RBCeq: A robust and scalable algorithm for accurate genetic blood typing

Sudhir Jadhao,[a,*] Candice L. Davison,[c] Eileen V. Roulis,[c,g] Elizna M. Schoeman,[c] Mayur Divate,[a] Mitchel Haring,[e] Chris Williams,[e] Arvind Jaya Shankar,[a] Simon Lee,[a] Natalie M. Pecheniuk,[f] David O Irving,[d] Catherine A. Hyland,[c,g] Robert L. Flower,[c,g] and Shivashankar H. Nagaraj,[a,b]

[a]Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, Queensland 4059, Australia
[b]Translational Research Institute, Brisbane, Australia
[c]Australian Red Cross Lifeblood Research and Development, Brisbane, Queensland, Australia
[d]Research and Development, Australian Red Cross Blood Service, Sydney, New South Wales, Australia
[e]Office of eResearch, Queensland University of Technology, Brisbane, Queensland 4059, Australia
[f]School of Biomedical Sciences, Queensland University of Technology, Brisbane, Australia
[g]Faculty of Health, Queensland University of Technology, Brisbane, Australia

EBioMedicine 2021;76: 103759
Published online xxx
https://doi.org/10.1016/j.ebiom.2021.103759

Summary

Background While blood transfusion is an essential cornerstone of hematological care, patients requiring repetitive transfusion remain at persistent risk of alloimmunization due to the diversity of human blood group polymorphisms. Despite the promise, user friendly methods to accurately identify blood types from next-generation sequencing data are currently lacking. To address this unmet need, we have developed RBCeq, a novel genetic blood typing algorithm to accurately identify 36 blood group systems.

Methods RBCeq can predict complex blood groups such as RH, and ABO that require identification of small indels and copy number variants. RBCeq also reports clinically significant, rare, and novel variants with potential clinical relevance that may lead to the identification of novel blood group alleles.

Findings The RBCeq algorithm demonstrated 99·07% concordance when validated on 402 samples which included 29 antigens with serology and 9 antigens with SNP-array validation in 14 blood group systems and 59 antigens validation on manual predicted phenotype from variant call files. We have also developed a user-friendly web server that generates detailed blood typing reports with advanced visualization (https://www.rbceq.org/).

Interpretation RBCeq will assist blood banks and immunohematology laboratories by overcoming existing methodological limitations like scalability, reproducibility, and accuracy when genotyping and phenotyping in multi-ethnic populations. This Amazon Web Services (AWS) cloud based platform has the potential to reduce pre-transfusion testing time and to increase sample processing throughput, ultimately improving quality of patient care.

Funding This work was supported in part by Advance Queensland Research Fellowship, MRFF Genomics Health Futures Mission (76,757), and the Australian Red Cross LifeBlood. The Australian governments fund the Australian Red Cross Lifeblood for the provision of blood, blood products and services to the Australian community.

Keywords: Alloimmunization; Blood groups; Next-generation sequencing; Population genomics; Red blood cell antigens; Single nucleotide polymorphism; Transfusion

Introduction

Red blood cells (RBC) are the most commonly transfused blood product. Globally, approximately 85 million RBC units/year are transfused to treat hematological conditions like severe anemia, leukemia, stem cell transplants, severe hemorrhage, and sickle cell disease (SCD).[1] While these transfusions are essential, patients requiring repetitive transfusion are at a high risk of alloimmunization that can lead to delayed or acute hemolytic transfusion reactions (HTRs), fetal anemia, and complications during pregnancy.[2] Currently, RBCs are matched for the ABO, RH and K blood group systems, consequently often resulting in sensitization to other non−self RBC antigens.[3] Other blood types can

*Corresponding author.
E-mail address: shiv.nagaraj@qut.edu.au (S. Jadhao).

Research in context

Evidence before this study

Blood transfusion is an important life-saving treatment in modern medicine. However, patients receiving multiple transfusions are at risk of alloimmunization. Next-generation sequencing (NGS)-based extended blood group profiling of blood donors and patients promises to facilitate extended matching for additional blood groups and reduce problems caused by alloimmunization. We have previously shown that the NGS data can be used to overcome serology limitations. We searched PubMed until January 2018, using the search terms "blood group genotyping", "blood group genetics", "red blood cell antigen", "transfusion genomics", and "blood group sequencing" and found two command line tools present called BOOGIE and bloodTyper that have been used to interpret blood groups from NGS data. We present a web-based user-friendly comprehensive tool capable of blood group profiling for users with limited computational background.

Added value of this study

Our novel algorithm, RBCeq, can rapidly analyze NGS data to provide an extended report of blood group phenotype and novel blood group alleles with potential clinical relevance. The tool provides a user-friendly platform for genotyping and phenotyping of blood groups in multi-ethnic populations. The algorithm was developed on a foundational, customised database of 36 blood groups consisting of 1502 alleles (+ two transcription factors consisting of 57 alleles) and validated on 100 MedSeq samples and 245 Australian Indigenous cohort with 99·07% accuracy. RBCeq reports blood groups with variants, coverage statistics and clinically significant annotations through advanced visualizations.

Implications of all the available evidence

RBCeq has the potential to improve transfusion safety for multi-ethnic population by assisting blood banks and immunohematology laboratories to access and extract predicted blood group profiles from NGS data.

be matched for if donor blood is extensively typed to minimize alloimmunization in transfusion dependant patients such as those with SCD or Thalassemia. Among the known 345 red cell antigens related to the 43 known blood group systems, many alloantibodies arise from other blood groups such as MNS, FY, JK VEL, LAN and AUG which are considered clinically significant.[4–6] Within these blood groups, antigens are population-specific, and are subject to varying rates of polymorphisms, further complicating transfusion safety efforts.[7,8] Clinical interpretation of genetic blood typing variants based upon the guidelines of the American College of Medical Genetics (ACMG) is recommended.[9]

RBC phenotyping by hemagglutination is the most used gold standard method for blood group identification and matching. Recently, DNA microarray platforms and single-nucleotide polymorphism (SNP) genotyping have enabled large scale blood group typing.[10] However, while effective for well-characterized and common variants, these traditional serological and/or molecular methods for blood group typing are insufficient for the full characterization of blood group antigens that are rare, weakly detectable, recombinant (RH, MNS), partial, or novel.[11] Next-Generation sequencing (NGS) technologies promise to overcome the limitations of serological and SNP based molecular techniques and have the capability to characterize blood group antigens in genetically diverse populations.[12–14] NGS technology is able to predict extended or complete blood group profiles without previous molecular background knowledge. It enables the description of all SNPs, indels, and large structural variants to identify rare, null, or novel genotypes.

Accurately predicting blood group phenotypes from NGS data requires detailed immuno-genetic knowledge. Multiple genotypes may result in the same phenotype (as with the ABO, MNS, LE, and XG groups), and not all blood group antigens direct primary gene products (including the ABO, LE, and H groups). Despite the clear value of applying NGS-based blood group characterization methods to blood bank services, at present there is a lack of user-friendly computational tools able to deliver accurate blood typing. Two command line tools are available at present: BOOGIE[15] and bloodTyper.[16] However, there's an unmet need for a user-friendly tool with serverless architecture for users with limited computational background. Thereby RBCeq was created to provide an accessible gateway to explore NGS data and characterize blood groups and novel alleles.

One of the major advantages to NGS-based blood typing is the potential to discover novel variants.[17] The clinical utility of integrating NGS based blood typing in a clinical immunohematology reference laboratory has been demonstrated to be beneficial at resolving complex serology problems arising from the novel or rare alleles altering or silencing blood group expression.[18] To efficiently analyze NGS data in the context of transfusion medicine applications, we have developed a novel algorithm called RBCeq and created a secure, comprehensive web platform with an easy-to-use interface. RBCeq can characterize not only known blood group alleles but also putative novel variants capable of reducing and silencing antigen expression. As a web server-based blood group genotyping software, RBCeq addresses both computational and storage challenges associated with the processing of large NGS datasets including whole genome sequences (WGS), whole exome sequences (WES) and targeted sequencing (TES) datasets.

Methods

Ethics

South East Queensland Indigenous TES data collection was conducted in collaboration with the Carbal Medical Services in Toowoomba (https://carbal.com.au/), in consultation with the Indigenous Community Advisory Committee and under the approval of the Australian Red Cross Lifeblood Human Research Ethics Committee (2018#17/ QUT Ethics 2,021,000,118).

Curation and construction of a new database of blood group antigen alleles

A comprehensive blood group allele database was created using multiple manually curated data sources such as the International Society of Blood Transfusion (ISBT),[19] Blood Group Antigen FactsBook,[20] Human Blood Groups,[21] Erythrogene,[22] and the RhesusBase.[23] The coordinates of known blood group antigen variants recorded in these databases were provided in the conventional cDNA reference sequence form. At the time of the design, the new database comprised 36 blood group genes and two transcription factors (TF). Corresponding blood group genotype coordinates for the hg19/GRCh37 genome were identified and validated using the Transvar,[24] NCBI Clinical Remap (https://www.ncbi.nlm.nih.gov/genome/tools/remap), Ensembl Variant Effect Predictor (VEP),[25] and UCSC[26] resources. During the curation process, alleles with inconsistencies between reference nucleotide change, positions and amino acid identities were detected (Supplementary Table 1). These alleles were manually curated and validated using NCBI (https://www.ncbi.nlm.nih.gov/) and UCSC to ensure non-redundancy of the data and uniformity of allele names and corresponding phenotypes. The database was further improved using the information from previous publications[17,27] and extensive literature mining. In total, the resultant backend database is representative of approximately 1502 alleles from 44 genes, and two transcription factors (GATA binding protein 1: GATA1 and Kruppel like factor 1: KLF1) with 57 alleles, encoding blood groups arising from 36 blood group systems recognized by the ISBT.

Variant calling

For variant calling, BAM files are processed in accordance with GATK4 best practices[28] which includes first pre-processing with BaseRecalibrator, ApplyBQSR, and then variant calling using HaplotypeCaller. Variant calling and haplotype phasing were conducted for a restricted set of 44 blood group genes and two transcription factors (GATA1 and KLF1) (Supplementary Table 2).

Characterization of novel blood group alleles with potential clinical relevance

The filtered high-quality variants that were not mapped to known blood group alleles were queried against the ClinVar database.[29] The mapped variants are reported as clinically significant variants, while the remaining variants were processed for rare variant analysis. RBCeq checks the frequency of the variant in the gnomAD database (~143,000 genomes, v2.1.1). In case the frequency is less than 0·05 (or less than a user-defined threshold) in any of the population groups (African, American, European, South Asian and East Asian) and the respective variant is nonsynonymous or a splice-site variant, then the variant was reported as a rare variant. The remaining variants that are not considered clinically relevant (listed by the ClinVar database) or rare will be processed for novel variant analysis. To predict novel SNVs, RBCeq uses six independent computational tools (SIFT, Polyphen2, MutationTaster2, FATHMM, PROVEAN, and CADD) to assess the impact of genetic variants on protein structure and function. If any one of these tools determined the variant to be deleterious, and it is nonsynonymous or a splice-site variant, then the variant was reported as a novel blood group variant with potential clinical relevance (Figure 1).

Datasets

To evaluate the overall performance of RBCeq, the following datasets were used as input data.

Proof of Principle Algorithm Development dataset: Initial development was undertaken on 18 well characterized samples for which serological, SNP-array and Targeted exome sequencing data have been previously reported[18,30] (Supplementary Table 3). In addition, 40 previously published complex blood group serology investigations with TES, from the Australian red cross lifeblood red cell reference laboratory, have been reported with prediction of phenotype from variant calls. (Supplementary Table 4).[17,18,30]

South East Queensland Indigenous TES data: This dataset comprised 244 targeted blood group exome sequencing samples with serologically validated phenotypes for ABO, D, C, c, E, e, K and k blood groups.[3]

MedSeq project: 110 whole-genome sequencing samples (30X) from the MedSeq Project randomized controlled trial (accession number phs000958) were accessed through dbGaP authorized access.

The 1000 genomes (1000 G) project: The 1000 G project includes 2504 whole-genome sequencing (WGS) samples from 26 population groups classified into five subpopulations. 2504 WGS bam files were accessed through the 1000 G project FTP server.

Erythrogene: Erythrogene is a database of the predicted blood group genotype information for 2504 WGS samples from the 1000 G Project. The blood group genotype information was accessed through

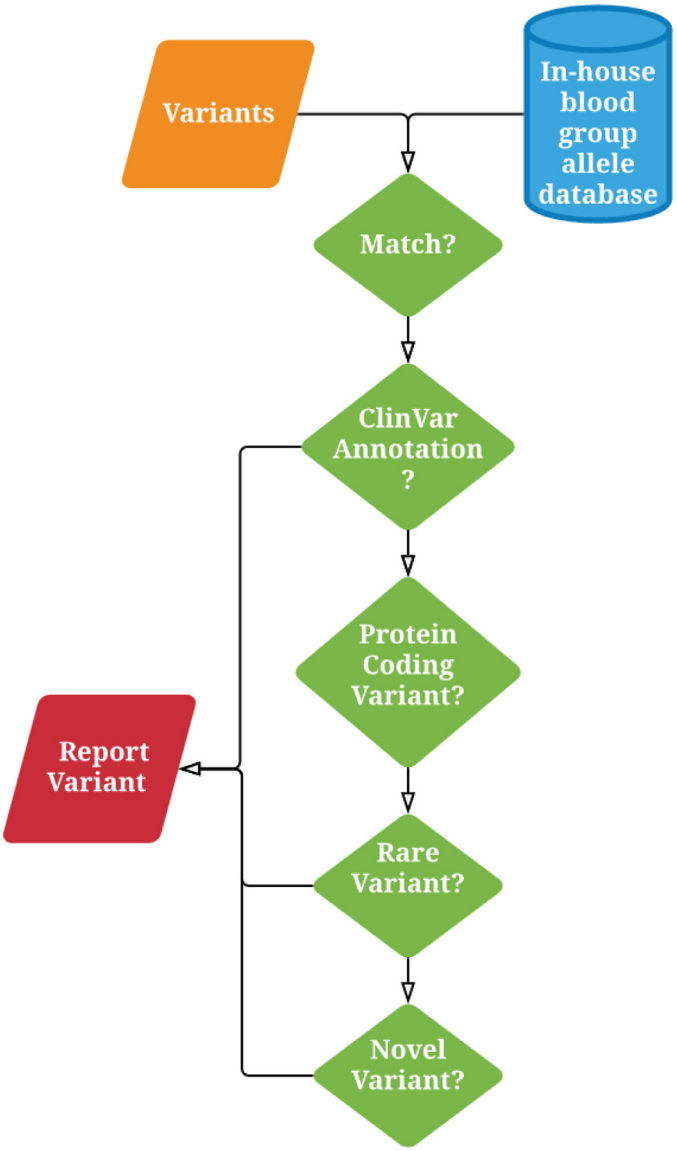


Figure 1. Workflow for the annotation of clinically significant, rare, and novel blood group variants.

www.erythrogene.com. The blood group genotype information from Erythrogene was used to compare the RBCeq genotype prediction from the 1000 G dataset.

Webserver

Implementation. We developed a webserver to enable the seamless analysis of blood group profiles from NGS data. The web server was developed using Apache and PHP and is hosted on Amazon Web Services (AWS), thus making it scalable across thousands of samples. The interactive visualizations are implemented through d3.js and c3.js libraries. RBCeq uses EC2 container service-Docker management on AWS to serve several different concurrent users/jobs. The user-based login allows users to obtain blood group profiles independently and seamlessly in parallel with many simultaneous sessions using an organized queue-based system. It also enables users to save their outputs and to access their project at a later date or from a different place, enabling user mobility and collaboration. User data uploaded into the server is only utilized for RBCeq analyses, and will be stored for six months, after which it will be erased. Users can also delete their data sooner or export the data from RBCeq in a user-friendly Excel spreadsheet containing run parameters, dates, times, input file names, uploaded data types, and complete blood group profiles for archiving purposes. Thorough website penetration testing and social engineering of

the RBCeq webserver has been undertaken to assess for vulnerabilities and potential security risks associated with the environment and the technology used in building the server and has been deemed secure according to OWASP (Open Web Application Security Project) guideline (https://owasp.org/www-project-top-ten/).

Web interface -Input files and analysis component. Once the user has logged in, interactions associated with job submission are facilitated by the "create job" tab which follows a relatively straightforward stepwise process based on the format of the input files for processing. The user has the option to upload a BAM file, or VCF and BAM files and define the run parameters for RBCeq file processing to support consistent and reproducible variant calling outputs (Allele Depth, Genotype Quality, MAF). To overcome large BAM file size and associated uploading issues, we have developed a standalone GUI tool named BAMTrimmer that can be downloaded directly from RBCeq and can run in Windows operating systems. The BAMTrimmer removes all unmapped and duplicate reads and trims the files with respect to blood group-associated genes. It thereby reduces BAM file size significantly, allowing users to upload these files to RBCeq rapidly. Once the trimmed BAM and compressed VCF (vcf.gz) are uploaded, the user can define run parameters and submit the job for processing. The uploaded file will be validated to ensure they adhere to their respective format guidelines.

The RBCeq analysis component is compartmentalized into three sequential parts: 1. variant calling, 2. known blood group profiling, and 3. annotation of non-ISBT variants with respect to clinical significance,

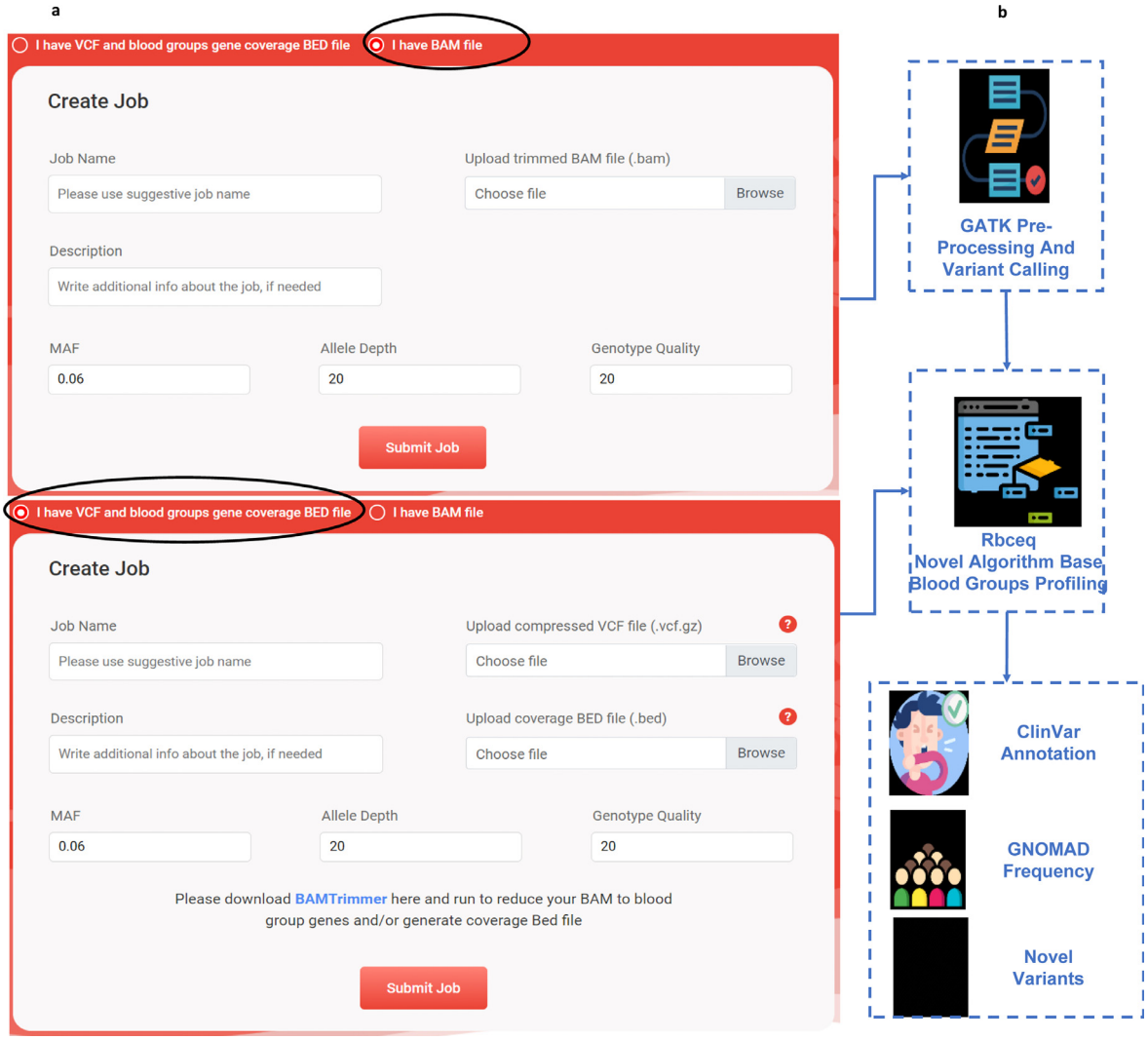


Figure 2. a. A screenshot illustrating the web interface used to upload different input file formats and the parameters required to run RBCeq. b: The workflow of analysis component of RBCeq.

population frequency distributions, and variant novelty (Figure 2).

Role of the funding source. The funding body had no part in study design, data collection, data analyses, interpretation, or writing of report. The authors Shivashankar H. Nagaraj and Robert L. Flower had full access to the data in parts.

Results

Development of RBCeq algorithm

RBCeq blood group database consists of 1559 known alleles covering 36 blood group systems, and phenotyping each one requires the mapping of multiple variants. The genotype and phenotype of each blood group is called separately because each blood group has different types of antigens and their regulation is driven by different types of genes (e.g. protein or carbohydrate or transcription factors encoding). However, the core algorithm remains the same. In the first step, each blood group is examined for reference/alternate alleles, and associated alleles are searched based on reference/alternate alleles zygosity (Supplementary Fig. 1A). In ABO blood group prediction, the frameshift (c.261delG) change present on exon 6 is scanned to differentiate *O* haplotype with *A/B*. The four nucleotide substitutions (rs7853989 [p.R176G], rs8176743 [p.G235S], rs8176746 [p.L266M], and rs8176747 [p.G286A]) in the coding sequence of exon 7 are scanned to differentiate the A and B haplotype.[31] To differentiate RHCe/CE alleles, eight nucleotides substitutions (rs586178 [p.W16C], rs1053344 [p.N68S], rs676785 [p.P103S], rs609320 [p.A226P], rs1818860403 [p.L60I], rs61777615 [intronic], rs1053343 [p.S67=], rs200955066 [p.V50=])[20] and the copy number variation ratio (2*[exon2 average coverage /average coverage of RHCE gene]) is used by algorithm.[16,30,32]

To explore all combinations of allele pairs for a particular blood group, each detected allele is paired (Sa1,2) against each other. Alleles with homozygous identical genotypes or heterozygous differing genotypes are paired (compatible [a1, a2]). The multiple allele pair ambiguities were detected due to cis-trans configuration of variants in the complex blood group systems. To overcome these ambiguities, we have developed the allele pair (AP) scoring system. For each AP, the scoring system first calculates the number of variants which are required to define an AP that are not present in the variant file (R-M); then, the algorithm calculates the number of known variants not required in the definition of allele (F-R) but present in the variant file. The addition of both variables ([R-M] + [F-R]) leads to the number of variants which is not getting utilized for that particular AP call. A lower score (minAPS[S]) means more variants were matched and less were missing an AP. The AP scoring model is a novel approach designed to prioritize optimal APs with the potential to alter gene phenotypes. If the algorithm fails to find an alternate allele for a particular blood group, then the reference genotype and phenotype will be reported (Supplementary Figure 1).

$$for\ a_1 \in \sum_{i=0}^{n} Ai : S_{a_1} = get_{SAS(a_1)}$$

$$for\ a_2 \in \sum_{i=i+1}^{n} Ai :$$

$$compatible(a_1, a_2) : \begin{cases} S_{a_{1,2}} = get_{APS(a_1,a_2)=\sum_{i=1}^{k}((R_i - M_i)+(F_i - R_i))} \\ pass \end{cases}$$

$$return : \ min_{APS(S)} : \begin{cases} a_1/a_2 \\ a_1/a_1\ \Delta\ a_2/a_2 \\ a_1/ref \end{cases}$$

n: the total number of detected alleles; Ai, scan each detected allele; APS, allele pair score; F: the total number of variants found in the sample file for that blood group; k, the total number of compatible allele pairs; M, the number of variants matched; R, the total number of variants required to define a blood group allele; ref, reference alleles; S, allele pair; SAS, single allele score.

Example of allele pair prediction using score calculation

For AP scoring, first the number of variants (R) required to define an AP is calculated (Figure 3: *ABO*A1.01/ABO*O.01.75*: 11; *ABO*A1.01/ABO*O.01.02*: 10). Then from the R, number of concordant variants (M) with input sample genotype (*ABO*A1.01/ABO*O.01.75*: 11; *ABO*A1.01/ABO*O.01.02*: 10) is calculated. Finally, the total number of known variants ([F: 11] ISBT or RBCeq database) present in the input VCF file for that particular blood group system is calculated. The subtraction of R-M, are the variants which are missing in the input sample genotype but required in the definition of the associated allele. In this case, both allele pairs have complete correspondence, so R-M is zero for both. The subtraction of F-R are the variants present in the sample genotype but not required in the definition of associated allele pair. The allele pair *ABO*A1.01/ABO*O.01.75* (F-R: 0) is defined by 11 variants and all of them are present in sample genotype whereas in the definition of *ABO*A1.01/ABO*O.01.02* AP c. 542G>A is not required (F-R: 1). The addition of ([R-M] + [F-R]) values for each AP indicates the number of variants present in the sample genotype not associated with the AP call (*ABO*A1.01/ABO*O.01.75*:0 *ABO*A1.01/ABO*O.01.02*: 1). As a result, the AP with the lowest number score is the one that leverages the maximum observed genotype (*ABO*A1.01/ABO*O.01.75*: 0) and vice versa (Figure 3).

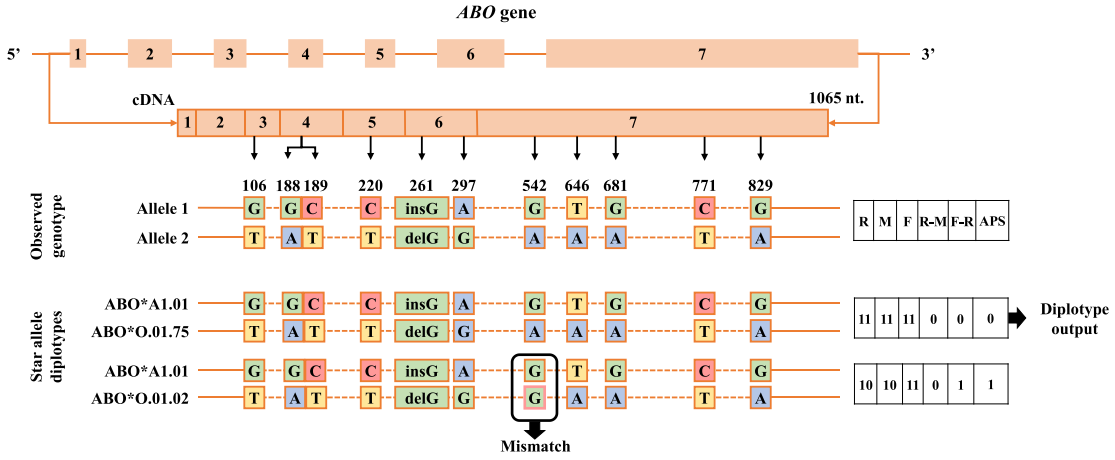


Figure 3. The breakdown of APS calculation and selection of best allele pair to predict ABO genotype in NA21112 sample from 1000 G dataset. The c.542G>A variant associated with the ABO blood group is present in the sample genotype but not needed for the definition of the allele ABO*0.01.02.

Algorithm validation

RBCeq was validated using serology, SNP array, and previously published TSO/Custom panel NGS data.[17,18,30] The RBCeq algorithm was validated on 402 samples which include 58 complex serology cases from Australian Red Cross LifeBlood, 100 samples from The MedSeq Project (phs000958) and a further 244 from Indigenous Australian participants. The algorithm was initially iteratively developed and validated on 18 samples for which serological data were available for 14 blood groups (+9 samples with SNParray) and remaining 21 blood groups with manual predicted phenotype from variant calls (Supplementary Table 3). In ABO, RHCE validation, cis-trans haplotype ambiguities were detected, we solved these ambiguities in most cases using the AP scoring system. We observed that certain haplotypes as defined in the ISBT database, did not require all variants as part of the allele definition to be present for a particular phenotype (e.g., *ABO*O.01.02*: c.829G>A). In this case, if partial allele matches are found, only nonsynonymous variants and their zygosity alone was used to determine the most likely genotype and phenotype. Previously published copy number variation (CNV) based genotyping approach for the RH blood group systems were adopted within RBCeq,[16,32,33] enabling the correct definition of the RHD, C/c variants on 18 training samples (Supplementary Table 3). We further validated the algorithm on 40 complex blood group serology cases from the Lifeblood red cell reference laboratory (Supplementary Table 4) with manually predicted phenotypes from the variant call.[17,18,30] The initial iteration of the algorithm predicted two discordant results, where it missed calling *ABO*O.02.01* allele which is associated with c.802G>A instead of c.261delG. Next, the low frequency alleles (e.g., *GYPB*06.01, LU*02.19, KEL*02.10*) of MNS, LU, and KEL blood groups were detected at the heterozygous level, but the algorithm missed predicting the low-frequency antigen phentoype. The algorithm was further improved to identify the heterozygous low frequency antigens phenotype change and the O genotyping with respect to both O alleles (*ABO*O.02.01* and *ABO*O.01.01*). Additionally, our algorithm was validated on The MedSeq Project (phs000958) comprising of 100 whole-genome sequencing and respective serology samples, and it achieved 100% accuracy across all the blood groups except for MNS (Supplementary Table 5). Due to read mis-mapping between highly homologous *GYPA* and *GYPE* genes, we found a poor call rate for the M/N antigen (*GYPA*) in 37 samples[34] (Supplementary Figure 2). To circumvent the problem of poor alignment, we re-run these samples with allele read depth ≥5 through which we achieved 100% concordance. To further develop and qualify the accuracy of RBCeq as a means of reporting blood groups through comparisons with serology results, we analyzed serological blood group data (ABO, RHCE and RHD) for 244 Indigenous Australian individuals to those predicted using RBCeq, again achieving 100% accuracy for ABO blood group profiles. In two samples, RBCeq predicted discordant calls for C antigen, which was due to inaccurate read mapping in the homologous region of exon 2 RHCE.

In summary, our validation approach successfully interpreted and called 97 antigens from which 29 were validated with serology and nine antigens with SNP-array comparison in 14 blood group system and 59 antigens were validated on manual predicted phenotype from variant call files in 29 blood group systems with an allele read depth and genotype quality of ≥ 15 (Supplementary Table 6). However, the NGS based blood group antigen prediction is highly dependent on

genome coverage and quality of sequence data,[35] and RBCeq allows user-supplied allele read depth and quality cut-off base variant selection to form blood group genotype. As such, users must exercise caution when defining run parameters and interpreting reported blood group profiles. The collective accuracy for 21 blood group systems in 402 samples was 99·07%, with discordant results only for RHCE (Supplementary Table 7).

Case study 1: 1000G-based blood group profiling

To further demonstrate the processing power and utility of RBCeq as a means of analyzing diverse whole genome sequence (WGS) datasets pertaining to five super population groups (AFR, AMR, EAS, EUR, SAS), we predicted the blood group profiles associated with the 1000 G dataset (2504 WGS samples).[36] The blood group genotype profiles corresponding to the 1000 G data have previously been published in the Erythrogene database.[22] A comparison at the allele frequency level was not possible because allele frequency assignments in Erythrogene are made with respect to extra novel alleles that are not reported in the ISBT blood group allele database and the sample-level blood group profile is not given by the Erythrogene database. RBCeq predicted 74 additional alleles than reported by the Erythrogene and has assigned revised comprehensive genotypes to the 1000 genome dataset (Supplementary Table 8). Overall, the Erythrogene database contains ~3340 alleles, of which 255 are known to the ISBT with a frequency greater than zero in the 1000 G samples. Of these 255 alleles, seven were missed by RBCeq calling, the variants defining the *KEL**02 M.*04, JK**01 N.*09, RHD**01EL.*36, RHD**01 W.*28*, and *RHD**59 alleles were not present in 1000 G dataset VCF files. The ISBT allele designation for the *AUG**02* allele is NM_001304463: c.1171G>A, whereas in the Erythrogene database it is associated with NM_001304463: c.1297G>A. With respect to ABO allele predictions, *ABO**O.01.58* allele variants were present in many samples and were checked in a reduced sample set where we found that our scoring system surpassed the prediction of *ABO**O.01.58* allele. For example, RBCeq predicted genotype is ABO*O.01.02/ABO*O.01.01 and phenotype is O for sample NA21086. A total of ten (c.106 G, c.188 G, c.189C, c.220C, c.261delG, c.297A>*G*, c.646T>A, c.681G>A, c.771C>T, c.829G>A) known variants were present in sample NA21086 that were associated with the ABO blood group (NM_020469.2), and all are covered in the *ABO**O.01.02/ABO**O.01.01* allele, whereas *ABO**O.01.58*/ABO*O.01.01 is defined by seven but only six of these variants (c.261delG, c.297A>G, c.646T>A, c.681G>A, sc.771C>T, c.829G>A,) are covered (Supplementary Table 9). As such, the *ABO**O.01.58*/ABO*O.01.01 AP did not maximally utilize the available variant data, whereas RBCeq selected the *ABO**O.01.02/ABO**O.01.01* AP utilized all present ABO blood group variants from the VCF file. RBCeq detected 74 different alleles that were not reported in Erythrogene. The 74 alleles are associated with the ABO, MNS, RH, LU, KEL, FY, JK, YT, H, and KN blood group systems. The details of each allele identified are provided in Supplementary Table 8. We were unable to ascertain why these alleles were not reported in Erythrogene, but as the study was published in 2016 after which many new alleles have been reported in the ISBT database, it may have been affected by their algorithm having also defined new alleles based on genotype.

We also detected 10 ClinVar (Supplementary Table 10), and 1042 rare (Supplementary Table 11) non-ISBT SNVs in 1000 G datasets. RBCeq algorithm used an average of 1·20 min and 15 Mb of memory to predict a blood group profile on each sample of 1000 G from a VCF file. This case study demonstrates the processing power and capability of RBCeq to manage analyzing large WGS datasets and providing additional insights into the blood group profiling of the extensively researched 1000 genome dataset. All 2504 sample blood group profile results for this case study are available on the RBCeq website (https://www.rbceq.org/).

Comparison with other blood typing tools

We found that BOOGIE[15] and bloodTyper[16] did not provide a pre-processing feature for variant calling, despite being the essential step prior to variant calling. The two tools did not report the capability to predict rare and novel variants with potential clinical relevance, which is one of the advantages of NGS-based blood typing. The RBCeq evaluation included 97 different blood group antigens from 29 blood group systems (Supplementary Table 6). Additionally, it classifies blood group gene variants as clinically significant (ClinVar annotation), rare (MAF $\leq$ 0.05 in genomAD) and novel with potential clinical relevance. RBCeq was also more efficient with processing time and memory, permitting streamlined analysis, blood group profile report and interactive visualization of sample data with an average of five million reads (BAM) in 15–20 min. The comparisons between RBCeq and the antigen prediction method are provided in Table 1 and are based on precision, data management and visualization capabilities and ease of use.

RBCeq webserver features and capabilities

Result visualization. RBCeq was developed with a specific emphasis on multiple approaches to visualize data by representing data using both interactive graphics and dynamic tables. Once jobs are completed the output can be visualized by the user by the "View Output" tab,

Tool	RBCeq	BOOGIE	bloodTyper
User interface	Web-based user friendly	Command line	Command line
Input	BAM &/or VCF	VCF file	BAM file
BAM pre-processing and variant calling	Yes	No	Only variant calling
# Samples tested	402 + (2504 WGS sample from 1000 G)	180 (69 WGS and 111 with SNP array) (serology for only ABO RHCE, and RHD)	310
Concordance with serological data	99·07% (97 antigens)	94%	99·2% (21 RBC antigens)
Tested on how many blood group systems	14 (Total 29 blood group system with manual predicted phenotype)	12	12
Clinical annotation	Yes	No	No
Classification of rare blood group variants	Yes	No	No
Prediction of novel blood group alleles with potential clinical relevance	Yes	No	No
Software installation	No installation	Installation	Installation
Output	HTML report Interactive plots, detailed QC and quantitative sequencing data statistics with an overview of results	Test file	HTML report

***Table 1*: Comparison of the features of RBCeq to those of existing tools.**

which will direct the user to the result page. Details pertaining to each results section are given below:

Job Summary: This section describes information pertaining to user-specified input file format and run parameters used, including input file job name, and RBCeq unique job ID number, run time and date, and provides an option to download completed results as an Excel file.

Overall Summary: This section includes a summary of variant annotation and average sequence read mapped coverage in blood group associated genes. The interactive pie chart serves as a visual reference for variant annotation distributions. The overall summary provides a quantitative overview of user-provided files, enabling a rapid bioinformatics quality check of inputted samples (Figure 4A).

Blood group change summary. This section summarizes the detected known blood groups alleles that are different from the reference together with information regarding the associated phenotype. Colours in the inner circle represent the blood group alleles that are changed relative to the reference, while the colours of the outer circle represent respective phenotypes. Users can interactively explore genotype and phenotype information pertaining to detected blood groups by moving the cursor on the plot (Figure 4B).

Per blood group variants and coverage statistics. This section provides an interactive graph that enables the quantitative analysis of gene coverage and detected variants for each blood group. The plots will help the user visualize correlations between read mapped coverage and detected variants for all blood group genes. Users can also zoom in or out on a particular gene by scrolling their mouse wheel over the plot (Figure 4C and D).

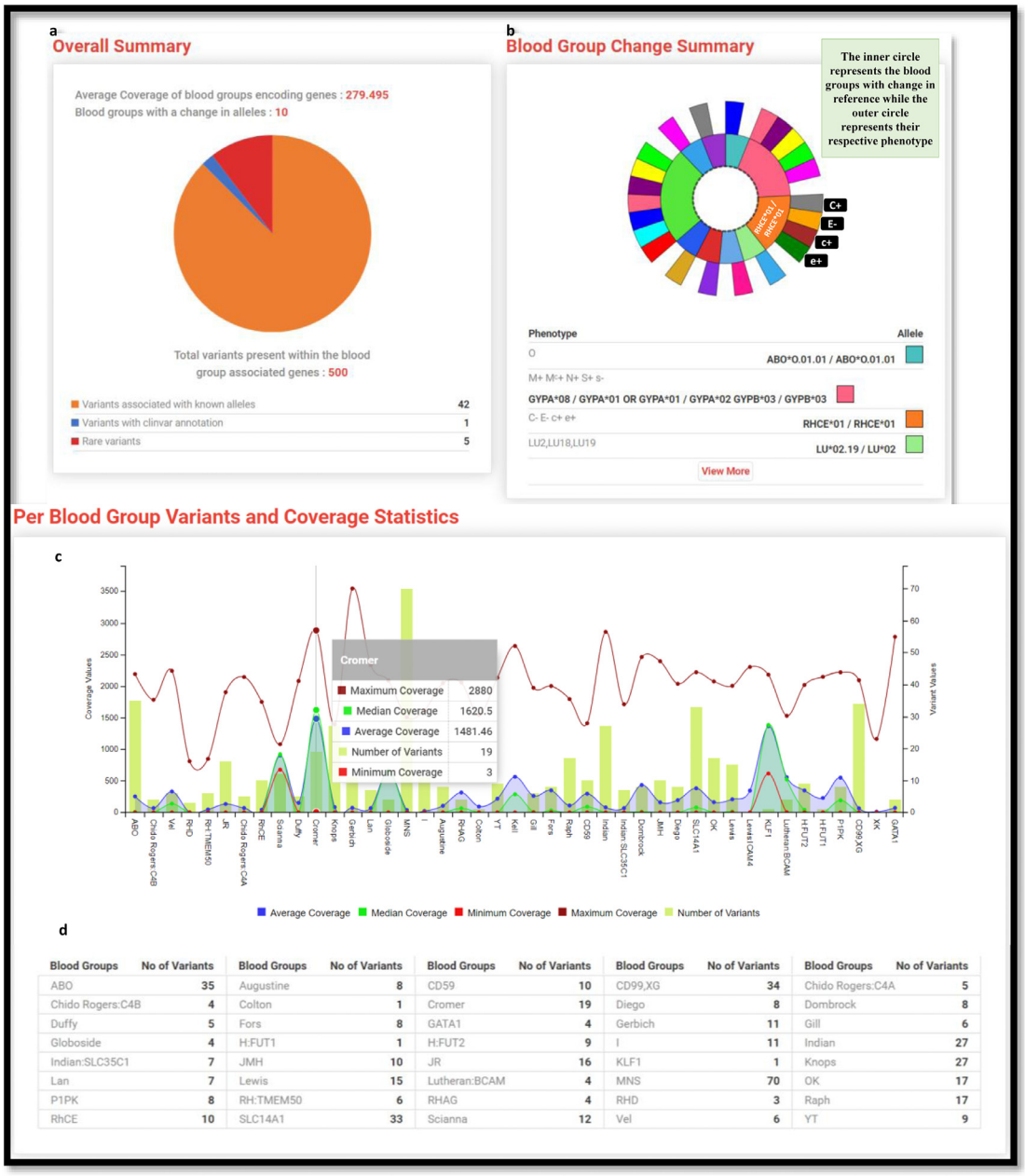
RH blood group system coverage statistics. This section describes the CNV calculator and predicted results for phenotyping/genotyping based on analysis coverage statistics extracted from the uploaded trimmed BAM file for the RH (*RHD* and *RHCE* genes) blood group system known for structural variation between the homologous genes. The CNV ratios and interpretation for exonic rearrangements or deletion/duplication/triplication are given (Figure 5A).

The known blood group allele table. A list of identified blood group alleles and predicted phenotype will be reported along with supporting information including variants, zygosity, allele depth, and allele frequency (Figure 5B).

Clinically significant, rare, and novel allele annotations. This section partitions and lists the variants into three categories (Figure 5C); according to the algorithm/workflow described in Figure 1: ClinVar, Rare, and Novel variants.

Discussion

RBCeq is the first scalable blood profiling software with an available web-based interface enabling users to



Blood Groups	No of Variants	Blood Groups	No of Variants	Blood Groups	No of Variants	Blood Groups	No of Variants	Blood Groups	No of Variants
ABO	35	Augustine	8	CD59	10	CD99,XG	34	Chido Rogers:C4A	5
Chido Rogers:C4B	4	Colton	1	Cromer	19	Diego	8	Dombrock	8
Duffy	5	Fors	8	GATA1	4	Gerbich	11	Gill	6
Globoside	4	H:FUT1	1	H:FUT2	9	I	11	Indian	27
Indian:SLC35C1	7	JMH	10	JR	16	KLF1	1	Knops	27
Lan	7	Lewis	15	Lutheran:BCAM	4	MNS	70	OK	17
P1PK	8	RH:TMEM50	6	RHAG	4	RHD	3	Raph	17
RhCE	10	SLC14A1	33	Scianna	12	Vel	6	YT	9

Figure 4. The summary of distribution and annotation of detected variants in the blood group antigen defining genes; a: includes average coverage of blood group defining genes; the number of blood groups alleles for which the reference is changed, and the distribution of total variants detected in blood group associated genes with blood group known, ClinVar and rare annotation; b: The interactive blood group change allele pie chart; c: The "Per Blood Group Variants and Coverage Statistics" graph is an interactive graph which gives the quantitative analysis with respect to blood group genes coverage and detected variants. The x-axis is the blood group antigen defining genes, the left y-axis represents the coverage of the genes, and the right y-axis represents the number of variants for each blood group antigen determining genes. Each line color represents a different coverage value. d: The table gives the number of variants detected in the input sample for each blood group antigen defining genes.

define analysis parameters, produce interactive visualizations, and archive their results. RBCeq takes four minutes on average to process a VCF file, and the time includes commencing a new instance on the AWS server, blood group prediction and report generation. Whereas reviewing and interpreting variants for a single blood group gene may take one hour for a simple analysis or up to one day for more complicated cases. RBCeq automation can significantly reduce both computational and hands-on analysis time of pre-transfusion blood

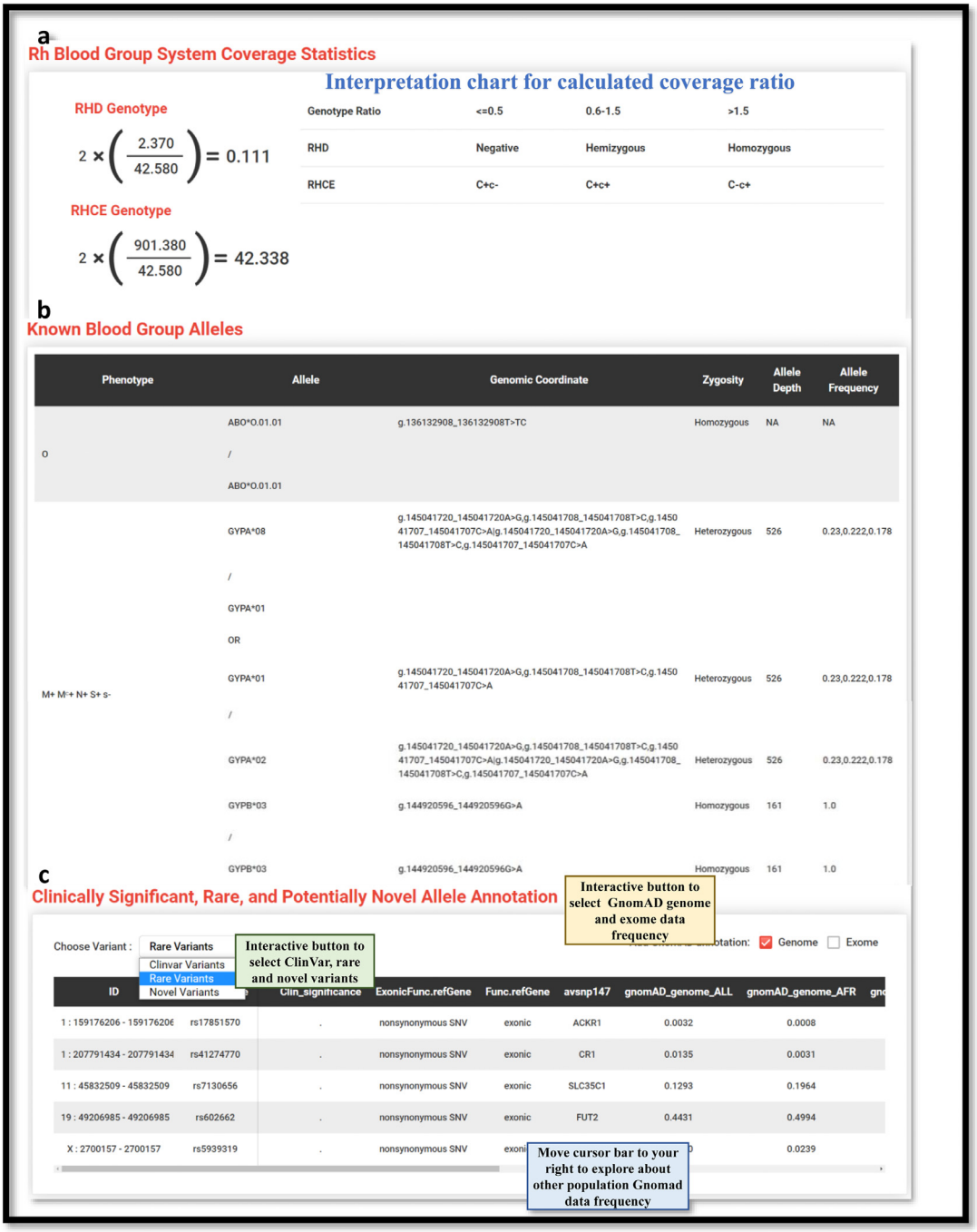


Figure 5. a: The CNV ratio formula (RHD=2* (*RHD* gene average Coverage/*RHCE* gene average Coverage and RHCE = 2* (*RHCE* gene exon2 average coverage/*RHCE* gene average coverage) and interpretation for exonic rearrangements or deletions/duplications/triplications; b: An excerpt of the Known Blood Group Allele table; The first column provides details pertaining to the blood group phenotype, while the second column includes allele information, the third column contains information regarding the variants responsible for defining the allele, the fourth column includes the zygosity of the allele (if one variant is heterozygous then the whole allele will be referred to as a heterozygous allele), the fifth column includes the average allele depth for that allele, and the sixth column includes the average allele frequency. For the complete table of 36 blood group profiles with two TF, please follow the tutorial page at https://www.rbceq.org/tutorial.php; c: The first option includes ClinVar Variants, which are variants that have ClinVar annotations. The second option includes Rare Variants that have a MAF less than the user-selected threshold value in the gnomAD database. The third option includes novel variants, that are predicted to be deleterious using *in silico* tools and with the exonic function that is predicted to be "nonsynonymous / frameshift /stop-gain/stop-loss/splicing". All lists are provided with known details pertaining to each variant, including dbsnpid, exonic function, refGene and gnomAD genome and exome frequencies in five different populations.

group typing using NGS data with increase sample processing throughput. RBCeq maintained accuracy of 99·07% for 29 blood group systems across 402 samples.

The advanced functionalities provided by RBCeq allows effective analysis of blood group profiles corresponding to 36 blood groups and two TFs based upon WGS, WES, or TES data, extracting non-ISBT variant information to gain a more profound understanding of the underlying data. Although BOOGIE[15] and bloodTyper[16] predicts blood group phenotypes from NGS data, they are mostly designed for bioinformatics experts. BOOGIE reports 34 blood group systems, and was validated for ABO and D blood groups with a serological concordance of 94%. The bloodTyper tool achieved a concordance of 99·2% for 21 antigens in 12 blood group systems (14 blood group genes). The performance for the RBCeq algorithm is comparable, achieving a concordance of 99·07% for 97 antigens in 29 blood group systems across 402 samples. RBCeq accurately reported complex and rare blood group phenotypes observed in a red cell reference laboratory setting such as Hy- and Jo(a-), which weaken the DO blood group expression, and which are only found in populations of African origin.[21] The Vel- phenotype defined by 17 bp deletion (rs566629828|) was also reported by RBCeq and was validated with the prediction of phenotype from variant calls data (Supplementary Table 4). The C/c antigen prediction concordance was significantly improved when we included the assessment of the intronic variant rs61777615. This variant abolishes the *Pst1* restriction site and in the presence of Ce or CE which signifies the presence of 109 bp insertion.[37,38]

RBCeq is designed to report the presence of variants that are on databases such as the ISBT or are novel but with a potential clinical significance. We previously reported on the clinical utility of TES for resolving the blood group status for samples presenting with complex or equivocal typing results(18). While TES provides a comprehensive blood group, platelet and neutrophil profile in a single test system the analysis and interpretation of the data is manual and time consuming. Subject to further validation RBCeq has the potential to be integrated into the algorithm for problem solving. For example, the prediction of novel variants provides the opportunity to identify blood group variants implicated in Haemolytic Disease of the Fetus and Newborn (HDFN). One prior complex HDFN analysis from a collaborative laboratory revealed, a SNV associated with a novel low-frequency antigens present in < 1% of the global population (SARA+/MNS:47 antigen).[39,40] The reagents necessary to detect these rare/novel blood types are available only in a small number of reference laboratories in the world.[12] RBCeq has the potential to assist in the analysis of data from such cases in the future.[39–41]

In future releases, we plan to incorporate a platelet antigen prediction function. Furthermore, in the time since work commenced on the development of the RBCeq platform, seven new blood group systems have been recognized by the ISBT Red Cell Immunogenetics and Blood Group Terminology Working Party (ISBT 037 KANNO, ISBT 038 SID, ISBT 039 CTL2, ISBT 040 PEL, ISBT 041 MAM, ISBT 042 EMM, and ISBT 043 ABCC1). These blood group systems will be included in future releases of RBCeq and studies are underway to validate RBCeq clinical application particularly for complex samples from the Red Cell Reference Setting. The CNV analysis is limited to the determination of *RHD* zygosity and to distinguish between the alleles for the C/c antigen. The automation of hybrid alleles (e.g., RHCE*ce-D[4–7] -ce and GPB[1–46] -A (47–118)) prediction, is highly dependent on statistical normalization and batch- corrections methodologies which are used in the read alignment tools. To eliminate coverage biases in homologous recombination regions, short-read base alignment methods are still struggling to obtain unique mapped reads.[42,43] Recent developments in the detection of structural/copy number variations are elucidating the challenges around complex genomics regions,[44–46] but approaches like alternative mapping locus and a combination of short and long-read sequencing data[43,47] will help to improve hybrid allele prediction. The RBCeq predicts novel variants with the potential of clinical relevance by applying the *in-silico* tools. Many novel or rare variants may require further study of potential structural changes and/or the ability to trigger the immune response. In future, the use of an artificial intelligence base model which will consider the protein-protein interaction, extracellular domains and other such information can lead to a more precise interpretation of the impact of variants.[48,49]

In conclusion, RBCeq is user-friendly, fast, accurate, and provides extended profiles of variants with no current known blood group phenotype association. RBCeq holds great promise for use in the automation of blood group antigen detection in personalized medicine. Given the constantly rising number of identified phenotypes, RBCeq can potentially aid global blood supply organizations in the computational genotyping of all clinically relevant blood groups. As personal genome sequencing is forecast to become increasingly prevalent in the near future, there is a clear need for the development of new bioinformatics algorithms for blood group characterization.

Contributors

SHN led the project and oversaw all activities of the project. Concept and design: SJ, CH, RF, SHN. Algorithm development: SJ, EMS, CLD, EVR and SHN. Software development and validation SJ, MD, MH, CW,

EMS, CLD, EVR, CH. Data visualization: SL. First draft of the manuscript: SJ and SHN. Manuscript revision: CLD, EVR, NMP, CH, RF and SHN. All authors have read and approved the final manuscript.

Data sharing statement

Proof of Principle Algorithm Development dataset and South East Queensland Indigenous TES data involves human participants. Due to ethical constraints, it cannot be made publicly available. However, researchers that meet the ethics requirements can request data from the authors (Robert Flower and Shivashankar H. Nagaraj). Data access requests for MedSeq (phs000958) can be made via dbGaP. The 1000 G project files are available via 1000 G project FTP server (http://ftp.1000 genomes.ebi.ac.uk/vol1/ftp/phase3/).

Declaration of interests

There are no conflicts of interest declared by authors.

Acknowledgement

We acknowledge and pay respects to the First Nations Peoples of Australia who were part of the Southeast Queensland Indigenous study to develop & validate RBCeq to predict their unique blood group profiles accurately. We are grateful for the opportunity to work together, and the trust placed in us to undertake research in a respectful and collaborative manner. The authors also thank Maree Perry (Nganyaywana (Anaiwan) and Wiradjuri), Aoibhe Mulcahy and Glenda Millard for their role in sample receipt, preparation, serological testing and sequencing (TES) of the Indigenous samples. The Australian governments fund the Australian Red Cross Lifeblood for the provision of blood, blood products and services to the Australian community.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103759.

References

1 Garcia-Roa M, Del Carmen Vicente-Ayuso M, Bobes AM, Pedraza AC, Gonzalez-Fernandez A, Martin MP, et al. Red blood cell storage time and transfusion: current practice, concerns and future perspectives. *Blood Transfus* 2017;**15**(3):222–31.
2 Fung MK, Eder A, Spitalnik SL, Westhoff CM. American association of blood B. technical manual. new york blood center. New York, NY.: Amer Assn of Blood Banks; 2017.
3 Gleadall NS, Veldhuisen B, Gollub J, Butterworth AS, Ord J, Penkett CJ, et al. Development and validation of a universal blood donor genotyping platform: a multinational prospective study. *Blood Adv* 2020;**4**(15):3495–506.
4 Helias V, Saison C, Ballif BA, Peyrard T, Takahashi J, Takahashi H, et al. ABCB6 is dispensable for erythropoiesis and specifies the new blood group system Langereis. *Nat Genet* 2012;**44**(2):170–3.
5 Colin Y, Rahuel C, London J, Romeo PH, d'Auriol L, Galibert F, et al. Isolation of cDNA clones and complete amino acid sequence of human erythrocyte glycophorin C. *J Biol Chem* 1986;**261**(1):229–33.
6 Storry JR, Castilho L, Chen Q, Daniels G, Denomme G, Flegel WA, et al. International society of blood transfusion working party on red cell immunogenetics and terminology: report of the Seoul and London meetings. *ISBT Sci Ser* 2016;**11**(2):118–22.
7 Komatsu F, Hasegawa K, Yanagisawa Y, Kawabata T, Kaneko Y, Watanabe S, et al. Prevalence of diego blood group Dia antigen in Mongolians: comparison with that in Japanese. *Transfus Apher Sci* 2004;**30**(2):119–24.
8 Grunbaum BW, Selvin S, Myhre BA, Pace N. Distribution of gene frequencies and discrimination probabilities for 22 human blood genetic systems in four racial groups. *J Forensic Sci* 1980;**25**(2):428–44.
9 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet Med* 2015;**17**(5):405–24.
10 Avent ND. Large-scale blood group genotyping: clinical implications. *Br J Haematol* 2009;**144**(1):3–13.
11 McBean RS, Hyland CA, Flower RL. Approaches to determination of a full profile of blood group genotypes: single nucleotide variant mapping and massively parallel sequencing. *Comput Struct Biotechnol J* 2014;**11**(19):147–51.
12 McBean R, Liew YW, Wilson B, Kupatawintu P, Emthip M, Hyland C, et al. Genotyping confirms inheritance of the rare at(a-) type in a case of haemolytic disease of the newborn. *J Pathol Clin Res* 2016;**2**(1):53–5.
13 Storry JR, Joud M, Christophersen MK, Thuresson B, Akerstrom B, Sojka BN, et al. Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nat Genet* 2013;**45**(5):537–41.
14 Cvejic A, Haer-Wigman L, Stephens JC, Kostadima M, Smethurst PA, Frontini M, et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat Genet* 2013;**45**(5):542–5.
15 Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, Tosatto SC. BOOGIE: predicting blood groups from high throughput sequencing data. *PLoS ONE* 2015;**10**(4):e0124579.
16 Lane WJ, Westhoff CM, Gleadall NS, Aguad M, Smeland-Wagman R, Vege S, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol* 2018;**5**(6):e241–e51.
17 Schoeman EM, Lopez GH, McGowan EC, Millard GM, O'Brien H, Roulis EV, et al. Evaluation of targeted exome sequencing for 28 protein-based blood group systems, including the homologous gene systems, for blood group genotyping. *Transfusion* 2017;**57**(4):1078–88.
18 Schoeman EM, Roulis EV, Liew YW, Martin JR, Powley T, Wilson B, et al. Targeted exome sequencing defines novel and rare variants in complex blood group serology cases for a red blood cell reference laboratory setting. *Transfusion* 2018;**58**(2):284–93.
19 International Society of Blood Transfusion (ISBT). Red cell immunogenetics and blood group terminology Internet. cited 2021, October 1. Available from: http://www.isbtweb.org/working-parties/red-cell-immunogeneticsand-blood-group-terminology/.
20 Reid C, Olsson ML. The blood group antigen factsbook, 3. London: Elsevier; 2012.
21 Daniels G. 3rd ed. Human blood groups, 3. Oxford: Wiley- Blackwell; 2013.
22 Moller M, Joud M, Storry JR, Olsson ML. Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Adv* 2016;**1**(3):240–9.
23 Wagner FF, Flegel WA. The Rhesus Site. Transfusion medicine and hemotherapy: offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie. 2014;41(5):357–63.
24 Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, et al. TransVar: a multilevel variant annotator for precision genomics. *Nat Methods* 2015;**12**(11):1002–3.
25 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. *Genome Biol* 2016;**17**(1):122.
26 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res* 2002;**12**(6):996–1006.

27 Lane WJ, Westhoff CM, Uy JM, Aguad M, Smeland-Wagman R, Kaufman RM, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* 2016;**56**(3):743–54.

28 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform* 2013;**43**:11.10.1–11.10.33.

29 Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**(D1):D862–8.

30 Roulis E, Schoeman E, Hobbs M, Jones G, Burton M, Pahn G, et al. Targeted exome sequencing designed for blood group, platelet, and neutrophil antigen investigations: proof-of-principle study for a customized single-test system. *Transfusion* 2020.

31 Yamamoto F, Cid E, Yamamoto M, Saitou N, Bertranpetit J, Blancher A. An integrative evolution theory of histo-blood group ABO and related genes. *Sci Rep* 2014;**4**:6601.

32 Chou ST, Flanagan JM, Vege S, Luban NLC, Brown RC, Ware RE, et al. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv* 2017;**1**(18):1414–22.

33 Baronas JJ, Westhoff CM, Vege S, Mah H, Aguad M, Smel R, et al. RHD zygosity determination from whole genome sequencing data. 2016;7:1–5.

34 Wheeler M, Frazar C, Lannert K, Fletcher SN, Huston H, Harris S, et al. Prediction of MNS Blood Group Antigens Using Next Generation Sequencing. *Blood* 2016;**128**(22).

35 Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* 2019;**9**(1):1784.

36 Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* 2015;**526**(7571):68–74.

37 Carritt B, Kemp TJ, Poulter M. Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. *Hum Mol Genet* 1997;**6**(6):843–50.

38 Poulter M, Kemp TJ, Carritt B. DNA-based rhesus typing: simultaneous determination of RHC and RHD status using the polymerase chain reaction. *Vox Sang* 1996;**70**(3):164–8.

39 Millard GM, McGowan EC, Wilson B, Martin JR, Spooner M, Morris S, et al. A proposed new low-frequency antigen in the Augustine blood group system associated with a severe case of hemolytic disease of the fetus and newborn. *Transfusion* 2018;**58**(5):1320–2.

40 McBean RS, Hyland CA, Hendry JL, Shabani-Rad MT, Flower RL. SARA: a "new" low-frequency MNS antigen (MNS47) provides further evidence of the extreme diversity of the MNS blood group system. *Transfusion* 2015;**55**(6 Pt 2):1451–6.

41 Khan J, Delaney M. Transfusion support of minority patients: extended antigen donor typing and recruitment of minority blood donors. *Transfus Med Hemother* 2018;**45**(4):271–6.

42 Wheeler MM, Lannert KW, Huston H, Fletcher SN, Harris S, Teramura G, et al. Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. *Genet Med* 2019;**21**(2):477–86.

43 Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;**13**(1):36–46.

44 Jenko Bizjan B, Katsila T, Tesovnik T, Sket R, Debeljak M, Matsoukas MT, et al. Challenges in identifying large germline structural variants for clinical use by long read sequencing. *Comput Struct Biotechnol J* 2020;**18**:83–92.

45 Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol* 2019;**20**(1):246.

46 Twyford AD, Ennos RA. Next-generation hybridization and introgression. *Heredity* 2012;**108**(3):179–89. (Edinb).

47 Fan X, Chaisson M, Nakhleh L, Chen K. HySA: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. *Genome Res* 2017;**27**(5):793–800.

48 Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun* 2021;**12**(1):510.

49 Callaway E. It will change everything': deepMind's AI makes gigantic leap in solving protein structures. *Nature* 2020;**588**(7837):203–4.