Transfusion Medicine and Hemotherapy

Review Article

Transfus Med Hemother 2014;41:346–351
DOI: 10.1159/000366108

Received: April 1, 2014
Accepted: May 19, 2014
Published online: September 15, 2014

BGMUT Database of Allelic Variants of Genes Encoding Human Blood Group Antigens

Santosh Kumar Patnaik[a] Wolfgang Helmberg[b] Olga O. Blumenfeld[c]

[a]Department of Thoracic Surgery, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY, USA;
[b]Department of Blood Group Serology and Transfusion Medicine, Medical University of Graz, Graz, Austria;
[c]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY, USA

Keywords
Allele · Antigen · Blood group · Database · DNA variation · Mutation

Summary
The Blood group antigen Gene MUTation (BGMUT) database documents variations in genes of human blood group systems. In March 2014, the database, accessible at *www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.cgi?cmd=bgmut*, listed 1,545 alleles of 44 genes of 34 blood group systems. Besides allelic information, the BGMUT resource also presents comprehensive and current information on blood group systems. This review describes the database and notes its utility for the transfusion medicine and human genetics communities.

Introduction

We assume that the readership is familiar with the language of transfusion medicine, and terms such as blood group system, antigen, and phenotype require no explanation. The key features of blood group antigens are inheritable polymorphism among humans and association with distinct immunophenotypes that allow for serology-based typing of the antigens. The diversity among the antigens is a result of sequence variations in genes that directly or indirectly affect expression or epitopic nature of the antigens on the surface of erythrocytes. For any blood group system, only one or a few common phenotypes predominate in the population, with a frequency distribution of variant phenotypes asymptotically approaching zero though exceptions have been documented [1]. However, the necessity of testing donated blood for compatibility for blood transfusion has led to the identification of even extremely rare variants of blood types. Alleles of blood group antigen genes have been and continue to be discovered through the elucidation of genetic variations underlying such phenotypic variations. The Blood group antigen Gene MUTation (BGMUT) database has been striving for the past 15 years as an information repository of these allelic variations. A concise description of BGMUT last appeared in 2012 [2]. That publication was the first update on the status of the database since 2004 [3].

Beginnings and Growth of BGMUT

BGMUT was created in 1999 by one of us (OOB) to collate data providing insight into the nature of sequence variation among alleles of the glycophorin genes of the MNS blood group system that had been discovered in her and others' laboratories. Technical help was provided by another author (SKP) who was a doctoral student at the time. The development of BGMUT as a gene locus-specific database followed the guidelines outlined by Dr. Richard Cotton of the Human Variation Genome Society [4]. Alleles of the genes of the RH and ABO blood group systems were soon added to the database following contributions from Drs. Cheng Huang [5] and Fumi-ichiro Yamamoto [6], who were studying the systems, at the New York Blood Center, New York, and the Burnham Institute in California, respectively. More systems were included over time as some of their key investigators contributed summarized information for specific systems and annotated lists of alleles discovered in their and other laboratories. The enthusiasm and cooperation shown by these contributors was nearly universal and most gratifying, and greatly facilitated the growth of the database. By 2004 BGMUT had 624

KARGER

Fax +49 761 4 52 07 14
Information@Karger.com
www.karger.com

1660–3796/14/0415–0346$39.50/0

Accessible online at:
www.karger.com/tmh

Olga Blumenfeld, PhD
Department of Biochemistry
Albert Einstein College of Medicine
1300 Morris Park Avenue, Bronx, NY 10461, USA
olga.blumenfeld@einstein.yu.edu

Table 1. The 34 blood group systems in the BGMUT database

System	Genes (chromosomal location)	Alleles[a]	Variations	Protein function
ABO	ABO (9q34)	335	mut, del, ins, rearrangement	galactosyltransferase or N-acetylgalactosaminyltransferase
Chido/Rodgers (CH/RG)	C4A (6p21.3), C4B (6p21.3)	3 C4A, 4 C4B	mut, dup, rearrangement	Complement factor
Colton (CO)	AQP1 (7p14)	12	mut, del, ins	water channel
Cromer (CROM)	CD55 (1q32)	15	mut	complement cascade regulator
Diego (DI)	SLC4A1 (17q21-22)	91	mut, del, ins	anion exchanger
Dombrock (DO)	ART4 (12p12-13)	20	mut, del	ADP ribosyltransferase
Duffy (FY)	DARC (1q21-22)	11	mut, del	chemokine receptor
Fors	GBGT1 (9q34.12-3)	2	mut	N-acetylgalactosyltransferase
Gerbich (GE)	GYPC (2q14-21)	10	mut, rearrangement	cytoskeletal element
Gill (GIL)	AQP3 (9p13)	8	mut	water channel
H	FUT1 (19q13.3), FUT2 and pseudogene (19q13.3)	50 FUT1, 63 FUT2	mut, del, ins, unequal hr	fucosyltransferase (all)
I	GCNT2 (6p24.2)	13	mut, del	N-acetylglucosaminyltransferase
Indian (IN)	CD44 (11p13)	4	mut	adhesion
John Milton Hagen (JMH)	SEMA7A (15q22.3-23)	12	mut	unknown; signaling?
Junior (JR)	ABCG2 (4q22.1)	23	mut, del, ins	ATP-binding cassette transporter
Kell (KEL)	KEL (7q33)	74	mut, del, ins	metalloendopeptidase
Kidd (JK)	SLC14A1 (18q11-12)	34	mut	urea transporter
Knops (KN)	CR1 (1q32)	32	mut, del, dup	receptor for complement activation
Kx (XK)	XK (Xp21.1)	35	del	unknown; membrane transport?
Lan	ABCB6 (2q36)	37	mut, dup, del, ins	ATP-binding cassette transporter
Landsteiner-Wiener (LW)	ICAM4 (19p13.2-cen)	3	mut, del	adhesion molecule
Lewis (LE)	FUT3, FUT6, FUT7 (all 19p13.3)	53 FUT3, 20 FUT6, 2 FUT7	mut, ins	fucosyltransferase (all)
Lutheran (LU)	BCAM (19q13.2)	20	mut, del	adhesion
MNS (glycophorin A/B/E)	GYPA (4q31.21), GYPB (4q31.21), GYPE (4q31.1)	48	mut, gene conversion, unequal hr	unknown; adhesion?
Ok (OK)	BSG (19p13.3)	5	mut	receptor
P1PK (globoside)	A4GALT (22qq13.2), B3GALNT1 (3q25)	52 A4GALT, 13 B3GALNT1	mut, del, ins	galactosyltransferase (A4GALT), N-acetylgalactosaminyltransferase (B3GALNT1)
Raph (RAPH)	CD151 (11p15.5)	4	mut	adhesion
Rh (RH)	RHCE (1p36.11), RHD (1p36.11)	127 RHCE, 255 RHD	mut, del, gene conversion, hr?	unknown; membrane transport?
Rh-associated glycoprotein (RhAG)	RHAG (6p21.1-11)	23	mut	ammonium or carbon dioxide transporter
Scianna (SC)	ERMAP (1p34.2)	9	mut	unknown; adhesion?
T/Tn	C1GALT1 (7p21.3), C1GALT1C1 (Xq24)	1 C1GALT1, 14 C1GALT1C1	mut	galactosyltransferase (C1GALT1), chaperone (C1GALT1C1)
Vel	SMIM1	2	del	unknown
Xg (XG)	CD99 (Xp22-32, Yp11.3), XG (Xp22.33)	1 CD99, 1 XG	unknown	unknown; adhesion?
Yt (YT)	ACHE (7q22)	4	mut, del	acetylcholinesterase

del = Deletion; dup =, duplication; hr = homologous recombination; ins = insertion; mut = mutation.
[a]Including the reference ('wild-type') allele; as on 24 March 2014.

alleles of 27 blood group systems [3]. As of March 24, 2014, it had 1,545 alleles of 34 systems, all except one (T/Tn) of which are recognized by the International Society Blood Transfusion (ISBT) (table 1).

In its early years, BGMUT was a collection of static but frequently updated web pages hosted and technically supported by the Department of Biochemistry of the Albert Einstein College of Medicine in New York. In 2006, BGMUT

Table 2. Information content of the BGMUT resource[a]

Gene alleles	amino acid sequence variation contributor examined gene regions[b] (67%) name and aliases nucleotide sequence variation phenotype (89%) prevalence in population (90%) published reference[c] (93%) sequence identifier (ID) in sequence repositories (43%) unique allele ID in BGMUT
Blood group systems	contributing experts disease association genes and proteins IDs for other databases like OMIM and dbSNP introduction nature of gene variation reference sequence IDs serological aspects
Other	carbohydrate blood group antigens like CAD and Sid links to resources orthologous ABO, MNS and RH systems non-erythroid RHAG gene homologs tutorials

[a]As on 24 March 2014.
[b]Percentage of non-reference alleles for which such information is available.
[c]94% of the publications are indexed in the PubMed database

was transferred to the US National Center for Biotechnology Information (NCBI) and became a component of its dbRBC database under the direction of Dr. Wolfgang Helmberg, an author of this article.

Content of BGMUT and Data Access, Curation and Submission

BGMUT uses a Microsoft® SQL Server relational database for data storage, and SQL (structured query language), C++ and XSLT (extensible structured language transformations) codes are used to generate BGMUT's web-based interface. A description of the core attributes of BGMUT as per the BioDBCore specifications [7] has been published [2].

As an informational resource, BGMUT has three types of content (table 2). The primary content is about alleles of genes of various blood group systems. Allelic information includes alternate names, nucleotide and deduced amino acid sequence variations, regions of the genes that were examined, associated serological phenotypes, prevalence frequencies in the population, identifiers (IDs) to look up the allelic sequences in sequence databases such as NCBI GenBank [8], and citations of publications in which the alleles were first identified. BGMUT uses gene symbols recommended by the HUGO Gene Nomenclature Committee (HGNC) [9], and follows commonly used guidelines for describing sequence variations [10]. Other content in the BGMUT resource includes comprehensive information on the various blood group systems, orthologs for some of the systems in other species, primarily primates, links to external resources of relevance, and a few tutorials (table 2). BGMUT data is accessible over the web with a browser. Allelic information can be searched for with plain text queries without wildcard characters through a web-based form, and raw database content can be downloaded for offline viewing and analyses. An online form is provided for researchers to submit new alleles for inclusion in the database. Direct search of allele sequences has yet to be implemented.

The database is curated manually. New alleles are added by the curators based on periodic searches for alleles recently published in literature and submissions from researchers. Special attention is given to the scientific quality of the work underlying the new allele's discovery. For submissions, accessibility to the new allele's sequence in a sequence repository such as GenBank or a publication is required. The sequence variation associated with an allele that is described in a publication or provided by a submitter is critically evaluated. This reduces the chance of an allele being added to the database with an incorrect annotation on sequence variation. Researchers who discovered the allele are contacted in case of a disagreement between the curators' interpretation and published or submitted data regarding sequence variation. If questions remain unresolved, a note to that effect is added to the database record for the allele. A problem occasionally encountered is the use of a reference sequence by the investigator

different from the one used in BGMUT. This can happen if the reference sequence for a gene has been updated. At least in one instance, the curators of the reference sequence database agreed to roll back the update to accommodate BGMUT. In another instance, the update was based on sequence data from individual(s) with an antithetical blood group phenotype. A plan to resolve this reference sequence-related problem is noted below.

Plans for BGMUT's Future

Over the period of BGMUT's existence, two significant problems have become apparent that likely affect other mutation databases as well: allele nomenclature and information on the extent to which gene sequences were determined for alleles. While BGMUT provides unique, immutable IDs for its alleles, naming of alleles has been a problem since the inception of the database and the haphazard naming of blood group systems has been a part of the problem. For an allele not christened by its discoverers, BGMUT tries to create a simple informative name that indicates the gene and the sequence variation. For example, in the Diego blood group system, the BGMUT-assigned allele name *SLC4A1 118A* designates the allele commonly known as *Diego Montefiore* and indicates that the *SLC4A1* nucleotide at position 118 along the protein-coding cDNA sequence is an A in the allele. For alleles already named by the research community, which is the case for many alleles of the ABO and RH systems, BGMUT creates a gene-based designation as an alias. The gene- and variation-based nomenclature system becomes cumbersome when many nucleotide sequence positions are altered in an allele. This is the case encountered, for instance, with the ABO system in which numerous alleles show multiple substitutions. As expected, the designation of alleles originating through gene rearrangements becomes difficult and requires a more wordy description. The resolution of this nomenclature problem is a difficult but urgent task.

During the years of the early growth of BGMUT, nucleotide sequencing was an effort-intensive procedure that was still carried out in many laboratories manually. The extent to which an individual's gene sequence was examined to characterize a new allele covered one or a few exons and occasionally an intron. Thus, the identification of a relatively significant number of alleles in BGMUT is based on partial sequencing of genes. Sequence information for many alleles is therefore likely to be deficient, as some DNA changes may have been missed because of incomplete coverage. Because of technological advances and reduced costs, gene sequences of alleles deposited in BGMUT in recent years have been examined more extensively, and in some instances, such as in the ABO system, information is available on the nature of an allele's sequence in intronic and untranslated regions. With genome and exome sequencing becoming more common [11, 12], we anticipate that in the not too distant future it will become possible to examine in a single individual the entire sequences of all the genes that encode for blood group antigens. It will be interesting to see if haplotype-like combinations of alleles from multiple blood group genes can be identified from such data. Attempts in that direction are currently underway [13].

In order to avoid issues arising from changes in reference sequences (noted in the last section), BGMUT is planning an effort to provide unchangeable reference sequences. These sequences, identical to the reference sequences currently used by BGMUT, will become part of the collection of *NG sequences* of the NCBI RefSeq database of reference sequences [14]. The sequences, which will be permanent and referred to as *locus reference genomic sequences* (LRGs), will be generated by using a genomic scaffold, replacing the corresponding sequences with whatever the current reference sequence contains. The LRGs, though artificial, will provide complete genomic coverage, with trailing sequences up- and downstream of each gene, and should allow for sequence-based search for alleles curated in BGMUT.

Some Observations on the Nature of Variations in Genes of Blood Group Systems

BGMUT and some other sequence variation databases such as the phenylalanine hydroxylase gene mutation database (PAHdb [15]) differ from others in that the selection of individuals whose DNA is used for study is based on a known variant phenotype. This has a bearing on observations that can be drawn from BGMUT by examining DNA changes across alleles. As noted above, BGMUT now has 1,545 alleles, with the number ranging from >300 for the ABO and RH to just 2 for the Xg and Vel systems (table 1). Clearly, alleles of ABO and RH systems are most numerous since typing for their common alleles has become routine throughout the world to avoid mismatch between donor and recipient for blood transfusion or between fetus and mother during gestation. Pre-transfusion typing to rule out MNS, KEL or H mismatches that may cause similar accidents is also common. In contrast, the lack of information on subjects carrying variant alleles of most other blood group systems reflects not a lack of interest but rather their exclusion from the typing routine due to their recognized immunological tolerance.

The accumulation of detailed knowledge about the molecular basis for such a high number of variant alleles of any gene occurs rarely in the current investigations of genomic diversity. The high numbers of alleles of *ABO*, *RHCE*/*RHD*, *GYPA*/*GYPB*, *FUT1*/*FUT2* and *KEL* genes compiled in BGMUT seem to be an exception. The availability of such high numbers and the molecular knowledge for their variation provides an opportunity to examine and compare patterns for the origin of sequence changes [16]. Thus, despite the caveat

that in a number of alleles of the above genes only partial exonic sequences are available for a closer scrutiny, a few challenging observations become apparent and are reinforced by examination of alleles of additional genes in the database.

The most salient observation is that the mechanism for DNA alterations is not identical for the different gene loci documented in BGMUT. All genes diversify by nucleotide substitutions (sense, missense or nonsense), deletions, or insertions; however, a significant number of alleles of *ABO*, *GYPA*/*GYPB,* and *RHCE*/*RHD* genes originate through gene recombination such as gene conversions or unequal crossovers. This is discussed in a number of articles [5, 16–18]. *GYPA*/*GYPB*, *RHCE*/*RHD,* and *FUT1*/*FUT2* form duplicated/multiple gene families, whereas *ABO* and *KEL* are single copy genes. *FUT1*, *FUT2,* and a *FUT2*-like pseudogene reside in close vicinity on chromosome 9, but, in contrast to the duplicated *GYPA*/*GYPB* and *RHCE*/*RHD* genes, *FUT1*/*FUT2* alleles resulting from gene recombination seem to occur much less frequently (4 of the 113 documented in BGMUT). In case of *ABO*, gene recombination most likely occurs at meiosis and this occasionally results in a change of the ABO serotype of the fetus causing a problem in paternity identification [18]. *KEL* alleles draw attention because the nucleotide changes in about 40% result in a K_0 or a null phenotype; two thirds of such mutations are nonsense or affect splicing. Such a high incidence of loss of a common cell surface epitope is also observed for a small number of other systems (Kidd, Junior, Kx, Lan, and Rh-associated glycoprotein), and is possibly explained by the relatively easier detection of a null phenotype..

The mechanism of gene recombinations in the *GYPA*/*GYPB*/*GYPE* family deserves special attention as it illustrates the complex origin of some sequence variations. The organization and sequence of *GYPA* and *GYPB* are nearly identical, the key difference being inactivation of exon 3 (named *pseudo-exon 3*) in *GYPB* because of a nucleotide substitution at the 5 splicing site. However, in several alleles of the MNS system, the splicing site is re-activated because of GYPA-GYPB gene recombination, most likely gene conversion with the 5 breakpoints occurring at different sites along the pseudo-exon. Thus, pseudo-exon 3 becomes a part of a true, hybrid exon made of a 5 *GYPB* and a 3 GYPA segment, and variation among the alleles results from incorporation of GYPA/GYPB segments of lengths that vary depending on the site of recombination [19].

Conclusions

One of the major advances in transfusion medicine over the last 15 years has been the elucidation of the molecular basis of human blood groups through identification of genes and their variations that are responsible for blood group antigens. This has clarified a field that was a black box for at least a century. Information on blood group gene variations gathered in databases such as BGMUT facilitates discovery of new alleles and assists the resolution of puzzling blood typing results. In addition, these databases are helping with the growth of genotyping as an adjunct to serology-based blood typing [20].

Because of the routine of blood typing which continually screens for new phenotypic variations, and a worldwide interest in research dealing with the molecular basis of such variations, relatively large numbers of alleles have been identified for these genes. This accumulation of knowledge on a large number of alleles with distinct phenotypes that can be seen for genes of blood group systems is rare for the human genome. That the mode of sequence diversification is not identical for all blood group genes can be seen in a comparison of the sequence changes among these alleles. Similarly, recurrent mutations can be seen among their alleles for only some of the genes [16]. These observations show the additional value of BGMUT in expanding our knowledge on genetic variation.

Acknowledgements

We thank Ray Dunivin and Douglas Hoffman at NCBI for technical maintenance of BGMUT. We also thank the blood group research community for supporting the database through data submission and usage. For funding support to OOB and WH, we thank the Department of Biochemistry of Albert Einstein College of Medicine and NCBI, respectively.

Disclosure Statement

The authors declare that they have no actual or potential conflict of interest in relation to this article.

References

1 Broadberry RE, Lin M: The distribution of the MiIII (Gp.Mur) phenotype among the population of Taiwan. Transfus Med 1996;6:145–148.

2 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:D1023–1029.

3 Blumenfeld OO, Patnaik SK: Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. Hum Mutat 2004;23:8–16.

4 Cotton RG, McKusick V, Scriver CR: The HUGO Mutation Database Initiative. Science 1998;279:10–11.

5 Huang CH, Liu PZ, Cheng JG: Molecular biology and genetics of the Rh blood group system. Semin Hematol 2000;37:150–165.

6 Yamamoto F, McNeill PD, Hakomori S: Genomic organization of human histo-blood group ABO genes. Glycobiology 1995;5:51–58.

7 Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, Bateman A, Blake JA, Bult CJ, Cherry JM, Chisholm RL, Cochrane G, Cook CE, Eppig JT, Galperin MY, Gentleman R, Goble CA, Gojobori T, Hancock JM, Howe DG, Imanishi T, Kelso J, Landsman D, Lewis SE, Mizrachi IK, Orchard S, Ouellette BF, Ranganathan S, Richardson L, Rocca-Serra P, Schofield PN, Smedley D, Southan C, Tan TW, Tatusova T, Whetzel PL, White O, Yamasaki C: Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res 2011;39:D7–10.
8 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. Nucleic Acids Res 2005; 33:D34–38.
9 Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H: The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 2001;109:678–680.
10 den Dunnen JT, Antonarakis SE: Nomenclature for the description of human sequence variations. Hum Gen 2001;109:121–124.
11 Matullo G, Di Gaetano C, Guarrera S: Next generation sequencing and rare genetic variants: from human population studies to medical genetics. Environ Mol Mutagen 2013;54:518–532.
12 Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: The next-generation sequencing revolution and its impact on genomics. Cell 2013; 155:27–38.
13 Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, Tosatto SCE: In silico genotyping from exome sequencing; in: Proceedings of the Workshop on Annotation, Interpretation and Management of Mutations (AIMM,2012). *http://ceur-ws.org/Vol-916/* (last accessed August 4, 2014).
14 Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014;42:D756–763.
15 Scriver CR, Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C: PAHdb 2003: what a locus-specific knowledgebase can do. Hum Mutat 2003;21:333–344.
16 Patnaik SK, Blumenfeld OO: Patterns of human genetic variation inferred from comparative analysis of allelic mutations in blood group antigen genes. Hum Mutat 2011;32:263–271.
17 Seltsam A, Hallensleben M, Kollmann A, Blasczyk R: The nature of diversity and diversification at the ABO locus. Blood 2003;102:3035–3042.
18 Suzuki K: ABO blood group alleles and genetic recombination. Legal Med 2005;7:205–212.
19 Blumenfeld OO, Huang CH: Molecular genetics of the glycophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification. Hum Mutat 1995;6:199–209.
20 Anstee DJ: Red cell genotyping and the future of pretransfusion testing. Blood 2009;114:248–256.