Blood Disorders & Transfusion

Baronas et al., J Blood Disord Transfus 2016, 7:5
DOI: 10.4172/2155-9864.1000365

Research Article OMICS International

RHD Zygosity Determination from Whole Genome Sequencing Data

John Baronas[1], Connie M Westhoff[2], Sunitha Vege[2], Helen Mah[1], Maria Aguad[1], Robin Smeland-Wagman[1], Richard M Kaufman[3,6], Heidi L Rehm[1,3,4,5], Leslie E Silberstein[6], Robert C Green[3,5,7] and William J Lane[1,3*]

[1]*Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA*

[2]*New York Blood Center, New York, USA*

[3]*Harvard Medical School, Boston, MA, USA*

[4]*Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Boston, MA, USA*

[5]*Partners Healthcare Personalized Medicine, Boston, MA, USA*

[6]*Department of Pathology, Division of Transfusion Medicine, Brigham and Women's Hospital, Boston, MA, USA*

[7]*Department of Medicine, Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA*

[*]**Corresponding author:** William J Lane, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Amory Lab Building 3rd Floor, Rm 3-117, 75 Francis Street, Boston, MA 02115, USA, Tel: 617-732-5469; E-mail: wlane@partners.org

Received date: Jul 27, 2016, **Accepted date:** Sep 15, 2016, **Publication date:** Sep 19, 2016

Abstract

In the Rh blood group system, the *RHD* gene is bordered by two homologous DNA sequences called the upstream and downstream Rhesus boxes. The most common cause of the D− phenotype in people of European descent is a deletion of the *RHD* gene region, which results in a hybrid combination of the two Rhesus boxes. PCR-based testing can detect the presence or absence of the hybrid box to determine *RHD* zygosity. PCR hybrid box testing on fathers can stratify risk for haemolytic disease of the fetus and new born in mothers with anti-D antibodies. Red blood cells and genomic DNA were isolated from 37 individuals of European descent undergoing whole genome sequencing as part of the MedSeq Project. A whole genome sequence-based *RHD* sequence read depth analysis was used to determine *RHD* zygosity (homozygous, hemizygous, or null states) with 100% agreement (n=37) when compared to conventional RhD serology and PCR-based hybrid box assay.

Keywords: RhD; *RHD* zygosity; Red blood cell; Blood group; Genomics; Whole genome sequencing; Next generation sequencing

Introduction

The Rh blood group system consists of the two homologous genes *RHD* and RHCE. The *RHD* gene controls the expression of the RhD protein, which contains numerous extracellular epitopes comprising the D antigen [1]. Individuals who lack the RhD protein and hence the D antigen can form anti-D antibodies. The *RHD* gene is surrounded by two homologous 9,000 bp regions called the Rhesus Boxes, which are 4,900 bp upstream and 104 bp downstream of *RHD* [2]. The most common cause of the D− phenotype in persons of European descent is deletion of the *RHD* gene due to a recombination event between of the upstream and downstream Rhesus boxes. This recombination results in a hybrid box sequence comprised of the 5' part of the upstream box and the 3' part of the downstream box [2].

The presence or absence of the hybrid box can be detected using DNA PCR methods. The hybrid box is present in individuals who are *RHD* hemizygous (e.g. deletion of one copy of the *RHD* gene) and *RHD* null (e.g. D− due to deletion of both copies of the *RHD* gene). Several different PCR-based assays are currently used to detect the hybrid box. Most of these methods depend on small differences between sequences of the 5' upstream region and 3' downstream region of the Rhesus boxes. However, the hybrid box PCR assay is not always reliable and requires that the hybrid box region does not subsequently mutate or recombine again. This assay can be falsely negative due to ethnic variations in the hybrid box sequences, or falsely positive due to other gene conversion events involving only one of the Rhesus boxes without *RHD* deletion (e.g. DAU-1; DIII type 4) [3].

Because of the potentially severe adverse consequences associated with hemolytic disease of the fetus and newborn (HDFN), it is important to be able to accurately determine the presence or absence of the *RHD* gene in pregnancies in which a D− woman has formed anti-D antibodies. If the father is D+, serologic testing cannot distinguish whether he is homozygous (two copies of the *RHD* gene), or hemizygous (only one copy). Children of *RHD* homozygous fathers have a 100% chance of being D+, but only 50% if the father is *RHD* hemizygous [2]. Therefore, *RHD* zygosity determination by DNA PCR testing of the father to predict probability of D+ fetus can help stratify the HDFN risk in D- mothers with anti-D antibodies [4].

The application of next generation sequencing (NGS) to clinical testing has enabled the field of precision medicine [5-8]. We recently demonstrated the feasibility of comprehensively predicting all molecularly understood RBC antigens using NGS based Whole Genome Sequencing (WGS) data [9] as part of the MedSeq Project [10]. It is well established that gene copy number can be calculated from WGS data using many different methods such as split read detection and read depth analysis [11]. However, we are unaware of any previously published studies describing the use of WGS analyses for determining *RHD* zygosity. Two main approaches split read detection and read depth analysis are used for zygosity determination by WGS. In split read detection, copy number changes are detected by looking for sequence reads in which one of the paired-end reads does not align to the nearby reference genome sequence, indicating the presence of a DNA recombination breakpoint. In read depth analysis,

the number of sequences that align to different regions in the reference genome are compared, looking for regions with increases or decreases of aligned reads, which can reflect changes in copy number. Here, we report that although it was not possible to find Rhesus box split reads in WGS of D– individuals, it was possible to use read depth analysis from WGS data to correctly determine *RHD* homozygous, hemizygous, and null states.

Materials and Methods

RBC serology

Traditional RBC serologic antigen testing was performed according to standard blood banking practices in the Brigham and Women's Hospital Blood Bank [9]. Peripheral blood from participants in the MedSeq study [10] was collected in EDTA tubes, and RBCs were isolated by centrifugation. A commercially available serologic monoclonal anti-D typing reagent [Bio-Rad, Hercules, CA], was used to type for the D antigen. Fifty microliter samples of washed RBCs were incubated with the typing reagent, centrifuged, and visually examined for agglutination.

Hybrid box assay

The hybrid box assay was performed according to previously published methods [12]. Briefly, allele-specific PCR was carried out using primers designed to amplify a product of 1,507 bp within the hybrid box sequence. PCR products were visualized by agarose gel electrophoresis with ethidium bromide staining.

Whole genome sequencing

WGS was performed by the CLIA-certified, CAP-accredited Illumina Clinical Services Laboratory (San Diego, CA) using paired-end 100 base pair (bp) reads on the Illumina HiSeq next generation sequencing (NGS) platform and sequenced to at least 30x mean coverage [13]. The genomic data from the MedSeq Project has been submitted to the dbGaP website. Sequence read data was aligned to the human reference sequence (GRCh37/hg19) using Burrows-Wheeler Aligner 0.6.1-r104 [14].

Whole Genome Sequence Based Rh Copy Number Determination

The Integrative Genomics Viewer [15] was used to visually inspect the WGS alignments for the Rhesus box regions expected to contain the hybrid box split read products and to visualize the read depth of coverage across the entire *RHD* gene and surrounding Rhesus boxes. Sequencing coverage values were extracted from the alignment file using BED Tools v2.17.0 [16]. Based on the breakpoint regions, the extracted coverage values were put into three bins upstream +10,000 bp upstream (chr1:25,580,018-25,593,111), RHD deletion region (chr1:25,593,112-25,660,343), and downstream (chr1:25,660,344-25,673,439). The average coverage value was then calculated for each bin followed by a calculation of the *RHD* copy number using:

$$RHD \text{ Copy number} = \frac{RHD \text{ Region Read Depth of Coverage Average}}{\text{Upstream and Downstream Average}} \times 2$$

On this scale, two copies of *RHD* should result in a value of 2(1.6-2.5), hemizygous a value of 1(0.6-1.5), and *RHD* null as a value of 0(0-0.5).

Results

Study overview

With approval from the Partners HealthCare Human Research Committee (IRB), samples for RBC and genomic DNA isolation were collected from 37 individuals of European descent who had WGS through participation in the MedSeq Project [10]. The WGS results were analyzed to see if it was possible to determine *RHD* zygosity using only the WGS data. Serologic D typing followed by PCR-based hybrid box assays were used to confirm the *RHD* zygosity as either homozygous, hemizygous, or null.

The serologic testing results were known at the time of WGS analysis for the first 15 cases (14 D+ and 1 D–), follow by blinded WGS analysis for the last 22 cases (12 D+ and 10 D–). The PCR hybrid box assay was performed after WGS analysis and was thus blinded in all cases. In all cases, the serologic testing, PCR hybrid box assay, and WGS analysis were performed by different individuals.

Hybrid box split read detection

As shown in Figure 1, the human reference genome coordinates of the Rhesus boxes and the hybrid box breakpoint were manually determined using the Integrative Genomics Viewer to identify the published box sequences (upstream box: AJ252311 and downstream box: AJ252312) [2].

This approach revealed a loss of sequence read depth of coverage right after the upstream Rhesus box and before the downstream Rhesus box in a D– individual (Figures 2A and 2B). However, there were no hybrid box split reads at either the upstream or the downstream breakpoint locations. This finding was later confirmed in three more D– individuals.

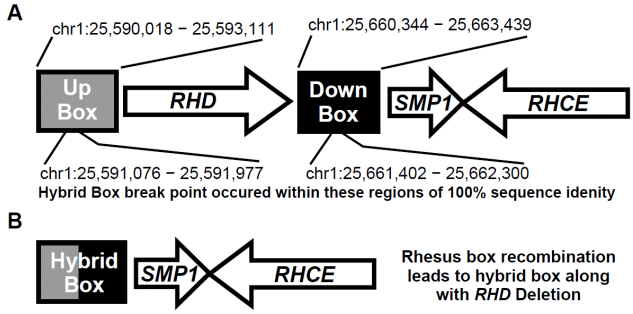

Figure 1: Rh Blood Group Gene Organization. The *RHD* and RHCE genes are oriented toward each other with a small interviewing region that contains *SMP1* gene. The *RHD* gene is surrounded by two highly homologous Rhesus box regions. The human reference genome (GRCh37/hg19) coordinates are shown for the Rhesus box regions as well as the ~900 bp region of 100% sequence identity in which the hybrid box break point occurred during *RHD* deletion.

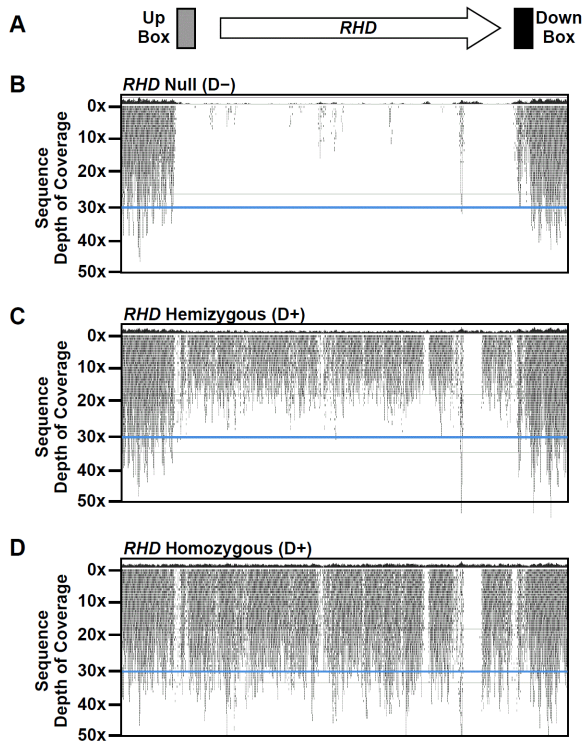

Figure 2: *RHD* and Rhesus Box WGS Sequence Read Depth of Coverage. Examples of WGS read depth of coverage screen shots from Integrated Genomics Viewer (IGV). (A) *RHD* gene and the surrounding Rhesus boxes. These regions are drawn to scale with the plots below. (B) D− (*RHD* null) individual with complete loss of coverage across the *RHD* gene. (C) D+ (*RHD* hemizygous) individual partial loss of coverage across the *RHD* gene. (D) D+ (*RHD* homozygous) individual with no loss of coverage across the *RHD* gene. To aid in comparison, the horizontal blue lines represent the average 30x coverage across this region as found in the *RHD* homozygous individual.

WGS based read depth coverage *RHD* zygosity determination

Since hybrid box split reads could not be directly found, sequence read depth of coverage analysis was performed. In the first 15 cases, it was found that D+ individuals displayed two distinct patterns of coverage over *RHD* and its upstream and downstream regions:

1. Those with noticeable but not complete absence in coverage (~15x) over *RHD* (Figure 2C) and,

2. Those with no change in coverage (~30x) over *RHD* (Figure 2D). Conventional hybrid box PCR confirmed that the difference between the two D+ coverage patterns reflected *RHD* copy number changes due to hemizygosity (Figure 2C) or homozygosity (Figure 2D).

To more rigorously evaluate the above coverage patterns, an average sequence depth of coverage was calculated for *RHD* and each Rhesus box region for all 37 individuals This revealed three patterns of coverage that were used to visually estimate RHD zygosity (Figure 3):

1. no *RHD* coverage (null),
2. partial loss of *RHD* coverage (hemizygous),
3. no change in *RHD* coverage (homozygous).

In order to automate the WGS-based *RHD* zygosity determination, the above read depth differences were used as the basis for an *RHD* copy number calculation that was performed on the 37 individuals (Figure 4). There was 100% agreement between the WSG read depth-based *RHD* copy number determination approach and the conventional RhD serology and hybrid box PCR.

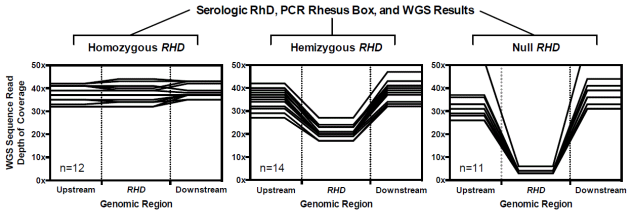

Figure 3: Calculation of *RHD* and Rhesus Box Coverage. The average sequence coverage values are shown for the Rhesus boxes (along with 10,000 bp extension away from *RHD*) and the *RHD* gene deletion region. The results are in three groups based on the *RHD* zygosity as determined by WGS analysis and confirmed by a combination of serology and PCR hybrid box assay. Hemizygous individuals seem to have lower sequencing depth of coverage in the *RHD* gene region and there is no coverage in *RHD* null individuals.

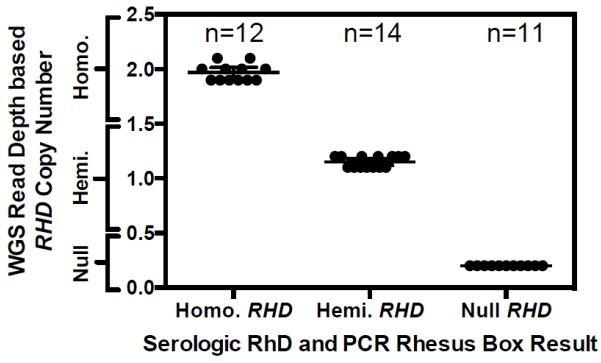

Figure 4: *RHD* Copy Number using WGS Read Depth of Coverage. The *RHD* copy number was calculated using the WGS data and normalized to the gene copy number such that the calculation is homozygous (copy number 1.6–2.5), hemizygous (copy number 0.6-1.5), 1, and null (copy number 0–0.5). The results are in three groups based on the *RHD* zygosity as determined WGS analysis and confirmed by a combination of serology and PCR hybrid box assay. Each dot represents a single patient with a line representing the mean and with bars showing the 95% confidence interval. There was statistical significance between the three groups with no outliers or overlapping copy number values.

Discussion

In 37 individuals, WGS-based hybrid box split read and sequence read depth of coverage analysis were evaluated for their ability to

determine *RHD* zygosity when compared to conventional serologic typing and hybrid Rhesus box PCR results.

We found that hybrid box split read analysis using WGS data failed to reliably allow zygosity determination of the *RHD* gene. This was not surprising given the known high degree of sequence homology between the Rhesus boxes. A consequence of this homology, is that with split read analysis it is impossible to precisely determine the exact breakpoint of the hybrid box; at best, this approach narrows the breakpoint location to within an approximately 900 bp region of 100% sequence identity. WGS data is comprised of short read (100 bp paired-end read) sequences. Thus, it is not possible to find a single hybrid sequence read with sequences specific to the upstream and downstream Rhesus boxes such a read would need to span the entire ~900 bp region. Conventional hybrid box PCR assays work because they can detect the presence of recombined upstream and downstream specific hybrid box sequences using PCR primers separated by 1,000-9,000 bp [4,12]. In the future, it might be possible to perform WGS-based detection of hybrid box split read sequences using 4th generation sequencing technologies that produce sequence reads that are several kilobases long [17].

However, even with short read sequences, sequence read depth of coverage evaluation was 100% accurate (n=37), and therefore provides a simple and easy method to determine *RHD* zygosity using WGS data. Conventional, hybrid box PCR based detection assays can be unreliable in individuals with variant Rhesus box sequences, most notably in those of African descent, but problems have also been reported in those of European descent [4]. In such situations, a more technically demanding hybrid box assay or quantitative *RHD* amplification compared to amplification of a control gene must currently be performed to properly determine zygosity [3,18]. Since the WGS sequence read depth of coverage approach is not based on detecting the hybrid box sequence, we speculate that it would still work in individuals with variant Rhesus boxes. In addition, the same WGS data should also be usable to detect other D– molecular mechanisms such as the *RHD* pseudogene in individuals of African descent. In summary, WGS based *RHD* evaluation has the potential to serve as a universal *RHD* zygosity determination assay in all ethnic groups.

In the future, it is likely that many individuals will have existing WGS as part of precision medicine initiatives. The *RHD* copy number calculation could routinely be used at low cost in anyone undergoing WGS. As such, a father's *RHD* zygosity would already be available to aid in the risk stratification and counseling for HDFN in sensitized D– mothers.

Members of the MedSeq Project

Members of the MedSeq Project are as follows: David W. Bates, MD, Alexis D. Carere, MA, MS, Allison Cirino, MS, Lauren Connor, Kurt D. Christensen, MPH, PhD, Jake Duggan, Robert C. Green, MD, MPH, Carolyn Y. Ho, MD, Lily Hoffman-Andrews, Joel B. Krier, MD, William J. Lane, MD, PhD, Denise M. Lautenbach, MS, Lisa Lehmann, MD, PhD, MSc, Christina Liu, Calum A. MacRae, MD, PhD, Rachel Miller, MA, Cynthia C. Morton, PhD, Christine E. Seidman, MD, Shamil Sunyaev, PhD, Jason L. Vassy, MD, MPH, SM, Rebecca Walsh, Brigham and Women's Hospital and Harvard Medical School; Sandy Aronson, ALM, MA, Ozge Ceyhan-Birsoy, PhD, Siva Gowrisankar, Ph.D., Matthew S. Lebo, PhD, Ignat Leschiner, PhD, Kalotina Machini, PhD, MS, Heather M. McLaughlin, PhD, Danielle R. Metterville, MS, Heidi L. Rehm, PhD, Partners Center for Personalized Genetic Medicine; Jennifer Blumenthal-Barby, PhD, Lindsay Zausmer Feuerman, MPH, Amy L. McGuire, JD, PhD, Ali Noorbaksh Jill Oliver Robinson, MA, Melody J. Slashinski, MPH, PhD, Julia Wycliff, Baylor College of Medicine, Center for Medical Ethics and Health Policy; Philip Lupo, PhD, MPH, Baylor College of Medicine, Department of Pediatrics; Stewart C. Alexander, PhD, Kelly Davis, Peter A. Ubel, MD, Duke University; Peter Kraft, PhD, Harvard School of Public Health; J. Scott Roberts, PhD, University of Michigan; Judy E. Garber, MD, MPH, Dana-Farber Cancer Institute; Tina Hambuch, PhD, Illumina, Inc.; Michael F. Murray, MD, Geisinger Health System; Isaac Kohane, MD, PhD, Sek Won Kong, MD, Boston Children's Hospital.

Acknowledgements

The authors thank the staff and participants of the MedSeq Project for their important contributions. The MedSeq Project is supported by the National Human Genome Research Institute HG006500. Drs. Green and Rehm are also supported by HD077671, HG006834 and HG008685. Additional funding was provided by the Brigham and Women's Hospital Pathology Department Stanley L. Robbins M.D.Memorial Research Fund Award.

References

1. Chou ST, Westhoff CM (2010) The Rh and RhAG blood group systems. Immunohematology 26: 178-186.
2. Wagner FF, Flegel WA (2000) *RHD* gene deletion occurred in the Rhesus box. Blood 95: 3662-3668.
3. Wagner FF, Moulds JM, Flegel WA (2005) Genetic mechanisms of Rhesus box variation. Transfusion 45: 338-344.
4. Dean L (2005) Blood Groups and Red Cell Antigens Bethesda (MD). National Center for Biotechnology Information, US.
5. Green RC, Rehm HL, Kohane IS (2012) Common terms and tools used to describe genome sequencing data. Genomic and Personalized Medicine.
6. Green ED, Guyer MS, National Human Genome Research Institute (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470: 204-213.
7. McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, Kohane IS, Krier J, et al. (2014) A systematic approach to the reporting of medically relevant findings from whole genome sequencing. BMC Med Genet 15: 134.
8. Chaitankar V, Karakülah G, Ratnapriya R, Giuste FO, Brooks MJ, et al. (2016) Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Prog Retin Eye Res.
9. Lane WJ, Westhoff CM, Uy JM, Aguad M, Smeland-Wagman R, et al. (2016) Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. Transfusion 56: 743-754.
10. Vassy JL, Lautenbach DM, McLaughlin HM, Kong SW, Christensen KD, et al. (2014) The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. Trials 15: 85.
11. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 Suppl 11: S1.
12. Murdock A, Assip D, Hue-Roye K, Lomas-Francis C, Hu Z, et al. (2008) *RHD* deletion in a patient with chronic myeloid leukemia. Immunohematology 24: 160-164.
13. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-539.
14. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

15. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.
16. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
17. Bayley H (2015) Nanopore sequencing: from imagination to reality. Clin Chem 61: 25-31.
18. Chiu RW, Murphy MF, Fidler C, Zee BC, Wainscoat JS, et al. (2001) Determination of RhD zygosity: comparison of a double amplification refractory mutation system approach and a multiplex real-time quantitative PCR approach. Clin Chem 47: 667-672.