

Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Acta Tropica

journal homepage: [www.elsevier.com/locate/actatropica](http://www.elsevier.com/locate/actatropica)

## Review

## varDB: A database of antigenic variant sequences—Current status and future prospects

Diego Diez<sup>a,1</sup>, Nelson Hayes<sup>a,1</sup>, Nicolas Joannin<sup>b,c</sup>, Johan Normark<sup>b,c</sup>, Minoru Kanehisa<sup>a</sup>, Mats Wahlgren<sup>b,c</sup>, Craig E. Wheelock<sup>a,d,\*</sup>, Susumu Goto<sup>a,\*\*</sup><sup>a</sup> Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan<sup>b</sup> Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, SE-17177 Stockholm, Sweden<sup>c</sup> Swedish Institute for Infectious Disease Control (SMI), SE-17182 Stockholm, Sweden<sup>d</sup> Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden

## ARTICLE INFO

## Article history:

Available online 17 June 2009

## Keywords:

Antigenic variation  
Hyper-variable sequence  
Database  
varDB  
Malaria  
Plasmodium

## ABSTRACT

Antigenic variation is a common mechanism employed by many pathogenic organisms to avoid recognition of surface proteins by the host immune system. The malaria parasite, *Plasmodium falciparum*, among many others, exploits this mechanism and manages to survive in an otherwise hostile environment. Although similarities in the mechanisms used among different species to generate antigenic variation are broadly recognized, there is a lack of studies using cross-species data. The varDB project (<http://www.vardb.org>) was created to study antigenic variation at a range of different levels, both within and among species. The project aims to serve as a resource to increase our understanding of antigenic variation by providing a framework for comparative studies. In this review we describe the varDB project, its construction, and the overall organization of information with the intent of increasing the utility of varDB to the research community. The current version of varDB supports 27 species involved in 19 different diseases affecting humans as well as other species. These data include 42 gene families that are represented by over 67,000 sequences. The varDB project is still in its infancy but is expected to continue to grow with the addition of new organisms and gene families as well as input from the general research community.

© 2009 Elsevier B.V. All rights reserved.

## Contents

1. Introduction .....	144
2. The varDB project .....	145
2.1. Data extraction from public databases .....	145
2.2. <i>P. falciparum</i> sequence content in varDB .....	146
2.3. Web site implementation .....	147
2.4. Data organization .....	147
2.5. The Search view .....	147
2.6. The Shopping cart .....	148
2.7. Sequence analysis tools .....	149
2.8. Tags .....	149
2.9. Documentation .....	149
3. Future directions .....	149
Acknowledgements .....	150
References .....	150

\* Corresponding author at: Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden.

\*\* Corresponding author at: Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan.

E-mail addresses: [craig.wheelock@ki.se](mailto:craig.wheelock@ki.se) (C.E. Wheelock), [goto@kuicr.kyoto-u.ac.jp](mailto:goto@kuicr.kyoto-u.ac.jp) (S. Goto).

<sup>1</sup> These authors contributed equally to this work.

## 1. Introduction

Many pathogens need to adapt to an extremely hostile environment: their vertebrate host. Indeed, vertebrates have developed an elaborate immune system that efficiently targets the non-self. One component of this system, unique to jawed vertebrates, is adaptive immunity, which is specific and long lasting (Cannon et

al., 2004). In order to avoid being destroyed by this host defense strategy, pathogens have evolved counter mechanisms to avoid it, including antigenic drift and variation. An important example can be observed in the malaria parasite, which makes extensive use of antigenic variation. *Plasmodium* species have a complex life cycle, shared between the mosquito vector and the vertebrate host, with different life stages experiencing distinct immune pressures (Rasti et al., 2004). After infecting the host, sporozoites first migrate to the liver where they undergo an initial proliferation phase in hepatocytes. Following extensive differentiation, merozoites are released into the host blood stream where they multiply asexually in red blood cells in the erythrocytic stage of the life cycle. Some merozoites may differentiate into gametocytes, which can be ingested by another mosquito to continue the sexual stage of parasite development. Within the vertebrate host, *Plasmodium* parasites are under constant immune pressure. The parasite evades detection by the immune system by hiding inside red blood cells. However, to avoid destruction by the spleen, it risks immune exposure by exporting adhesive proteins to the red blood cell surface. Although the function of these proteins is poorly understood, they are involved in parasite sequestration and rosetting, adaptations that prevent the infected erythrocytes from reaching the spleen. Though less well understood, the gametocytes also experience immune pressure while undergoing sexual reproduction in the mosquito vector.

All *Plasmodium* species surveyed at the genome-wide scale have revealed the existence of multi-copy gene families with suspected involvement in antigenic variation, most often located in the genetically unstable region near the telomeres (Carlton et al., 2002, 2008; Gardner et al., 2002; Barry et al., 2003; Pain et al., 2008). These gene families have been most extensively studied in *Plasmodium falciparum*, specifically the PfEMP1 virulence factor (Fig. 1). However, sequence analysis of antigenic variant proteins presents many challenges due to the large number of gene copies. Some studies have focused on designing novel tools for the analysis of an almost unlimited number of variant sequences (Normark et al., 2007; Bull et al., 2008), but novel specialized methods are still needed to fully understand these data.

The study of antigenic variant sequences can be categorized into three broad groups, which sometimes overlap: (1) regulation of

gene transcription (Cunningham et al., 2005; Ralph et al., 2005; Fonager et al., 2007; Tham et al., 2007; Dzikowski and Deitsch, 2008), (2) protein diversity (Fischer et al., 2003; Oliveira et al., 2006; Bull et al., 2008; Frank et al., 2008; Joannin et al., 2008) and (3) functional/phenotypic analysis (Springer et al., 2004; Pettersson et al., 2005; Baratin et al., 2007; Bertonati and Tramontano, 2007; Normark et al., 2007; Andersen et al., 2008; Klein et al., 2008; Vigan-Womas et al., 2008). Some attempts at cross-species analyses have been made (Janssen et al., 2004; Korir and Galinski, 2006; Bockhorst et al., 2007), but they are still uncommon. These studies generate large amounts of sequence and phenotypic data, which are deposited in repositories such as GenBank (Benson et al., 2009) and specialized databases such as ClinMalDB (<http://clinmaldb.usp.br>) or quite often only as supplemental data to their respective articles. Therefore, these data are often difficult to access and/or compile, which represents a significant hindrance to understanding the relationship between sequence and phenotype or clinical manifestation of disease.

The varDB project aims to produce a comprehensive platform focused on issues specific to the study of antigenic variation that are common across gene families (Hayes et al., 2008). The project was initiated by compiling existing sequence information pertaining to specific gene families into a common database. The database is now in the process of being further augmented by complementing this information with associated phenotypic and clinical data linked to these sequences. In addition, the varDB platform is intended to serve as a workspace, comprising analytical tools and sequence management facilities, with various options for downloading and uploading individual datasets into private accounts for logged-in users. In this review, we describe the current database framework, our data mining method and the tools already available in varDB. We then provide a vision for the continued evolution of the varDB project focusing on the malaria parasite as an example.

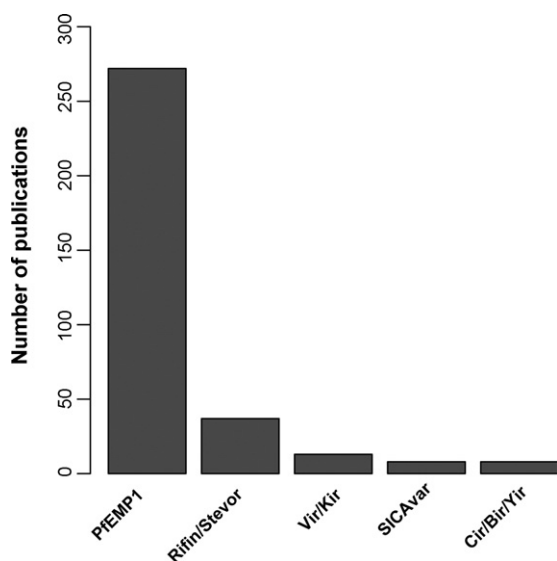
## 2. The varDB project

The varDB project components can be divided into three general categories: data acquisition, data representation, and analysis tools.

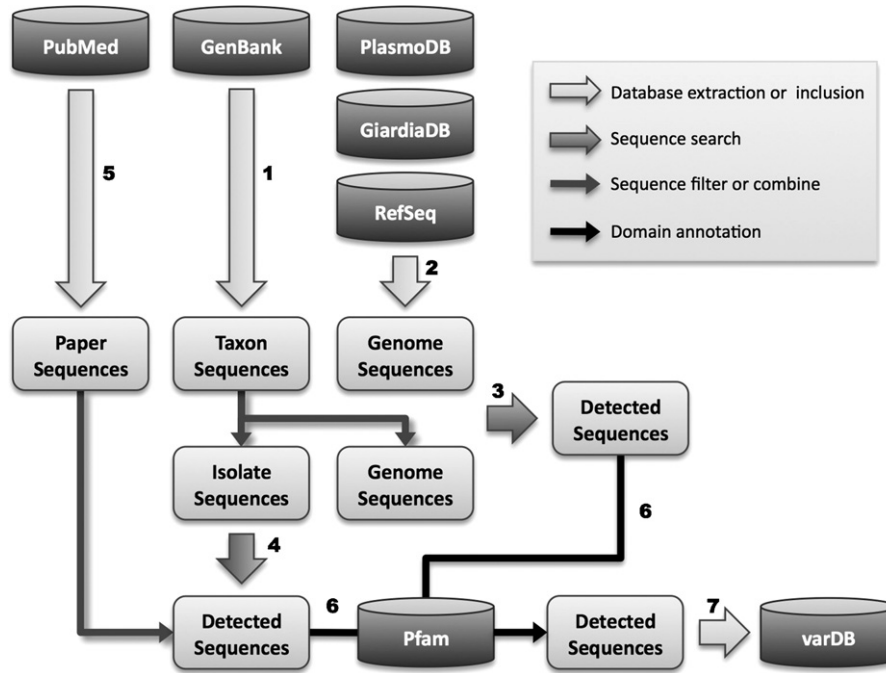
### 2.1. Data extraction from public databases

VarDB aims to serve as a repository for antigenic variant sequences. Therefore, our aim is to detect and collect all available sequences, including those from genome projects (genome sequences) as well as sequences obtained directly from patient samples (isolate sequences). This task has proven to be challenging due to a number of distinct obstacles. GenBank is the central repository where most sequences are initially submitted prior to publication; therefore it is the most important primary source for sequences. Additionally, sequences in GenBank format contain extensive information about sequence features, links to other databases, references to publications, etc., making it a very useful source for extracting sequence annotation. However, eukaryotic genome assemblies and gene predictions and annotations evolve very quickly after initial publication and submission to GenBank. In addition, they are often maintained by independent database projects that do not frequently update the original GenBank submissions. Consequently, our approach to address this problem is to use specialized database projects (e.g. PlasmoDB, GiardiaDB, EuPathDB) for sources of eukaryotic genomes when available and then to use GenBank for other sequences.

A framework has been developed to automate tasks such as downloading and preprocessing data, annotating sequences, and integrating data from different sources. The sequence detection pipeline is shown in Fig. 2 and the process can be outlined as



**Fig. 1.** Number of publications available for several gene families involved in antigenic variation in different *Plasmodium* species. There is a bias in the number of studies, mainly focused on the *P. falciparum* var gene family, followed by *rif/stevor*. Significantly fewer studies are available for antigenic variation gene families belonging to different *Plasmodium* species.



**Fig. 2.** Data mining pipeline used to obtain the sequences that populate the varDB database. A description of the methods used for each numerical point (1–7) is provided in the text.

follows: (1) in the first step, sequences from specific taxonomic groups are downloaded from GenBank (from the Nucleotide Core and ESTdb repositories) and separated into genome and isolate sequences. (2) In parallel, genome sequences are downloaded from external projects (e.g., PlasmODB). Out of date or duplicate GenBank data from step (1) are discarded so that only the most current data are used. To enable automated analysis, all sequences are processed to conform to a common format (FASTA format for sequence data, GFF3 format for annotations). (3) Gene families are detected in genomic data using *HMMER* (Eddy, 1998) for protein sequences and *GENEWISE* (Birney et al., 2004) for nucleotide sequences, using a single profile domain to detect family membership. For this purpose, protein domain profiles are extracted from the Pfam database (version 23). (4) *PSI-BLAST* (Altschul et al., 1997) is used to detect gene families in the isolate sequences, enabling the detection of sequence fragments that are common in isolate studies and may be missing the domains used for genomic sequence detection. In the current version of varDB, *PSI-BLAST* models are derived from selected genomes (e.g. *P. falciparum* 3D7 for the *var* gene family) using the best hit from the *HMMER* search as the sequence seed. For both *HMMER* and *PSI-BLAST* searches, sequences are considered as significant hits when the  $E_{value} \leq 0.01$ . (5) In parallel to this similarity search pipeline, sequences are downloaded from publications describing and analyzing antigenic variant sequences and combined with the sequences from step (4). Sequences found in step (5), but not in (4), are tagged as non-detected, so that they may be excluded from queries if desired. (6) All sequences are scanned for Pfam domains using *HMMER* and the domain architecture (i.e. the composition of different domains) is computed. (7)

Finally, sequences are integrated in the varDB database, and made available from the varDB web site.

2.2. *P. falciparum* sequence content in varDB

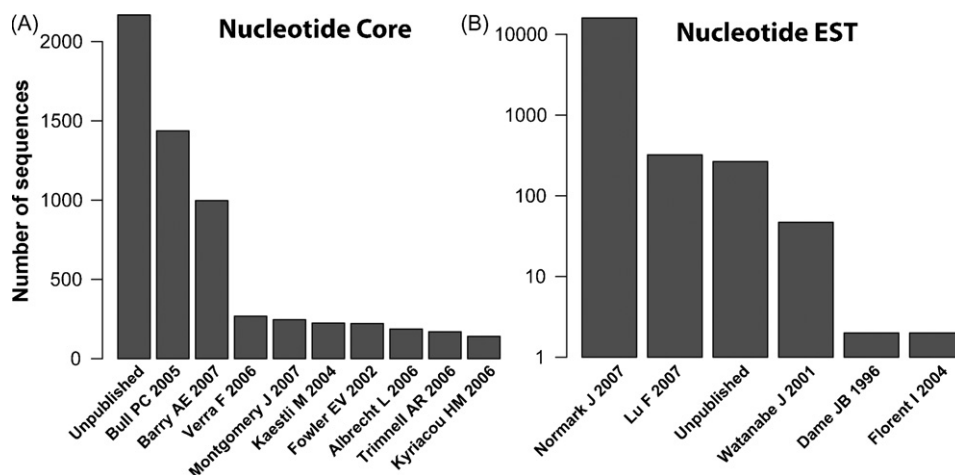
At the time of writing, the varDB project supports 27 organisms spanning 42 gene families, 26,361 protein sequences and 67,650 DNA sequences. This number increases periodically, as new organisms and gene families are being included on a regular basis. The malaria parasite *P. falciparum* is extensively covered, with sequences from three different strains (3D7, Hb3 and Dd2), as well as isolates from the GenBank Nucleotide Core and ESTdb repositories (Table 1). Two antigenic variable gene families are included: *var* and *rif/stevor*. Overall, for *P. falciparum*, 28,501 sequences were detected: 27,304 *var* and 1,197 *rif/stevor* sequences.

The main contribution comes from isolate sequences, and genomic sequences constitute only a small fraction of the total. The isolate data includes both non-processed and processed sequence material, in that individual read sequences are present as well as assembled contigs. An example of this is the collection of approximately 16,000 DBL1-alpha reads from the Normark et al. (2007) project. The presence of read sequences allows researchers to generate new assemblies with the parameters of their choosing. However, sequences of this class are tagged automatically so they can be easily excluded from down-stream analyses. Additionally, Fig. 3 reveals a significant number of GenBank entries that do not have any formal publications attached to them. This suggests that either there is occasionally a delay in the sequence-publication association system in GenBank or that these sequences have been submitted but never

**Table 1**  
Number of PfEMP1 and Rifin/Stevor sequences from *P. falciparum* in varDB<sup>a</sup>.

	Nuc. Core	Nuc. EST	Strain 3D7	Strain DD2	Strain HB3	Total
var	9,500	17,537	60	47	74	27,304
rif/stevor	519	24	223	177	156	1,197
Total	10,019	17,561	283	224	230	28,501

<sup>a</sup> Sequences were drawn from the GenBank Nucleotide Core and ESTdb repositories and from the genome sequencing projects for the 3D7, Dd2 and Hb3 strains.



**Fig. 3.** Distribution of GenBank *var* sequences among publications for (A) Nucleotide Core (shows the top 10 publications with the most sequences out of 139) and (B) Nucleotide ESTdb repositories.

published, providing an important source of variability information that may not have been used in previous analysis.

### 2.3. Web site implementation

One of the most severe drawbacks of working with a web-based tool is the limitations of the user interface. With so many antigenic variant sequences already in varDB, and many expected to come, it is challenging to have an overview of the existing data. In this situation, dynamic and interactive data representation becomes a key asset in efficient information querying and sorting and improves the overall user experience. The web site makes extensive use of AJAX and the Ext JS JavaScript library to provide an interactive and responsive user interface. For example, tables are visualized in a paged, sortable grid, where columns can be resized, rearranged, or hidden. Through the use of asynchronous methods and compression, the initial page load time is reduced and data can be loaded in the background as needed.

The varDB site (<http://www.vardb.org>) is constructed as a Java web application based on the *Spring Framework*, built on top of a *PostgreSQL* relational database and running under JBoss and Apache on a Linux server. Following best practices, varDB was designed according to a modular three-tiered architecture. The data access layer is comprised of wrappers and adapters for external tools and uses the *Hibernate* object-relational mapping tool to map database tables to Java objects. The services layer analyzes, manages, and aggregates data from different sources, employing *Biojava*, the KEGG API, NCBI *E-Utils*, *BLAST*, *MAFFT*, and other tools for sequence analysis and management. The presentation layer converts URLs into function calls, manages users and security, and prepares web-friendly page views. Web content is generated using the *FreeMarker* template engine, and some graphical content is generated using *Google Maps* and the *Google Chart API*. The web site is tested and evaluated using *JUnit*, *Xenu*, *YSlow*, and *JSLint*.

### 2.4. Data organization

The varDB website is constructed in a way that facilitates viewing and organizing sequences and related data. A number of different organizational structures are employed to arrange and categorize the sequence data. One way of grouping sequences is by homology. Sequences are organized by gene family, and gene families are grouped into ortholog groups. Because sequences are detected in part by their Pfam domain structure, sequences are also organized by Pfam family, and these are further grouped into

their corresponding Pfam clans. In addition each gene family page contains a figure showing the major representative domain architectures found in that family along with a list of Pfam domains found in sequences belonging to that family. Each gene family is grouped by the pathogen in which it occurs, and pathogens are grouped hierarchically based upon the NCBI Taxonomy database and divided into protists, fungi, bacteria, and viruses. Bacteria are further divided by Gram staining and shape, and viruses are organized based on Baltimore classification. Sequences belonging to the same genome project or chromosome are also grouped together. Sequences can also be organized based on the date or region in which they were collected or based on the publication in which they were reported. All of these predefined categories assist in quickly identifying common sequence sets for a particular application.

### 2.5. The Search view

Although the default data categories described above make it easier to retrieve common sequence sets, more fine-grained searches may be needed to selectively include or exclude sequences based on other criteria. The *Search view* was created to enable users to quickly find the sequences they are interested in based on sequence attributes and associations. During the sequence extraction process a number of fields are indexed and are searchable in the database (Table 2). For example, sequences can be queried and sorted based upon fields such as gene product, isolation source, host, locus tag, UniProt ID, isolate, molecule type, sequence length, etc. In viruses, additional fields such as strain, serotype, and segment may be useful to compare sequences known to be antigenically distinct. This view also provides a powerful Boolean query language allowing the user to combine different search fields in a complex way.

To assist in the construction of specific queries, the input box suggests potential keywords, and the *Add Term* button allows the user to select from a list of the available fields and displays the unique values for each field along with the number of sequences that match each value. When the query is executed, a summary table is presented showing the number of sequences belonging to each gene family, pathogen, ortholog group, reference, etc. Individual values can be selected or deselected to filter the result set, and sequences can also be filtered based on date range or annotation quality, and pseudogenes and truncated sequences can be excluded. Sequences can be selected by clicking the checkboxes in the first column and downloaded or exported. For more fine-grained con-

**Table 2**  
Summary of the fields available for searching and filtering of sequences in the Search view.

Fields	Description
Pathogen/taxon/division	Sequences belonging to a particular pathogen or that are found in the same taxonomic group
Genome/chromosome/scaffold	Sequences that occur on the same chromosome/scaffold or in the same genome
Family/ortholog	Sequences that belong to the same gene/protein family or belong to families in the same ortholog group
Pfam domain/clan	Sequences that have a given Pfam domain or that share domains belonging to the same Pfam clan
Strain/serogroup/serotype	Virus or bacteria sequences in the same serotypic group
Region/country/subregion	Sequences collected in the same geographic region, country, or part of a country
Publication	Sequences that are referred to in the same publication

trol of individual sequences, search results can be loaded into the *Shopping cart* for further analysis.

2.6. The Shopping cart

The *Shopping cart* offers a flexible workspace to organize and explore sequence diversity (Fig. 4). Sequences can be added to the cart from a number of sources throughout the web site as well as from the *BLAST* and search pages. Additional sequences can also be queried and added to the workspace from within the *Shopping cart*, and uninteresting sequences can be removed. The *Quick cart* provides a convenient popup interface to the *Shopping cart* from anywhere on the site. In the full cart, the default table can be split into any number of subsets, and sequences can be added or removed from subsets as well as copied or moved between them. Each subset is displayed in a separate tab with a query form from which sequences can be quickly filtered by family, genome, ortholog group, reference, etc. Selected sequences can be downloaded, and multiple alignments can be generated using *MAFFT* and displayed

in a custom alignment viewer. Logged-in users can also upload their own sequences into the database to create a private virtual database for secure individual online analysis. Uploaded sequences are marked as private and only the logged-in user can access them, but in every other respect the uploaded sequences can be sorted, searched, and organized in the same way as any other database sequences. In the case of logged-in users, search history, results of analyses, and uploaded data are stored in the database between sessions. All user data uploaded to varDB is treated as the confidential and exclusive property of the submitting user and can be permanently removed by the user at any time. No user data is used for any other purpose than private analysis and is not viewed by varDB staff except by request or in the event of an error related to the submitted data.

While all user-submitted data are private by default, we also welcome and solicit community contributions to varDB in the form of sequence and/or phenotypic data, annotations, corrections, comments, and suggestions. A feedback form and comment fields are provided for this purpose and additional community annotation

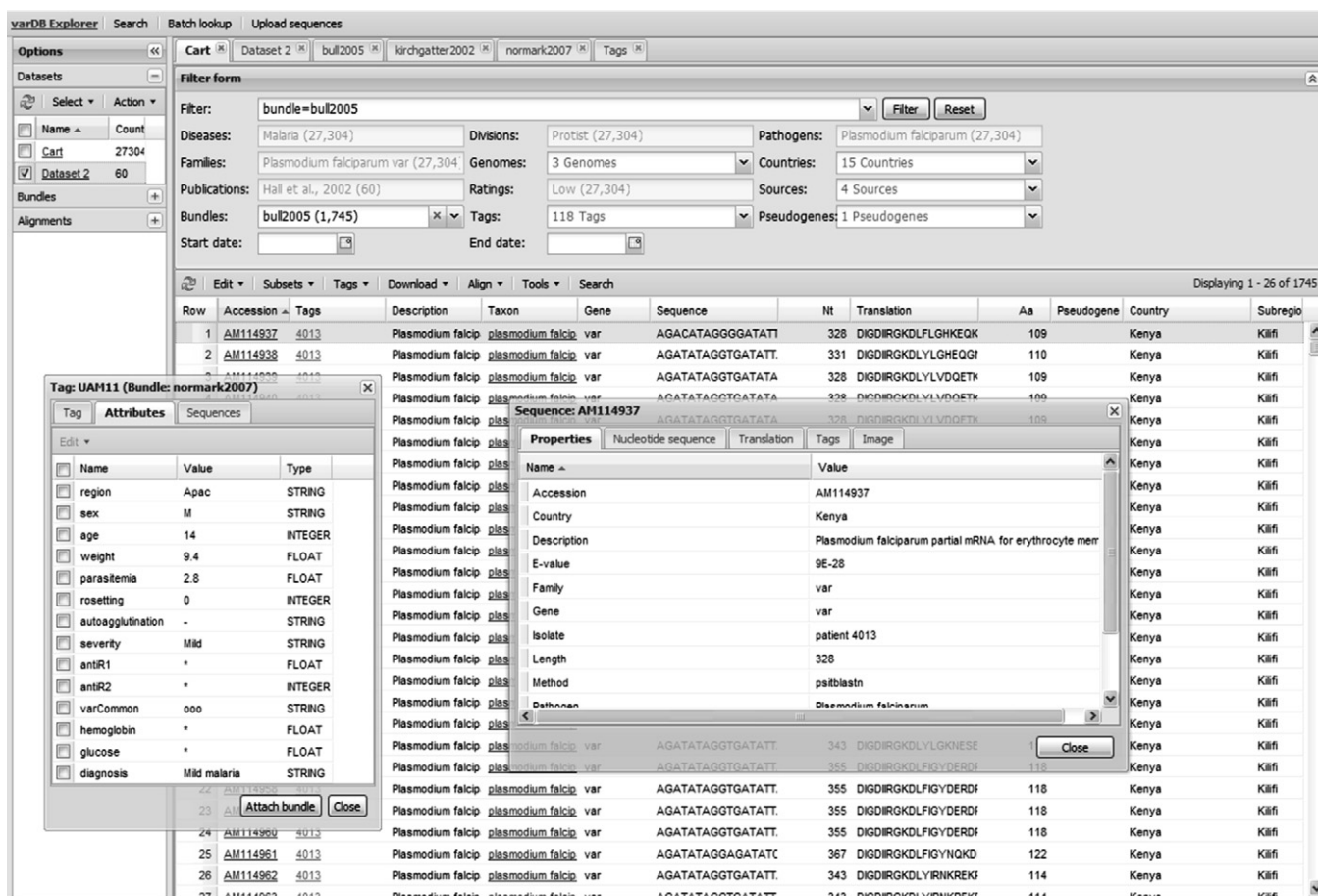


Fig. 4. The *Shopping cart* is the most important tool in varDB that can be used to organize, tag and analyze a set of sequences.

tools are under development. In order for the varDB project to continue to expand, it is vital that input is provided by the general research community. Towards that end, we highly encourage the submission of additional sequence and/or phenotypic data for inclusion in varDB as well as any and all suggestions on improvement and modification.

### 2.7. Sequence analysis tools

As more sequences are added to the database, it is increasingly important to be able to search and organize sequences based on sequence similarity. In addition, several simple analysis tools may be needed to link sequence features to phenotypic information. Towards this aim, a number of specific tools and applications have been added to the varDB web site.

*BLAST* (Altschul et al., 1997) is a general purpose tool that can be used to search for sequences having local or global sequence similarity to a query sequence, an approach used frequently to infer sequence homology. In the varDB project, a local *BLAST* database is provided to search the protein and nucleotide sequences by similarity. Results are presented in a format similar to the *Search view* and can be imported into the *Shopping cart* for further processing. Using the history function, search results can be easily combined, for example, to find all sequences found in both the *BLAST* search and in one or more keyword searches.

While *BLAST* is useful for searches detecting sequence homologs, sometimes it is instructive to search for a simple sequence motif using patterns in the form of regular expressions. VarDB provides several types of pattern searches: (1) exact matches of a sequence fragment or primer sequence, (2) “classic” Perl-like regular expressions, (3) PROSITE-like patterns, and (4) a preconfigured glycosaminoglycan (GAG) binding site search tool based on well known patterns involving alternation of basic and non-polar residues.

Multiple sequence alignment provides a useful method to discover conserved domains, deletions, insertions, point mutations, etc. in a set of related sequences. However, typical alignment tools underperform when applied to antigenically variant sequences, mainly because these contain large stretches of hyper-variable regions. An alignment tool based on the *MAFFT* (Katoh et al., 2005) program is provided that tries to improve the quality of the alignment by isolating hyper-variable regions and aligning them independently using different parameters. This alignment tool can be accessed from either the main menu or from the *Shopping cart*. Alignments can be downloaded or viewed using *Jalview* (Clamp et al., 2004) or a lightweight online alignment viewer.

### 2.8. Tags

Many Web 2.0 web sites (e.g. Facebook, Flickr) allow users to assign tags to information such as emails, photographs, and links to help organize them in a flexible, customized way. Tag clouds offer a simple but powerful visual summary of a data set by weighting the font size of the tag by its frequency in the data set. This approach also provides a convenient way to link sequences with clinical data and disease states to facilitate more fine-grained association studies between sequence variation and disease/parasite features. Sequence analysis can be used to characterize the variation among sequences, revealing, for example, sites of diversifying or purifying selection that can be used to predict the antigenic stability of potential vaccine targets. When this type of sequence variation can be correlated with disease outcomes, functionally important residues can be determined and potential drug targets can be identified.

Researchers often use a database such as MySQL, Access or the VLOOKUP function in Excel to link clinical data with sequence data, but this process is rigid and error prone. VarDB provides a simpler

way to accomplish this by using tagging. For example, Normark et al. (2007) collected blood samples from Ugandan children suffering from mild or severe forms of malaria and sequenced the DBL1-alpha region of *var* transcripts in an attempt to pinpoint patterns associated with clinical outcome. Their data set consists of PfEMP1 sequences, clinical data for each patient, and a list of dominant transcripts found in each patient. Using varDB, they could upload their sequences in FASTA format to create a private dataset in the *Shopping cart*. To associate sequences with patient data, varDB uses collections of related tags sharing a common data format, which are referred to as bundles. The patient data can be uploaded as a bundle by organizing it in a spreadsheet-friendly tab-delimited format with one row for each patient and each column containing a different type of patient information, such as age, weight, region, outcome, rosetting rate, etc. In this case, each patient represents a tag and the different types of patient information represent attributes. When the patient data is uploaded, varDB attempts to determine the data type of the attribute (numeric, text, true/false), but the attribute data types can be changed at any time, and tags and attributes can be added, changed, and removed on the fly using the web interface. Finally, the sequences are tagged with the ID of the patients from which they were collected, either one at a time or using a simple bulk tagging approach.

With this relational structure in place, it is simple to perform queries to retrieve all sequences found in a particular patient or all patients in which a particular sequence was isolated. More complex and informative queries can be performed to retrieve, for example, all sequences found in patients with a high rosetting rate and not found in patients with only mild malaria symptoms. A regular expression search can be used to find sequences with predicted glycosaminoglycan binding sites, which are associated with rosetting, and a user-defined tag such as “GAG” can be applied to these sequences. Then additional queries can be used to select sequences with potential GAG sites to analyze the relationship between glycan binding site frequency and rosetting rate, perhaps to identify potential drug or vaccine targets to disrupt rosetting or to target motifs associated with severe malaria. The Normark data set and several other published clinical data sets are included as reference data in varDB, and others will be added as they become available. A step-by-step tutorial based on the above example is under development. By tagging and classifying sequences based on a combination of standard and user-defined tags, we hope to provide a foundation for cross-referencing and data mining within an increasingly large and diverse pool of sequences and clinical data.

### 2.9. Documentation

The varDB web site contains a number of reference materials related to antigenic variation. In addition to general information on the pathogens, diseases, and gene families included in varDB, the site also contains descriptions of antigenic variation, a glossary of terms related to antigenic variation and a brief description of the pipeline used to extract the sequences from other databases. In addition, we provide tutorials explaining basic functionality of the varDB project to assist users in getting started with the varDB web-site. These tutorials include demonstrations on how to select, filter and download a set of sequences with the *Search view*, and show the basic functionality of the *Quick cart* and *Shopping cart* interfaces, etc. Additional tutorials will be added as further functionalities are developed.

## 3. Future directions

Although it is generally acknowledged that antigenic variation is a mechanism common to many pathogens, it has traditionally been

studied without regard for convergent patterns among organisms. VarDB is an attempt to solve this gap by providing the foundation for a common resource where different antigenically variant gene families can be obtained and analyzed simultaneously. The project has already drawn attention from the research community, as evidenced by the recent publication of a commentary based on the first release of varDB (Allred et al., 2009). This paper provided a valuable critique of the varDB project and gave several concrete steps that can be taken to improve database utility. A key point was the need to work closely with the scientific community to develop a platform of broad use to researchers studying a variety of different organisms and diseases. The challenges are many, and collecting different sequences from diverse databases is only a first step. Based upon this framework, the varDB project aims to evolve into a resource where the majority of the currently time-consuming analysis can be easily performed. It is expected that cross-species studies will increase our understanding of the mechanisms by which these pathogens have evolved and shed light on the continually evolving pathways used to evade the host's immune system. To achieve these aims, it is vital that the varDB project receive input from the general research community on the utility of the database. In particular, input is needed from experts in the individual organisms to provide feedback on specific antigenic gene families. A long-term goal of the project is to involve the general research community in curating the sequences in the database as a way of providing (1) extremely accurate sequence sets and (2) so-called expert alignments of sets of antigenic sequences as this process is often non-trivial. Towards this end, varDB includes a feedback system (in the *Contact us* link) whereby users can submit comments and suggestions regarding the project. Another long-term goal is to develop a user-forum on antigenic variation and varDB to address many of the issues that are specific to working with these types of proteins. Taken together, the tools and resources included in varDB should be a useful asset to the general research community in working with these exciting, yet challenging, sequences in a wide range of pathogens.

## Acknowledgements

This work was supported by a Vinnova-SSF-JST Multidisciplinary BIO Grant, grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, The STINT Foundation, PREG-VAX, FP7-Health-2007-A-201588, Marie Curie IIF Fellowship EUFP6 (02154), and the Swedish Royal Academy of Sciences. C.N.H. and D.D were supported by post-doctoral fellowships from The Japanese Society for the Promotion of Science and C.E.W. was supported by The Centre for Allergy Research.

## References

- Allred, D.R., Barbet, A.F., Barry, J.D., Deitsch, K.W., 2009. varDB: common ground for a shifting landscape. *Trends Parasitol.* 25, 249–252.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Andersen, P., Nielsen, M.A., Resende, M., Rask, T.S., Dahlback, M., Theander, T., Lund, O., Salanti, A., 2008. Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA. *PLoS Pathog.* 4, e42.
- Baratin, M., Roetyncck, S., Pouvelle, B., Lemmers, C., Viebig, N.K., Johansson, S., Bierling, P., Scherf, A., Gysin, J., Vivier, E., Ugolini, S., 2007. Dissection of the role of PfEMP1 and ICAM-1 in the sensing of *Plasmodium falciparum*-infected erythrocytes by natural killer cells. *PLoS One* 2, e228.
- Barry, J.D., Ginger, M.L., Burton, P., McCulloch, R., 2003. Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.* 33, 29–45.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009. GenBank. *Nucleic Acids Res.* 37, D26–D31.
- Bertonati, C., Tramontano, A., 2007. A model of the complex between the PfEMP1 malaria protein and the human ICAM-1 receptor. *Proteins* 69, 215–222.
- Birney, E., Clamp, M., Durbin, R., 2004. GeneWise and genomewise. *Genome Res.* 14, 988–995.
- Bockhorst, J., Lu, F., Janes, J.H., Keebler, J., Gamain, B., Awadalla, P., Su, X.Z., Samudrala, R., Jojic, N., Smith, J.D., 2007. Structural polymorphism and diversifying selection on the pregnancy malaria vaccine candidate VAR2CSA. *Mol. Biochem. Parasitol.* 155, 103–112.
- Bull, P.C., Buckee, C.O., Kyes, S., Kortok, M.M., Thathy, V., Guyah, B., Stoute, J.A., Newbold, C.I., Marsh, K., 2008. *Plasmodium falciparum* antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol. Microbiol.* 68, 1519–1534.
- Cannon, J.P., Haire, R.N., Rast, J.P., Litman, G.W., 2004. The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. *Immunol. Rev.* 200, 12–22.
- Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S.V., Merino, E.F., Amedeo, P., Cheng, Q., Coulson, R.M., Crabb, B.S., Del Portillo, H.A., Essien, K., Feldblyum, T.V., Fernandez-Becerra, C., Gilson, P.R., Gueye, A.H., Guo, X., Kang'a, S., Kooij, T.W., Korsinczyk, M., Meyer, E.V., Nene, V., Paulsen, I., White, O., Ralph, S.A., Ren, Q., Sargeant, T.J., Salzberg, S.L., Stockert, C.J., Sullivan, S.A., Yamamoto, M.M., Hoffman, S.L., Wortman, J.R., Gardner, M.J., Galinski, M.R., Barnwell, J.W., Fraser-Liggett, C.M., 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763.
- Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Perlea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallow, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaihi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J., Carucci, D.J., 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419, 512–519.
- Clamp, M., Cuff, J., Searle, S.M., Barton, G.J., 2004. The Jalview Java alignment editor. *Bioinformatics* 20, 426–427.
- Cunningham, D.A., Jarra, W., Koernig, S., Fonager, J., Fernandez-Reyes, D., Blythe, J.E., Waller, C., Preiser, P.R., Langhorne, J., 2005. Host immunity modulates transcriptional changes in a multigene family (*yir*) of rodent malaria. *Mol. Microbiol.* 58, 636–647.
- Dzikowski, R., Deitsch, K.W., 2008. Active transcription is required for maintenance of epigenetic memory in the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* 382, 288–297.
- Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763.
- Fischer, K., Chavchich, M., Huestis, R., Wilson, D.W., Kemp, D.J., Saul, A., 2003. Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of *Plasmodium chabaudi*, a malaria species that undergoes antigenic variation in the laboratory mouse. *Mol. Microbiol.* 48, 1209–1223.
- Fonager, J., Cunningham, D., Jarra, W., Koernig, S., Henneman, A.A., Langhorne, J., Preiser, P., 2007. Transcription and alternative splicing in the *yir* multigene family of the malaria parasite *Plasmodium y. yoelii*: identification of motifs suggesting epigenetic and post-transcriptional control of RNA expression. *Mol. Biochem. Parasitol.* 156, 1–11.
- Frank, M., Kirkman, L., Costantini, D., Sanyal, S., Lavazec, C., Templeton, T.J., Deitsch, K.W., 2008. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. *Int. J. Parasitol.* 38, 1099–1109.
- Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallow, S.J., Suh, B., Peterson, J., Angiuoli, S., Perlea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.L., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511.
- Hayes, C.N., Diez, D., Joannin, N., Honda, W., Kanehisa, M., Wahlgren, M., Wheelock, C.E., Goto, S., 2008. varDB: a pathogen-specific sequence database of protein families involved in antigenic variation. *Bioinformatics* 24, 2564–2565.
- Janssen, C.S., Phillips, R.S., Turner, C.M., Barrett, M.P., 2004. *Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res.* 32, 5712–5720.
- Joannin, N., Abhiman, S., Sonhammer, E.L., Wahlgren, M., 2008. Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. *BMC Genom.* 9, 19.
- Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518.
- Klein, M.M., Gittis, A.G., Su, H.P., Makobongo, M.O., Moore, J.M., Singh, S., Miller, L.H., Garboczi, D.N., 2008. The cysteine-rich interdomain region from the highly variable *Plasmodium falciparum* erythrocyte membrane protein-1 exhibits a conserved structure. *PLoS Pathog.* 4, e1000147.
- Korir, C.C., Galinski, M.R., 2006. Proteomic studies of *Plasmodium knowlesi* SICA variant antigens demonstrate their relationship with *P. falciparum* EMP1. *Infect. Genet. Evol.* 6, 75–79.
- Normark, J., Nilsson, D., Ribacke, U., Winter, G., Moll, K., Wheelock, C.E., Bayarugaba, J., Kironde, F., Egwang, T.G., Chen, Q., Andersson, B., Wahlgren, M., 2007. PfEMP1-DBL1alpha amino acid motifs in severe disease states of *Plasmodium falciparum* malaria. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15835–15840.
- Oliveira, T.R., Fernandez-Becerra, C., Jimenez, M.C., Del Portillo, H.A., Soares, I.S., 2006. Evaluation of the acquired immune responses to *Plasmodium vivax* VIR variant antigens in individuals living in malaria-endemic areas of Brazil. *Malar. J.* 5, 83.

- Pain, A., Bohme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, M.R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M., Rajandream, M.A., Kocken, C.H., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455, 799–803.
- Pettersson, F., Vogt, A.M., Jonsson, C., Mok, B.W., Shamaei-Tousi, A., Bergstrom, S., Chen, Q., Wahlgren, M., 2005. Whole-body imaging of sequestration of *Plasmodium falciparum* in the rat. *Infect. Immun.* 73, 7736–7746.
- Ralph, S.A., Bischoff, E., Mattei, D., Sismeiro, O., Dillies, M.A., Guigon, G., Coppee, J.Y., David, P.H., Scherf, A., 2005. Transcriptome analysis of antigenic variation in *Plasmodium falciparum*—var silencing is not dependent on antisense RNA. *Genome Biol.* 6, R93.
- Rasti, N., Wahlgren, M., Chen, Q., 2004. Molecular aspects of malaria pathogenesis. *FEMS Immunol. Med. Microbiol.* 41, 9–26.
- Springer, A.L., Smith, L.M., Mackay, D.Q., Nelson, S.O., Smith, J.D., 2004. Functional interdependence of the DBLbeta domain and c2 region for binding of the *Plasmodium falciparum* variant antigen to ICAM-1. *Mol. Biochem. Parasitol.* 137, 55–64.
- Tham, W.H., Payne, P.D., Brown, G.V., Rogerson, S.J., 2007. Identification of basic transcriptional elements required for rif gene transcription. *Int. J. Parasitol.* 37, 605–615.
- Vigan-Womas, I., Guillotte, M., Le Scanf, C., Igonet, S., Petres, S., Juillerat, A., Badaut, C., Nato, F., Schneider, A., Lavergne, A., Contamin, H., Tall, A., Baril, L., Bentley, G.A., Mercereau-Puijalon, O., 2008. An in vivo and in vitro model of *Plasmodium falciparum* rosetting and autoagglutination mediated by varO, a group A var gene encoding a frequent serotype. *Infect. Immun.* 76, 5565–5580.