



A whole genome approach for discovering the genetic basis of blood group antigens: independent confirmation for P1 and Xg^a

William J. Lane ^{1,2} Maria Aguad,¹ Robin Smeland-Wagman,¹ Sunitha Vege,³ Helen H. Mah,¹ Abigail Joseph,¹ Carrie L. Blout,⁴ Tiffany T. Nguyen,⁴ Matthew S. Lebo,^{1,2,5,6} Manpreet Sidhu,³ Christine Lomas-Francis,³ Richard M. Kaufman,^{1,2} Robert C. Green ^{2,4,6,7}
 Connie M. Westhoff,³ for the MedSeq Project[‡]

BACKGROUND: Although P1 and Xg^a are known to be associated with the *A4GALT* and *XG* genes, respectively, the genetic basis of antigen expression has been elusive. Recent reports link both P1 and Xg^a expression with nucleotide changes in the promoter regions and with antigen-negative phenotypes due to disruption of transcription factor binding.

STUDY DESIGN AND METHODS: Whole genome sequencing was performed on 113 individuals as part of the MedSeq Project with serologic RBC antigen typing for P1 (n = 77) and Xg^a (n = 15). Genomic data were analyzed by two approaches, nucleotide frequency correlation and serologic correlation, to find *A4GALT* and *XG* changes associated with P1 and Xg^a expression.

RESULTS: For P1, the frequency approach identified 29 possible associated nucleotide changes, and the serologic approach revealed four among them correlating with the P1 +/P1– phenotype: chr22:43,115,523_43,115,520AAAG/delAAAG (rs66781836); chr 22:43,114,551C/T (rs8138197); chr22:43,114,020 T/G (rs2143918); and chr22:43,113,793G/T (rs5751348). For Xg^a, the frequency approach identified 82 possible associated nucleotide changes, and among these the serologic approach revealed one correlating with the Xg(a+)/Xg(a–) phenotype: chrX:2,666,384G/C (rs311103).

CONCLUSION: A bioinformatics analysis pipeline was created to identify genetic changes responsible for RBC antigen expression. This study, in progress before the recently published reports, independently confirms the basis for P1 and Xg^a. Although this enabled molecular typing of these antigens, the Y chromosome PAR1 region interfered with Xg^a typing in males. This approach could be used to identify and confirm the genetic basis of antigens, potentially replacing the historical approach using family pedigrees as genomic sequencing becomes commonplace.

There are over 350 RBC antigens divided into 36 blood group systems. The molecular mechanisms of antigen expression are understood for the vast majority, with more than 2000 alleles

ABBREVIATIONS: NT = nucleotide; SNPs = single nucleotide polymorphisms; WGS = whole genome sequencing.

From the ¹Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; ²Harvard Medical School, Boston, Massachusetts; ³New York Blood Center, New York City, New York; ⁴Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; ⁵Laboratory for Molecular Medicine, Boston, Massachusetts; ⁶Partners Personalized Medicine, Boston, Massachusetts; and the ⁷Broad Institute of MIT and Harvard, Boston, Massachusetts.

Address correspondence to: William J. Lane, MD, PhD, Brigham and Women's Hospital and Harvard Medical School, Pathology Department, Hale Building for Transformative Medicine, Room 8002L, 60 Fenwood Rd, Boston, MA 02115; e-mail: wlane@bwh.harvard.edu

[‡]Additional members of the MedSeq Project are listed in the end of the paper.

The MedSeq Project was supported by the National Human Genome Research Institute U01-HG006500. RCG is supported by grant funding from NIH, the Broad Institute and the Department of Defense. RCG receives compensation for advising the following companies: AIA, Helix, Ohana, OpraHealth, Prudential and Veritas; and is cofounder of Genome Medical, Inc, a nationwide telemedicine service providing expert advice in genetics. WJL was additionally supported by the Brigham and Women's Hospital Pathology Department Stanley L. Robbins MD Memorial Research Fund Award.

Received for publication September 10, 2018; revision received October 30, 2018; and accepted November 10, 2018.

doi:10.1111/trf.15089

© 2018 AABB

TRANSFUSION 2018;99:999;1–8

carrying nucleotide (NT) changes (allelic variations) defined in 45 genes. Despite the advancement in the genotype-to-phenotype relationship of blood group antigens, the genetic basis of several remain elusive. However, this gap is quickly closing as bioinformatics technologies are applied. For example, recent work has provided evidence for the role of the transcription factors RUNX1 in the expression of P1 antigen¹ and of GATA-1 in the expression of Xg^a.^{2,3}

The P1 antigen was discovered in 1927 by injecting rabbits with human RBCs and screening for antibodies by testing the sera against human RBCs from different individuals.⁴ There are ethnic differences in frequency of P1 antigen expression on RBCs of individuals of European ancestry and African ancestry with RBCs from 79% European ancestry versus 94% African ancestry typing as P1+ (P₁ phenotype) and 21% of European ancestry vs 6% of African ancestry typing as P1- (P₂ phenotype). In 2001, it was shown that *A4GALT* encodes a 4- α -galactosyltransferase enzyme, which adds α -galactose to paragloboside to create the P1 antigen.⁵ In 2011 and 2014, it was reported that the single-nucleotide polymorphisms (SNPs) designated rs8138197,⁶ rs2143918,⁷ and rs5751348⁷ correlated with P1+/P1- expression, but the mechanism remained unknown. There was uncertainty as to the actual SNP responsible, but recently Westman et al.¹ showed that SNP rs5751348 (NT G > T) is located in a RUNX1 transcription factor binding region that controls the expression of P1 with chromosomal location chr22:43,113,793G associated with the P1+ phenotype and chr22:43,113,793 T with P1-.

The Xg^a antigen was first described in 1962 in a multiply transfused male of European ancestry whose serum reacted at different frequencies to RBCs from males and females, indicating that the expression of Xg^a was X-linked, with a sex-biased distribution: Xg(a+) 66% males/89% females, and Xg(a-) 34% males/11% females.⁸ In 1994, it was shown that Xg^a was expressed on the protein product of *PBDX* (renamed *XG*) in a manner that suggested that antigen expression was controlled by the presence or absence of Xg protein,⁹ but the mechanism remained unknown. Recently, Moller et al.² and Yeh et al.³ showed that the SNP rs311103 (NT G > C) is located in a GATA-1 transcription factor binding region in intron 1 of *XG*, and controls the expression of the protein and the Xg^a antigen with genomic coordinate chrX:26,66,384G associated with Xg(a+) and chrX:26,66,384C associated with a Xg(a-) phenotype.

We recently created and validated an automated blood group antigen typing software (bloodTyper, Partners HealthCare) for translating whole genome sequencing (WGS) data to predict RBC and platelet antigen phenotypes and validated the performance of whole genomes with conventional serologic and SNP typing for the common antigens.¹⁰⁻¹² As part of those analyses, we also performed P1 and Xg^a serologic typing for a subset of samples with the goal of identifying the basis of P1 and Xg^a expression by correlating the serologic typing with whole genome data. Here, we present a bioinformatics analysis of P1 and Xg^a expression. We also describe updates to our automated typing algorithm to type for the P1

and Xg^a antigens, including limitations that make it possible to molecularly type Xg^a from females only.

MATERIAL AND METHODS

Serologic typing

With approval from the Partners HealthCare Human Research Committee (Institutional Review Board) and informed consent from participants, blood samples for RBC isolation were collected in ethylenediaminetetraacetic acid and RBC serologic antigen typing was performed according to standard tube methods,¹³ and as previously described.¹⁰⁻¹² Commercially available serologic typing reagents were used to type for P1 (Bio-Rad), and human source anti-Xg^a was used to type for Xg^a.

Whole genome sequencing

With approval from the Partners HealthCare Human Research Committee (Institutional Review Board) and informed consent from participants, blood samples for DNA isolation were collected in blood collection tubes (PAXgene, PreAnalytiX GmbH) and genomic DNA was isolated from WBCs by standard methods. For quality control, a genotyping array was performed in parallel to confirm identity and lack of sample inversion during WGS workflow. Another blood sample was also genotyped to serve as an independent verification of identity.

Polymerase chain reaction free WGS was performed by the Clinical Laboratory Improvement Amendments-certified, College of America Pathologists-accredited Illumina Clinical Services Laboratory (San Diego, CA) using paired-end 100-base pair reads of DNA fragments with an average length of 300 base pairs on the Illumina HiSeq platform and sequenced to at least 30 \times average depth of coverage.¹⁴ Sequence read data was aligned to the human reference sequence (GRCh37/hg19) using Burrows-Wheeler Aligner 0.6.1-r104.¹⁵ Data from sequencing was analyzed, interpreted, reported, and returned to participants as part of the MedSeq Project, a randomized clinical trial of WGS in primary care practice.¹⁶⁻¹⁹

Whole genome based blood typing

P1 and Xg^a antigen NT changes were added to our allele database (<http://bloodantigens.com>) and the custom typing software (bloodTyper)¹² was used to predict P1 and Xg^a phenotypes from the whole genomes analyzed as part of the MedSeq Project. In brief, variant calls for *XG* and *A4GALT* genes and promoter regions were made using Genomic Analysis Tool Kit version 2.3-9-gdcdccbb (Broad Institute) and saved as a variant calling format file (.vcf) showing differences between the WGS data and the reference genome.²⁰ Sequencing coverage was extracted from the alignment file using BEDTools version 2.17.0.²¹ A high-performance visualization tool (Integrative Genomics Viewer, Broad Institute)²² was used as needed to verify coverage and sequence identity. Antigen typing using bloodTyper was performed at the relevant genome positions using a 4 \times sequence read depth of coverage calling cutoff.

Software and data availability

The MedSeq Project genomes are available through dbGaP under study accession phs000958. The curated RBC antigen allele database used in this study is available at <http://bloodantigens.com>. The code used to search the variant calling files for the antigen-associated changes using both the frequency and serology approaches is available at <http://lanelab.org/data>.

RESULTS

Genetic search approaches

The genetic basis for P1 and Xg^a expression were identified using 113 whole genomes from the MedSeq Project with paired serologic typing (Fig. 1). Two different but complementary approaches (frequency and serology) were used to search the the *A4GALT* and *XG* genes, including coding exons, introns, and upstream promoter regions for single-NT changes and small insertions/deletions that correlated with antigen expression.

In the frequency approach, heterozygous and homozygous NT variations over the *A4GALT* and *XG* gene and promoter regions were identified in the genomes. Each change was then individually evaluated by simulating the sample antigen type using the following rules: antigen negative if homozygous for the change or antigen positive if heterozygous for this

change or homozygous for another change. The resulting antigen-positive and antigen-negative population frequency was then calculated for each change identified. The NT changes were filtered to include only those in the range of the known antigen frequency.

In the serologic approach, serologic typing results were correlated with NT changes over the *A4GALT* and *XG* gene and promoter regions, independent of the frequency approach described above. By assuming that antigen-negative individuals resulted from the same recessive homozygous change, a starting list of possible NT changes was created by identifying common homozygous changes. These were then filtered by removing any homozygous changes also present in antigen-positive individuals. To account for the possibility of a serologic typing error or other sample-specific issues, the searches were performed multiple times with each sample excluded from the analysis. Therefore, if a serologic typing were incorrect, there would be a search in which this erroneous information would not adversely affect the results.

Identification of the basis of P1 from whole genomes

Using a frequency approach, the genetic basis of P1 expression was searched in 113 whole genomes over a 29-kb region (chr22:43,088,126 – chr22:43,117,175) including the *A4GALT* promoter and gene. The possible P1 associated

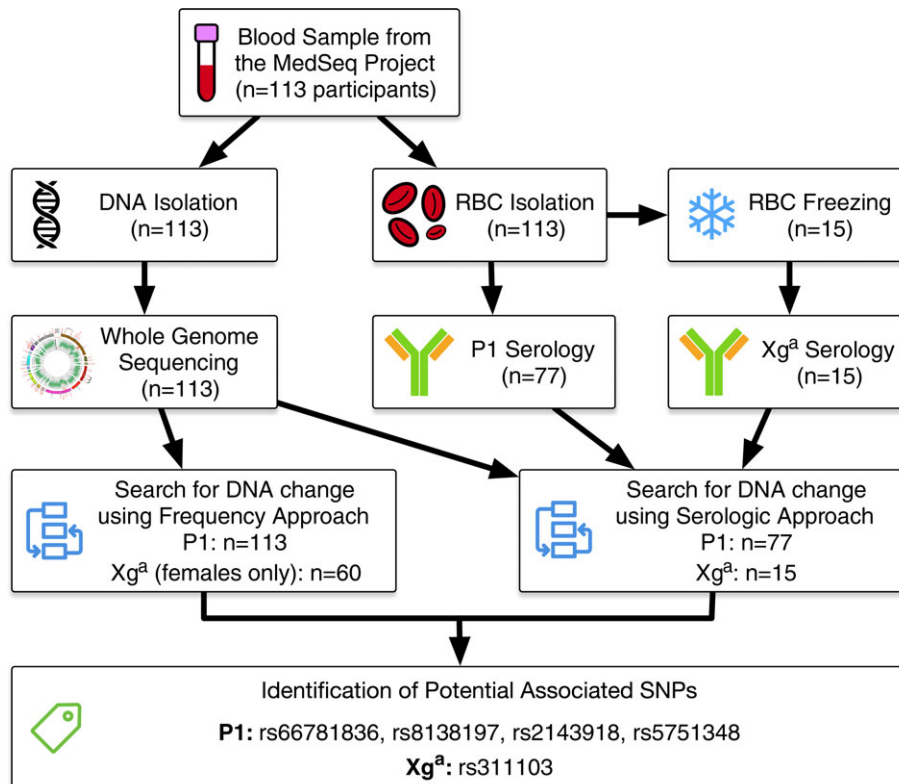


Fig. 1. WGS based approach for determination of molecular basis for P1 and for Xg^a. [Color figure can be viewed at wileyonlinelibrary.com]

changes were filtered to include only those with a frequency between 15% and 25%, close to the known P1- antigen frequency of 21% in individuals of European ancestry as the recorded ancestry for most MedSeq samples. This reduced

the number of possible responsible NT positions from 1196 to only 29 NT changes (Fig. 2A).

Using a serologic approach, P1 expression was correlated between genomic nucleotide changes over the same

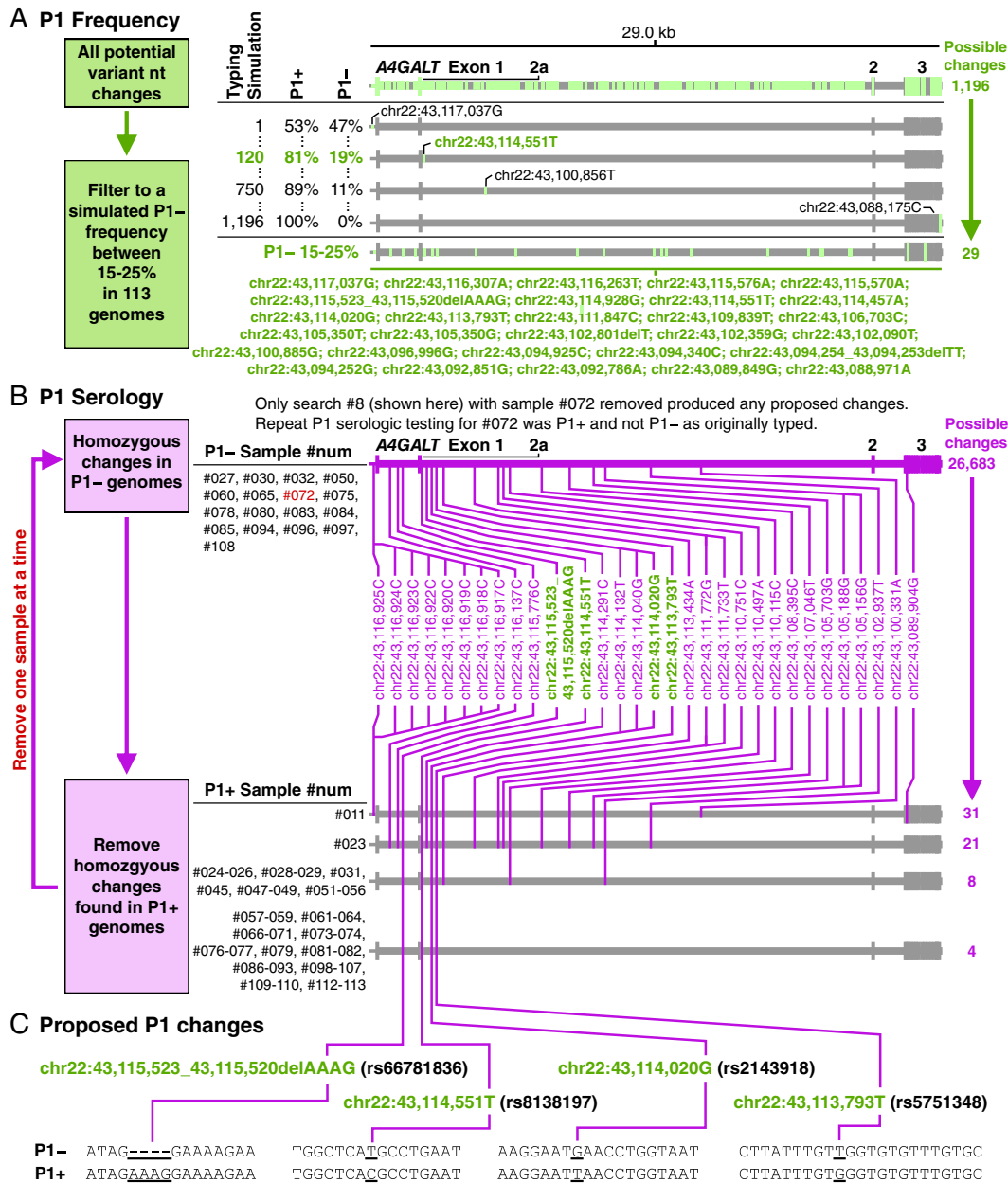


Fig. 2. Identification of the P1 genetic basis. Two approaches used to identify *A4GALT* changes associated with P1 expression. (A) The frequency approach used whole genome sequences of the *A4GALT* region to correlate NT changes with expected P1 antigen frequency. The percentages represent the antigen phenotype frequencies for each NT change when used to simulate typing for the antigen from whole genomes. (B) The serology approach used whole genome sequences of the *A4GALT* region to correlate NT changes with serologic typing. Changes also found in the frequency approach shown in (A) are colored green. (C) Potential NT changes responsible for P1 expression designated by chromosomal coordinates and by rs database numbering. Note: *A4GALT* is transcribed in reverse to its direction in the reference genome, as such larger genomic positions correspond to smaller gene positions. The NT bases have been reverse complimented from the human reference genome. [Color figure can be viewed at wileyonlinelibrary.com]

region as above with the RBC phenotypes of 77 individuals (17 P1- and 60 P1+). Analysis of homozygous NT changes common to P1- individuals identified 26,683 possible NT variants, which were then filtered to remove homozygous changes also found in P1+ individuals (Fig. 2B). This initial analysis did not result in any correlated NT changes. However, this could have occurred due to a P1 serologic phenotype error in one or more samples, which could have caused the actual P1-associated NT changes to be incorrectly filtered. As mentioned above, to account for the possibility of serologic testing or transcription error, multiple searches were performed in which each sample was excluded from the analysis. When P1- participant 072 was excluded from the analysis in this manner, four potential P1-associated changes were identified (Fig. 2C): chr22:43,115,523_43,115,520AAAG/delAAAG (rs66781836); chr 22:43,114,551C/T (rs8138197); chr22:43,114,020 T/G (rs2143918); and chr22:43,113,793G/T (rs5751348). All four of these sites were also identified in the frequency approach above. A follow-up sample from participant 072 was obtained and the RBCs typed as P1+, confirming that the initial P1- serologic typing for that individual was incorrect.

Identification of the basis of Xg^a in whole genomes

Because males have only one X chromosome, sequence analysis to identify hemizygous NT changes could theoretically lead to quicker identification of the associated Xg^a genetic change than analysis of female genomes. Unfortunately, the XG promoter and part of the XG gene are located in the homologous PAR1 region shared by both the X and Y chromosomes.²³ This causes male Y chromosome sequences to misalign to the X chromosome, and thus male samples contained regions of misplaced reads that interfered with analysis. In particular, some NT positions incorrectly appeared heterozygous where misplaced Y sequences contain SNPs that differ from the X chromosome. Therefore, only males presenting with homozygous NT positions and females were considered in the serologic approach.

Using the frequency approach, the genetic basis of Xg^a expression was examined in 60 female genomes over a 75.2-kb region including the XG gene and promoter region (chrX:2,659,351 – chrX:2,734,541). The potential Xg^a nucleotide associations were filtered to include those that would result in a frequency close to the known 11% female Xg(a-) phenotype frequency, which reduced the number of possible NT variants from 1917 to only 82 changes (Fig. 3A).

Using the serologic approach, the genetic basis for Xg^a expression was correlated between nt changes over the same region indicated above with the RBC phenotypes of 15 individuals, two Xg(a-) females and 12 Xg(a+) males and females were included in the analysis. One Xg(a-) male was excluded as uninformative for the reasons described above. Analysis of homozygous NT changes common to Xg(a-) individuals identified 74,791 possibilities (Fig. 3B), which were filtered to

remove homozygous changes present in Xg(a+) individuals. This identified just one potential variation associated with Xg(a-) phenotype, chrX:2,666,384C (rs311103), located in a GATA-1 binding region (Fig. 3C). This NT change was also identified in the frequency approach.

To further investigate and differentiate true X chromosome NT variants and misplaced Y chromosome NT variation, the serologic typing of male samples was correlated with the DNA sequence aligned to position chrX:2,666,384 (Fig. 3D). The analysis shows that the Y chromosome PAR1 region sequences indeed misaligned to the X chromosome Xg^a associated region and can differ in sequence (Fig. 3D, #075, 110, 112). This confounds molecular typing in male samples that appear to be heterozygous. Hence, molecular typing of Xg^a in males is not readily possible unless they are homozygous at this position.

bloodTyper

We updated our curated antigen allele database (<http://bloodantigens.com>) with the newly confirmed Xg^a and P1 alleles, and verified that bloodTyper could correctly type P1 and Xg^a from the whole genomes analyzed within the Med-Seq Project with paired serology. As discussed above, bloodTyper restricts Xg^a typing to females.

DISCUSSION

We have demonstrated an extensible and reusable framework for antigen allele discovery and verification from whole genomes. As a proof of principle, we used WGS data and limited paired serologic typing to identify NT changes correlated with P1 and Xg^a expression. This analysis was done prior to the recent publications detailing the role of RUNX1¹ GATA-1^{2,3} and transcription factor binding in expression of these antigens, and is independent confirmation of those reports. Importantly, since the entire genomic sequence was available for analysis, we could rule out the possibility of other common NT changes controlling P1 and Xg^a expression in the *A4GALT* and *XG* genes and upstream promoter regions. As the genetic changes associated with P1 and Xg^a antigen expression are located outside of the coding exons, a whole genome-based approach was particularly important in determining their identity. By automating the search analysis, we could quickly perform multiple iterative searches with each sample excluded from the analysis to successfully account for any sample discrepancy. We have now updated our curated antigen-allele database and companion typing software to provide whole genome-based typing for Xg^a and P1.

Although the combined frequency and serologic approach for P1 expression identified four NT changes, this could simply be due to linkage disequilibrium, which, in theory, could be addressed with a larger sample size. Indeed, a recent publication identified these same four NT changes associated

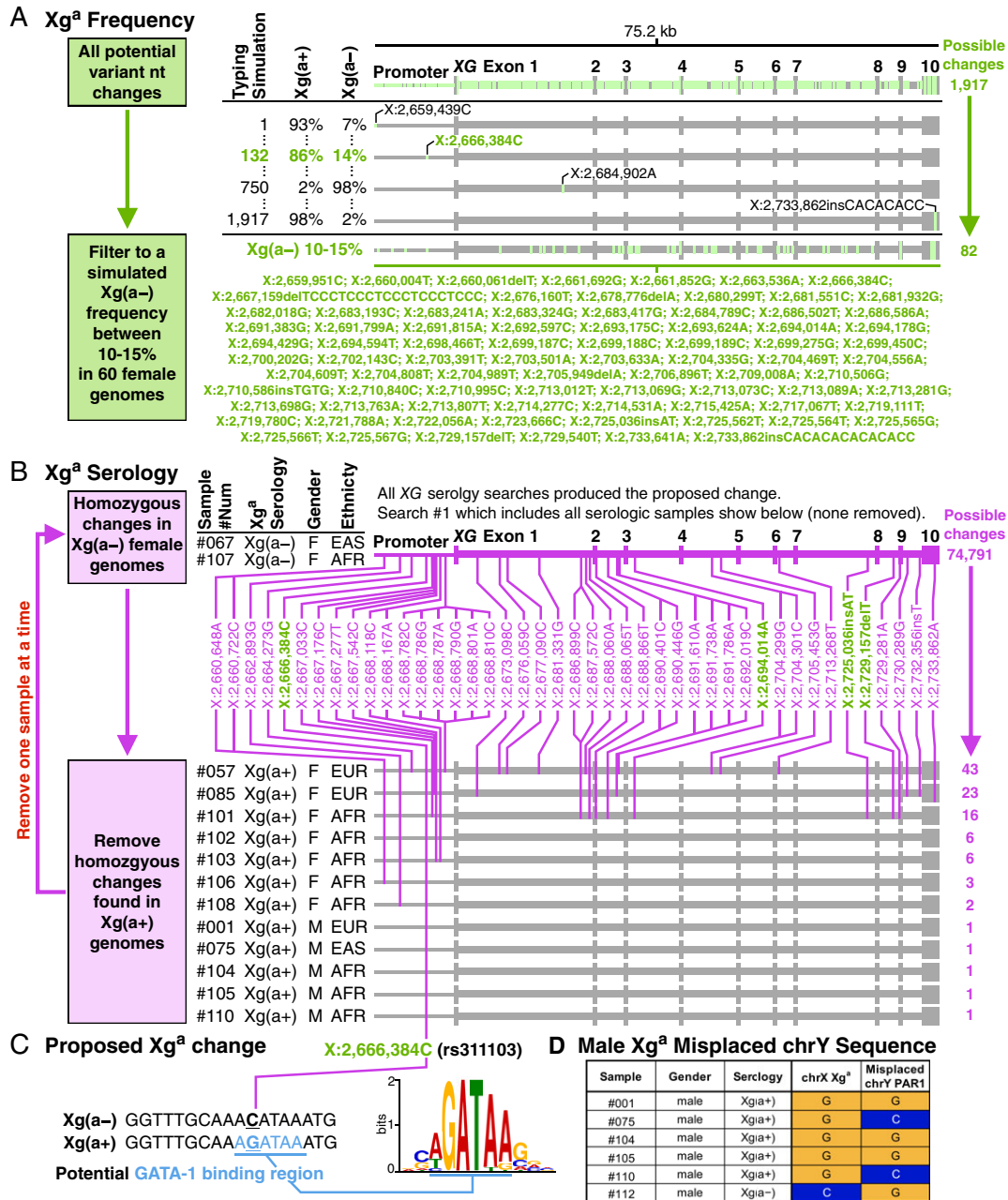


Fig. 3. Identification of the Xg^a genetic basis. Two approaches used to identify XG changes associated with Xg^a expression. (A) The frequency approach used whole genome sequences of the XG region to correlate nt changes with expected Xg^a antigen frequency. The percentages represent the antigen phenotype frequencies for each NT change when used to simulate typing for the antigen from whole genomes. (B) The serology approach used whole genome sequences of the XG region to correlate NT changes with serologic typing. Changes also found in the frequency approach shown in (A) are colored green. (C) Potential NT change responsible for Xg^a expression. (D) Xg^a NT changes in males and the corresponding misplaced chromosome Y PARI NT changes. [Color figure can be viewed at wileyonlinelibrary.com]

with P1, but one (rs66781836) was later excluded in an uncommon sample with haplotype recombination between the possible NT changes.⁷ To optimize exclusion among strongly linked SNPs, mining of large-scale genomic data sources for examples where one or more of the linked NT changes occur independently and performing targeted follow-up serologic typing

could be employed. Importantly, the potential P1 associated NTs found here include rs5751348, which has recently been reported to contain a RUNX1 transcription factor binding region controlling P1 expression,¹ along with three other NT changes located outside of the RUNX1 binding region, but in linkage disequilibrium with it.

The combined frequency and serologic approach was successful in determining the molecular basis of Xg^a expression as it identified just one potential NT change, rs311103. This is the same change recently reported to contain a GATA-1 transcription factor-binding region controlling Xg^a expression.^{2,3} We determined that genotyping of Xg^a from males will prove difficult due to the homologous PAR1 region shared by chromosomes X and Y. Determination of Xg^a in males by genomic methods will require long read sequencing or long-range polymerase chain reaction of a 33.2-kb region spanning the Xg^a SNP position (chrX:2,666,384) and chromosome X positions located outside of the PAR1 region (around chrX:2,699,625).

Historically, identifying the molecular basis of new blood groups was accomplished using genetic segregation of NT changes with antigen phenotype in related individuals. In this study, we present a new paradigm using whole genomes with paired, and often limited, serologic typing data as illustrated by identification the basis of Xg^a with less than 14 typed individuals combined with sequence data of 77 unrelated individuals. Although the frequency approach was only able to filter to 29 *A4GALT* and 82 XG NT change, even this level of enrichment could be of value. For example, if RBCs had not been available for serologic typing, the NT changes identified by the frequency approach could have been used to design inexpensive targeted SNP assays and tested on samples with paired RBC serologic testing. This type of analysis will accelerate as large-scale genomic sequencing becomes commonplace, by allowing sequence data analysis from tens of thousands of individuals at present, and millions within the next decade.

As new antigens are described, it is critical to have adequate evidence for allele-antigen associations, but this is challenging because expression systems are not readily available for many blood groups. In addition, bioinformatic analysis can be used to determine if potential molecular changes are located in putative transcription factor-binding regions or other regulatory regions along with evaluation of expression using predictive models or large-scale experimental protein expression data sets. This would provide a potential robust hypothesis for association of the SNP with the phenotype; however, direct experimental confirmation should be done when possible. As such, large-scale sequencing data sets with paired serologic typing could provide a powerful tool to rapidly identify the effect of novel allelic changes on serologic phenotype.

Because it is already common for large biobanks to archive DNA and serum, they should partner with blood bank and donor centers to freeze RBCs for future antigen-allele association discovery and confirmation. For example, the Xg^a analysis performed here was possible only because a subset of genomic sequencing participants had RBC samples frozen for future serologic typing. If frozen RBCs for serologic typing were available from a sufficiently large genomic data set, it might be possible to find the molecular basis of a low- or high-frequency antigen if even just one

high-frequency antigen-negative or low-frequency antigen-positive individual could be identified.

In summary, we have developed an analysis pipeline to identify the NT changes responsible for RBC antigen expression from whole genomes with paired serologic typing. This approach independently confirms the recently reported NT changes responsible for P1 and Xg^a antigen expression. Additionally, this approach should allow for rapid and comprehensive antigen discovery as large-scale genomic sequencing data sets are created over the coming years.

MEMBERS OF THE MEDSEQ PROJECT (INCLUDING AUTHORS LISTED ABOVE)

Members of the MedSeq Project are as follows: David W. Bates, MD; Carrie Blout, MS, CGC; Kurt D. Christensen, PhD; Allison L. Cirino, MS; Robert C. Green, MD, MPH; Carolyn Y. Ho, MD; Joel B. Krier, MD; William J. Lane, MD, PhD; Lisa S. Lehmann, MD, PhD, MSc; Calum A. MacRae, MD, PhD; Cynthia C. Morton, PhD; Denise L. Perry, MS; Christine E. Seidman, MD; Shamil R. Sunyaev, PhD; Jason L. Vassy, MD, MPH, SM; Erica Schonman, MPH; and Tiffany Nguyen, Eleanor Steffens, and Wendi Nicole Betting, Brigham and Women's Hospital and Harvard Medical School; Samuel J. Aronson, ALM, MA; Ozge Ceyhan-Birsoy, PhD; Matthew S. Lebo, PhD; Kalotina Machini, PhD, MS; Heather M. McLaughlin, PhD; Danielle R. Azzariti, MS; Heidi L. Rehm, PhD; Ellen A. Tsai, PhD, Partners Healthcare Personalized Medicine; Jennifer Blumenthal-Barby, PhD; Lindsay Z. Feuerman, MPH; Amy L. McGuire, JD, PhD; Kaitlyn Lee, Jill O. Robinson, MA; Melody J. Slashinski, MPH, PhD, Baylor College of Medicine, Center for Medical Ethics and Health Policy; Pamela M. Diamond, PhD, University of Texas Houston School of Public Health; Kelly Davis and 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; and Isaac Kohane, MD, PhD, and Sek Won Kong, MD, Boston Children's Hospital.

ACKNOWLEDGMENTS

The authors thank the staff and participants of the MedSeq Project, as well as the staff of the Brigham and Women's Blood Bank.

CONFLICT OF INTEREST

The authors have disclosed no conflicts of interest.

REFERENCES

1. Westman JS, Stenfelt L, Vidovic K, et al. Allele-selective RUNX1 binding regulates P1 blood group status by transcriptional control of *A4GALT*. *Blood* 2018;131:1611-6.

2. Moller M, Lee YQ, Vidovic K, et al. Disruption of a GATA1-binding motif upstream of XG/PBDX abolishes Xg(a) expression and resolves the Xg blood group system. *Blood* 2018;132:334-8.
3. Yeh CC, Chang CJ, Twu YC, et al. The molecular genetic background leading to the formation of the human erythroid-specific Xg(a)/CD99 blood groups. *Blood Adv* 2018;2:1854-64.
4. Landsteiner K, Levine P. Further observations on individual differences of human blood. *Exp Biol Med* 1927;24:941-2.
5. Steffensen R, Carlier K, Wiels J, et al. Cloning and expression of the histo-blood group Pk UDP-galactose: Galbeta-4G1cbeta1-cer alpha1, 4-galactosyltransferase. Molecular genetic basis of the p phenotype. *J Biol Chem* 2000;275:16723-9.
6. Thuresson B, Westman JS, Olsson ML. Identification of a novel A4GALT exon reveals the genetic basis of the P1/P2 histo-blood groups. *Blood* 2011;117:678-87.
7. Lai YJ, Wu WY, Yang CM, et al. A systematic study of single-nucleotide polymorphisms in the A4GALT gene suggests a molecular genetic basis for the P1/P2 blood groups. *Transfusion* 2014;54:3222-31.
8. Mann JD, Cahan A, Gelb AG, et al. A sex-linked blood group. *Lancet* 1962;1:8-10.
9. Ellis NA, Tippett P, Petty A, et al. PBDX is the XG blood group gene. *Nat Genet* 1994;8:285-90.
10. Lane WJ, Westhoff CM, Uy JM, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* 2016;56:743-54.
11. Baronas J, Westhoff CM, Vege S, et al. RHD zygosity determination from whole genome sequencing data. *J Blood Disord Transfusion* 2016;7:365.
12. Lane WJ, Westhoff CM, Gleadall NS, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol* 2018;5:e241-51.
13. Fung MK, Eder AF, Spitalnik SL, et al., editors. Technical manual. 19th ed. 19th ed. Bethesda (MD), American Association of Blood Banks (AABB); 2017.
14. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53-9.
15. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589-95.
16. Vassy JL, Lautenbach DM, McLaughlin HM, et al. The MedSeq project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* 2014;15:85.
17. McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, et al. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Med Genet* 2014;15:134.
18. Vassy JL, Christensen KD, Schonman EF, et al. The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. *Ann Intern Med* 2017;167:159-69.
19. Christensen KD, Vassy JL, Phillips KA, et al. Short-term costs of integrating whole-genome sequencing into primary care and cardiology settings: a pilot randomized trial. *Genet Med* 2018.
20. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-303.
21. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841-2.
22. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178-92.
23. El-Mogharbel N, Graves JA. X and Y chromosomes: homologous regions. Hoboken, NJ: John Wiley & Sons, Ltd; 2006. 