medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A machine-learning method for biobank-scale genetic prediction of blood group antigens

Hyvärinen K[1*], Haimila K[2], Moslemi C[3,4], Blood Service Biobank[5], Olsson ML[6,7], Ostrowski SR[8,9], Pedersen OB[3,9], Erikstrup C[10], Partanen J[1], Ritari J[1]

1. *Research and Development, Finnish Red Cross Blood Service, Helsinki, Finland*
2. *Blood Group Unit, Finnish Red Cross Blood Service, Vantaa, Finland*
3. *Department of Clinical Immunology, Zealand University Hospital,* Køge, *Denmark*
4. *Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark*
5. *Finnish Red Cross Blood Service, Vantaa, Finland*
6. *Department of Laboratory Medicine, Lund University, Lund, Sweden*
7. *Department of Clinical Immunology and Transfusion Medicine, Office for Medical Services, Region Skåne, Sweden*
8. *Department of Clinical Immunology, Copenhagen University Hospital, Rigshopitalet, Copenhagen, Denmark*
9. *Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark*
10. *Department of Clinical Immunology, Aarhus University Hospital, Skejby, Denmark*

Short title: Blood group prediction method

*Corresponding author:
Kati Hyvärinen, PhD, Associate professor
Finnish Red Cross Blood Service
Research and Development
Biomedicum Helsinki 1, Haartmaninkatu 8, 00290 Helsinki, FINLAND
Phone +358 40 920 3083
kati.hyvarinen@bloodservice.fi

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ABSTRACT

Blood transfusion is a life-saving medical procedure performed routinely worldwide. A key element for successful transfusion is compatibility of the patient and donor red blood cell (RBC) antigens. Precise antigen matching reduces the risk for immunization and other adverse transfusion outcomes. RBC antigens are encoded by specific genes, which allows developing computational methods for determining antigens from genomic data.

We describe here a classification method for determining RBC antigens from genotyping array data. Random forest models for 39 RBC antigens in 14 blood group systems and for human platelet antigen (HPA)-1 were trained and tested using genotype and RBC antigen and HPA-1 typing data available for 1,192 blood donors in the Finnish Blood Service Biobank. The algorithm and models were further evaluated using a validation cohort of 111,667 Danish blood donors.

In the Finnish test data set, the median (interquartile range [IQR]) balanced accuracy for 39 models was 99.9 (98.9–100)%. We were able to replicate 34 out of 39 Finnish models in the Danish cohort and the median (IQR) balanced accuracy for classifications was 97.1 (90.1–99.4)%. When applying models trained with the Danish cohort, the median (IQR) balanced accuracy for the 40 Danish models in the Danish test data set was 99.3 (95.1–99.8)%.

The RBC antigen and HPA-1 prediction models demonstrated high overall accuracies suitable for probabilistic determination of blood groups and HPA-1 at biobank-scale. Furthermore, population-specific training cohort increased the accuracies of the models. This stand-alone and freely available method is applicable for research and screening for antigen-negative blood donors.

Keywords

Red blood cell antigen, blood group, random forest, classification, biobank, blood donor

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

INTRODUCTION

Blood transfusion is a life-saving procedure performed widely in treating various medical conditions. Despite routine practices, the safety of transfusions remains a major concern[1]. Exposure to foreign RBC antigens may result in alloantibody formation and hemolytic transfusion reactions. Additionally, sensitization to non-self RBC antigens and human platelet antigens (HPAs) can also occur via pregnancy and cause fetal morbidity and mortality[2,3]. The current general practice of matching the recipient and blood donor for ABO and RhD antigens is inadequate to prevent sensitization to other antigens. Extended matching could reduce the risk of alloimmunization and adverse events, which are especially pronounced among patients receiving regular transfusions[4,5].

Blood group typing of blood donors has been conventionally performed by serotyping and is still the main method used in blood centers. To overcome limitations regarding low throughput and lack of valid reagents for all clinically relevant antigens, numerous DNA-based genotyping and sequencing methods have emerged within the last decades[6–10]. This development has been enabled by the accumulating knowledge about the genetic basis of the blood groups[11,12] and the rapid evolution of molecular methodology. However, the systematically extended blood group typing of blood donors and, even more so, the recipients, remains sparse. Economic feasibility has been a major restraint to the progress. The development of genotyping array technologies has promoted high-throughput and cost-effective genetic studies in many fields and, in 2020, Gleadall *et al*.[13] introduced a microarray platform for RBC antigen, human leukocyte antigens (HLA), and HPA typing for precision matching of blood.

While accurate blood group typing is obligatory for safe transfusions, an initial screening for potential donors could be achieved using less stringent procedures. In the last decade, the development of machine learning approaches for high-dimensional data has provided new opportunities for exploitation of expanding genetic data. In 2015, Giollo *et al*.[14] presented BOOGIE, an RBC antigen predictor based on Boolean rules and *k*-nearest neighbor (*k*-NN) algorithm. Decision tree -based methods, including bootstrap aggregation[15] and random forest[16], have been utilized for imputation of HLA alleles[17,18] and killer cell immunoglobulin-like receptor (KIR) copy number[19] and gene content[20]. To our knowledge, these methods

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

have not yet been implemented on RBC antigen and HPA screening. The analysis of high-dimensional data with computational performance suitable for large-scale analyses may be implemented using "RANdom forest GEneRator" software R package[21]. The execution is feasible in the local computing environment and sensitive data uploads are not required.

Here we describe a stand-alone and freely available random forest classification method and models for determining RBC antigens and HPA-1 from array technology-based genotyping data. We investigate the performance of models trained with Finnish blood donor biobank data and further validate the method with a Danish cohort. Our results suggest that the method is applicable for biobank-scale probabilistic determination of RBC antigens and HPA-1, and could facilitate research and screening for antigen-negative blood donors.

STUDY SUBJECTS AND METHODS

Study cohorts and design

The Finnish study cohort consists of 1,192 blood donors belonging to the Blood Service Biobank, Helsinki, Finland (https://www.veripalvelu.fi/en/biobank/). Genotype and blood group phenotype data were obtained from the Blood Service Biobank. The study (biobank decision 002-2018) conforms to the principles of the Finnish Biobank Act (688/2012) and the participants have given written informed consent to the Blood Service Biobank.

The Danish validation cohort consists of 111,667 participants of the Danish Blood Donor Study (DBDS) Genomic Cohort expanding on the Danish blood bank system[22,23]. The genetic studies in DBDS have been approved by the Danish Data Protection Agency (P-2019-99) and the Scientific Ethical Committee system (NVK-1700407).

Genotyping and genotype imputation

The genotyping and genotype imputation of the Finnish cohort have been performed originally as a part of FinnGen project (https://www.finngen.fi/en). Biobank samples were genotyped using FinnGen ThermoFisher Axiom custom array v2 (Thermo Fisher Scientific,

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Santa Clara, CA, USA) and imputed using the population-specific Sisu v3 imputation reference panel with Beagle 4.1. Detailed description of the procedures is available at https://finngen.gitbook.io/documentation/v/r4/methods/genotype-imputation and the marker content of the custom array v2 is downloadable at https://www.finngen.fi/en/researchers/genotyping. The phased genotypes were filtered for the imputation INFO-score >0.6 and were in vcf format.

In the Danish cohort, the genotyping was performed using Illumina's Infinium Global Screening Array and imputed using the deCODE genetics' (Reykjavik, Iceland) North European reference sequence panel. Unphased genotypes were filtered for the imputation INFO-score >0.75, minor allele frequency >0.01, Hardy–Weinberg equilibrium P-values <1 $\times 10^{-4}$, and samples for missingness per individual <3%.

RBC antigen and HPA typing

The RBC antigen and HPA-1 phenotypic information for the Finnish and Danish cohorts is presented in **Table 1**. The availability of the phenotype data varied in a wide range depending on the antigen due to the different testing criteria practices. In the Finnish cohort, RBC antigen and HPA-1 typing was performed at the FRCBS Blood Group Unit by routine methods and the results were obtained using validated serological and genotyping techniques.

The sources for RBC antigen and HPA-1 typing results were the Danish electronic blood bank systems and the typing was performed using serological methods, except for Vel-status, which was determined using polymerase chain reaction technique.

Classification random forest models

An overview of the study design is depicted in **Figure 1**. RBC antigen and HPA-1 coding genes and the genetic regions used in the models are presented in **Supplementary Table 1**. The models for the antigens were generated separately using the same hyperparameters. Only antigens having at least four cases in each respective typing data class were included, resulting altogether in 39 models. For the Finnish reference data set, SNVs in RBC antigen

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

and HPA-1 coding genetic regions ± 2,000 bp flanking regions were utilized in dosage format. **Table 2** presents the number of SNVs available for each model. Only samples having full dosage data were used. The genetic and antigen typing information were combined into a single full data set and divided randomly 1:1 into train and test data sets.

R v4.3.0 environment[24] was used for the implementation and classification random forest models were created using the R package ranger v0.13.1[21]. The number of trees was 2,000 and split criteria based on node impurity measured by the Gini index. Class weights were applied due to unbalanced outcome classes. Number of variables to possibly split at each node (mtry) was number of SNVs divided by 2 and the variable importance was determined by permutation. Feature selection was based on variable importance >0 and the model was re-fitted using these important SNVs only. The number of important variables and prediction errors for each antigen model are presented in the **Table 2**. Prediction error was determined as misclassification frequency obtained from out-of-bag data and prediction on the test set. The important variables for Finnish models are listed in **Supplementary Data 2**. The full data set was used in fitting the final models.

Model evaluation metrics

The model accuracy was evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and balanced accuracy. The data was wrangled using tidyverse v1.3.1 package[25] and the evaluation metrics were derived using caret v6.0-92[26]. For each model, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were determined. Sensitivity was defined as TP / (TP + FN), specificity as TN / (TN + FP), PPV as TP / (TP + FP), NPV as TN / (TN + FN)[27]. Balanced accuracy accounts for imbalanced classification and was defined as (sensitivity + specificity) / 2.

Validation of the Finnish models and the random forest method for generating the models

The models obtained using the Finnish data set were applied to the Danish cohort. The implementation required imputed genotype data in vcf or PLINK format. The Danish allele

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

dosage data was harmonized by naming and allele orientation for compatibility with the Finnish models and the dosage data for the missing important variables was imputed using mean values.

The model-generating method was further validated by fitting the models on the Danish data set to create models specific to the Danish cohort. In the Danish data set, the percentage of missing genotypes was on average 5% depending on the genetic region of the blood group/HPA system. Missing allele dosage values were imputed separately for train and test data sets using mean values before classification random forest step. Characteristics of the Danish models are presented in **Supplementary Table 2**. The important variables for the Danish models are listed in **Supplementary Data 3**. The evaluation metrics for both prediction and modelling were defined as depicted in the "*Model evaluation metrics*" section.

The significance of variation of balanced accuracies was analyzed using Mann-Whitney-Wilcoxon Test implemented with R v3.6.1.

Data availability

Genotyping and RBC antigen/phenotype and HPA-1 typing data for the Finnish cohort are stored in the Blood Service Biobank, Helsinki, Finland. Researchers may apply for access to data (https://www.veripalvelu.fi/en/biobank/for-researchers/). Due to privacy laws, the Danish genetic data and phenotypes are only available to DBDS researchers and blood banks.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

RESULTS

Evaluation of the Finnish classification models

In the Finnish cohort, the genotype data was accessible for 1,192 blood donors and the RBC antigen typing data was available for 39 antigens representing 15 blood group systems. The blood group typing frequency varied greatly depending on RBC antigen/phenotype, being at the lowest 5% for HPA-1b and at the highest 100% for A, B, AB, O, K, D, C, c, E, and e (**Table 1**).

After data partitioning, the number of study subjects in the test data set was 596. The median (interquartile range [IQR]) balanced accuracy for 39 models was 99.9 (98.9–100)% in the test data set and accuracy metrics for all models are presented in the **Table 3**. The models for antigen/phenotype positivity of AB, B, A_1, A_2, Yt^b, Co^a, Do^a, Do^b, Fy^a, HPA-1b, K, Kp^a, Ul^a, Jk^a, Lu^a, S, and s reached balanced accuracy of 100%. For other models, the balanced accuracy was ≥98.0%, except 83.3% for Ls^a, 94.0% for Le^b, 95.0% for HPA-1a, and 96.0% for hr^S. Accuracy metrics for the test and full data sets are presented in the **Supplementary Tables 3** and **4**, respectively. **Figure 2** illustrates the confusion matrices for classification models in the Finnish test data set. The number of false negative plus false positive (FN + FP) samples out of all samples was low, ranging from 0 to 1% in all models, except 2% for hr^S. Confusion matrices for the test and full data sets are presented in **Supplementary Figures 1** and **2**, respectively. The median (IQR) prediction error, determined as misclassification frequency obtained from out-of-bag data, of the Finnish models was 1.6×10^{-3} (1.9×10^{-4}–7.0×10^{-3}) (**Table 2.**).

The distributions of posterior probabilities (PP) in the test data set are depicted in **Figure 3**. The samples having PP >0.5 were classified as antigen positive and ≤0.5 as antigen negative. The majority of the PPs were close to 1 for the antigen typing positive samples and close to 0 for the antigen typing negative samples. The Co^a-negative samples (only two samples in the test data set) were classified correctly but the PPs were closer to 0.5 than to 0. One of the three Ls^a-positive samples were misclassified and the PPs for the other two were closer to 0.5 than to 1 (specificity 66.7%). The spectrum of PP distribution with some misclassifications was observed for Co^b, Le^b, M, N, C, C^w, D, and hr^S. **Supplementary Figures 3** and **4** depict the distributions of PPs in the test and full data sets.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Validation of the Finnish classification models in the Danish cohort

The Danish validation cohort had genotype and phenotype data for 34 out of the 39 Finnish classification models. Antigen/phenotype typing data varied from 433 for A_2 to ~111,000 for A, AB, B, O, and D (**Table 1**). Due to missing Finnish model variables in the Danish genotype data, the Danish allele dosage data was harmonized using mean imputation before applying the Finnish models.

The median (IQR) balanced accuracy for classifications was 97.1 (90.1–99.4)% and all the evaluation metrics are presented in **Supplementary Table 4**. The balanced accuracies were >98.0% for 14 models including antigen/phenotype positivity of A, AB, B, O, Yt^b, Do^a, Do^b, HPA-1a, Jk^a, Le^a, S, s, E, and e. Models for antigen/phenotype positivity of A_1, Co^b, Fy^a, Fy^b, HPA-1b, K, Kp^a, Lu^a, M, N, and C^w had balanced accuracy ranging from 91.6 to 98.0%. Six models, A_2, Co^a, Jk^b, Le^b, D, and C had balanced accuracy ranging from 64.6 to 89.4%. The Finnish models for LW^b, P1, and c failed classification in the Danish cohort.

Validation of the classification model algorithm in the Danish cohort

The RBC antigen/phenotype and HPA-1 typing and genotype data available for the Danish cohort enabled implementation of 40 Danish classification models representing 15 blood group systems. Due to missing genotypes (approximately 5%), missing allele dosage values were imputed separately for train and test data sets using mean values.

Median (IQR) balanced accuracy for the 40 Danish models in the Danish test data set was 99.3 (95.1–99.8)%. The evaluation metrics for test data set are available in **Table 4** and for the train and full data sets in **Supplementary Tables 6–7**, respectively. Majority (23/40) of the Danish models reached balanced accuracy of ≥99.0% including models for antigen/phenotype positivity of A, AB, B, O, Yt^a, Yt^b, Do^a, Do^b, Fy^a, HPA-1a, HPA-1b, Jk^a, Jk^b, M, N, S, s, C, c, D, E, Le^a, and Kn^b. Balanced accuracies for A_1, Co^b, Fy^b, K, Kp^a, Lu^a, C^w, e, and P1 models ranged from 94.4 to 98.1%, and for A_2, Co^a, k, Kp^b, Lu^b, Vel, and Le^b from 70.0 to 89.3%. Danish model for Kn^a failed classification due to too low number of Kn^a-negative samples in the test data set. Confusion matrices for the Danish models in the Danish train,

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

test and full data sets depict the distribution of TN, FN, TP, and FP samples and are illustrated in **Supplementary Figures 5–7**, respectively**.** The median (IQR) prediction error of the Danish models was 2.3 x 10^{-3} (9.3 x 10^{-4} –7.1 x 10^{-3})% (**Supplementary Table 5**).

Comparison of the Finnish and Danish classification models

Assembly of the balanced accuracies for Finnish and Danish models in the Finnish and Danish full data sets is presented in **Table 5**. When analyzing the shared 33 models, the Finnish models predicted the blood groups of the Finnish cohort more accurately than the blood groups of the Danish cohort (median [IQR] balanced accuracy 99.9 [98.8–100]% vs. 97.1 [91.6–99.5], p = 1.15e-06). The Danish models were performing better than the Finnish models in the blood group classification of the Danish cohort (median [IQR] balanced accuracy 99.5 [96.5–99.8]% vs. 97.1 [91.6–99.5]%, p = 0.006).

The number of genetic variants available for the Finnish random forest modelling ranged from 35 to 688 depending on the blood group/HPA system and number of the important variables selected by the classifier for the final models ranged from 12 to 214 (**Table 2**). In the Danish genotyping data set, the number of variants varied from 42 to 766 and the final models utilized 20–743 variants (**Supplementary Table 5**).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

DISCUSSION

Our study introduces random forest classification models for predicting RBC antigens/phenotypes and HPA-1 from array-based genotyping data. The method and models were generated utilizing blood group typing data from Finnish blood donors and further validated using a large Danish blood donor cohort. The results demonstrate high overall accuracy, and the method is suitable for biobank-scale screening and analysis of HPA-1 and RBC antigens.

Blood transfusion is one of the most common clinical procedures in the hospitals and the key element for safe transfusion is compatibility between the recipient and donor RBC antigens[1]. Although transfusion-related severe outcomes are rare, the prominent risk of sensitization and further alloimmunization affects especially patients dependent on recurrent transfusions [4,5]. Extended blood group typing has proven to be beneficial by reducing the incidence of alloantibody formation[28,29]. Additionally, studies have shown that the extended genotyping of blood donors markedly increases the number of suitable donors for immunized recipients [13] and enhances the supply of antigen-negative blood[6].

At present, preventive matching strategies are implemented only for specific patient groups and, despite the obvious advantages of the extended genotyping of donors, the procedure has not been considered feasible covering all blood donors. Over the last decades, the genotyping of different populations has expanded widely. Using machine learning approaches to screen blood donor and research biobank genotyping data may provide a cost-effective solution for enlarging the pool of antigen-negative blood donors. Our random forest classification method infers RBC antigens and HPA-1 from genotype-imputed microarray data. The R package ranger performed fast and handled the dimensionality of input data without problems[21]. The obtained results demonstrated high balanced accuracies both in the Finnish discovery cohort (median 99.8% for the 39 Finnish models) and in the Danish validation cohort (median 99.3% for the 40 Danish models) (**Table 5**). The performance was not affected by nearly a 100-fold size difference between the Finnish and the Danish cohorts (~1,200 vs. ~111,000, respectively).

Rh and MNS blood group system antigens have been challenging to determine by sequencing due to complex genetic variation and gene rearrangements[12,30]. We observed

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

reduced balanced accuracy in the Finnish model for hr^S (93.3%) and the Danish model for C^w (95.3%). However, the other Rh and MNS antigen models, including clinically significant E, e, C, c, S, and s, performed accurately. The balanced accuracies for clinically significant antigens in other systems, including K, Jk^a, Jk^b, Fy^a, and Fy^b, ranged from 95.6% to 100% (**Table 5**).

The BOOGIE method for prediction of RBC antigens was published in 2015[14]. It builds on 1-NN algorithm and implementation requires genotype sequencing data and curated haplotype tables for the RBC antigen phenotypes. When compared, the Finnish models for ABO and RhD performed better than the BOOGIE method (median balanced accuracy for the Finnish ABO models 100% vs. BOOGIE ABO accuracy 94.2%; balanced accuracy for the Finnish RhD model 98.8% vs. BOOGIE RhD accuracy 94.2%). The observed differences in accuracies could be explained by the potentially limited haplotype tables utilized by BOOGIE. Additionally, the reported results of BOOGIE are based on low number of samples.

When applying the Finnish models to the Danish cohort, the observed decrease in balanced accuracies was expected because of the evident genetic, genotyping, and imputation differences between the Finnish and the Danish cohorts (**Table 5**). The Finnish cohort was imputed using population-specific imputation reference panel having no missingness per individual. On the contrary, the Danish cohort was imputed using the North European reference sequence panel resulting in an average missingness of 5%. As random forest is not able to handle missing input data and the important variables of the Finnish models were not fully present in the Danish data, we were obliged to use mean imputation for missing variant dosage data. It is obvious that this approach also introduces errors to the data, which may partly explain the reduced accuracy. The better performance of the Danish models in Danish cohort underlined the benefit of the population-specific training cohort.

To our surprise, the Finnish genotyping data had only one variant in the *RHD* region. Nonetheless, the Finnish model for RhD performed with sufficient balanced accuracy in the Finnish cohort (98.8%). Our method combines *RHD* and *RHCE* region variants for the modelling and the high linkage disequilibrium may have supported the classification (**Table 2**). However, the Finnish model for RhD worked poorly in the Danish cohort (78.4%), which may be attributed to the mean imputation of missing values.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The present modelling method is restricted to the RBC antigen typing data available for the training and test data sets, which can be considered as a major limitation because the data for some RBC antigens are scarce. RBC antigens have demonstrated significant diversity among populations and rare blood group variants may not be discovered without substantially large typing numbers. The Danish model for Kn^a failed because of lacking Kn^a-negative samples in the test data set and we were not able to create Finnish models for e.g., Vel, k, Kp^b, Lu^a, and LW^a.

In the future, comprehensive donor and recipient typing and precision matching are likely to increase. A recent publication by van Sambeeck *et al*.[31] demonstrated the feasibility of preventive matching for all genotyped recipients and donors. Our method is suitable for initial screening for antigen-negative donors at biobank-scale, presenting a cost-effective solution for the extended blood group and HPA-1 typing. Additionally, successful prediction of polygenic blood groups may facilitate the research of desease associations in large biobanks.

Scripts for random forest modelling and for applying the tested 39 Finnish models are freely available in the GitHub. The implementation is possible in the local computing environment without sensitive data uploads and requires only a moderate level of bioinformatic skills.

AUTHOR CONTRIBUTION

K. Hyvärinen, K. Haimila, J. Partanen, and J. Ritari designed the study. The Blood Service Biobank provided the samples and genomic data. K. Haimila supervised blood group typing and provided blood group expertise. K. Hyvärinen and J. Ritari scripted the random forest method, carried the analyses out in the Finnish cohort, and interpret the results. O.B. Pedersen and C. Erikstrup established the Danish cohort. C. Moslemi executed the replications analysis in the Danish cohort. M.L. Olsson, S.R. Ostrowski. and O.B. Pedersen supervised and interpreted the analysis performed in the Danish cohort. K. Hyvärinen drafted the manuscript. All authors contributed to the final version of the manuscript.

ACKNOWLEDGEMENTS

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

We want to thank Dr. Satu Pastila and Ms. Ritva Toivanen at the FRCBS for the collaboration with the blood group typing data. We are also grateful for Ms. Birgitta Rantala, Mr. Petteri Vaskin, Ms. Katariina Karjalainen, Ms. Nina Nikiforow, Ms. Jonna Clancy, and Dr. Mikko Arvas and Dr. Tiina Wahlfors at the Blood Service Biobank for their help in handling the data and samples, and Dr. Jaana Mättö and the personnel at the FRCBS Blood Group Unit for blood group typing analyses.

From Denmark, we wish to thank the Danish blood donors and deCODE Genetics for genotyping the Danish cohort.

DISCLOSURE AND FUNDING

M.L. Olsson is an inventor on patents about Vel blood group genotyping (unrelated to the methods and models presented in this study) and owns 50% each of the shares in BLUsang AB, an incorporated consulting firm, which receives royalties for said patents. The other authors declare no conflicts of interest.

The study was partially supported by funding from the Government of Finland VTR funding, the Academy of Finland, the Finnish Cancer Association, and Business Finland. The study of the Danish cohort was supported by Independent Research Fund Denmark (project number 0214-00127B), Bloddonornes forskningsfond, and A.P Møller Fonden. M.L. Olsson is a Wallenberg Clinical Scholar funded by Knut and Alice Wallenberg Foundation.

SUPPORTING INFORMATION STATEMENT

Supplementary Information is online.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES

1. Goel R, Tobian AAR, Shaz BH. Noninfectious transfusion-associated adverse events and their mitigation strategies. *Blood*. 2019;133(17):1831-1839. doi:10.1182/blood-2018-10-833988

2. Hendrickson JE, Delaney M. Hemolytic Disease of the Fetus and Newborn: Modern Practice and Future Investigations. *Transfus Med Rev*. 2016;30(4):159-164. doi:10.1016/j.tmrv.2016.05.008

3. Bussel JB, Vander Haar EL, Berkowitz RL. New developments in fetal and neonatal alloimmune thrombocytopenia. *Am J Obstet Gynecol*. 2021;225(2):120-127. doi:10.1016/j.ajog.2021.04.211

4. Hendrickson JE, Tormey CA, Shaz BH. Red blood cell alloimmunization mitigation strategies. *Transfus Med Rev*. 2014;28(3):137-144. doi:10.1016/j.tmrv.2014.04.008

5. Evers D, Middelburg RA, de Haas M, et al. Red-blood-cell alloimmunisation in relation to antigens' exposure and their immunogenicity: a cohort study. *Lancet Haematol*. 2016;3(6):e284-e292. doi:10.1016/S2352-3026(16)30019-9

6. Flegel WA, Gottschall JL, Denomme GA. Implementing mass-scale red cell genotyping at a blood center. *Transfusion*. 2015;55(11):2610-2615. doi:10.1111/trf.13168

7. Cone Sullivan JK, Gleadall N, Lane WJ. Blood Group Genotyping. *Clin Lab Med*. 2022;42(4):645-668. doi:10.1016/J.CLL.2022.09.016

8. 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(6):e241-e251. doi:10.1016/S2352-3026(18)30053-X

9. Veldhuisen B, Van Der Schoot CE, De Haas M. Blood group genotyping: From patient to high-throughput donor screening. *Vox Sang*. 2009;97(3):198-206. doi:10.1111/j.1423-0410.2009.01209.x

10. Hyun J, Oh S, Hong YJ, Park KU. Prediction of various blood group systems using Korean whole-genome sequencing data. *PLoS One*. 2022;17(6). doi:10.1371/JOURNAL.PONE.0269481

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

11. Möller M, Jöud 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):240-249. doi:10.1182/bloodadvances.2016001867

12. ISBT Blood Group Allele Tables. https://www.isbtweb.org/isbt-working-parties/rcibgt/blood-group-allele-tables.html#blood group allele tables

13. Gleadall NS, Veldhuisen B, Gollub J, et al. Development and validation of a universal blood donor genotyping platform: A multinational prospective study. *Blood Adv*. 2020;4(15):3495-3506. doi:10.1182/bloodadvances.2020001894

14. Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, Tosatto SCE. BOOGIE: Predicting blood groups from high throughput sequencing data. *PLoS One*. 2015;10(4):1-15. doi:10.1371/journal.pone.0124579

15. Breiman L. Bagging predictors. *Mach Learn*. 1996;26(2):123-140.

16. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324/METRICS

17. Zheng X, Shen J, Cox C, et al. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J*. 2014;14(2):192-200. doi:10.1038/tpj.2013.18

18. Ritari J, Hyvärinen K, Clancy J, FinnGen, Partanen J, Koskela S. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics Bioinforma*. 2020;2(2):lqaa030. doi:10.1093/nargab/lqaa030

19. Vukcevic D, Traherne JA, Næss S, et al. Imputation of KIR Types from SNP Variation Data. *Am J Hum Genet*. 2015;97(4):593-607. doi:10.1016/j.ajhg.2015.09.005

20. Ritari J, Hyvärinen K, Partanen J, Koskela S. KIR gene content imputation from single-nucleotide polymorphisms in the Finnish population. *PeerJ*. 2022;10(e12692). doi:10.7717/peerj.12692

21. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Softw*. 2017;77(1):1–17. doi:10.18637/jss.v077.i01

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

22. Pedersen OB, Erikstrup C, Kotzé SR, et al. The Danish Blood Donor Study: A large, prospective cohort and biobank for medical research. *Vox Sang*. 2012;102(3):271. doi:10.1111/j.1423-0410.2011.01553.x

23. Hansen TF, Banasik K, Erikstrup C, et al. DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open*. 2019;9(6). doi:10.1136/bmjopen-2018-028401

24. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2023. https://www.r-project.org/

25. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686

26. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28(5):1–26. doi:10.18637/jss.v028.i05

27. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina (Kaunas)*. 2021;57(5). doi:10.3390/medicina57050503

28. Lasalle-Williams M, Nuss R, Le T, et al. Extended red blood cell antigen matching for transfusions in sickle cell disease: a review of a 14-year experience from a single center (CME). *Transfusion*. 2011;51(8):1732-1739. doi:10.1111/J.1537-2995.2010.03045.X

29. Schonewille H, Honohan Á, Van Der Watering LMG, et al. Incidence of alloantibody formation after ABO-D or extended matched red blood cell transfusions: a randomized trial (MATCH study). *Transfusion*. 2016;56(2):311-320. doi:10.1111/TRF.13347

30. Zhang Z, An HH, Vege S, et al. Accurate long-read sequencing allows assembly of the duplicated RHD and RHCE genes harboring variants relevant to blood transfusion. *Am J Hum Genet*. 2022;109(1):180-191. doi:10.1016/j.ajhg.2021.12.003

31. van Sambeeck JHJ, van der Schoot CE, van Dijk NM, Schonewille H, Janssen MP.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Extended red blood cell matching for all transfusion recipients is feasible. *Transfus Med*. 2022;32(3):221-228. doi:10.1111/TME.12831

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

FIGURE LEGENDS

Figure 1. Study design. Random forest classification models were generated using Finnish reference data set (n = 1,192). Allele dosages of genes determining RBC antigens/phenotypes and HPA-1 were combined with the antigen typing data. The dataset was divided randomly to train and test data sets. Random forest modelling was executed in the training data set (n = 596) and the important variables were selected using permutation. The models were evaluated in the test data set (n = 596) for prediction accuracy and errors. The final models were fitted using the full data set and both models and the method were validated in the Danish cohort (n = 111,677).

Figure 2. Confusion matrices of the Finnish models in the Finnish test data set. Confusion matrices for the Finnish antigen classification models in the Finnish test data set are presented in alphabetical order of the blood group systems. The RBC antigen/phenotype and HPA-1 typing results are on the x-axis and the model predictions on the y-axis. The antigen-negative samples are denoted by 0 and the antigen-positive samples by 1 on both axes. The numbers of true positive and true negative samples are depicted in the green boxes and the numbers of false positive and false negative samples in the red boxes.

Figure 3. Posterior probability boxplots of the Finnish models in the Finnish test data set. Posterior probability boxplots for the Finnish antigen classification models in the Finnish test data set are presented in alphabetical order of the blood group systems. The RBC antigen/phenotype and HPA-1 typing results are on the x-axis and the antigen-negative samples are denoted by 0 and the antigen-positive samples by 1. The posterior probabilities for samples range from 0 to 1 and are presented on the y-axis. Samples are depicted as open circles.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Blood group/HPA-1 antigen typing information of the Finnish and Danish cohorts.

Blood group/HPA system	Antigen[a]	Finnish cohort(n)	Antigen positivity in the Finnish cohort(%)	Danish cohort(n)	Antigen positivity in the Danish cohort(%)
ABO	A	1,192	40.9	111,656	41.2
ABO	A_1	70	62.9	17,541	75.3
ABO	A_2	70	38.6	433	27.0
ABO	AB	1,192	8.0	111,656	4.4
ABO	B	1,192	8.2	111,656	10.2
ABO	O	1,192	42.9	111,656	44.2
Cartwright	Yt^a	NA	NA	7,640	99.6
Cartwright	Yt^b	1,160	6.2	7,362	7.3
Colton	Co^a	1,162	99.8	7,490	98.5
Colton	Co^b	1,164	7.3	8,277	9.2
Dombrock	Do^a	1,162	53.1	7,352	62.9
Dombrock	Do^b	1,162	90.7	7,352	84.6
Duffy	Fy^a	1,177	67.0	77,920	67.0
Duffy	Fy^b	1,175	75.1	71,096	80.8
Gerbich	Ls^a	190	3.2	NA	NA
HPA-1	HPA-1a	232	91.4	518	97.1
HPA-1	HPA-1b	62	43.5	518	30.7
Kell	K	1,192	4.7	96,497	7.8
Kell	k	NA	NA	14,332	99.0
Kell	Kp^a	1,177	2.6	22,497	2.4
Kell	Kp^b	NA	NA	10,350	99.9
Kell	Ul^a	219	15.5	NA	NA
Kidd	Jk^a	1,177	72.6	77,986	76.6
Kidd	Jk^b	1,177	68.5	71,840	72.7
Knops	Kn^a	NA	NA	3,462	99.9
Knops	Kn^b	NA	NA	3,462	6.8
Landsteiner-Wiener	LW^b	327	5.2	NA	NA
Lewis	Le^a	253	10.2	9,220	17.4
Lewis	Le^b	253	80.2	8,925	56.5
Lutheran	Lu^a	1,164	3.6	9,615	8.6
MNS	M	1,177	87.7	29,642	78.1
MNS	N	1,177	59.0	14,837	71.6
MNS	S	1,177	54.2	28,949	50.4
MNS	s	1,177	87.6	24,291	90.9
P1PK	P1	253	75.9	7,465	77.4
Rh	C	1,192	54.4	44,451	63.3
Rh	c	1,192	79.4	42,968	82.2
Rh	C^w	1,177	3.2	40,720	3.3
Rh	C^x	337	12.2	NA	NA
Rh	D	1,192	71.6	111,667	79.5
Rh	E	1,192	19.9	89,289	28.1
Rh	e	1,192	96.2	82,467	97.3

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Rh	hr^B	1,162	96.4	NA	NA
Rh	hr^S	1,162	97.3	NA	NA
Vel	Vel	NA	NA	11,755	99.9

NA, data not available.

[a] O, A_1, and A_2 and in this column refer to phenotype.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Characteristics of the Finnish classification models.

Blood group/HPA system	Antigen[a]	Genes analyzed	n(variants available)	n(model variants)	Prediction error[b]
ABO	ABO	*ABO*	496	164	8.68E-04
ABO	A_1	*ABO*	496	18	1.29E-03
ABO	A_2	*ABO*	496	43	2.20E-02
Cartwright	Yt^b	*ACHE*	62	12	0.00E+00
Colton	Co^a	*AQP1*	108	15	4.26E-04
Colton	Co^b	*AQP1*	108	39	2.03E-03
Dombrock	Do^a	*ART4*	109	17	6.97E-05
Dombrock	Do^b	*ART4*	109	17	0.00E+00
Duffy	Fy^a	*ACKR1*	57	25	6.80E-05
Duffy	Fy^b	*ACKR1*	57	26	4.32E-03
Gerbich	Ls^a	*GYPC*	342	81	7.71E-03
HPA-1	HPA-1a	*ITGB3*	386	24	1.74E-03
HPA-1	HPA-1b	*ITGB3*	386	73	2.06E-02
Kell	K	*KEL*	127	33	2.93E-05
Kell	Kp^a	*KEL*	127	26	8.56E-05
Kell	Ul^a	*KEL*	127	21	2.01E-04
Kidd	Jk^a	*SLC14A1*	622	18	1.62E-04
Kidd	Jk^b	*SLC14A1*	622	178	8.14E-04
Landsteiner-Wiener	Lw^b	*ICAM4*	35	19	4.27E-03
Lewis	Le^a	*FUT2, FUT3*	199	45	4.13E-03
Lewis	Le^b	*FUT2, FUT3*	199	88	1.56E-02
Lutheran	Lu^a	*BCAM*	124	14	9.47E-04
MNS	M	*GYPA, GYPB, GYPC*	688	132	6.80E-03
MNS	N	*GYPA, GYPB, GYPC*	688	214	1.18E-02
MNS	S	*GYPA, GYPB, GYPC*	688	44	1.54E-05
MNS	s	*GYPA, GYPB, GYPC*	688	49	1.19E-04
P1PK	P1	*A4GALT, B3GALNT1*	368	135	1.73E-02
Rh	C	*RHCE, RHD*	266	108	9.23E-03
Rh	c	*RHCE, RHD*	266	68	3.23E-03
Rh	C^w	*RHCE, RHD*	266	53	2.80E-03
Rh	C^x	*RHCE, RHD*	266	12	4.70E-03
Rh	D	*RHCE, RHD*	266	113	7.47E-03
Rh	E	*RHCE, RHD*	266	80	1.07E-03
Rh	e	*RHCE, RHD*	266	23	1.33E-03
Rh	hr^B	*RHCE, RHD*	266	20	1.39E-03
Rh	hr^S	*RHCE, RHD*	266	22	7.71E-03

[a] O, A_1, and A_2 and in this column refer to phenotype.

[b] Misclassification frequency obtained from out-of-bag data.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3. Accuracy metrics for the Finnish models in the Finnish test data set.

Blood group/HPA system	Antigen[a]	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Balanced accuracy
ABO	A	0.997	1.000	1.000	0.996	0.999
ABO	A_1	1.000	1.000	1.000	1.000	1.000
ABO	A_2	1.000	1.000	1.000	1.000	1.000
ABO	AB	1.000	1.000	1.000	1.000	1.000
ABO	B	1.000	1.000	1.000	1.000	1.000
ABO	O	1.000	0.996	0.997	1.000	0.998
Cartwright	Yt^b	1.000	1.000	1.000	1.000	1.000
Colton	Co^a	1.000	1.000	1.000	1.000	1.000
Colton	Co^b	1.000	0.952	0.996	1.000	0.976
Dombrock	Do^a	1.000	1.000	1.000	1.000	1.000
Dombrock	Dob	1.000	1.000	1.000	1.000	1.000
Duffy	Fy^a	1.000	1.000	1.000	1.000	1.000
Duffy	Fy^b	0.979	1.000	1.000	0.993	0.990
Gerbich	Ls^a	1.000	0.667	0.989	1.000	0.833
HPA-1	HPA-1a	0.900	1.000	1.000	0.991	0.950
HPA-1	HPA-1b	1.000	1.000	1.000	1.000	1.000
Kell	K	1.000	1.000	1.000	1.000	1.000
Kell	Kp^a	1.000	1.000	1.000	1.000	1.000
Kell	Ul^a	1.000	1.000	1.000	1.000	1.000
Kidd	Jk^a	1.000	1.000	1.000	1.000	1.000
Kidd	Jk^b	0.995	1.000	1.000	0.998	0.997
Landsteiner-Wiener	LW^b	0.994	1.000	1.000	0.889	0.997
Lewis	Le^a	0.991	1.000	1.000	0.929	0.996
Lewis	Le^b	0.880	1.000	1.000	0.971	0.940
Lutheran	Lu^a	1.000	1.000	1.000	1.000	1.000
MNS	M	0.972	0.994	0.959	0.996	0.983
MNS	N	0.992	0.986	0.980	0.994	0.989
MNS	S	1.000	1.000	1.000	1.000	1.000
MNS	s	1.000	1.000	1.000	1.000	1.000
P1PK	P1	1.000	0.979	0.938	1.000	0.990
Rh	C	0.978	1.000	1.000	0.982	0.989
Rh	c	1.000	0.994	0.976	1.000	0.997
Rh	C^w	0.996	1.000	1.000	0.905	0.998
Rh	C^x	0.986	1.000	1.000	0.909	0.993
Rh	D	0.970	0.993	0.982	0.988	0.982
Rh	E	1.000	0.975	0.994	1.000	0.987
Rh	e	1.000	0.998	0.957	1.000	0.999
Rh	hr^B	1.000	0.998	0.955	1.000	0.999
Rh	hr^S	0.933	0.986	0.636	0.998	0.960

[a] O, A_1, and A_2 and in this column refer to phenotype.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4. Accuracy metrics for the Danish models in the Danish test data set.

Blood group/HPA system	Antigen[a]	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Balanced accuracy
ABO	A	0.999	0.998	0.998	0.999	0.998
ABO	A_1	0.971	0.992	0.974	0.991	0.981
ABO	A_2	0.975	0.810	0.933	0.922	0.893
ABO	AB	1.000	0.996	1.000	0.997	0.998
ABO	B	1.000	0.998	1.000	0.998	0.999
ABO	O	0.999	0.997	0.998	0.999	0.998
Cartwright	Yt^a	1.000	1.000	0.933	1.000	1.000
Cartwright	Yt^b	0.999	1.000	1.000	0.985	0.999
Colton	Co^a	0.500	1.000	0.750	0.999	0.750
Colton	Co^b	0.993	0.950	0.995	0.928	0.971
Dombrock	Do^a	1.000	0.999	0.999	1.000	1.000
Dombrock	Do^b	0.996	1.000	0.998	0.999	0.998
Duffy	Fy^a	0.995	0.998	0.996	0.998	0.997
Duffy	Fy^b	0.963	0.999	0.995	0.991	0.981
HPA-1	HPA-1a	1.000	1.000	1.000	1.000	1.000
HPA-1	HPA-1b	0.994	1.000	1.000	0.988	0.997
Kell	K	0.996	0.906	0.992	0.956	0.951
Kell	k	0.681	1.000	0.942	0.997	0.840
Kell	Kp^a	0.999	0.888	0.997	0.956	0.944
Kell	Kp^b	0.400	1.000	1.000	0.999	0.700
Kidd	Jk^a	0.996	0.998	0.995	0.999	0.997
Kidd	Jk^b	0.996	0.999	0.997	0.998	0.997
Knops	Kn^a	0.000	1.000	NA	0.999	0.500
Knops	Kn^b	1.000	1.000	1.000	1.000	1.000
Lewis	Le^a	0.996	0.986	0.997	0.981	0.991
Lewis	Le^b	0.499	0.979	0.948	0.717	0.739
Lutheran	Lu^a	0.997	0.931	0.994	0.971	0.964
Lutheran	Lu^b	0.750	1.000	0.923	0.999	0.875
MNS	M	0.989	0.998	0.993	0.997	0.994
MNS	N	0.981	0.990	0.975	0.993	0.986
MNS	S	0.997	0.997	0.997	0.997	0.997
MNS	s	0.985	0.999	0.995	0.998	0.992
P1PK	P1	0.939	0.996	0.986	0.982	0.968
Rh	C	0.994	0.999	0.999	0.997	0.997
Rh	c	0.997	0.999	0.995	0.999	0.998
Rh	C^w	0.998	0.898	0.997	0.952	0.948
Rh	D	0.994	0.999	0.997	0.998	0.997
Rh	E	0.999	0.989	0.996	0.998	0.994
Rh	e	0.962	1.000	0.993	0.999	0.981
Vel	Vel	0.568	1.000	0.955	0.999	0.784

[a] O, A_1, and A_2 and in this column refer to phenotype.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 5. Balanced accuracies for the Finnish and Danish models in full data sets.

Blood group/HPA system	Antigen[a]	Balanced accuracy		
		Finnish full data set Finnish models	Danish full data set Finnish models	Danish full data set Danish models
ABO	A	0.999	0.995	0.999
ABO	A_1	1.000	0.980	0.984
ABO	A_2	0.981	0.894	0.891
ABO	AB	1.000	0.998	0.999
ABO	B	1.000	0.987	0.999
ABO	O	0.999	0.995	0.998
Cartwright	Yt^a	NA	NA	1.000
Cartwright	Yt^b	1.000	0.998	0.998
Colton	Co^a	1.000	0.833	0.833
Colton	Co^b	0.988	0.926	0.963
Dombrock	Do^a	1.000	0.999	0.999
Dombrock	Do^b	1.000	0.999	0.999
Duffy	Fy^a	1.000	0.979	0.997
Duffy	Fy^b	0.993	0.971	0.980
Gerbich	Ls^a	0.917	NA	NA
HPA-1	HPA-1a	1.000	1.000	1.000
HPA-1	HPA-1b	0.986	0.939	0.999
Kell	K	1.000	0.916	0.956
Kell	k	NA	NA	0.838
Kell	Kp^a	1.000	0.935	0.961
Kell	Kp^b	NA	NA	0.750
Kell	Ul^a	1.000	NA	NA
Kidd	Jk^a	1.000	0.995	0.997
Kidd	Jk^b	0.999	0.662	0.997
Knops	Kn^a	NA	NA	0.750
Knops	Kn^b	NA	NA	1.000
Landsteiner_Wiener	LW^b	0.998	0.500	NA
Lewis	Le^a	0.998	0.981	0.989
Lewis	Le^b	0.968	0.729	0.736
Lutheran	Lu^a	0.988	0.956	0.964
Lutheran	Lu^b	NA	NA	0.828
MNS	M	0.984	0.970	0.993
MNS	N	0.988	0.947	0.987
MNS	S	1.000	0.995	0.997
MNS	s	1.000	0.988	0.993
P1PK	P1	0.984	0.500	0.965
Rh	C	0.989	0.646	0.997
Rh	c	0.997	0.500	0.997
Rh	C^w	0.999	0.918	0.953
Rh	C^x	0.997	NA	NA
Rh	D	0.988	0.784	0.997
Rh	E	0.998	0.992	0.995
Rh	e	0.988	0.982	0.984
Rh	hr^B	0.988	NA	NA
Rh	hr^S	0.933	NA	NA
Vel	Vel	NA	NA	0.811

NA, data not available.

[a] O, A_1, and A_2 and in this column refer to phenotype.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

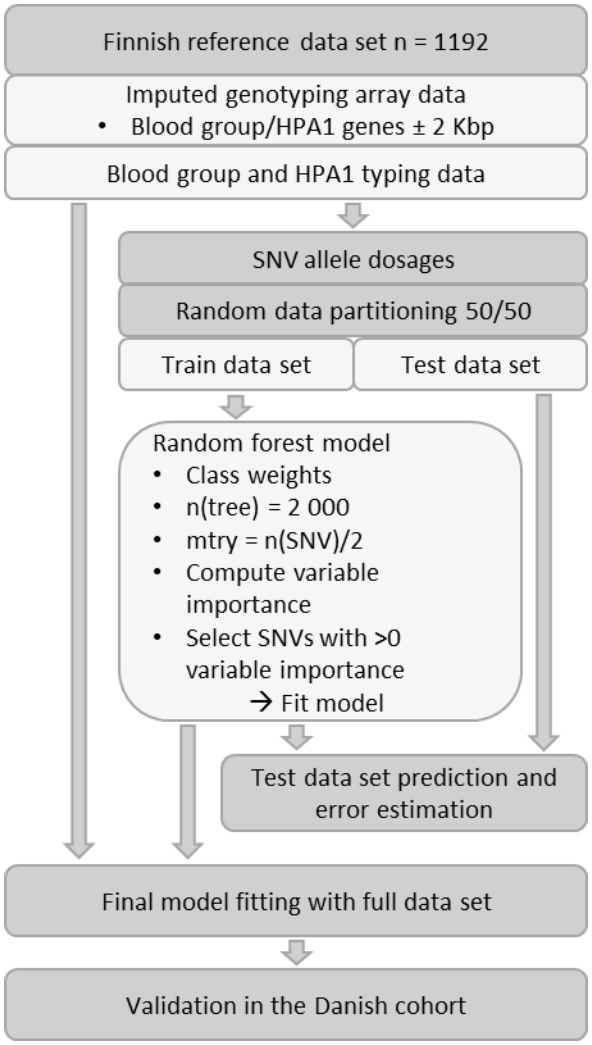

Figure 1. Study design.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

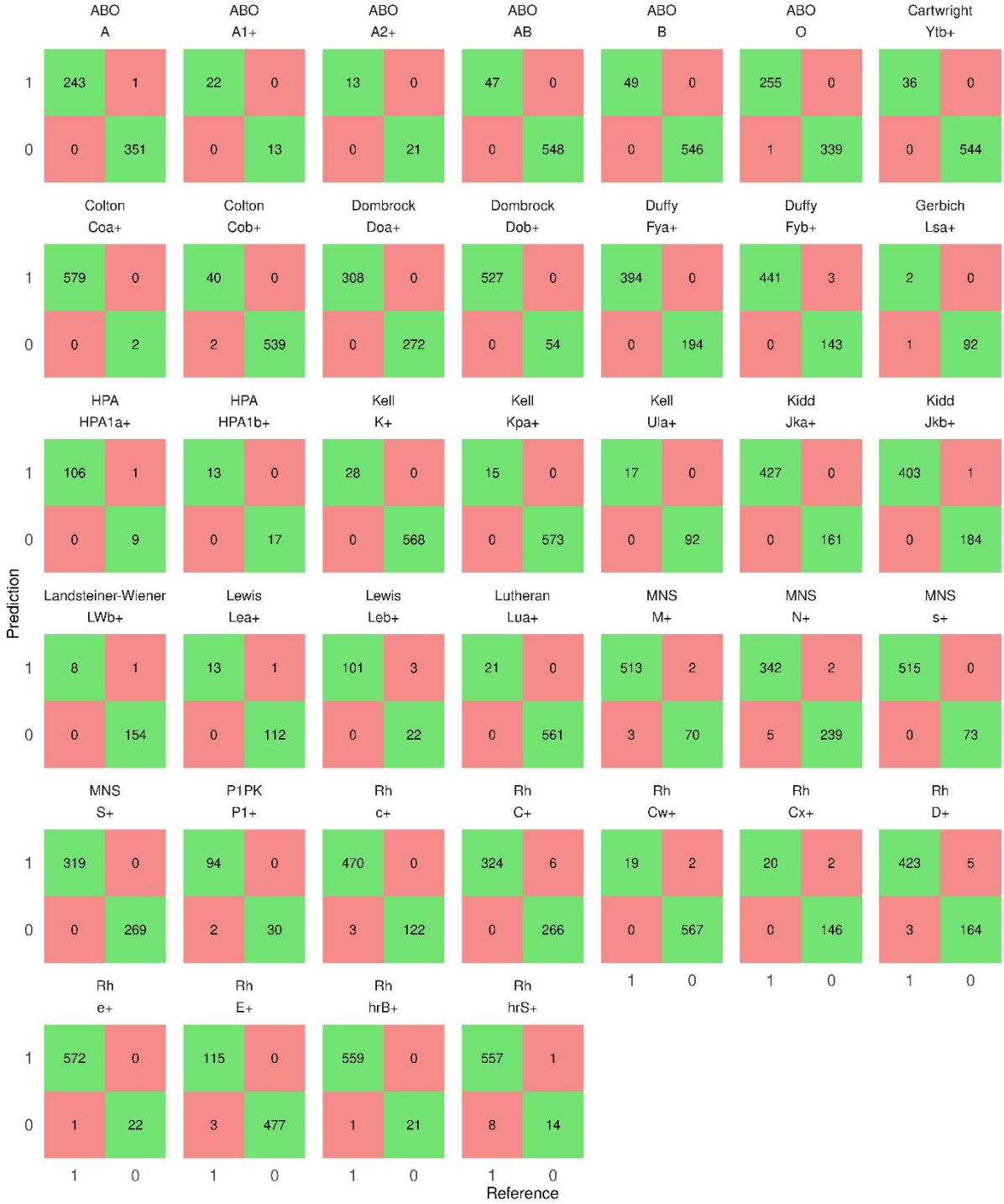

Figure 2. Confusion matrices of the Finnish models in the Finnish test data set.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.18.23295700; this version posted September 19, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

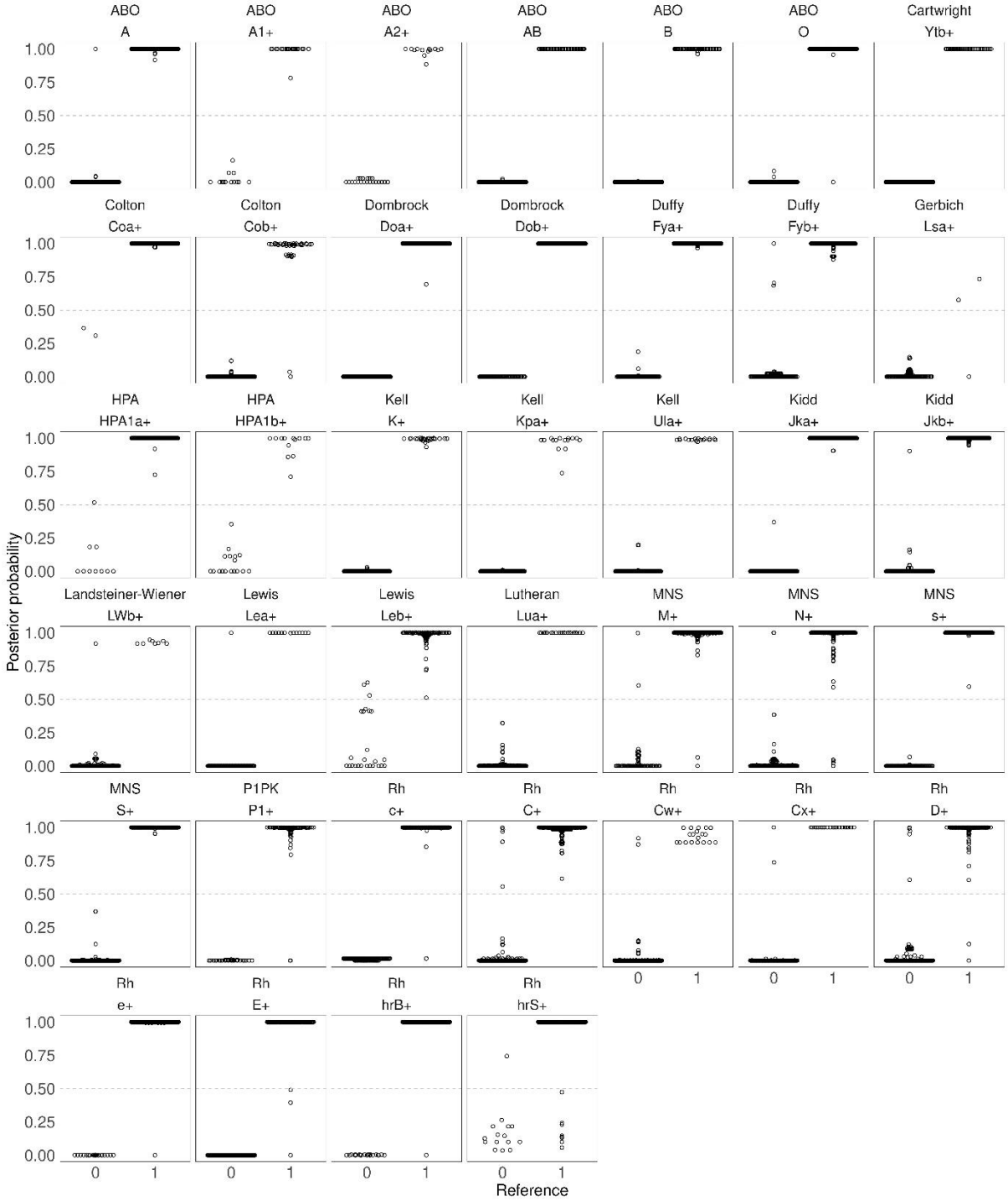

Figure 3. Posterior probability boxplots of the Finnish models in the Finnish test data set.