In silico blood genotyping from exome sequencing data

Manuel Giollo[1,2], Giovanni Minervini[1], Marta Scalzotto[1], Emanuela Leonardi[1], Carlo Ferrari[2] and Silvio C E Tosatto[1*]

[1]Department of Biology, University of Padova, Viale G. Colombo 3, 35131 Padova, Italy

[2]Department of Information Engineering, University of Padova, Via Gradenigo 6, 35121 Padova, Italy

Email: silvio.tosatto@unipd.it

[*]Corresponding author

Abstract

Background: Over the last decade, we have witnessed an incredible growth in the field of exome and genome sequencing. This information can be used to predict phenotypes for a number of traits of medical relevance. Here, we have focused on the identification of blood cell traits, developing BOOGIE, a tool recognizing relevant mutations through genome analysis and interpreting them in several blood traits important for transfusions.

Results: In our method, we extract relevant mutation data and annotate a genome with ANNOVAR. These variants are then directly compared with our knowledge base, containing association rules between mutations and phenotypes for the ten major blood groups: ABO, Rh, Duffy, Kell, Diego, Kidd, Lewis, Lutheran, MNS and Bombay. Whenever a match is found, it is used to predict the related phenotype and list causative mutations. The decision process is implemented as an expert system, automatically performing the logical reasoning connected to the genome variants. Interactions with other proteins and enzymes are easily kept into account during the full process, e.g. for the Bombay phenotype. This rare and easily misclassified genetic trait involves three blood groups, making blood donations potentially lethal.

Conclusions: BOOGIE was tested on Personal Genome Project (PGP) data. The blood traits for genomes with available ABO and Rh annotation were correctly predicted in between 86% and 100% of cases. The analysis is very efficient, making it suitable for genome scale diagnostic applications in personalized medicine. The versatility and simplicity of the analysis make it easily interpretable and allows easy extension of the protocol towards other blood related traits.

Background

Genome analysis problem

Advances in genome sequencing over the last years have detected a huge amount of new Single Nucleotide Polymorphisms (SNPs) [1], producing a tremendous growth of mutation databases. As the understanding of these variants is still far from being comprehensive, several bioinformatics tools like SIFT [2] and PolyPhen [3] were developed. Despite the relatively good performance of these methods [4], predicting the loss of activity of a single protein is not sufficient to explain a phenotype. The importance of considering several genetic loci and corresponding mutations to determine a phenotype has led to Genome Wide Association Studies (GWAS), focusing mainly on the multifactorial nature of traits [1]. As an example, consider the Bombay phenotype, a blood group for which expression of the trait depends on the ABO, FUT1 and FUT2 genes [5]. Finding genotype-phenotype correlations is a critical topic in personalized medicine, as personal genome sequencing is expected to become increasingly common over the next few years [6]. One of the most interesting developments in this field is the Personal Genome Project (PGP), collecting genome sequences and clinical phenotypes of participants who have signed an informed consent [7]. The goal is to make freely available for research the genome information for thousands of participants (PGP1K) [8].

Here, we describe the prototype of a new back-end tool, BOOGIE (BlOOd group Genome predIction Expert), for predicting the existence of antigens related to ten different blood groups through genome analysis based on SNP evaluation. Such a tool can assume strong relevance in blood transfusions, where only few blood systems are regularly considered in order to determine compatibility between donor and receiver [9].

Biological background

Blood groups are determined by the presence of specific proteins on the surface of red blood cells and body fluids [10]. These proteins act as antigens and can cause severe immune reactions whenever the immune system recognizes exogenous red blood cells. Usually, the antigenic determinants are oligosaccharides located on glycoproteins and glycolipids expressed on erythrocytes and tissue cells [11]. This carbohydrate component is then selectively modified by enzymes that are expressed by the same genes that determine the inclusion in a blood group system. A brief introduction to the most clinically relevant blood groups studied in this work follows and is also summarized in Table 1.

ABO group: The ABO blood type is the most important blood group system in medicine. Its antigenic determinants are oligosaccharides located on erythrocytes and tissue cells glycoproteins. Four phenotypes are related to the ABO system: A, B, AB and O. The ABO gene codes for the glycosyltransferases that transfer specific sugar residues to H substance. Depending on the transferred sugar, two different antigens A or B are obtained [11].

Rh group: The Rhesus blood group is the second most important blood system in humans. The Rh blood group system is highly polymorphic, consisting of over 45 independent antigens. Clinically, the correct

recognition of the Rh factor is important in blood transfusion and in the prevention and diagnosis of erythroblastosis fetalis disease [12].

Duffy group: This blood system, also known as the Duffy antigen receptor for chemokines (DARC), is actively expressed by erythrocytes and endothelial cells. This antigen is of a certain importance in patients who receive regular blood transfusions such as hemophiliacs [13].

Kell group: Kell antigen is a glycoprotein expressed on red blood cell membranes encoded by Kel gene, homologous to zinc-binding family metalloendopepsidases. The Kell system seems to be involved in alloimmunization in thalassemic patiens and in hemolytic disease in newborns [14].

Diego group: This system consists of 21 antigens. The antithetic couple Di^a/Di^b and Wr^a/Wr^b are considered the most common. Other 17 antigens are poorly distributed and considered local variation [15].

Kidd group: The Kidd blood type is one of the not Rh-dependent causes of newborn hemolytic disease. It remains difficult to detect due to its high serologic variability and weak *in vitro* expression [16].

Lewis group: Lewis antigens include type 1 (Lewis a and b) and type 2 (Lewis X and Y) carbohydrates. Lewis X and Y were recently identified as tumor-associated markers [17].

Lutheran group: The Lutheran gene (Lu) encodes for a glycoprotein of the Ig transmembrane receptor superfamily (IgSF). Lutheran includes Lua and Lub antigens, with the latter being very rare [18].

MNS group: The MNS system is the second blood group system discovered. It includes 46 antigens and at least 16 result from genetic recombination. MNS mismatch causes hemolytic newborn disease [19].

Bombay group: The Bombay phenotype (hh) is a severe mutation that causes silencing of the gene encoding for the H antigen present in blood group ABO. As a result, Bombay phenotypes result unable to produce either A or B antigen on red blood cells [5].

Expert systems

In order to predict the blood phenotypes induced by genome variants, there are many explicit rules in the literature that can be considered for prediction purposes. This knowledge takes the form of “IF – THEN” sentences, e.g. “IF there is a total RHD deletion in the genome THEN the patient has D- phenotype in the RH blood system”. A large amount of explicitly coded relationships is available in databases (DBs) such as BGMUT [20]. This suggests the use of an inference chaining procedure as sufficient to decide for a given phenotype from the entire genome. Hence, we chose to exploit the principles of expert systems [21]. The idea behind this kind of predictors is simple. Known facts can be iteratively used by inference rules for finding new facts, and eventually decide about the problem of interest. This kind of system emulates part of the decision process taken by a human expert, since the program considers the known facts about a given domain of knowledge. Another interesting point about expert systems is their ability to exploit human intuition by means of the so-called conversational process [21], where machine and expert user interact to solve situations that are too complex for automatic computation.

Methods

BOOGIE is designed as an in-house expert system returning blood trait predictions from a genome sequence input. The overall workflow is shown in Figure 1. The genome sequence is first reduced to a subset of SNPs on genes relevant for blood phenotypes using ANNOVAR [22]. This tool uses various filters to identify variants likely to have a functional impact, effectively reducing the SNPs to be analyzed by several orders of magnitude. It also adds useful information on gene position and SNP type while supporting the most common genome file formats. In our implementation, the entire genome is represented as a tree structure with four levels, as shown in Figure 2. Trees are the most suitable representation, since genetic data is fixed and the analysis puts more effort on the localization of relevant mutations (i.e. like the offline search problem). Hence, search time complexity in this context drops from O(n) to O($log\ n$), where n is the number of mutations. Inference is used to make decisions from the annotated data. To predict the blood group of a patient (say, the Rh blood trait), the genome data is loaded into the tree-like structure. Then, taking advantage of a newly developed knowledge base, we check if the preconditions of each rule are satisfied according to the tree, and predict the phenotype arising from the genome. The knowledge base for instance knows that "IF L245V and G336V mutations occur in the RHCD gene THEN phenotype is *c e*". Simply put, we look for the existence of these mutations in the patient genome, and if they both appear, the user is notified about evidence for the *c* and *e* traits. Information on heterozygosity and dominance is taken into account in the process through proper definition of conditional statements in the knowledge base. A simple conversational process is currently implemented by providing the user with alternative interpretations for contradictory data. The method is very efficient, as it can represent an entire exome in just 20 MB of memory, and can load its data and analyse it in seconds on a HP G62 laptop.

Results

Knowledge base

BOOGIE relies heavily on the availability of known mutations and their phenotypes. All the data was manually gathered by a human annotator, and stored in a knowledge base. Data extracted from BGMUT [20], dbSNP [23], Uniprot [24], OMIM [25], and PubMed [26] as of May 2012 was manually curated to remove redundancy and resolve possible contradictions. Blood trait allelic data was stored in text files with the chromosome of interest, gene, locus, amino acid and nucleotide change, metabolic pathways involved, DB references, gene ontology annotations, known blood traits, population distribution and correlation with other phenotypes. A total of 580 rules were derived for usage in the identification of traits during this data retrieval step. It should be noted that building such a knowledge base is not a straightforward process [27]. The exomes under analysis may be built on different reference genomes, with sequencing tools generally using either NCBI build 36 or 37, e.g. the PGP data so far uses both. This is a key issue, requiring a conversion step to properly interpret the meaning of a variation. On the other hand, published mutations may assume a different reference genome. A simple example is the

ABO gene: the A blood-group is used as reference in the literature, whereas NCBI genome build 37 defines the 0 blood group as reference. This may be a tricky problem, since most of the known SNPs have to be transformed. E.g. 297G>A, known to lead to 0 phenotype [20], will be found in a genome in the form of 296A>G, leading to an A blood group due to the inverted nucleotides. Finally, whenever few mutations express weak phenotypes, uncertainty in the decision is higher, such as T1136C in the RH gene. This is the main reason suggesting a conversational process in the expert system.

Patient data and classification

Testing BOOGIE is severely hampered by the scarcity of available genotype-phenotype pairs. To the best of our knowledge, only the PGP project readily provides this data. The first 10 PGP participants, which have the most extensive annotation, were selected as test case for BOOGIE. The method predicts blood groups for all tested genomes and provides additional information not immediately available through standard clinical trials such as classification in subgroups and presence of alleles suggesting a weak antigenic response. Representative results can be seen in Table 2. BOOGIE predicts correctly the ABO type for all 9 cases with known blood phenotype (PGP3 contains no information). The Rh blood type is predicted correctly in 8 out of 9 cases. For PGP4, BOOGIE predicts an intermediate condition, potentially dangerous in case of pregnancy, which was interpreted as Rh+. The available genomic data is apparently not sufficient to explain the Rh- phenotype. Overall BOOGIE accuracy for PGP10 is 100% for ABO and ca. 88% for Rh. No experimental validation is possible for the other blood groups considered in this research, at present. In view of the performance for the ABO and Rh groups, we expect a reasonable accuracy for these, although validation is still pending.

A second test was performed on participants from PGP1K. This dataset differs from PGP10 as microarray data on a panel of known SNPs is provided instead of genome sequences. The available data is also much sparser, allowing us to select a total of 22 participants with known blood groups (as of August 2012). Results for this dataset, shown in Figure 3, broadly confirm that BOOGIE is able to correctly classify most of the participants for ABO, but has lower accuracy for Rh-. This is mostly the case for participants with incomplete SNP coverage, where missing information on a few critical SNPs hampers correct classification. The overall accuracy for both blood groups nevertheless remains ca. 86% (19 out of 22).

Discussion

In this work we have developed BOOGIE, a back-end expert system for the prediction of phenotypes related to blood cell traits. The application is particularly suitable for analysing a large amount of data, due to its efficiency both from the time complexity and memory management point of view. This is possible due to the known data necessary for a decision is stored in the knowledge base. Decisions taken by BOOGIE are easy to interpret by a human expert, as predictions rely on chaining of well-known facts available from the literature. This is an appealing aspect, which differs significantly from the largest part

of recent bioinformatics tools, which mostly use complex decision methods and learning principles that cannot be directly interpreted by an end-user. A key aspect is the flexibility of the system. Adding more data to the knowledge base will allow us to tackle similar genotype to phenotype problems, an important step towards predicting the likelihood of disease in personalized medicine. The approach is related to the simple univariate analysis used in the last years by life scientists for phenotype explanation. Several rules in our knowledge base consider only one SNP at a time to reach reasonable decisions for most blood traits. In order to strengthen the system, we plan to add optional modules for multivariate analysis of GWAS genomes in the future. This will be essential for the prediction of complex diseases like Asthma or Crohn's disease, where most genotype-phenotype relations are still unclear, so an inference procedure relying on rules from the literature cannot be used. This improvement can also be of great utility for the conversational process.

BOOGIE is also a powerful tool for the analysis of minor blood groups, such as those systems for which experimental techniques for antigen detection are poorly sensitive, e.g. the KELL system. In these cases, analysis at the genome level can be a significant step forward. In addition to blood transfusion targets, the platform can also be a valuable tool for the study of population. Some anthropological marker genes are important due to ethnicity-specific polymorphisms of certain human populations. The versatility of the tool allows us to imagine different scenarios where similar methods will be used for detection of rare diseases or, in forensic medicine, to solve paternity assignment disputes.

Acknowledgements

The works is supported by Cariplo grant 2011/0724 and FIRB Futuro in Ricerca grant RBFR08ZSXY to

S.T. G.M. is an AIRC research fellow.

References

1. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L *et al*: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**(7311):52-58.

2. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863-874.

3. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.

4. Thusberg J, Vihinen M: **Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods**. *Hum Mutat* 2009, **30**(5):703-714.

5. Dipta TF, Hossain AZ: **The Bombay blood group: are we out of risk?** *Mymensingh Med J* 2011, **20**(3):536-540.

6. von Bubnoff A: **Next-generation sequencing: the race is on**. *Cell* 2008, **132**(5):721-723.

7. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, Angrist M, Bhak J, Bobe J, Callow MJ *et al*: **A public resource facilitating clinical use of genomes**. *Proc Natl Acad Sci U S A* 2012, **109**(30):11920-11927.

8. **Personal Genome Project** [http://www.personalgenomes.org/]

9. Anstee DJ: **Red cell genotyping and the future of pretransfusion testing**. *Blood* 2009, **114**(2):248-256.

10. Denomme GA: **Molecular basis of blood group expression**. *Transfus Apher Sci* 2011, **44**(1):53-63.

11. Seltsam A, Hallensleben M, Kollmann A, Blasczyk R: **The nature of diversity and diversification at the ABO locus**. *Blood* 2003, **102**(8):3035-3042.

12. Avent ND, Reid ME: **The Rh blood group system: a review**. *Blood* 2000, **95**(2):375-387.

13. Zhao Y, Mangalmurti NS, Xiong Z, Prakash B, Guo F, Stolz DB, Lee JS: **Duffy antigen receptor for chemokines mediates chemokine endocytosis through a macropinocytosis-like process in endothelial cells**. *PLoS One* 2011, **6**(12):e29624.

14. Yang MH, Li L, Kuo YF, Hung YS, Yu LC, Hung CS, Tsai SJ, Lin KS, Chu DC: **Genetic and functional analyses describe a novel 730delG mutation in the KEL gene causing K0 phenotype in a Taiwanese blood donor**. *Transfus Med* 2011, **21**(5):318-324.

15. Xu XG, He J, He YM, Tao SD, Ying YL, Zhu FM, Lv HJ, Yan LX: **Distribution of Diego blood group alleles and identification of four novel mutations on exon 19 of SLC4A1 gene in the Chinese Han population by polymerase chain reaction sequence-based typing**. *Vox Sang* 2011, **100**(3):317-321.

16. Liu JC, Wang Y, Liu FP, He YS: **The manual Polybrene test has limited sensitivities for detecting the Kidd blood group system**. *Scand J Clin Lab Invest* 2009, **69**(7):797-800.

17. Soejima M, Koda Y: **Molecular mechanisms of Lewis antigen expression**. *Leg Med (Tokyo)* 2005, **7**(4):266-269.

18. Kikkawa Y, Miwa T, Tohara Y, Hamakubo T, Nomizu M: **An antibody to the lutheran glycoprotein (Lu) recognizing the LU4 blood type variant inhibits cell adhesion to laminin alpha5**. *PLoS One* 2011, **6**(8):e23329.

19. Heathcote DJ, Carroll TE, Flower RL: **Sixty years of antibodies to MNS system hybrid glycophorins: what have we learned?** *Transfus Med Rev* 2011, **25**(2):111-124.

20. Patnaik SK, Helmberg W, Blumenfeld OO: **BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems**. *Nucleic Acids Res* 2012, **40**(Database issue):D1023-1029.

21. Russell SJ, Norvig P: **Artificial Intelligence: A Modern Approach**, 3rd edn: Pearson Education; 2009.

22. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**(16):e164.

23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

24. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource**. *Nucleic Acids Res* 2011, **39**(Database issue):D214-219.

25. **Online Mendelian Inheritance in Man, OMIM** [http://www.ncbi.nlm.nih.gov/omim]

26. **PubMed** [http://www.ncbi.nlm.nih.gov/pubmed/]

27. Thomas PE, Klinger R, Furlong LI, Hofmann-Apitius M, Friedrich CM: **Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers**. *BMC Bioinformatics* 2010, **12 Suppl 4**:S4.

Figures

Figure 1 – Overview of the BOOGIE workflow.

The genome data is annotated by ANNOVAR, and the resulting information is used to infer phenotypes. Optionally, the user can interact with the system whenever the produced output do not satisfy the expectation (e.g. BOOGIE cannot decide between two phenotypes due to unavailable data). b) Overview of the tree-like representation of the genome

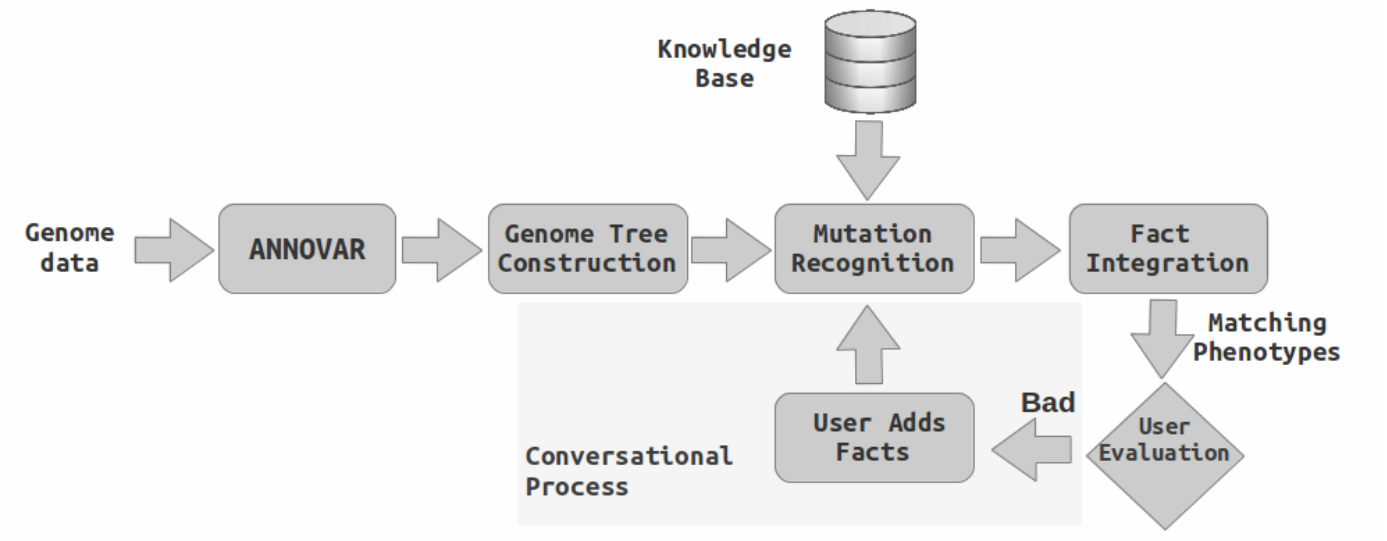


Figure 2 – BOOGIE tree-like genome representation levels.

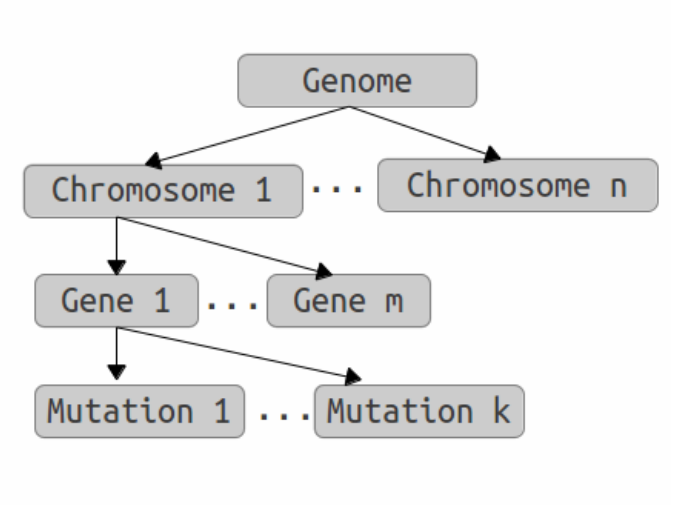
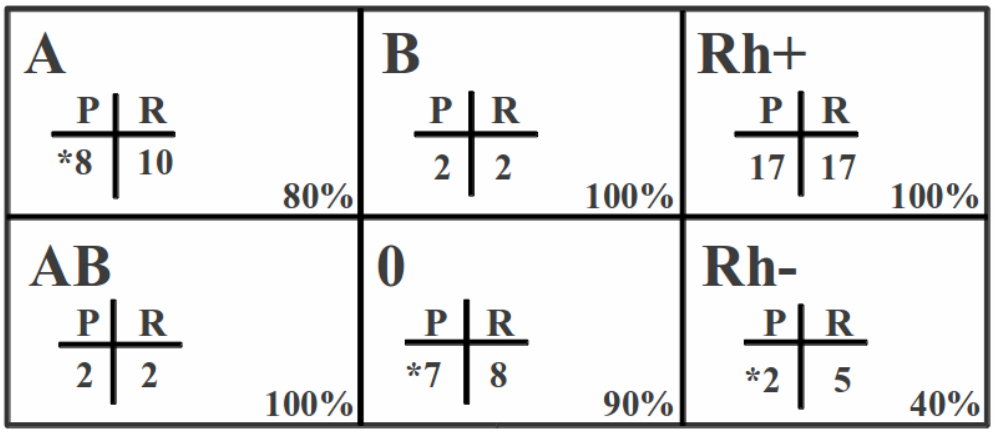


Figure 3 – *BOOGIE performance on the PGP1K dataset*

Benchmarking results on the 22 PGP1K participants with publicly available ABO and Rh phenotypes. The comparison between predicted (P) and real (R) data is shown for each phenotype (A, B, AB, 0, Rh+ and Rh-) with the respective accuracy in the lower right corner. Note that this dataset contains micro array observations rather than genome data, making it intrinsically incomplete. Incorrect predictions marked with (*) have missing data for some important SNPs used by BOOGIE.



Tables

Table 1 – Overview of the ten used blood systems.

Blood	*Relevant genes*	*Possible antigens*
ABO	ABO	A, B, O
RH	RHCE, RHD	D, E, e, C, c plus 50 minor antigens
Duffy	DARC	FY(a), FY(b)
Kell	KEL	K1, K2, plus 23 minor antigens
Diego	SLC4A1	Di^a, Di^b, Wr^a/Wr^b
Kidd	SLC14A1	Jk(a), Jk(b)
Lewis	FUT3	a, b
Lutheran	BCAM	Lu(a), Lu(b) plus 15 minor antigens
MNS	GYPA, GYPB, GYPE	M, N, S, s plus 40 minor antigens
Bombay	FUT1, FUT2	H, secretor

Table 2 – Predicted phenotypes for selected PGP10 participants.

Results are shown for three representative participants, listing the known phenotype and BOOGIE predictions for sub-groups in each phenotype. In this table Rh+ assignments are a consequence of predicted "c; e; weakD" antigenes. The predictions are correct except for PGP4, erroneously predicted to be Rh+.

	PGP1	**PGP4**	**PGP8**
Known	O +	A -	B +
ABO	O	A	B
Rh	c; e; weak D	c; e; weak D	c; e; weak D
DUFFY	FY(a+); FY(b-)	FY(a-); FY(b+)	FY(a-); FY(b+)
KELL	K2; K21+; K4-; K3-; K11; K17; K14,; K24; K6+; K7-	K2; K21+; K4-; K3-; K11; K17; K14; K24; K6+; K7-	K2; K21+; K4-; K3-; K11; K17; K14; K24; K6+; K7-
Diego	Dib; Memph neg	Dib; Memph neg	Dib; Memph neg
KIDD	Jk(a-); Jk(b+)	Jk(a-); Jk(b+)	Jk(a+); Jk(b-)
Lewis	negative	negative	negative
Lutheran	Lu(a-); Lu(b+); Lu6+; Lu9-; Lu4; Lu8+; Aua+; Aub-	Lu(a-); Lu(b+); Lu6-; Lu9+; Lu4-; Lu8+; Aua-; Aub+	Lu(a-); Lu(b+); Lu6+; Lu9-; Lu4-; Lu8+; Aua+; Aub-
MNS	M; S	M; s	M,s
Bombay	H+; secretor	H+; secretor	H+; secretor