Editorial

Considerations on terminology and database organization for blood group genotyping data

Franz F. Wagner[1,2]

[1]Institute Springe, Red Cross Blood Service NSTOB, Springe, Germany; [2]MVZ am Clementinenkrankenhaus, Springe, Germany

Correspondence to: Franz F. Wagner. DRK Blutspendedienst NSTOB, Institut Springe, Eldagsener Str. 38, 31832 Springe, Germany. Email: franz.wagner@bsd-nstob.de.

Received: 01 June 2021; Accepted: 26 September 2021.

doi: 10.21037/aob-21-42

View this article at: https://dx.doi.org/10.21037/aob-21-42

Terminologies and databases are necessary to cope with the increasing number of blood group alleles. Currently, the international society of blood transfusion (ISBT) allele terminology is widely used but it has weaknesses in special situations due to the linkage of allele name and serology. This commentary discusses the current status of nomenclatures and derives conclusions for database organization for blood group genotyping.

Background

In 1986, for the first time the nucleotide sequence of a gene underlying a blood group system, glycophorin A, was identified (1). Interest in molecular blood group determination really started off when the nucleotide structures of the genes underlying the ABO (2) and RH system (3,4) were identified in 1990. Within a few years after the identification of the main alleles, the molecular basis of variant alleles like A2 (5), A3 and B3 (6) for ABO and D category VI (7) for RH were published. In the following years, the number of known variant alleles exploded. Several databases were developed to present the accrued data on these alleles, and recently the ISBT decided to develop a database of blood group alleles. In this commentary, both the need for such databases as well as possible limitations and arising problems are discussed.

The need for a standardized terminology

Initially, authors usually named the alleles discovered by them according to the phenotype of the samples investigated, often complemented by a sample code. If several different alleles were associated with the same phenotype, they started to number them. Obviously, such procedure was likely to generate ambiguities, because different authors could use the same designation for different alleles: For example, the terms "h1" to "h5" to describe non-functional FUT1-alleles were both used by Wagner and Flegel (8) in the March 1997 issue of *Transfusion* and by Kaneko *et al.* (9) in the July 1997 issue of *Blood*. Both publications appeared almost simultaneous, and the same designations indicated different alleles depending on the publication (*Table 1*). An even more confusing example is the naming of alleles underlying the D category V phenotype, because in this case two allele listings accessible in the internet—BGMUT (11) and the Human RhesusBase (12,13)—used different numberings: for example, the allele originally described as DVa (Hus) (14) was dubbed DVa 4 in BGMUT and DV type 2 in RhesusBase.

The ISBT allele nomenclature

To prevent such confusion, the ISBT decided to devise a blood group allele nomenclature. The Red Cell Immunogenetics and Blood Group Terminology working party was mandated to define rules for naming and establish listings for each blood group system. In 2010, a draught of the nomenclature rules was presented in the Berlin meeting (15), and the first lists of alleles were published. Key features of this nomenclature included (16):

- The allele names started with a gene symbol derived from the blood group name (with exceptions for blood group systems based on more than one gene), followed by an asterisk.
- The names were grouped starting with more general information to allow designations for partially characterized alleles.

Table 1 Parallel description of FUT1 alleles "h1" to "h5"

Original designation	Nucleotide change	Protein change	Current ISBT name	Reference
h1	c.1047G > C	W349C	FUT1*01N.17	(8)
h1	c.695G > A	Trp232Ter	FUT1*01N.08	(9)
h2	c.[461A>G;474A>G;954T>A]	Tyr154Cys	FUT1*01N.02	(8)
h2	c.990delG	Pro331GlnfsTer6	FUT1*01W.20	(9)
h3	c.776T > A	Val259Glu	FUT1*01N.10	(8)
h3	c.721T > C	Tyr241His	FUT1*01W.14	(9)
h4	c.513G > C	Trp171Cys	FUT1*01N.04	(8)
h4	c.442G > T	Asp148Tyr	FUT1*01W.04	(9)
h5	c.944C > T	Ala315Val	FUT1*01N.14	(8)
h5	c.[460T>C;1042G>A]	Tyr154His, Glu348Lys	FUT1*01W.05.02	(9); allele described previously (10)

ISBT, international society of blood transfusion.

- The nomenclature focused on alleles relevant to transfusion medicine and did not necessarily include alleles irrelevant to transfusion medicine.
- Phenotypes and alleles could be listed in more than one place.
- Silent changes of the nucleotide sequence were disregarded.
- The phenotype encoded by the alleles was the key determinate of the naming, for example ABO null alleles were listed as O alleles even if their molecular structure was more similar to A than to the frequent O alleles.
- References were not included in the published allele listings, because these were linked in dbRBC.

In the following years, the ISBT allele nomenclature gained general acceptance and is currently regarded as gold standard for naming alleles.

Limitations of the ISBT allele nomenclature

While the ISBT allele nomenclature was devised in a way that was adapted to the state of molecular analysis at that time, some of the decisions turned out to have a downside, as discussed in the following paragraphs:

Antigen recombination impedes meaningful antigen-based allele names

The intention to allow naming of the alleles with incomplete data works just for the main SNV defining the "major" antithetical antigens of a blood group system. If there are more SNV of relevance, naming is getting complicated or even impossible.

For example, in the Kell blood group system, there are three pairs of antithetical antigens of general interest: K/k, Kp^a/Kp^b and Js^a/Js^b. These are defined by SNV at positions c.578, c.841 and c.1790, respectively. Since K/k is the most important antigen pair, all K positive alleles are designated as KEL*01.x while the k positive alleles are listed as KEL*02.x (17). Obviously, testing the SNV at c.578 will discriminate KEL*01 from KEL*02 alleles. Antigen Kp^a (K3) usually occurs on a k background in the KEL*02.03 allele, sometimes the same antigen is found in a K context defining the KEL*01.03 allele. As a result, demonstrating c.841T indicates that the allele is Kp^a positive but leaves it undecided whether the allele is KEL*02.03 or KEL*01.03. Happily, Js^a (K6) is usually found in a k-context in the KEL*02.06 allele. Thus, correlating antigen and allele numbers is challenging but works as long as there are not too much recombinations.

A much more difficult example is the Knops blood group system. Here, the "main" allele KN*01 carries the Kn^a (Kn1), McC^a (Kn3), Kn4, Kn8 and probably DACY antigens defined by p.Val1561, p.Lys1590, p.Arg1601, p.Ser1610 (together with p.Arg1601) and p.His1208, respectively (18). Alleles with the antithetical antigens are designated KN*02 (encoding for Kn^b instead of Kn^a), KN*01.06 (McC^b instead of McC^a), KN*01.07 (KN7 rather than KN4), KN*01.–08 (lacking Kn8), KN*01.–09 (lacking Kn9) and KN*01.–05 (lacking Yk^a = Kn5 defined by p.Thr1408). However, only

the positions 1561 to 1610 are closely linked, while there are recombinations between this block and position 1408 and between positions 1208 and 1408 (19). The p.Thr1408Met substitution defining the Kn:–5 phenotype usually occurs on a standard KN*01 background leading to the KN*01.–05 allele, but sometimes it occurs in a KN*01.–08 background leading to a yet unnamed allele. Furthermore, the KN*01.–05 allele occurs both in conjunction with p.1208His and p.1208Arg. The p.1208 polymorphism is the target of a relevant number of antibodies in the Knops blood group system (19), but obviously it is difficult to find useful allele names for samples in which only this polymorphism has been tested.

Serology as primary key of the allele name is sensitive to serologic uncertainties

The primary focus on transfusion relevance, antigens and phenotype induces difficulties to name alleles with inconsistent or incomplete serologic information. For example, the RHD alleles include separate lists for normal and partial phenotypes (RHD*01= standard RHD to RHD*62), alleles with weak D expression (RHD*01W.01 to RHD*01W.145), DEL phenotype (RHD*01EL.01 to RHD*01EL.46), D negative phenotypes (RHD*01N.01 to RHD*01N.78) (20). With the increasing use of molecular methods, the serologic description of many alleles is "D positive yet peculiar phenotype". Should such alleles get an RHD*x, RHD01W.x or RHD01EL.x designation? Even with better characterization, borderline alleles may appear as weak D or DEL depending on the methods used; and sometimes the partial D phenotype is disputed. Since the numbers are counted separately for each phenotype, moving an allele from one list to another is impossible without assigning a new number. A similar situation exists in the ABO system, the c.802G>A substitution vastly disrupts A transferase activity leading to an "O" phenotype (21) with a minimal residual A activity that can be demonstrated in expression models and is testified by an often weak anti-A isoagglutinine (22). This mutation is present in ABO*O02.x alleles and in ABO*Aw08 (as well as ABO*Bw18). The classification of these alleles as Aw or O may be a matter of taste.

Silent polymorphisms may be important

While the omission of silent changes is correct regarding the expected antigens, sometimes such changes are of relevance for the clustering of alleles. For example, a silent SNV at position c.357 (23) is of importance for grouping the FUT2 alleles in several populations. However, as silent SNV it is disregarded in the ISBT allele names for FUT2.

Users need references

The omission of references temporarily became a problem when dbRBC was taken offline. However, this limitation is easily overcome by adding this information.

Possible solutions for the limitations

Silent polymorphisms could easily be integrated without changing the whole naming process, and the references are already added in the new allele tables. In contrast, the problems with antigen recombination and discrepant serologic descriptions are inherent to the allele naming approach selected by the ISBT and cannot be overcome without abandoning the current ISBT allele names. Most likely, such step would incur more confusion than the current limitations. Nevertheless, it is obvious that any attempt to group allele names by criteria not derived from the molecular structure may lead to difficulties. Therefore, in a database the primary key should not reflect serology but such data should be added as additional fields.

The importance of allele databases.

In 1999, the first listings of blood group alleles appeared in the Blood Group Mutation database (now offline) (11), and the Human RhesusBase (www.rhesusbase.info) (12,13). More than ten years later, the first allele tables were shown at the ISBT website (https://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology). Since then, many resources list blood group alleles: Erythrogene (www.erythrogene.com/) (24), BloodAntigens (bloodantigens.com/cgi-bin/a/a.fpl) (25); Rheference (www.rheference.org) (26). In addition, the National Center for Blood Group Genomics maintains RHCE tables (www.bloodgroupgenomics.org/rhce/rhce-table). These databases get increasingly important for automated analysis of sequencing or next generation sequencing data, as several tools are available: bloodtyper (25,27), BOOGIE (28), RBCeq (29).

What is an allele?

Obviously, the central part of a blood group allele database

is a table with alleles. While this answer seems obvious, the question what defines an allele is far from trivial. In early days of molecular analysis, characterization was sometimes based on testing a few SNV only. For many years, the standard of characterizing ABO alleles consisted of sequencing exons 6 and 7 only, disregarding any changes in exons 1 to 5. While such characterization would be considered incomplete nowadays, the information on the relation with the phenotypes is out there, and the balancing act is how to deal with such incomplete yet important information.

As silent changes are neglected by the ISBT nomenclature, the ISBT alleles are effectively defined by the protein sequence. Most current publications include the full exon sequence, often with exon-near intron sequences, and possibly most current transfusion specialists would consider this characterization as a characterization of an allele. However, full genomic sequences are increasingly becoming available, and it is likely that in the future a full allele characterization is expected to include the full genomic sequence of the allele (30-32), or in case of RH and MNS, even of the full haplotype.

Obviously, it is easy to derive the less detailed descriptions from the more detailed ones. However, many data have been accumulated based on less detailed descriptions, and it is a balancing act for a database to deal with such data while allowing analyses based on the most detailed allele definition available. Possibly, the data model should include a hierarchical approach dealing with several levels of details and defining a possible inheritance between the more and less detailed descriptions.

Possible scope of allele information

One major intent of an allele database is the possibility to identify an allele and to get its name or designation. For many years, the ISBT allele listing just gave this information: the mutations, the name and a very imprecise phenotype description. Nevertheless, this bundled information was extremely important to enhance the communication between scientists who could now indicate alleles in a readable way rather than referring to long molecular structures or self-invented names.

The next important information are the references, which were initially omitted in the ISBT allele listings but are currently added. The availability of references is extremely helpful for scientists, as they can easily access additional information on a defined allele. Even the oldest allele listings like the Blood Group Mutation database (11) and the Human RhesusBase (12) included references for these alleles.

Other information of pertinent interest includes a phenotypic characterization of the allele regarding different antigens (normal, weakened or absent expression), the frequency of this allele in different populations, clinical observations (immunization, disease association) correlated with the allele. Often, direct evidence is lacking, but structural models may predict them.

Web frontend and querying

Usually, the users communicate with an allele database by a web frontend. Older listings were simple lists of alleles (e.g., the "old" ISBT allele tables), or listings of alleles linked to descriptions of the alleles (e.g., the Human RhesusBase). For several blood group systems, the Blood Group Mutation database allowed to depict the alleles as excel files giving a rapid overview. More modern frontends as used in Erythrogene and Rheference allow searching for alleles by name, SNV or gene. The detailed information is often enriched with information on phenotype, frequency, or haplotype association. Very detailed information is given by Rheference, the Human RhesusBase and the Blood group genomics RHCE tables.

These features are becoming increasingly important, and the choice of a suitable programming framework for the frontend is an important for the programming efforts needed to generate a user-friendly frontend. Likewise, the expectations and needs of the users might differ between user groups and applications (identification of alleles of known serology, prediction of serology from sequencing data, analysis of results obtained with commercial test kits) and should be investigated, e.g., by surveys.

Data entry

Another crucial topic of an allele database is data entry. While the quality of this feature is invisible for the user, it is essential for the quality of the data in the database. Information in publications is usually not structured, and artificial intelligence analysis of publications are not yet in use. When the first alleles were described, they were usually not submitted to a nucleotide database but just mentioned in the publication. For some blood group systems, the counting of the nucleotides has been changed making the correct alignment of such descriptions to

modern sequences laborious. Even if sequences have been deposited in nucleotide databases, the description of serologic features is insufficiently standardized. Sometimes, sequence information has been found to be incorrect most likely because authors did not submit original sequences but edited a template, leading to series of incorrect alleles if the template contained an error.

Nowadays, the number of alleles described is painstakingly high, and the time needed to enter the data into the database is a limiting factor, because the specialists are often involved in other projects and have only limited time for data entry. Thus, when a new blood group allele database is designed, its ultimate fate will most likely depend on versatile options helping the curators to spend their time on important decisions rather than manually entering the data.

Summary and conclusions

A standardized blood group allele nomenclature is necessary to prevent ambiguities. For an allele database, the primary key should not depend on serology. A major challenge for the data model is the different level of allele characterization in different publications that might best be coped with by a hierarchical allele model. Besides information structuring, a user-friendly web front-end and a curator-friendly data entry support will be key factors for the success of a blood group allele database.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editor (Frederik Banch Clausen) for the series "Blood Group Genotyping" published in *Annals of Blood*. The article has undergone external peer review.

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/aob-21-42). The series "Blood Group Genotyping" was commissioned by the editorial office without any funding or sponsorship. The author had royalties from patents on the molecular biology of RH (weak D and RhD negative) and is member of the ISBT working party on red cell immunogenetics and blood group terminology and of the steering committee for the ISBT allele database. The author administrates the Human RhesusBase. The author has no other conflicts of interest to declare.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Siebert PD, Fukuda M. Isolation and characterization of human glycophorin A cDNA clones by a synthetic oligonucleotide approach: nucleotide sequence and mRNA structure. Proc Natl Acad Sci U S A 1986;83:1665-9.
2. Yamamoto F, Clausen H, White T, et al. Molecular genetic basis of the histo-blood group ABO system. Nature 1990;345:229-33.
3. Chérif-Zahar B, Bloy C, Le Van Kim C, et al. Molecular cloning and protein structure of a human blood group Rh polypeptide. Proc Natl Acad Sci U S A 1990;87:6243-7.
4. Le van Kim C, Mouro I, Chérif-Zahar B, et al. Molecular cloning and primary structure of the human blood group RhD polypeptide. Proc Natl Acad Sci U S A 1992;89:10925-9.
5. Yamamoto F, McNeill PD, Hakomori S. Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. Biochem Biophys Res Commun 1992;187:366-74.
6. Yamamoto F, McNeill PD, Yamamoto M, et al. Molecular genetic analysis of the ABO blood group system: 1. Weak subgroups: A3 and B3 alleles. Vox Sang 1993;64:116-9.
7. Mouro I, Le Van Kim C, Rouillac C, et al. Rearrangements of the blood group RhD gene associated with the DVI category phenotype. Blood 1994;83:1129-35.

8. Wagner FF, Flegel WA. Polymorphism of the h allele and the population frequency of sporadic nonfunctional alleles. Transfusion 1997;37:284-90.
9. Kaneko M, Nishihara S, Shinya N, et al. Wide variety of point mutations in the H gene of Bombay and para-Bombay individuals that inactivate H enzyme. Blood 1997;90:839-49.
10. Wang B, Koda Y, Soejima M, et al. Two missense mutations of H type alpha(1,2)fucosyltransferase gene (FUT1) responsible for para-Bombay phenotype. Vox Sang 1997;72:31-5.
11. Patnaik SK, Helmberg W, Blumenfeld OO. BGMUT Database of Allelic Variants of Genes Encoding Human Blood Group Antigens. Transfus Med Hemother 2014;41:346-51.
12. Wagner FF, Flegel WA. The Rhesus Site. Transfus Med Hemother 2014;41:357-63.
13. Wagner FF. Getting comfortable with RH blood group system terminologies and databases. ISBT Science Series 2019;14:24-31.
14. Rouillac C, Colin Y, Hughes-Jones NC, et al. Transcript analysis of D category phenotypes predicts hybrid Rh D-CE-D proteins associated with alteration of D epitopes. Blood 1995;85:2937-44.
15. Storry JR, Castilho L, Daniels G, et al. International Society of Blood Transfusion Working Party on red cell immunogenetics and blood group terminology: Berlin report. Vox Sang 2011;101:77-82.
16. Available online: https://www.isbtweb.org/fileadmin/user_upload/files-2015/red%20cells/blood%20group%20allele%20terminology/ISBT%20Guidelines%20Naming%20Blood%20Group%20Alleles%20v2.0%20110914.pdf
17. Available online: https://www.isbtweb.org/fileadmin/user_upload/_ISBT_006__KEL_blood_group_alleles_v5.0_01-MAR-2020.pdf
18. Available online: https://www.isbtweb.org/fileadmin/user_upload/Working_parties/WP_on_Red_Cell_Immunogenetics_and/022_KN_Alleles_v3_0_160704.pdf
19. Grueger D, Zeretzke A, Habicht CP, et al. Two novel antithetical KN blood group antigens may contribute to more than a quarter of all KN antisera in Europe. Transfusion 2020;60:2408-18.
20. Available online: https://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology
21. Yamamoto F, McNeill PD, Yamamoto M, et al. Molecular genetic analysis of the ABO blood group system: 4. Another type of O allele. Vox Sang 1993;64:175-8.
22. Seltsam A, Das Gupta C, Wagner FF, et al. Nondeletional ABO*O alleles express weak blood group A phenotypes. Transfusion 2005;45:359-65.
23. Kudo T, Iwasaki H, Nishihara S, et al. Molecular genetic analysis of the human Lewis histo-blood group system. II. Secretor gene inactivation by a novel single missense mutation A385T in Japanese nonsecretor individuals. J Biol Chem 1996;271:9830-7.
24. Möller M, Jöud M, Storry JR, et al. Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. Blood Adv 2016;1:240-9.
25. Lane WJ, Westhoff CM, Gleadall NS, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. Lancet Haematol 2018;5:e241-51.
26. Floch A, Téletchéa S, Tournamille C, et al. A Review of the Literature Organized Into a New Database: RHeference. Transfus Med Rev 2021;35:70-7.
27. Lane WJ, Vege S, Mah HH, et al. Automated typing of red blood cell and platelet antigens from whole exome sequences. Transfusion 2019;59:3253-63.
28. Giollo M, Minervini G, Scalzotto M, et al. BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. PLoS One 2015;10:e0124579.
29. Jadhao S, Davison C, Roulis EV, et al. RBCeq: An Integrated Bioinformatics Algorithm Designed to Improve Blood Type Compatibility Testing. bioRxiv 2021. doi: https://doi.org/10.1101/2021.01.13.426510.
30. Fichou Y, Berlivet I, Richard G, et al. Defining Blood Group Gene Reference Alleles by Long-Read Sequencing: Proof of Concept in the ACKR1 Gene Encoding the Duffy Antigens. Transfus Med Hemother 2020;47:23-32.
31. Srivastava K, Wollenberg KR, Flegel WA. The phylogeny of 48 alleles, experimentally verified at 21 kb, and its application to clinical allele detection. J Transl Med 2019;17:43.
32. Srivastava K, Fratzscher AS, Lan B, et al. Cataloguing experimentally confirmed 80.7 kb-long ACKR1 haplotypes from the 1000 Genomes Project database. BMC Bioinformatics 2021;22:273.

doi: 10.21037/aob-21-42
Cite this article as: Wagner FF. Considerations on terminology and database organization for blood group genotyping data. Ann Blood 2021.