Published in final edited form as:

Trends Parasitol. 2009 June ; 25(6): 249–252. doi:10.1016/j.pt.2009.03.003.

varDB: common ground for a shifting landscape

David R. Allred^{1,2,3}, Anthony F. Barbet^{1,2,3}, J. David Barry⁴, and Kirk W. Deitsch⁵

¹Department of Infectious Diseases and Pathology, University of Florida, Gainesville, FL 32611, USA ²Department of Pathology, Immunology, and Laboratory Medicine, University of Florida, Gainesville, FL 32610, USA ³Genetics Institute, University of Florida, Gainesville, FL 32610, USA ⁴Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow, G12 8TA, Scotland ⁵Department of Microbiology and Immunology, Weill Medical College, Cornell University, New York, NY 10021, USA

Abstract

Antigenic variation is a phylogenetically widespread phenomenon thought to lead to survival benefits for the pathogen. Although governed by genetic mechanisms, antigenic variation is ultimately manifested in variant proteins. The varDB database is an attempt to gain an overview of common structures and functions of variant proteins related to enhanced survival. varDB provides a wealth of sequence data and several tools to facilitate their analysis, but current limitations preclude achievement of its full promise. A critique of this database and how it could serve the scientific community is provided here.

The problem

Many infectious agents have the capacity to establish persistent infections, even in the face of ongoing immunity; one strategy used by infectious agents of diverse phylogenetic origins is antigenic variation. Without bias toward specific mechanisms, antigenic variation can be broadly defined as the rapid alteration in structure and antigenicity of specific pathogen components, enabling the organism to avoid unfavorable interactions with the host immune system or providing novel advantageous phenotypes. Antigenic variation *sensu stricto* occurs rapidly enough to assist in the long-term survival of a pathogen within a single host, typically enhancing the probability of transmission to a naïve host, but probably impacts upon transmission among immune individuals as well [1,2]. The study of antigenic variation has resulted in identification of several mechanistic themes underlying the phenomenon, yet the ramifications of component variation for host– pathogen interactions is poorly understood. The varDB database (<http://www vardb.org vardb/>) is a joint venture created by researchers from the Karolinska Institute in Sweden and Kyoto University in Japan to help fill that knowledge gap [3]. The stated aim of varDB is ‘...to facilitate large-scale examination of the evolution of antigenic gene families and help provide insight into how a conserved function can be maintained despite extreme sequence diversity’.

The approach

varDB brings together sequence data on proteins encoded by 39 gene families from 25 pathogen species (11 bacteria, ten protozoa, two fungi, one metazoan and HIV) causing 17 different diseases. The sequences are accompanied by sequence alignments, secondary structure predictions and information on predicted functional Pfam domains of the various

proteins. The information is organized the same way, making it easy to find much of the information even when approaching a question from different angles. Links to published references are provided for sequence data sources. Tools are provided within varDB that enable one to develop and quickly test hypotheses in silico about relationships between proteins, their potential function(s) and their interactions with the host. In general, the varDB website is logically and consistently organized and user friendly.

The sequences

Key criteria for the value of such a database include the completeness and accuracy of the information contained. In that regard, users should note some deficiencies that somewhat diminish the robustness of analyses initiated from varDB. These deficiencies seem to have arisen largely from automated acquisition of data from general databases, an approach that excludes the essential specialist annotation that has been applied within each pathogen community. For example, the var genes of *Plasmodium falciparum* are all compiled into a single classification. However, var genes are, in fact, divisible into multiple subclasses [4,5], and expression of genes from these different subclasses has direct implications for the phenotypes of infected cells and for pathogenicity [6–8]. In compiling the ves multigene family of *Babesia bovis*, full-length ves1a and ves1b genes from different isolates were pooled together with cDNAs derived from each branch, as well as partial sequences derived by reverse transcriptase PCR [9–11]. The same holds for *Trypanosoma*, in which VSG genes of proven function (cDNA) are combined with silent genes not yet demonstrated to be functional [12,13]. Jumbling all these sequences together without regard for origin, completeness or key characteristics could result in serious misinterpretations of structure and potential function and creates additional work in identifying appropriate sequences for comparison. The available data are also far from complete for some species, such as *Anaplasma* and *Ehrlichia*. For these genera, only data acquired from complete genomes seem to have been incorporated, excluding studies describing the sequence of the same gene family or locus in additional isolates or species [14,15]. This approach also excludes analyses of variation of individual paralogs in different isolates, such as MAP1 paralogs in *Ehrlichia ruminantium* [16,17] and outer membrane protein sequences from *Ehrlichia chaffeensis* [18–20]. Finally, the creators sometimes seem to have interpreted sequences differently from the original sources. For example, in *Anaplasma phagocytophilum*, 113 p44 genes were described in the original publication of the genome. By contrast, varDB includes only 109 p44 genes, derived through the omission of 12 of the original sequences and the inclusion of eight new sequences. Similarly, only half of the ves genes identified in the *B. bovis* genome were included and atypical ves genes were not included [11]. The discrepancies in trypanosome VSG sequences are even greater. Well over 1000 VSG sequences exist (www.VSGdb.net) [21], but varDB includes only 221, of which many are related proteins, such as transferrin receptor subunits [22], and at least 14% have no homology with VSG. Similar shortcomings might exist for sequences from other species. As with several other pathogens, including *Anaplasma*, trypanosomes can assemble functional mosaic antigen genes from silent pseudogenes [12], but there is no allusion to this phenomenon, representing an important gap. Despite any shortcomings, varDB represents a wealth of sequence data available in one location, with the common thread of involvement in antigenic variation, and stands to save cautious users the drudgery of identifying and downloading numerous individual sequences from other sources. Any shortcomings noted by users can be reported via a feedback function, but the curators are encouraged to expand this into a formal community annotation capability (Box 1).

The tools

The creators of varDB have provided a set of useful sequence analysis tools and have even done some of the work for the user. Predeveloped alignments are provided for the polypeptides encoded by members of each gene family, along with a fasta file of the alignment and two different viewers: the 'varDB alignment viewer' and 'Jalview'. Each enables selective colorization of residues to emphasize structural propensities. Jalview also enables a certain amount of alignment editing, which is a real benefit when dealing with highly variant sequences. Jalview is faster (under Windows XP) and offers more options for manipulation of the alignment and, in general, seems the more useful of the two. The major disappointment comes when attempting to perform a new alignment. The provided alignment tool, MAFFT, requires an existing alignment in fasta format as input and is limited to protein sequences. Until these limitations are alleviated, users will have to download sequences of interest and take advantage of alternative opportunities for sequence alignment. In that regard, varDB provides links to several alignment tools at other sites (Muscle, ClustalW, TCOFFEE and fully functional MAFFT), so the limited functionality of the varDB tool should not be a serious hindrance to initiating an analysis.

A Basic Local Alignment Search Tool (BLAST) search interface is provided and easy to use, enabling the researcher to find sequences related to those of interest. Several functions are provided, including BLASTN, TBLASTN, BLASTP, BLASTX and TBLASTX searches and the ability to modify search parameters. Missing are Gapped BLAST, PHI-BLAST and PSI-BLAST functions, which are immensely useful when comparing sequences of hypervariable proteins. The larger concern is that the BLAST tool provided searches only the varDB database, thus restricting what one might find to those sequences identified by the creators as sequences of interest. This seems an unnecessarily restrictive and counterproductive approach. Perhaps the most useful tool within varDB is the provision of Pfam search results for all proteins within which domains could be identified (Pfam is a curated collection of alignments, hidden Markov models and three-dimensional structural information on characterized protein domains [23]). The variety of recognizable domains represented is impressive, illustrating the significance of variant proteins. Having all this homework done is potentially a great time-saver for the investigator in the initiation of a project. Although the Pfam results are useful, accessing them is awkward because they are organized by domain. To find out which proteins contain each domain, one must go through each domain one at a time, which is cumbersome. It would be helpful if these results were cross-referenced by protein family so that one could quickly find all the predicted domains within a specific family (Box 2).

Concluding remarks

Given the ambitious aims of varDB, one must ask how close this database comes to achieving those aims. Much useful information is contained within the database, and the creators have done a tremendous amount of work in capturing and organizing the sequence data, even providing limited structural information and tools for sequence analysis. Regardless, it must be concluded that the current iteration (December 2008) of varDB does not yet fulfill those goals. However, it is much easier to critique another's efforts than it is to be the creative individual, and our comments should not be construed as criticism of the ideas behind varDB, which are laudable. It should also be remembered that this database is early in its development. If a mechanism were included for community annotation of the sequences, it could relieve some of the burden of curation and would result in an overall more accurate and information-dense resource. With some minor reorganization, more careful annotation and additional tools, varDB could prove to be a very valuable resource for

those intrepid souls wishing to embark upon projects to study the basis for antigenic variation and its roles in pathogenesis.

Acknowledgments

The authors were supported by grants #R01 AI055864 (D.R.A.), #R01 GM081714 (A.F.B.) and #R01 AI052390 (K.W.D.). J.D.B. is a Wellcome Trust Principal Research Fellow.

References

1. Borst, P. Mechanisms of antigenic variation: an overview. In: Craig, A.; Scherf, A., editors. *Antigenic Variation*. Academic Press; 2003. p. 1-15.
2. Futse JE, et al. Superinfection as a driver of genomic diversification in antigenically variant pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 2008; 105:2123–2127. [PubMed: 18252822]
3. Hayes CN, et al. varDB: a pathogen-specific sequence database of protein families involved in antigenic variation. *Bioinformatics*. 2008; 24:2564–2565. [PubMed: 18776192]
4. Lavstsen T, et al. Sub-grouping of *Plasmodium falciparum* 3D7 var genes based on sequence analysis of coding and non-coding regions. *Malar. J.* 2003; 2:1–14. [PubMed: 12605724]
5. Kraemer SM, Smith JD. Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum* var gene family. *Mol. Microbiol.* 2003; 50:1527–1538. [PubMed: 14651636]
6. Kaestli M, et al. Virulence of malaria is associated with differential expression of *Plasmodium falciparum* var gene subgroups in a case-control study. *J. Infect. Dis.* 2006; 193:1567–1574. [PubMed: 16652286]
7. Jensen ATR, et al. *Plasmodium falciparum* associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes. *J. Exp. Med.* 2004; 199:1179–1190. [PubMed: 15123742]
8. Salanti A, et al. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. *Mol. Microbiol.* 2003; 49:179–191. [PubMed: 12823820]
9. Allred DR, et al. The ves multigene family of *B. bovis* encodes components of rapid antigenic variation at the infected erythrocyte surface. *Mol. Cell.* 2000; 5:153–162. [PubMed: 10678177]
10. Al-Khedery B, Allred DR. Antigenic variation in *Babesia bovis* occurs through segmental gene conversion of the ves multigene family, within a bidirectional site of transcription. *Mol. Microbiol.* 2006; 59:402–414. [PubMed: 16390438]
11. Brayton KA, et al. Genome sequence of *Babesia bovis* and comparative analysis of Apicomplexan hemoprotozoa. *PLOS Pathog.* 2007; 3:e148.1–e148.13.
12. Marcello L, Barry JD. Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.* 2007; 17:1344–1352. [PubMed: 17652423]
13. Hutchinson OC, et al. VSG structure: similar N-terminal domains can form functional VSGs with different types of C-terminal domain. *Mol. Biochem. Parasitol.* 2003; 130:127–131. [PubMed: 12946849]
14. Ohashi N, et al. Analysis of transcriptionally active gene clusters of major outer membrane protein multigene family in *Ehrlichia canis* and *E. chaffeensis*. *Infect. Immun.* 2001; 69:2083–2091. [PubMed: 11254561]
15. Zhang C, et al. Identification of 19 polymorphic major outer membrane protein genes and their immunogenic peptides in *Ehrlichia ewingii* for use in a serodiagnostic assay. *Clin. Vaccine Immunol.* 2008; 15:402–411. [PubMed: 18094116]
16. Reddy GR, et al. Sequence heterogeneity of the major antigenic protein 1 genes from *Cowdria ruminantium* isolates from different geographic areas. *Clin. Diagn. Lab. Immunol.* 1996; 3:417–422. [PubMed: 8807206]

17. Allsopp MTEP, et al. Ehrlichia ruminantium major antigenic protein gene (map1) variants are not geographically constrained and show no evidence of having evolved under positive selection pressure. *J. Clin. Microbiol.* 2001; 39:4200–4203. [PubMed: 11682561]
18. Yu X-J, et al. Genetic diversity of the 28-kilodalton outer membrane protein gene in human isolates of Ehrlichia chaffeensis. *J. Clin. Microbiol.* 1999; 37:1137–1143. [PubMed: 10074538]
19. Reddy GR, Streck CP. Variability in the 28-kDa surface antigen protein multigene locus of isolates of the emerging disease agent Ehrlichia chaffeensis suggests that it plays a role in immune evasion. *Mol. Cell Biol. Res. Commun.* 1999; 1:167–175. [PubMed: 10425222]
20. Cheng C, et al. Molecular heterogeneity of Ehrlichia chaffeensis isolates determined by sequence analysis of the 28-kilodalton outer membrane protein genes and other regions of the genome. *Infect. Immun.* 2003; 71:187–195. [PubMed: 12496165]
21. Marcello L, et al. VSGdb: a database for trypanosome variant surface glycoproteins, a large and diverse family of coiled coil proteins. *BMC Bioinformatics.* 2007; 7:143.1–143.8. [PubMed: 17474977]
22. Carrington M, Boothroyd J. Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol. Biochem. Parasitol.* 1996; 81:119–126. [PubMed: 8898328]
23. Finn RD, et al. The Pfam protein families database. *Nucleic Acids Res.* 2008; 36:D281–D288. [PubMed: 18039703]
24. Carlton JM, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature.* 2008; 455:757–763. [PubMed: 18843361]
25. Balaji S, et al. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* 2005; 33:3994–4006. [PubMed: 16040597]
26. Freitas-Junior LH, et al. Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell.* 2005; 121:25–36. [PubMed: 15820676]
27. Duraisingh MT, et al. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. *Cell.* 2005; 121:13–24. [PubMed: 15820675]
28. Figueiredo LM, et al. A histone methyltransferase modulates antigenic variation in African trypanosomes. *PLoS Biol.* 2008; 6:1539–1548.

Box 1. What's missing?

A brief overview of the general structure and function of each antigen would help the user to understand what might be conserved or variable among systems. Some opportunities for inclusion of interesting sequences were overlooked. An example of this is the small open reading frame (SmORF) proteins and genes recently discovered during sequencing of the *B. bovis* genome [11]. To be fair, no demonstration has yet been made of antigenic variation in the SmORF protein. However, there are 44 members of the smorf gene family in the genome, the members are polymorphic in both size and sequence, and cDNAs derived from this gene family only occasionally matched genomic copies [11]. All these characteristics suggest strongly that the SmORF protein undergoes antigenic variation. Similar uncharacterized gene families that seem to be antigenically variant exist within the *Plasmodium vivax* genome [24]. It might be useful if varDB were to include a category of genes and proteins considered highly likely to be players in the process of antigenic variation, to take advantage of such situations. This is where direct engagement of specialists from the different pathogen communities could be invaluable.

Additional sequence information that could be of considerable use includes those proteins involved in the regulation of variant protein gene expression. Examples include the ApiAP2 family of transcription factors [25], Sir2 [26,27], and DAM methyltransferase and other histone methyltransferases [28]. These proteins are often much more conserved in sequence and function than the variant proteins are, and comparisons across phylogenetic lineages could be powerful. It would also be useful to include a tool illustrating the genomic context of a gene entry, such as Artemis (Sanger Institute). For example, is a gene in tandem arrays or dispersed? Similarly, to the extent possible, links to three-dimensional crystal structure data and software to enable superimposing aligned sequences onto the scaffold of a characterized variant protein, or structurally related invariant protein, could be of enormous utility in understanding the functions and evolution of individual gene families and proteins.

Box 2. Other assets

In addition to sequences and sequence analysis tools, varDB provides cursory and rudimentary information about each of the pathogens included and the diseases they cause. Unfortunately, the references provided refer back to sequence studies rather than authoritative references on organism biology or disease. Similarly, a brief discussion on antigenic variation that would be of use to someone unfamiliar with the phenomenon is provided, but it, too, is inadequately referenced. It also lacks any discussion of segmental gene conversion, an omission compounded by the exclusion of trypanosome pseudogenes, atypical babesial genes and perhaps others from the database, even though these can or might be major sources of coding information. More useful, however, is the inclusion of a small glossary in which terms associated with antigenic variation are defined but with which many might not be familiar. Finally, a world map, which conveys the numbers of variant protein sequences obtained from each country, is provided on the opening page. This makes for an interesting graphic but seems of marginal utility. More useful would be inclusion of a phylogenetic tree demonstrating evolutionary relationships and distances among variant proteins from pathogens, and related proteins and/or domains uninvolved in antigenic variation.