved accuracy of ~99.5%, a substantial improvement from less than 90% average accuracy in the raw data. Our study demonstrated the strength of long and accurate reads in variant calling and phasing of large regions, which are particularly valuable for assembly of *RHD-RHCE* or other regions with long repeated sequences and frequent structural variations.

While mismatches between *RHD* and *RHCE* are natural assembly markers to phase the genes, variants at those loci can cause ambiguity in read-to-gene mapping. For example, there are 1,118 *RHD* and 1,081 *RHCE* known SNPs, but these overlap with the 1,630 base differences between the two genes. Furthermore, the alternative bases of ~80% of the SNPs reported in *RHD* or *RHCE* are due to gene conversion and so are the same as the reference base for the other gene. We found that current average consensus read length of ~2.5 kb cannot guarantee full assembly of this region for subjects with too few variable markers. In general, Africans have higher genetic variation than other populations, leading to an increased likelihood of successful

Table 3. Association of four nearby SNPs to rs586178 (*RHCE* c.48G>C) in multiple populations

RSID	Variant	R2_ACB	R2_ASW	R2_LWK	R2_YRI	R2_CHB	R2_JPT	R2_CEU	R2_FIN
rs1883427	g.25417977C>T	0.9793	0.9062	0.7740	0.9449	0.9545	0.9533	0.9220	0.9208
rs4649083	g.25422304G>A	0.9793	0.9365	0.9604	0.9636	0.9545	0.9533	0.9220	0.9208
rs12402120	g.25425141C>T	0.9793	0.9062	0.9400	0.9636	0.9545	0.9533	0.9220	0.9208
rs932372	g.25428523G>A	0.9793	0.9365	0.9604	0.9636	0.9545	0.9533	0.9220	0.9208

R square values are correlation coefficients calculated by previous GWASs. All four variants and rs586178 itself are among a set of seven assembly marks linked to each other in 11 subjects of these studies and having similar minor allele frequencies across global populations. The other two SNPs, rs2072932 and rs2281179, are too close to rs586178 and were not measured by the GWASs. ACB, African Caribbeans in Barbados; ASW, Americans of African Ancestry in SW USA; LWK, Luhya in Webuye, Kenya; YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; CEU, Utah Residents with Northern and Western European Ancestry; and FIN, Finnish in Finland.

assembly. Nonetheless, the 5′ end of *RHD* remained a challenge for several samples because of a relatively low frequency of SNPs in that region.

Previous studies have identified a few intronic variants associated with certain Rh phenotypes.[34,35] By phasing long-read sequences, we were able to directly phase exonic, intronic, and intergenic variants and further validate their linkage with public resources. Our analysis suggested that haplotypes of SNPs, which could be global or population specific, are common across the *RHD-RHCE* region (Figure 4). Identification and documentation of variant haplotypes not only provides insight into the evolution of *RH* alleles but helps to categorize genetic variations and their association to clinical phenotypes. One method to sequence the complete *RHD* gene with overlapping long-range PCR amplicons has been described, but no similar method for *RHCE* or the entire *RH* locus has been produced.[35]

Long-read sequencing also allowed us to determine more precise translocation breakpoints for structural variants (Figure 5). Unlike short-read NGS technologies that depend on read depth of coverage or copy number variation within translocated regions to determine structural variants and predict breakpoints,[15,18,36,37] long-read sequencing does not rely on copy number variation. As a result, translocation breakpoints can be narrowed down to the individual base as opposed to a range within hundreds of bases as determined by sequence-specific PCR or predicted by various algorithms with short-read NGS assays.[18,29,37–39] One such example is the segment of *RHD* translocated into *RHCE* responsible for C expression. In 1997, Carritt et al. reported a 4.26 kb translocation by using sanger sequencing,[29] similar to the 4.1 kb translocation we identified. However, NGS-based assays predicted larger translocations of 5,216 ± 796 bp and 4,953 ± 238 bp for Africans and Asian/Native American samples, respectively.[36] For the hybrid allele *RHD*DIIIa-CEVS(4-7)-D*, sequence-specific PCR described similar breakpoint regions but NGS-based assays predicted larger translocations.[18,39,40] For both sequence-specific PCR and short-read NGS assays, micro-translocations could confound the breakpoint prediction depending on primer position. We demonstrated long-read sequencing can identify these micro-translocations without compromising the identification of breakpoints for large translocations.

The successful assembly of the entire *RH* region was limited to highly heterozygous samples. A remaining bottleneck to full assembly of *RH*, *HLA*, and similar genes is high accuracy reads long enough to bridge large assembly gaps in all populations. While other sequencing platforms such as Oxford Nanopore Technologies (ONT) long or ultra-long reads have been shown to generate raw reads that are tens of kilobases to megabases in length, the accuracy of the raw reads are on average 87% to 98% and can be as low as 69%.[6,41,42] The average lengths of the consensus sequences are also ~2 kb,[43,44] though the accuracy of the consensus sequences from ONT is 97% to 98%.[45,46] One goal of future studies is to develop new strategies to incorporate DNA enrichment protocols with more advanced sequencing technologies, such as next-generation PacBio HiFi sequencing.

While short-read NGS technologies have led to a dramatic reduction in sequencing cost, more advanced sequencing technology and increased average read length is required for the comprehensive assembly of duplicated gene families such as *RH*. Long-read technologies can enable improved resolution for more distant nucleotide variations, which can be crucial for diagnostic purposes, especially for compound heterozygosity, which is a commonly observed phenomenon underlying recessive Mendelian disorders. Combined with customized bioinformatic tools such as PAClindrome, improved long-read NGS technology may allow individual fully phased *de novo* assembled genomes in the near future.

Data code and availability

Documentation and custom code of bioinformatics analysis are available at GitHub (https://github.com/zhezhangsh/PAClindrome). Raw sequencing data are archived at the SRA database (Bioproject ID PRJNA775954).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.12.003.

Acknowledgments

This work was supported by the National Heart, Lung, and Blood Institute R01 HL147879-01 (S.T.C. and C.M.W.) and U01

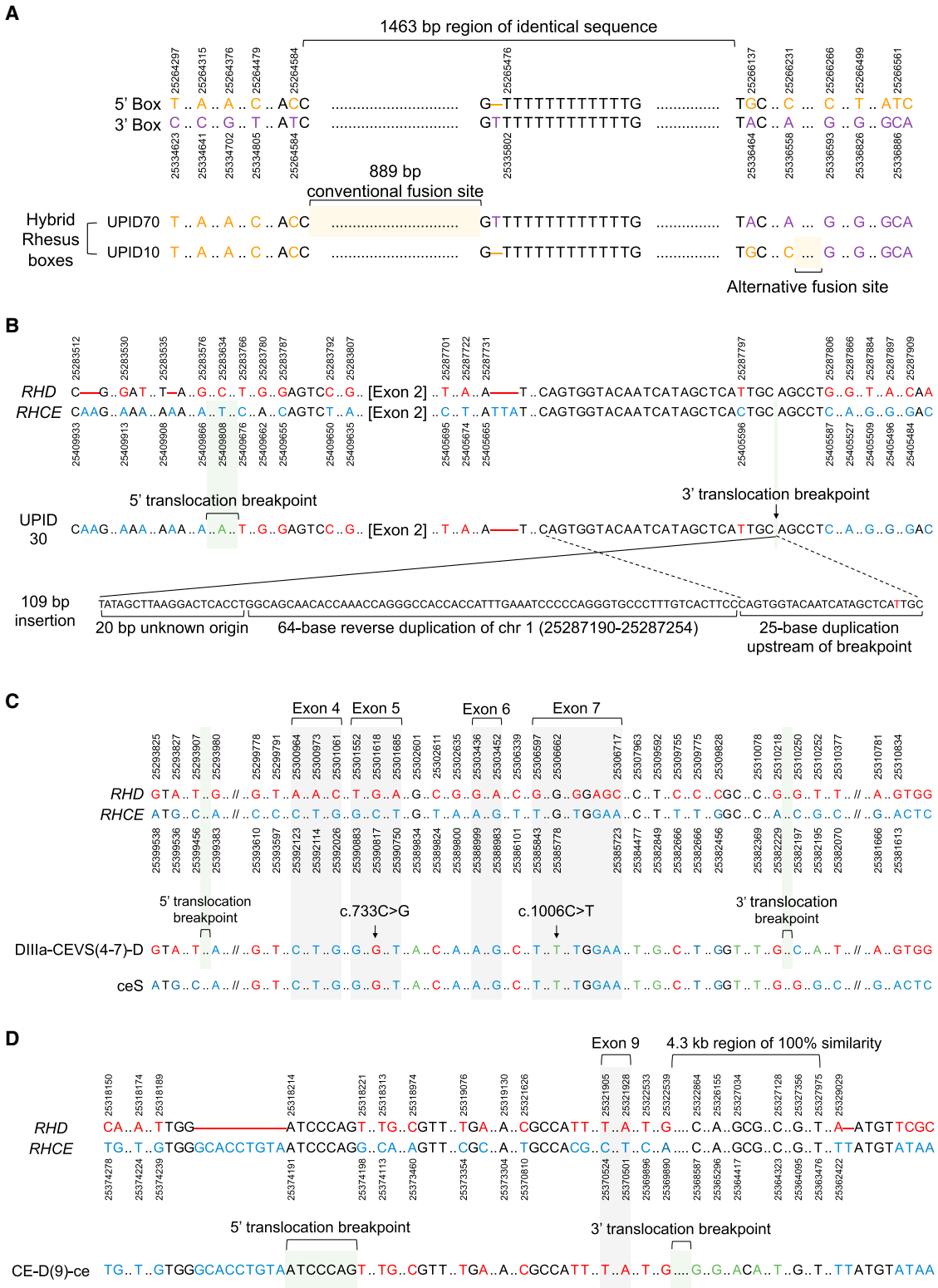

Figure 5. Fusion sites and breakpoints of known major structural *RH* variations
(A) *RHD* deletion caused by Rhesus box fusion. Nucleotides specific to 5′ or 3′ Rhesus boxes are shown in orange and purple type, respectively. Conventional and alternative fusion sites are shaded yellow for each hybrid box found in samples here.
(B–D) Translocation of *RHD* exon 2 to *RHCE* with a 109-base insertion (B), translocation of *RHCE* exon 4–7 into *RHD* (C), and translocation of *RHD* exon 9 to *RHCE* (D). *RHD* and *RHCE* reference sequence and genomic positions on GRCh38 are shown at the top of each panel, and the sequences of study subjects are shown below. Nucleotides shared by both *RHD* and *RHCE* are shown in black type, nucleotides that do not resemble either *RHD* or *RHCE* are shown in green type, and nucleotides specific to *RHD* or *RHCE* are shown in red and blue type, respectively. Regions containing 5′ and 3′ translocation breakpoints are shaded green. Exonic SNPs are shaded gray.

HL134696 (S.T.C. and C.M.W.), a Catalent grant (S.T.C.), and a generous donation from the DiGaetano family (S.T.C.). We would like to thank members of the University of Delaware DNA Sequencing and Genotyping Center for their assistance.

Declaration of interests

The authors declare no competing interests.

Received: July 26, 2021
Accepted: December 7, 2021
Published: December 29, 2021

Web resources

AnthOligo, https://pubmed.ncbi.nlm.nih.gov/32484858/
BLASR (The PacBio long-read aligner), https://github.com/PacificBiosciences/blasr

References

1. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.
2. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front. Genet. *10*, 426.
3. Delaneau, O., Howie, B., Cox, A.J., Zagury, J.F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. Am. J. Hum. Genet. *93*, 687–696.
4. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet. *53*, 779–786.
5. Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics *13*, 278–289.
6. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.
7. Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
8. Westhoff, C.M. (2019). Blood group genotyping. Blood *133*, 1814–1820.
9. Peyrard, T., and Wagner, F. (2020). The Rh System. In AABB Technical Manual (Bethesda, MD: AABB).
10. Wagner, F.F., and Flegel, W.A. (2000). RHD gene deletion occurred in the Rhesus box. Blood *95*, 3662–3668.
11. Colin, Y., Chérif-Zahar, B., Le Van Kim, C., Raynal, V., Van Huffel, V., and Cartron, J.P. (1991). Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis. Blood *78*, 2747–2752.
12. Chou, S.T., Evans, P., Vege, S., Coleman, S.L., Friedman, D.F., Keller, M., and Westhoff, C.M. (2018). *RH* genotype matching for transfusion support in sickle cell disease. Blood *132*, 1198–1207.
13. Chou, S.T., Jackson, T., Vege, S., Smith-Whitley, K., Friedman, D.F., and Westhoff, C.M. (2013). High prevalence of red blood cell alloimmunization in sickle cell disease despite transfusion from Rh-matched minority donors. Blood *122*, 1062–1071.
14. Sippert, E., Fujita, C.R., Machado, D., Guelsin, G., Gaspardi, A.C., Pellegrino, J., Jr., Gilli, S., Saad, S.S., and Castilho, L. (2015). Variant RH alleles and Rh immunisation in patients with sickle cell disease. Blood Transfus. *13*, 72–77.
15. Lane, W.J., Westhoff, C.M., Gleadall, N.S., Aguad, M., Smeland-Wagman, R., Vege, S., Simmons, D.P., Mah, H.H., Lebo, M.S., Walter, K., et al. (2018). Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. Lancet Haematol. *5*, e241–e251.
16. Lane, W.J., Westhoff, C.M., Uy, J.M., Aguad, M., Smeland-Wagman, R., Kaufman, R.M., Rehm, H.L., Green, R.C., Silberstein, L.E.; and MedSeq Project (2016). Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. Transfusion *56*, 743–754.
17. Chou, S.T., Flanagan, J.M., Vege, S., Luban, N.L.C., Brown, R.C., Ware, R.E., and Westhoff, C.M. (2017). Whole-exome sequencing for *RH* genotyping and alloimmunization risk in children with sickle cell anemia. Blood Adv. *1*, 1414–1422.
18. Chang, T.C., Haupfear, K.M., Yu, J., Rampersaud, E., Sheehan, V.A., Flanagan, J.M., Hankins, J.S., Weiss, M.J., Wu, G., Vege, S., et al. (2020). A novel algorithm comprehensively characterizes human RH genes using whole-genome sequencing data. Blood Adv. *4*, 4347–4357.
19. Dapprich, J., Ferriola, D., Mackiewicz, K., Clark, P.M., Rappaport, E., D'Arcy, M., Sasson, A., Gai, X., Schug, J., Kaestner, K.H., and Monos, D. (2016). The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. BMC Genomics *17*, 486.
20. Jayaraman, P., Mosbruger, T., Hu, T., Tairis, N.G., Wu, C., Clark, P.M., D'Arcy, M., Ferriola, D., Mackiewicz, K., Gai, X., et al. (2020). AnthOligo: automating the design of oligonucleotides for capture/enrichment technologies. Bioinformatics *36*, 4353–4356.
21. ISBT. Red Cell Immunogenetics and Blood Group Terminology. https://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology.
22. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.
23. Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. *7*, 19.
24. Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA *99*, 5261–5266.
25. Tu, J., Guo, J., Li, J., Gao, S., Yao, B., and Lu, Z. (2015). Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. PLoS ONE *10*, e0139857.
26. Tu, J., Lu, N., Duan, M., Huang, M., Chen, L., Li, J., Guo, J., and Lu, Z. (2017). Hotspot Selective Preference of the Chimeric Sequences Formed in Multiple Displacement Amplification. Int. J. Mol. Sci. *18*, 492.
27. Singleton, B.K., Green, C.A., Avent, N.D., Martin, P.G., Smart, E., Daka, A., Narter-Olaga, E.G., Hawthorne, L.M., and

Daniels, G. (2000). The presence of an RHD pseudogene containing a 37 base pair duplication and a nonsense mutation in africans with the Rh D-negative blood group phenotype. Blood *95*, 12–18.
28. Wagner, F.F., Moulds, J.M., and Flegel, W.A. (2005). Genetic mechanisms of Rhesus box variation. Transfusion *45*, 338–344.
29. Carritt, B., Kemp, T.J., and Poulter, M. (1997). Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. Hum. Mol. Genet. *6*, 843–850.
30. Faas, B.H., Beckers, E.A., Wildoer, P., Ligthart, P.C., Overbeeke, M.A., Zondervan, H.A., von dem Borne, A.E., and van der Schoot, C.E. (1997). Molecular background of VS and weak C expression in blacks. Transfusion *37*, 38–44.
31. Sabina, J., and Leamon, J.H. (2015). Bias in Whole Genome Amplification: Causes and Considerations. Methods Mol. Biol. *1347*, 15–41.
32. Warris, S., Schijlen, E., van de Geest, H., Vegesna, R., Hesselink, T., Te Lintel Hekkert, B., Sanchez Perez, G., Medvedev, P., Makova, K.D., and de Ridder, D. (2018). Correcting palindromes in long reads after whole-genome amplification. BMC Genomics *19*, 798.
33. Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. *27*, 677–685.
34. Wagner, F.F., Frohmajer, A., and Flegel, W.A. (2001). RHD positive haplotypes in D negative Europeans. BMC Genet. *2*, 10.
35. Tounsi, W.A., Madgett, T.E., and Avent, N.D. (2018). Complete *RHD* next-generation sequencing: establishment of reference *RHD* alleles. Blood Adv. *2*, 2713–2723.
36. Wheeler, M.M., Lannert, K.W., Huston, H., Fletcher, S.N., Harris, S., Teramura, G., Maki, H.J., Frazar, C., Underwood, J.G., Shaffer, T., et al. (2018). Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. Genet. Med. *21*, 477–486.
37. Halls, J.B.L., Vege, S., Simmons, D.P., Aeschlimann, J., Bujiriri, B., Mah, H.H., Lebo, M.S., Vijay Kumar, P.K., Westhoff, C.M., and Lane, W.J. (2020). Overcoming the challenges of interpreting complex and uncommon RH alleles from whole genomes. Vox Sang. *115*, 790–801.
38. Silvy, M., Tournamille, C., Babinet, J., Pakdaman, S., Cohen, S., Chiaroni, J., Galactéros, F., Bierling, P., Bailly, P., and Noizat-Pirenne, F. (2014). Red blood cell immunization in sickle cell disease: evidence of a large responder group and a low rate of anti-Rh linked to partial Rh phenotype. Haematologica *99*, e115–e117.
39. Flegel, W.A., and Wagner, F.F. (2014). Two molecular polymorphisms to detect the (C)ce(s) type 1 haplotype. Blood Transfus. *12*, 136–137.
40. Silvy, M., Granier, T., Beley, S., Chiaroni, J., and Bailly, P. (2013). Identification of novel polymorphism restricted to the (C)ces type 1 haplotype avoids risk of transfusion deadlock in SCD patients. Br. J. Haematol. *160*, 863–867.
41. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.
42. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet. *21*, 597–614.
43. Li, C., Chng, K.R., Boey, E.J., Ng, A.H., Wilm, A., and Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. Gigascience *5*, 34.
44. Wilson, B.D., Eisenstein, M., and Soh, H.T. (2019). High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets. Anal. Chem. *91*, 6783–6789.
45. Rang, F.J., Kloosterman, W.P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. *19*, 90.
46. Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. *20*, 129.