Analysis of exome sequencing data sets reveals structural variation in the coding region of *ABO* in individuals of African ancestry

Keolu Fox,[1] Jill M. Johnsen,[2,3] Bradley P. Coe,[1] Chris D. Frazar,[1] Alexander P. Reiner,[4,5] NHLBI Exome Sequencing Project, Minority Health-GRID Network, Evan E. Eichler,[1,6] and Deborah A. Nickerson[1]

BACKGROUND: ABO is a blood group system of high clinical significance due to the prevalence of ABO variation that can cause major, potentially life-threatening, transfusion reactions.

STUDY DESIGN AND METHODS: Using multiple large-scale next-generation sequence data sets, we demonstrate the application of read-depth approaches to discover previously unsuspected structural variation (SV) in the *ABO* gene in individuals of African ancestry.

RESULTS: Our analysis of SV in the *ABO* gene across 6432 exomes reveals a partial deletion in the *ABO* gene in 32 individuals of African ancestry that predicts a novel O allele.

CONCLUSION: Our study demonstrates the power that analyses of large-scale sequencing data, particularly data sets containing underrepresented populations, can provide in identifying novel SVs.

The *ABO* gene commonly encodes two different forms of a glycosyltransferase defined by the addition of either the A or B sugars (*N*-acetylgalactosamine [GalNAc] for A or galactose [Gal] for B) in an α-1,3 configuration to the H antigen substrate (Fucα1-2Galβ1-3GalNAc).[1,2] Single-nucleotide variants in the *ABO* gene affect the efficiency and specificity of this

ABBREVIATIONS: AA(s) = African American(s); EA = European American; ESP = Exome Sequencing Project; RD = read-depth; SD = segmental duplication; SV(s) = structural variation(s); XHMM = eXome Hidden Markov Model.

From the [1]Department of Genome Sciences and the [2]Hematology Division, University of Washington, School of Medicine; [3]Bloodworks Northwest; the [4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center; [5]Department of Epidemiology, University of Washington, School of Public Health; and the [6]Howard Hughes Medical Institute, Seattle, Washington.

Address reprint requests to: Deborah A. Nickerson, Department of Genome Sciences, School of Medicine, University of Washington, Seattle, WA 98195; e-mail: debnick@uw.edu.

This work was supported by the NHLBI Grand Opportunity Exome Sequencing Project, the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). This work was also supported by a National Center on Minority Health and Health Disparities grant (RC4MD005964), an NHGRI training grant (HG-00035), and an NHLBI F31 (HL-122119).

Received for publication March 14, 2016; revision received July 6, 2016; and accepted July 7, 2016.

doi:10.1111/trf.13797

TRANSFUSION 2016;56;2744–2749

enzyme for these sugars.[3] Loss-of-function variation in the form of a single base pair frame shift accounts for the majority of variation resulting in the O phenotype.[4] Characterizing variation in *ABO* is important in transfusion and transplantation medicine because variants in *ABO* have significant consequences with regard to recipient compatibility. Additionally, variation in the *ABO* gene has been associated with cardiovascular disease risk (e.g., myocardial infarction) and quantitative blood traits (e.g., von Willebrand factor, Factor VIII, intercellular adhesion molecule 1, E-cadherin, and P-selectin).[5] Relating *ABO* genotypes to blood group antigen phenotypes requires the analysis of haplotypes. These haplotypes are composed of both common and rare variants that can affect the structure and/or function of the *ABO* glycosyltransferase.[6] Located on the distal end of chromosome 9 (9q34.2), *ABO* represents an interesting candidate gene for structural variation (SV) discovery as this region has a higher than average recombination rate relative to more proximal genes on the p arm of Chromosome 9.[7] Recent strategies utilizing polymerase chain reaction (PCR)-specific primers have identified a noncoding 5.8-kb deletion in Intron 1 of the *ABO* gene.[8] Yet despite decades of studies of the *ABO* gene, including in large-scale population specific screens allowing for rare variant discovery (i.e., exome and whole genome sequencing), an SV has not been previously reported in the coding portion of the *ABO* locus. Here we apply two read-depth (RD) approaches to the analysis of a large exome data set, and we report the discovery of SV in the *ABO* gene.

MATERIALS AND METHODS

We used *ABO* coding sequence data derived from the NHLBI Exome Sequencing Project (ESP) and the Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID).[9,10] Through the ESP, 15,336 genes were sequenced at high coverage (median depth > 100x) in a total of 6515 unrelated European American (EA; n = 4298) and African American (AA; n = 2217) individuals from 19 different cohorts.[11] Through MH-GRID, exome sequences at similar high coverage (median depth > 100x) were obtained from a total of 1313 unrelated AA individuals. The library construction, exome capture, sequencing, mapping, calling, and filtering were carried out as described previously.[12] Exome sequence data were aligned to NCBI human reference GRCh37.

Sequence data were aligned to *ABO* reference sequence obtained from GenBank (NG_006669.1). Mean sample RD for the *ABO* gene was 77x (ranging from 10 to 374x) and includes the entire coding sequence and exon–intron boundaries. Of the 4298 EA ESP participants, 3405 had minimum coverage of 50x across the entire targeted *ABO* region. Additionally, 3027 self-identified AAs from both the ESP and MH-GRID were included in our *ABO* analysis bringing our total to 6432 participants, roughly divided equally between EA (53%) and AA (47%).

For SV discovery, we applied two algorithms optimized for exome data sets. The first, eXome Hidden Markov Model (XHMM),[13] uses principal component analysis and a hidden Markov model to detect and genotype structural variants normalized for RD data across samples. Applying this method to the RD obtained with the genome analysis toolkit, we called small (1-100 kb) SV in our combined set of 6432 participants.[14] We also implemented an orthogonal RD based algorithm (CoNIFER) to call SV.[15] CoNIFER uses singular value decomposition (SVD) to detect rare SV and genotype these from exome sequencing data. Both algorithms were run with default variables. As SV callers are subject to false positives,[16] we explored only SV events found in two or more participants and filtered on allele balance (70% cutoff), while allowing for a 50% overlap between unique independent SVs (Fig. 1A).

Putative novel SVs (identified by both approaches) were further confirmed by TaqMan copy number assays.[17] To accomplish this, a probe-specific to *ABO* Exon 7 (assay ID Hs01862499_cn, ThermoFisher Scientific) was used to evaluate the copy number of genomic DNA targets using Applied Biosystems real-time system and the ribonuclease P RNA component H1 (RPPH1) gene as the control assay. To assess copy number, a 2.0-μL sample of DNA (5 μg/mL), 5.0 μL of copy number–specific TaqMan master mix, 0.2 μL of *ABO* Exon 7 probe (assay ID Hs01862499_cn), 0.5 μL of copy number–specific TaqMan reference assay (RPPH1), and 2 μL of PCR quality water (10.0 μL per sample) were used. AA samples with and without the predicted SV in Exon 7 and a HapMap reference sample (NA12878) without the *ABO* SV were used as controls. Copy number was then calculated with computer software (CopyCaller, Version 2.0, Applied Biosystems). We surveyed the region surrounding *ABO* using the UCSC hg19 genome browser for nearby duplicated elements, which could be involved in SV formation.[18] Segmental duplication (SD) and repeat masker tracks were assessed for elements that might associate with the regions of the identified exon deletion breakpoints discovered in both the exome sequence data set and an orthogonal low coverage (approx. 6x) whole genome sequence data set: the 1000 Genomes project. After identifying putative breakpoints with both the repeat masker and the SD tracks in the 1000 Genomes data set, we designed a long-range PCR assay spanning the predicted breakpoint regions for the *ABO* deletion, which includes Exons 5, 6, and 7.

RESULTS

With XHMM, 16 distinct SVs were predicted in the coding sequence of *ABO* in 6432 individuals (12,864 chromosomes). Five predicted *ABO* SVs were identified in two or more individuals; four deletions were detected in 32

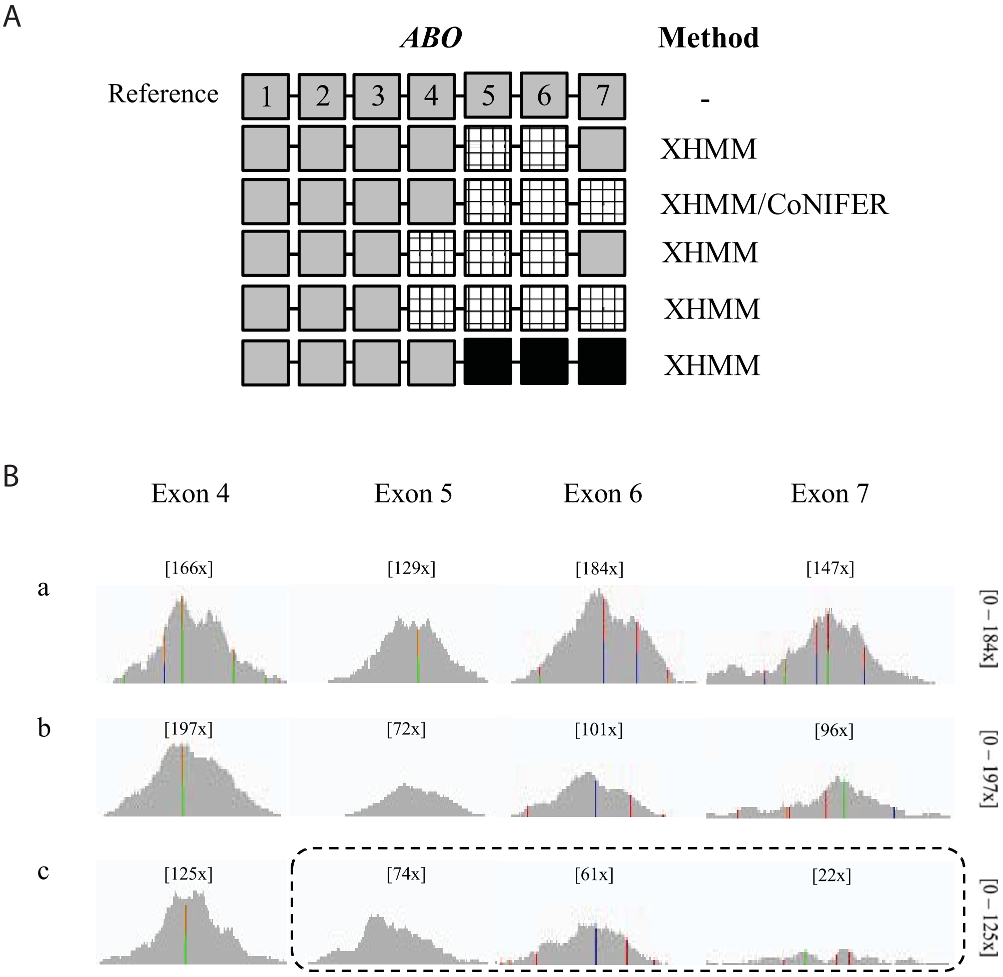

Fig. 1. (A) Putative SV discovered in ABO in 6432 exomes using XHMM and CoNIFER. Putative SV discovered in ABO in 2+ individuals using two RD methods are shown; each gray box represents intact (two copies of) exons in the ABO gene. The ABO gene is shown left to right in a 5′ to 3′ orientation. Gridded boxes represent exons predicted to be deleted using both XHMM and CoNIFER. Black boxes represent ABO duplications identified using XHMM. Both RD algorithms putatively predicted deletion of [E5-7]. (B) Examples of RD data in support of predicted SV in ABO. Integrative genomics viewer (IGV) RD comparisons are shown for ABO Exons 4 to 7. Colored vertical bars in IGV indicate the number of reads supporting each allele for single nucleotide variants in each exon (i.e., allele balance). Loss of heterozygosity and lower RD (i.e., monocolored vertical line or solid gray vertical line) is consistent with a deletion in a given exon, versus heterozygosity and higher RD (i.e., a multicolored vertical line) of a SNV is consistent with two copies of an exon (e.g., Exon 4). (a) An individual with mean RD at ABO (Exons 4, 5, 6, and 7). (b) An individual predicted to carry a deletion of [E5-7] by both XHMM and CoNIFER. (c) An individual predicted to harbor a putative deletion of [E5-6] and not Exon 7 by XHMM. However, visual inspection of RD and allele balance clearly show low coverage consistent with a deletion of Exon 7 as well. Dashed region represents minimal size of deletion.

individuals, and one duplication was detected in two individuals. Eleven predicted SVs were found only in a single individual, that is, singletons (Table S1 [available as supporting information in the online version of this paper] and Fig. 1). Using CoNIFER, we identified three predicted *ABO* SVs (a single deletion discovered in five individuals) and two singletons (both deletions including Exon 7; Table S1). We focused on SVs found in more than one individual since singletons have a higher rate of false-positive detection.[19] The five predicted SVs found in 2+ individuals were overlapping and included a putative deletion of *ABO* Exons 5 and 6, a deletion of *ABO* Exons 4 to 7, a deletion of *ABO* Exons 5 to 7 (Fig. 1A), and a potential duplication of *ABO* Exons 5 to 7 identified by XHMM. Visual inspection of the RD and allele balance revealed that the putatively identified deletion of *ABO* Exons 5 to 6 also included deletion of Exon 7 (Fig. 1B) and that the deletions of *ABO* Exons 4 to 7 likely only affected *ABO* Exons 5 to 7. Thus, putative *ABO* exon deletions detected by XHMM appeared similar to the *ABO* Exons 5 to 7 deletion identified by CoNIFER, highlighting the importance of manual curation and the benefit of applying more than one RD based algorithm in screening for SV in next-generation sequencing data sets.

Using a TaqMan probe in *ABO* Exon 7, we tested 16 individuals predicted to carry one copy (deletion), two individuals predicted to have three copies (duplication), and eight individuals predicted to have two copies of Exon 7 (control) in the *ABO* gene. While a deletion affecting Exon 7 was confirmed in all individuals (n = 16), the *ABO* duplication in Exon 7 [E5-7] could not be confirmed and, therefore, represents a false positive (Fig. S1, available as supporting information in the online version of this paper). Because whole exome sequencing only targets exonic sequence (excluding intronic regions), our validation experiments indicated that one of the events identified in 2+ individuals was larger than the length predicted from our exome-based XHMM analysis (i.e., [E5-6] deletion). By performing a combined analysis (integrated SV call set) with visual inspection, we were able to refine our call boundaries and robustly predict and validate SV in the *ABO* gene. In addition to the recurrent deletions, we observed rare lower confidence overlapping singleton deletions encompassing Exons 5 to 7 with variable proximal and distal breakpoints. As DNA was not available for validation, we have excluded these events from our allele frequency estimate for the AA *ABO* deletions (Fig. S2, available as supporting information in the online version of this paper).

DISCUSSION

To our knowledge, SV in the coding portion of the *ABO* gene has not been previously suspected or reported. Interestingly, the partial *ABO* gene deletion we identify has a greater than 50% reciprocal overlap with an event identified in the third and final stage of the 1000 Genomes Project reference panel.[20] In further support of our discovery of SV in *ABO*, our three-exon deletion ([E5-7] deletion) including the distal Exon 5 breakpoint (Fig. 1B) was identified in both our cross-sectional analysis of a combined exome data set and the Phase III 1000 Genomes Project.[20] Finally in both the 1000 Genomes Project and our combined exome data set, the three-exon *ABO* deletion was exclusively found in individuals of both African and African Admixed ancestry with a similar allele frequency; our combined exome data set (AA 0.00469) and 1000 Genomes Project Phase II (African [0.0129], Admixed American [0.004]). The slight differences detected in SV allele frequency likely reflect the differences in historical admixture in African, AA, and Admixed American populations.[21]

Further examination of the recurrent deletions identified a SD associated with the distal breakpoint region. Previous studies have suggested that SDs are capable of driving increased copy number variation mutation rates, not only through flanking direct orientation repeats (nonallelic homologous recombination), but also through a replication or repair-based mechanism in regions adjacent to duplications.[22] To this end we examined all nearby duplicated elements annotated in the UCSC hg19 genome browser SD and repeat masker track that associated with the theorized Exon 5 to 7 deletion breakpoints discovered in the 1000 Genomes Project Phase III data. Both breakpoints reported in the Phase III 1KG data were mapped to intronic regions (proximal breakpoint between Exons 4 to 5 and distal breakpoint between the end of Exon 7 and *ABO*'s neighboring gene, *OBP2B*). Through our repeat masker analysis we identified two flanking Alu retrotransposons within the same subfamily (AluSx1) with direct orientation on either side of the three-exon deletion suggesting that these AluSx1 mobile elements may be the mechanism responsible for these deletion alleles. We designed long-range PCR primers with approximately 200-bp flanking buffer from our theorized breakpoints in an effort to confirm the presence of the three exon deletion in AA samples identified via our combined RD analysis. Our results confirm both our theorized breakpoints and the three-exon deletion in *ABO* (Fig. S3, available as supporting information in the online version of this paper) supporting a deletion mechanism likely involving the breakpoint flanking AluSx1 elements.

By using sequence data from thousands of AA exomes, we were able to characterize, validate, and provide population frequency estimates for an SV, which results in a partial *ABO* gene deletion. This finding strongly supports the utility of exploring all types of genetic variation (e.g., SV) in large emerging data sets. Moreover, our study shows it is possible to identify previously unrecognized sources of genetic variation even in well-studied genes. The observation of this SV, which impacts the coding regions of *ABO*, is of value in that this allele would be predicted to encode a nonfunctional (blood group O) glycosyltransferase. Conventional *ABO* genotyping methods are insensitive to SV, and thus this allele would be expected to confound *ABO* zygosity determinations and could impact the overall interpretation of *ABO* genotyping results. This work also sets a precedence for finding other as-yet-unidentified alleles harboring SV (i.e., unidentified deletions, duplication, or more complex variation) in diverse populations. By focusing on underrepresented populations (AA), we have identified novel loss of function SV in one of the most well-studied genes in the human genome. A comprehensive understanding of variation in blood group genes in minority populations has the potential to 1) refine our knowledge of genotype–phenotype correlation in intermediary biomarkers associated with cardiovascular disease and 2) create a more comprehensive understanding of health disparity in the United States. Although we were focused on the identification of previously undetected SV in the *ABO* gene, the approaches presented here can be readily applied to many other genes that direct expression of blood cell antigen systems, including those with known SV (e.g., Rhesus blood groups, MNS [glycophorins], major histocompatibility complex antigens, minor histocompatibility antigens, killer cell immunoglobulin receptor genes) and those not known previously to harbor SV, such as we now report for *ABO*.[23,24]

ACKNOWLEDGMENTS

We thank numerous colleagues who collaborated through both the Minority Health GRID (MH-GRID) and the NHLBI Exome Sequencing Project. We also thank everyone involved in the Nickerson lab at the University of Washington, Department of Genome Sciences. Special thanks to Adam Gordon, Cindy Shephard, Eric Johanson, Xander Nuttle, Stuart Cantsilieris, Colleen Davis, Jason Underwood, Marsha Wheeler, and Riley Taitingfong.

CONFLICT OF INTEREST

The authors have disclosed no conflicts of interest.

REFERENCES

1. Zhang W. Structural modification of H histo-blood group antigen. Blood Transfus 2015;13:143-9.
2. Stanley P. Cummings R. Structures common to different glycans. Chapter 13. In: Essentials of glycobiology. 2nd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009.
3. Yamamoto F. Molecular genetic basis of the histo-blood group ABO system. Nature 1990;345:229-33.
4. Evans SV. The structural basis for specificity in human ABO(H) blood group biosynthesis. Nature Struct Biol 2002;9: 685-90.
5. Zhang H. ABO blood groups and cardiovascular diseases. Int J Vasc Med 2012;2012:641917.
6. Yip SP. Sequence variation at the human ABO locus. Ann Hum Genet 2002;66:1-27.
7. Wang J. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell 2012;150:402-12.
8. Sano R. A 3.0-kb deletion including an erythroid cell-specific regulatory element in intron 1 of the ABO blood group gene in an individual with the Bm phenotype. Vox Sang 2015;108: 310-3.
9. NHLBI Grand Opportunity Exome Sequencing Project (NHLBI-ESP) [Internet]. Bethesda (MD): NHLBI; [cited 2015 May 11]. Available from: https://esp.gs.washington.edu/drupal/.
10. Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) [Internet]. Atlanta (GA): Moorehouse School of Medicine; 2015 [cited 2015 Jan 11]. Available from: https://rcenterportal.msm.edu/index.php/2015-09-29-06-34-56/minority-health-grid/.
11. Fu W. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 2013;493: 216-20.
12. Johnsen JM. Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. Blood 2013;122:590-7.
13. Fromer M. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 2012;91:597-607.
14. McKenna A. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
15. Krumm N. Copy number variation detection and genotyping from exome sequence data. Genome Res 2012;22: 1525-32.
16. Teo SM. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 2012;28:2711−8.
17. Dennis MY. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell 2012;149: 912-22.
18. Bailey JA. Recent segmental duplications in the human genome. Science 2002;297:1003-7.
19. Mills RE. Mapping copy number variation by population-scale genome sequencing. Nature 2011;470:59-65.
20. Sudmant PH. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526:75-81.
21. Parra E. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 1998; 63:1839-51.
22. Itsara A. De novo rates and selection of large copy number variation. Genome Res 2010;20:1469-81.
23. Storry JR. The ABO blood group system revisited: a review and update. Immunohematology 2009;25:48-59.
24. Mullalay A, Ritz J. Beyond HLA: the significance of genomic variation for allogeneic hematopoietic stem cell transplantation. Blood 2007;109:1355-62.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's Web site:

Fig. S1. Validation of structural variation in *ABO* exon 7. Vertical black lines indicate rtPCR error. a.) A control predicted to have two copies of the *ABO* gene by XHMM and CoNIFER. b.) An individual predicted by both XHMM and CoNIFER to have an *ABO* [E5-7] deletion confirmed by TaqMan. c.) An individual predicted by XHMM to have a deletion of only exons 5 and 6. However, manual curation indicated exon 7 was deleted, which was confirmed by TaqMan; d.) An individual predicted by XHMM to have a putative duplication of *ABO* [E5-7]. However, TaqMan detected two copies of

ABO, indicating a false positive duplication call. These validation data highlight the importance of a combined call-set to robustly predict SVs in the *ABO* gene. In total we validated 16 of the 32 events predicted to have the [E5-7] deletion.

Fig. S2. Putative SVs discovered in *ABO* in 6,432 exomes using XHMM and CoNIFER (singletons). Both the *ABO* and *OBP2B* genes are shown left to right in a 5' to 3' orientation. Putative SV discovered in *ABO* and neighboring downstream gene *OBP2B* using two read depth methods are shown. Each gray box represents intact versions of 1-7 exons in the *ABO* and *OBP2B* gene. Dotted boxes represent *ABO* singleton deletions identified using CoNIFER (two events). Black boxes represent *ABO* duplications identified using XHMM (eight events). Cross-hatched boxes represent exons predicted to be deleted using XHMM (three events).

Fig. S3. Detection of a ~5,800 bp *ABO* deletion spanning exons 5, 6, and 7 in AA samples. Using a long-range Takara PCR strategy spanning the predicted *ABO* deletion (products shown on a 10% agarose gel), we demonstrate the presence of the predicted wild type amplified PCR product (~8,000 bp) in our control samples (columns 2 and 4) and the presence of both the ~8,000bp product and a ~2,000 bp product in samples with a predicted partial *ABO* gene deletion (column 3, 5, and 6). This is consistent with heterozygosity for a ~5,800 bp deletion and validates the presence of *ABO* SV.

Table S1. Summary of SV discovery using multiple read depth based algorithms