TOOLS FOR INVESTIGATING MECHANISMS OF ANTIGENIC VARIATION: NEW EXTENSIONS TO VARDB

C. NELSON HAYES [1]
nelson@kuicr.kyoto-u.ac.jp

DIEGO DIEZ [1]
diez@kuicr.kyoto-u.ac.jp

NICOLAS JOANNIN [2,3]
nicolas.joannin@ki.se

MINORU KANEHISA [1]
kanehisa@kuicr.kyoto-u.ac.jp

MATS WAHLGREN [2,3]
mats.wahlgren@ki.se

CRAIG E. WHEELOCK [1,4*]
craig.wheelock@ki.se

SUSUMU GOTO [1*]
goto@kuicr.kyoto-u.ac.jp

[1] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan*
[2] *Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, SE-17177 Stockholm, Sweden*
[3] *Swedish Institute for Infectious Disease Control (Smittskyddsinstitutet), SE-17182 Stockholm, Sweden*
[4] *Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden*
* *To whom correspondence should be addressed*

The varDB project (`http://www.vardb.org`) aims to create and maintain a curated database of antigenic variation sequences as well as a platform for online sequence analysis. Along with the evolution of drug resistance, antigenic variation presents a moving target for public health endeavors and greatly complicates vaccination and eradication efforts. However, careful analysis of a large number of variant forms may reveal structural and functional constraints that can be exploited to identify stable and cross-reactive targets. VarDB attempts to facilitate this effort by providing streamlined interfaces to standard tools to help identify and prepare sequences for various forms of analysis. We have newly implemented such tools for codon usage, selection, recombination, secondary and tertiary structure, and sequence diversity analysis. Just as the adaptive immune system encodes a mechanism for dynamically generating diverse receptors instead of encoding a receptor for every possible epitope, many pathogens take advantage of heritable diversity generating mechanisms to produce progeny able to evade immune recognition. Instead of merely cataloging the observed variation, a major goal of varDB is to characterize and predict the potential range of antigenic variation within a pathogen by investigating the mechanisms by which it attempts to expand its implicit genome. We believe that the new sequence analysis tools will improve the usefulness and range of varDB.

Keywords: antigenic variation, hyper-variable sequence, codon usage, selection.

1. Introduction

Following infection, cells of the innate immune system rapidly detect the pathogen intruder and present fragments of the pathogen to surveillance cells of the adaptive immune system. This interaction triggers a cascade in which individual B and T

cells that possess high affinity for the target antigen undergo clonal selection to initiate a highly effective and specific immune response [9]. The overwhelming immune response that follows is usually sufficient to clear the pathogen, but the time required to orchestrate clonal selection and affinity maturation affords antigenically distinct subpopulations more time to replicate and disperse before an immune response against the new targets can be generated. Many of the most severe public health problems, including malaria and HIV/AIDS, are caused by pathogens that have evolved mechanisms allowing them to maintain an edge in the arms race with the immune response by stochastically varying their antigenic profiles. This process of antigenic variation is encountered in a vast range of organisms including viruses, bacteria, fungi, and protists [15]. A major goal of public health is to develop preventative vaccines able to thwart this mechanism by training the immune system to recognize conserved epitopes in one or a few common variants so that the pathogen can be recognized and cleared before the pathogen is able to establish immunogenically variable subpopulations.

The success of this approach depends in part on the way in which variants are generated, and in many cases the mechanisms are incompletely understood or vary greatly by species. Viruses such as HIV rely on a high mutation rate coupled with a high reproductive rate [3], whereas bacteria may use mechanisms based on recombination, gene conversion, or lateral transfer [15]. Eukaryotes may employ these or other still more complex mechanisms involving coordinated expression of multi-gene families with multiple levels of control [15]. VarDB was developed both as a repository for antigenic variation data as well as a platform for comparative analysis of mechanisms of antigenic variation [16, 21]. It consists of a database of sequences belonging to antigenic variation gene families as well as an integrated suite of tools to simplify analysis of sequence data and estimate the range and nature of sequence variation. Recently, we have greatly expanded the sequence analysis tools in varDB to include codon usage, selection, recombination, secondary and tertiary structure, and sequence diversity analyses, in addition to standard tools such as BLAST and HMM search. Here we describe the new extensions to varDB and discuss how they can be used to examine patterns of sequence variability within the context of antigenic variation.

2. Tools for Investigating Mechanisms of Antigenic Variation

The current varDB release contains more than 62,000 sequences from known or suspected antigenic variation gene families in 30 organisms. Sequence data is collected from genome sequencing projects and field isolate submissions using an HMM-based data mining pipeline described previously [16]. Because of the large number of sequence types, varDB provides a range of search and filtering methods to quickly extract datasets and prepare them for analysis. Once a working set is obtained, interfaces to multiple external analysis tools are provided to estimate sequence variability in a streamlined fashion. While some of these tools represent novel im-

plementations or interfaces, the major focus of varDB is not so much to develop new tools as it is to simplify and integrate usage of standard tools already in wide use by the research community. Many of these programs are platform-dependent, have strict input requirements or limited output options, or are available only as separate command line or web-based tools. Therefore, varDB attempts to bridge these differences and help integrate diverse resources by using a flexible and consistent Ajax-based interface. The application of some of these tools for estimating sequence variability is described below.

2.1. *Sequence Selection and Preparation*

One of the challenging aspects of studying antigenic variation is compiling and formatting an appropriate data set for analysis. To simplify this process, sequences in varDB can be searched using a flexible query language. Matching sequences are returned as a sortable, pageable grid view with a panel showing the number of sequences belonging to each of a number of categories, including Pfam domain, gene family, genome project, source (e.g., PlasmoDB, Broad Institute, GenBank Core or dbEST), etc., which can then be used to further filter the result set.

2.2. *Generating a Codon Alignment*

To analyze codon usage and selection, it is helpful to first create an alignment from coding sequences. However, this procedure often requires multiple steps and is sometimes complicated by problems in matching the nucleotide sequence with the correct translation, particularly in eukaryotes with complex splicing rules or when using sequences from unfinished genome sequencing projects. While nucleotide and amino acid sequences can be downloaded separately in varDB, a new feature in the database allows pre-processed pairs of corresponding protein and nucleotide sequences to be downloaded together, along with a table of sequence annotations. In this step introns and terminal stop codons are removed, and sequences are verified to ensure that there are no internal stop codons and that sequence lengths are divisible by three. Although nucleotide sequences can be aligned directly, protein sequence alignments are often used to generate more accurate codon-based alignments, especially among highly variable sequences in which synonymous substitutions and amino acid substitution matrix scores provide additional information to help resolve alignment ambiguities. Given a pre-computed protein alignment, codon alignments can be generated within varDB using Pal2Nal [38] or RevTrans [41] to align coding sequences using the protein alignment as a template.

2.3. *Analyzing Codon Usage*

A great deal of information can be extracted from codon alignments by examining variation in the way alternative codons are used to code for the same amino acid. The amino acid code is degenerate, using 64 codons to code for 20 amino acids and

three stop codons. Therefore most amino acids are coded by at least two or as many as six different codons, usually differing only in the third position. In principle, synonymous codons are interchangeable and should be invisible to selection, but in reality codon usage is often biased towards a subset of optimal codons [23]. This is due in part to variation in the overall GC content, which constrains the frequency of codons that contain predominantly G/C or A/T nucleotides. Genome-wide codon usage may also depend on the relative frequency of corresponding tRNAs during translation [23], whereas variation in codon usage within a gene may reflect functional constraints at the DNA level [14].

VarDB provides both standard and novel utilities to assist in codon usage analysis. When a codon alignment is selected, the codon usage tool displays GC content and GC3skew as well as tables of nucleotide, codon, and amino acid frequencies (Fig. 1). Codons can be grouped by synonymous amino acid in descending order based on within-group relative codon usage or other measures [13]. For comparison with background frequencies, pathogen-specific codon usage statistics are provided for each pathogen based on the Codon Usage Database (`http://www.kazusa.or.jp/codon/`).

Although codon usage can be analyzed using unaligned sequences, variation in codon usage by alignment position may be useful for detecting nucleotide patterns involved in generating new variants. VarDB displays the number and type of codon at each position using different colors to indicate non-synonymous codons (Fig. 2). This view is intended to highlight unusual codon usage patterns by alignment column.

A separate view shows codon usage for cysteine residues alone. Due to their high degree of conservation and structural role in disulfide bond formation, certain cysteine residues may be maintained under purifying selection. Even in the face of balancing selection for cysteine conservation, the two codons coding for cysteine should occur at their respective background frequencies. Therefore, bias in codon usage that varies by position might suggest selection acting at the nucleotide level. Conticello et al. demonstrated that snail conopeptides show exclusive usage of only one of the two codons at certain key positions immediately upstream of hypervariable sites, leading them to propose that clusters of cysteine codons may cause stalling of the replication fork and temporary recruitment of an error prone DNA polymerase, thereby increasing the mutation rate in a predictable and highly localized fashion [14]. This idea is similar in principle to the mechanism used by the vertebrate adaptive immune response during targeted somatic hypermutation, which also shows significant codon usage bias by location, particularly in the case of serine [40]. The potential importance of this type of mechanism is that it provides a means for an organism to efficiently explore sequence space without disrupting essential structural or functional motifs required by selectively varying the mutation rate throughout the sequence [11].

Cysteines are likely to play an important role in the evolution of antigenic variation proteins because they provide a rigid super-structure that can accommodate

Codon Usage

| GC content | **Codons** | Amino acids | Amino acid groups |

Amino acid	Percent	Codon 1	Codon 2	Codon 3	Codon 4	Codon 5	Codon 6
Lysine	10.84 %	AAA 1.59	AAG 0.41				
Glycine	8.97 %	GGT 1.85	GGA 1.29	GGG 0.59	GGC 0.26		
Aspartic Acid	7.63 %	GAT 1.71	GAC 0.29				
Leucine	7.40 %	TTG 2.21	TTA 2.06	CTT 0.99	CTA 0.51	CTG 0.15	CTC 0.08
Glutamic Acid	7.19 %	GAA 1.59	GAG 0.41				
Asparagine	6.91 %	AAT 1.68	AAC 0.32				
Serine	6.77 %	AGT 2.86	TCT 1.11	TCA 1.08	AGC 0.43	TCG 0.39	TCC 0.14
Alanine	6.46 %	GCA 2	GCT 1.06	GCC 0.52	GCG 0.43		
Arginine	5.23 %	AGA 2.41	CGT 1.38	CGA 1.06	AGG 0.72	CGC 0.38	CGG 0.06
Isoleucine	4.85 %	ATA 1.73	ATT 1.12	ATC 0.15			
Threonine	4.81 %	ACA 1.79	ACT 1.32	ACG 0.64	ACC 0.24		
Tyrosine	4.09 %	TAT 1.74	TAC 0.26				
Valine	3.65 %	GTG 1.38	GTA 1.33	GTT 1.09	GTC 0.2		
Cysteine	3.01 %	TGT 1.34	TGC 0.66				
Histidine	2.90 %	CAT 1.68	CAC 0.32				
Glutamine	2.70 %	CAA 1.56	CAG 0.44				
Proline	2.68 %	CCA 1.74	CCT 1.16	CCG 0.92	CCC 0.18		
Phenylalanine	2.43 %	TTT 1.6	TTC 0.4				
Methionine	1.38 %	ATG 1					
Tryptophan	0.11 %	TGG 1					

Fig. 1.   Relative synonymous codon usage in *P. falciparum* var genes.

short insertions and deletions in variable regions. Selection may favor particular combinations of codons that help to increase the probability of mutations occurring where they are likely to be favorable or decrease the probability of mutations where they may disrupt sensitive regions [10]. To facilitate testing of these hy-

Fig. 2.   Codon usage by column in *P. falciparum* var genes.

potheses, varDB provides a tool to compare codon usage between conserved and variable blocks of an alignment. The tool uses Gblocks [39] to split the alignment into conserved and variable regions and then compares codon usage and amino acid composition between them. Histograms are generated to help visualize the distribution of the variable region sequence lengths. Similarly, the programs seg and dust can be used within varDB to mask low complexity regions for amino acid and nucleotide sequences, respectively [42], making it easier to detect conserved flanking regions. *Plasmodium falciparum*, in particular, is unusual for the presence of long insertions comprised of low complexity regions [32].

2.4. *Nucleotide Repeats and DNA Secondary Structure*

Conserved nucleotide motifs independent of codon boundaries may also play a role in generating diversity. These may be found by examining consensus sequences or by using motif detection tools [5]. The codon usage tool in varDB contains a view that shows the relative frequency of each nucleotide for each column. Repetitive patterns in a sequence are often biologically relevant, and short repeat elements have been implicated in slippage of the replication machinery, which may introduce variation by causing frame shift mutations, turning gene expression on or off, or changing the rate of gene expression by altering the affinity of transcription factor binding sites [29]. Similarly, inverted repeats and pseudo-palindromes may form secondary stem-loop structures that are thought to interfere with transcription/replication or that may be "repaired" to form complete palindromes [10, 11, 34]. Moreover, triplet

repeats may lead to in-frame insertions and deletions that could substantially alter a protein's antigenic signature without disrupting translation [29]. To help identify this kind of short direct or inverted repeat elements, VarDB provides a web service interface for the Oligo Repeat Finder web site (`http://wwwmgs.bionet.nsc.ru/mgs/programs/oligorep/InpForm.htm`).

2.5. *Mutation Hotspot Motifs*

Analysis of mutation spectra has identified a number of DNA motifs associated with an increased probability of mutation based on sequence context [36]. Notably, RGYW/WRCY motifs have been implicated in immunoglobulin somatic hypermutation, in which the mutation rate is six orders of magnitude higher than surrounding regions as part of a mechanism to generate and select high affinity variants via C→U deamination and error prone repair using DNA polymerase η [31, 35]. Consequently, it may be helpful to examine antigenic variation sequences for the presence of known and potential mutagenic motifs. VarDB provides flexible query tools to search for these patterns using the IUPAC extended genetic alphabet (e.g., R for purines and Y for pyrimidines), or sequences can be searched using Perl-like regular expressions. Alternatively, PHI-BLAST can be used to restrict matches to those that also share high overall sequence homology to the query sequence to reduce the rate of false positives when searching for short motifs in a large data set [4].

2.6. *Recombination*

One of the ways in which pathogens generate antigenic diversity involves gene duplication, in which paralogous gene families, ranging from four or five members in some bacteria to over 1000 in *Trypanosoma brucei*, are coordinately expressed only one member at a time, causing an antigenic shift and triggering a new immune challenge [8]. This strategy is effective because members of these gene families are often located in the rapidly evolving subtelomeric regions near the ends of chromosomes [6] (Fig. 3). In spite of high sequence variability, these genes often contain conserved motifs, possibly in the upstream and downstream flanking regions, that provide targets for non-homologous recombination and gene conversion [12]. This process can help to rapidly diversify subtelomeric gene families in a parallel fashion, such that over time strains contain fewer and fewer identical genes [25]. In some cases gene conversion using only partial genes or pseudogenes can generate novel mosaic genes [27].

VarDB provides several utilities to aid investigation of recombination patterns. Wherever possible, sequences from multiple genome sequencing projects are included, providing a way to estimate the degree of variation within and among strains. A simple custom genome browser is provided to visualize the location and genomic context of genes from sequencing projects (Fig. 4). Because flanking regions may provide important information for identifying potential recombination sites, arbitrarily long regions upstream or downstream of a given feature in a genome can
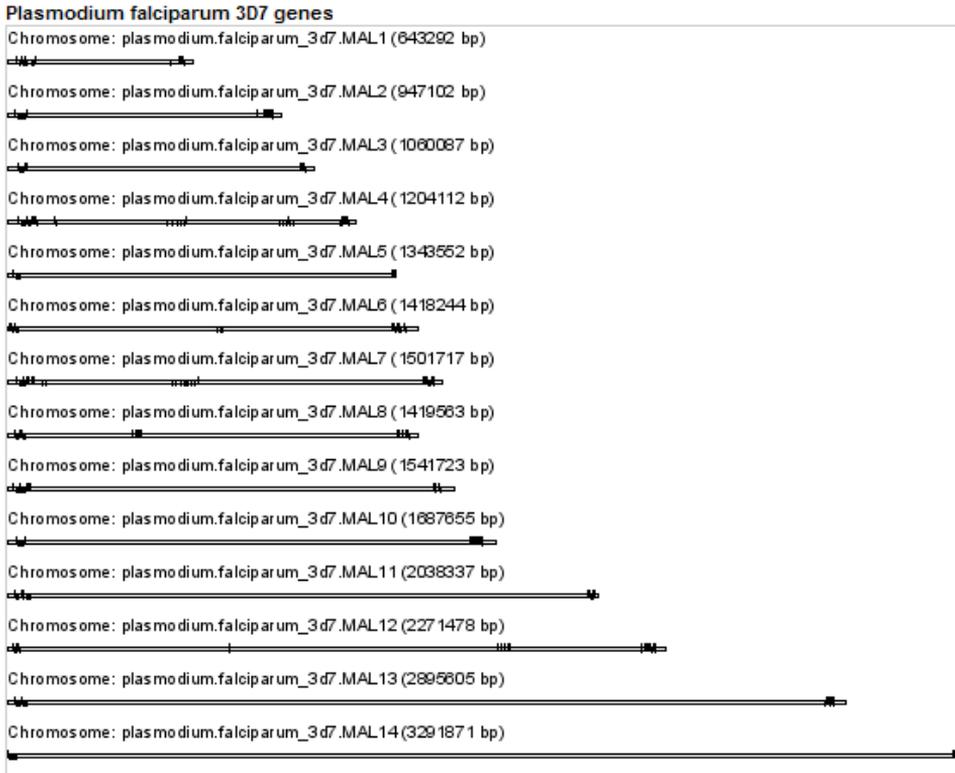
Plasmodium falciparum 3D7 genes

Chromosome: plasmodium.falciparum_3d7.MAL1 (643292 bp)

Chromosome: plasmodium.falciparum_3d7.MAL2 (947102 bp)

Chromosome: plasmodium.falciparum_3d7.MAL3 (1060087 bp)

Chromosome: plasmodium.falciparum_3d7.MAL4 (1204112 bp)

Chromosome: plasmodium.falciparum_3d7.MAL5 (1343552 bp)

Chromosome: plasmodium.falciparum_3d7.MAL6 (1418244 bp)

Chromosome: plasmodium.falciparum_3d7.MAL7 (1501717 bp)

Chromosome: plasmodium.falciparum_3d7.MAL8 (1419563 bp)

Chromosome: plasmodium.falciparum_3d7.MAL9 (1541723 bp)

Chromosome: plasmodium.falciparum_3d7.MAL10 (1687655 bp)

Chromosome: plasmodium.falciparum_3d7.MAL11 (2038337 bp)

Chromosome: plasmodium.falciparum_3d7.MAL12 (2271478 bp)

Chromosome: plasmodium.falciparum_3d7.MAL13 (2895605 bp)

Chromosome: plasmodium.falciparum_3d7.MAL14 (3291871 bp)

Fig. 3. Subtelomeric clusters of var and rif/stevor antigenic variation genes are found in all 14 chromosomes of the *P. falciparum* 3D7 genome.

be downloaded in bulk by uploading an accession list. BLAST and PSI-BLAST are useful for finding stretches of homologous DNA that may provide a target for recombination [4]. Antigenic variation sequence data can also be exported for analysis in external programs such as RDP3, a standalone tool that attempts to infer the recombination history of a set of sequences [28].

2.7. *Variability and Immune Selection*

Ideally vaccines targeting invariant protein structures could be developed against some of these recalcitrant pathogens, but realistically most vaccines must protect against a wide range of potential conformations that may be encountered during infection [7]. Despite these sophisticated antigenic variation generating mechanisms, however, it may be possible to find a subset of evolutionarily constrained structural motifs that are conserved among multiple variants [20]. Although only a few solved crystal structures are available for antigenic variation proteins, varDB uses BLAST to map sequences to available structures and Jmol (`http://www.jmol.org/`) to visualize the location of matches on the structure in order to provide a rough sense
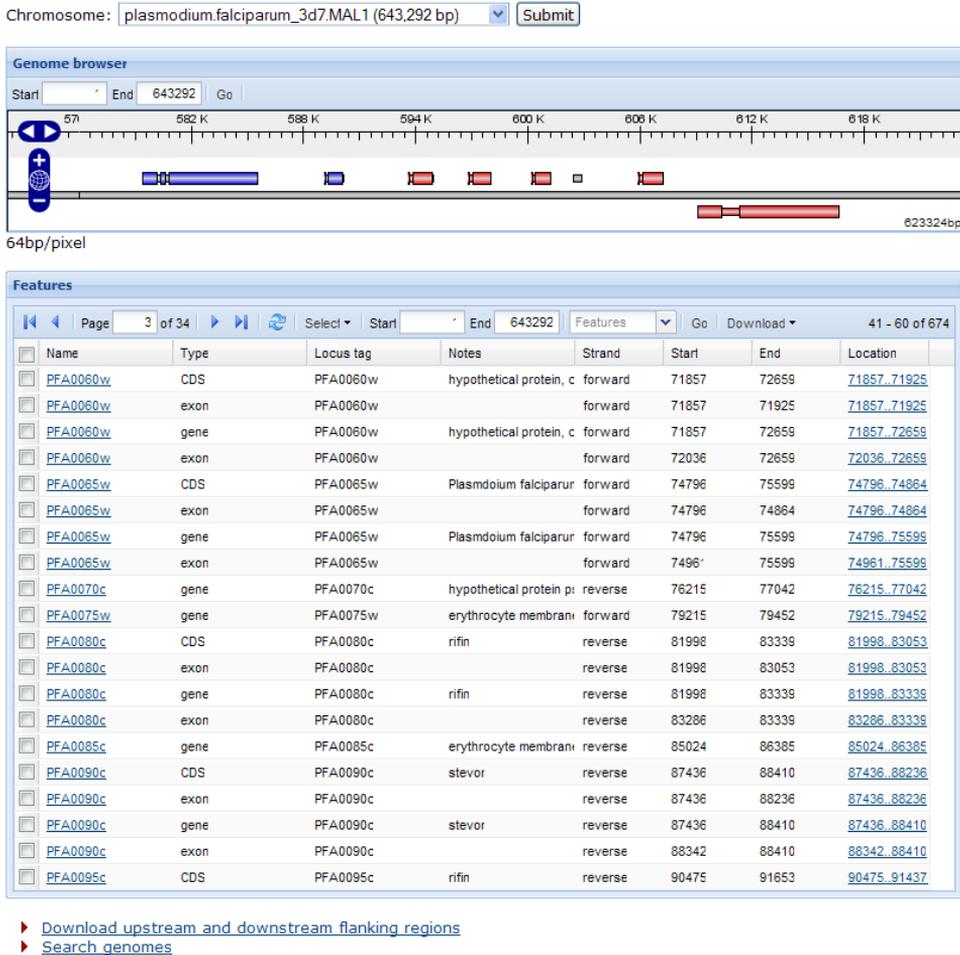
Fig. 4.   Simple genome browser showing orientation of genes and a pseudogene in the subtelomeric region of *P. falciparum* 3D7 chromosome 1.

of the spatial orientation of residues (Fig. 5). A more powerful alternative would be to create homology models based on existing structures using external software such as MODELLER [17]. VarDB does not include an interface to MODELLER due to licensing restrictions, but references to published homology models are included where available. Given the degree of sequence variation and the limited number of crystal structures currently available, however, the accuracy of homology models for antigenic variation proteins may be low, particularly within the variable loop regions that are of most interest. Nonetheless, as additional crystal structures become available, this approach should become increasingly useful for visualizing the effects of sequence variation on 3D structure and epitope conformation. VarDB also contains links to a number of tools to predict T and B cell epitopes that may be

Fig. 5.   Sequence from the Trypanosoma brucei TREU927 genome project mapped onto the two chains of PDB structure 1VSG using bl2seq and Jmol.

helpful in predicting which residues are accessible to immune receptors or whether structural variants are likely to be antigenically distinct [18].
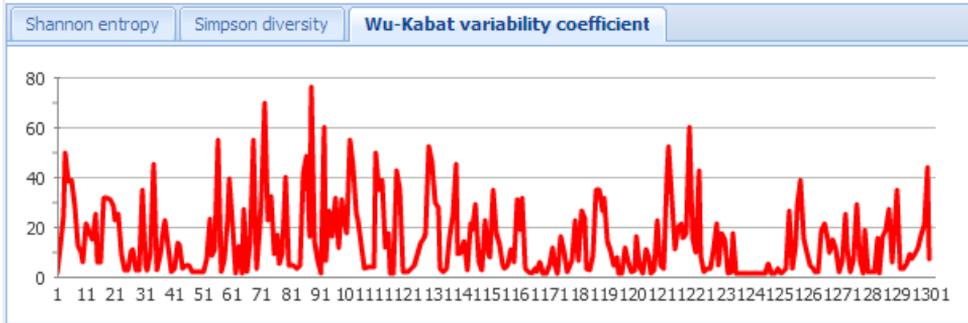
Fig. 6.   Wu-Kabat variability by column in a *P. falciparum* PfEMP1 protein multiple sequence alignment.

In the absence of a crystal structure, one approach to identifying potential drug or vaccine targets is to create a protein alignment and analyze the information content at each position. VarDB implements many of the methods available in the Protein Variability Server [19], including Shannon Entropy, the Simpson Diversity Index, and the Wu-Kabat Variability Coefficient, each of which characterizes a different aspect of amino acid variability (Fig. 6). Positions with high amino acid diversity may reflect residues exposed to the immune system and undergoing diversifying selection. Evidence for diversifying selection at the nucleotide level can be inferred by an excess of non-synonymous changes per non-synonymous site (Ka) relative to the number of synonymous substitutions per synonymous site (Ks) [22]. VarDB can calculate the Ka/Ks ratio along a sliding window, and sequences can be formatted for more extensive analysis using external tools such as DataMonkey [33] or Selecton [37].

3. Conclusions

VarDB provides a streamlined interface and a number of utilities for compiling and analyzing variation among sequences. Combining data from multiple such methods should provide insight into which residues are essential for structural or functional integrity and which tend to serve as immunodominant decoys. Vaccines which are able to target the former and avoid the latter are likely to be more effective in preempting immune evasion, thereby reducing the length of infection and the ability of the pathogen to reproduce and disperse.

Much attention has recently been focused on heritable diversity generating mechanisms, mutator phenotypes, and the concept of an implicit genome [11]. As microorganisms rapidly evolve and freely trade genetic material able to confer resistance or evade immune recognition, the effectiveness of existing treatments will continue to erode, and formerly innocuous organisms will continue to acquire virulence traits and become emerging threats. While Lynch [26] and others [30] caution against invoking potentially adaptationist arguments in interpreting the evolution of genomic

architecture, many pathogens have converged on strikingly similar patterns of antigenic variation while facing the common threat of their host immune systems [15]. Therefore, much can be learned by comparing strategies employed to great effect by distantly related pathogens. VarDB has already drawn attention from the research community [2], and it is our hope that we will be able to provide a useful resource for investigating antigenic variation in the effort to control infectious diseases that continue to evade not only our immune defenses but also, increasingly, existing front line drugs and vaccines.

Acknowledgments

References

[1] Allred, D. R., A. F. Barbet, J. D. Barry, K. W. Deitsch, C. L. Althaus and R. J. De Boer, Dynamics of immune escape during HIV/SIV infection, *Trends Parasitol*, 4(7): e1000103, 2008.

[2] Allred, D. R., A. F. Barbet, J. D. Barry and K. W. Deitsch, varDB: common ground for a shifting landscape, *Trends Parasitol*, 25(6): 249-52, 2009.

[3] Althaus, C. L. and R. J. De Boer, Dynamics of immune escape during HIV/SIV infection, *PLoS Comput Biol*, 4(7): e1000103, 2008.

[4] Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 25(17): 3389–402, 1997.

[5] Bailey, T. L., N. Williams, C. Misleh and W. W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, *Nucleic Acids Res*, 34(Web Server issue): W369–73, 2006.

[6] Barry, J. D., M. L. Ginger, P. Burton and R. McCulloch, Why are parasite contingency genes often associated with telomeres?, *Int J Parasitol*, 33(1): 29–45, 2003.

[7] Belli, S. I., R. A. Walker and S. A. Flowers, Global protein expression analysis in apicomplexan parasites: current status, *Proteomics*, 5(4): 918–24, 2005.

[8] Berriman, M., E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D. C. Bartholomeu, N. J. Lennard, E. Caler, N. E. Hamlin, B. Haas, U. Bohme, L. Hannick, M. A. Aslett, J. Shallom, L. Marcello, L. Hou, B. Wickstead, U. C. Alsmark, C. Arrowsmith, R. J. Atkin, A. J. Barron, F. Bringaud, K. Brooks, M. Carrington, I. Cherevach, T. J. Chillingworth, C. Churcher, L. N. Clark, C. H. Corton, A. Cronin, R. M. Davies, J. Doggett, A. Djikeng, T. Feldblyum, M. C. Field, A. Fraser, I. Goodhead, Z. Hance, D. Harper, B. R. Harris, H. Hauser, J. Hostetler, A. Ivens, K. Jagels, D. Johnson, J. Johnson, K. Jones, A. X. Kerhornou, H. Koo, N. Larke, S. Landfear, C. Larkin, V. Leech, A. Line, A. Lord, A. Macleod, P. J. Mooney, S. Moule, D. M. Martin, G. W. Morgan, K. Mungall, H. Norbertczak, D. Ormond, G. Pai, C. S. Peacock, J. Peterson, M. A. Quail, E. Rabbinowitsch, M. A. Rajandream, C. Reitter, S. L.

Salzberg, M. Sanders, S. Schobel, S. Sharp, M. Simmonds, A. J. Simpson, L. Tallon, C. M. Turner, A. Tait, A. R. Tivey, S. Van Aken, D. Walker, D. Wanless, S. Wang, B. White, O. White, S. Whitehead, J. Woodward, J. Wortman, M. D. Adams, T. M. Embley, K. Gull, E. Ullu, J. D. Barry, A. H. Fairlamb, F. Opperdoes, B. G. Barrell, J. E. Donelson, N. Hall and C. M. Fraser, The genome of the African trypanosome Trypanosoma brucei, *Science*, 309(5733): 416–22, 2005.

 [9] Cannon, J. P., R. N. Haire, J. P. Rast and G. W. Litman, The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms, *Immunol Rev*, 200: 12–22, 2004.

[10] Caporale, L. H., Natural selection and the emergence of a mutation phenotype: an update of the evolutionary synthesis considering mechanisms that affect genome variation, *Annu Rev Microbiol*, 57: 467–85, 2003.

[11] Caporale, L. H., *The implicit genome. Oxford ; New York*, Oxford University Press, Oxford ; New York, 2006.

[12] Centurion-Lara, A., R. E. LaFond, K. Hevner, C. Godornes, B. J. Molini, W. C. Van Voorhis and S. A. Lukehart, Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of Treponema pallidum during infection, *Mol Microbiol*, 52(6): 1579–96, 2004.

[13] Charif, D., J. Thioulouse, J. R. Lobry and G. Perriere, Online synonymous codon usage analyses with the ade4 and seqinR packages, *Bioinformatics*, 21(4): 545–7, 2005.

[14] Conticello, S. G., Y. Gilad, N. Avidan, E. Ben-Asher, Z. Levy and M. Fainzilber, Mechanisms for evolving hypervariability: the case of conopeptides, *Mol Biol Evol*, 18(2): 120–31, 2001.

[15] Deitsch, K. W., S. A. Lukehart and J. R. Stringer, Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens, *Nat Rev Microbiol*, 7(7): 493–503, 2009.

[16] Diez, D., N. Hayes, N. Joannin, J. Normark, M. Kanehisa, M. Wahlgren, C. E. Wheelock and S. Goto, varDB: A database of antigenic variant sequences-Current status and future prospects, *Acta Trop*, 2009.

[17] Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper and A. Sali, Comparative protein structure modeling using MODELLER, *Curr Protoc Protein Sci*, Chapter 2: Unit 2 9, 2007.

[18] Flower, D. R., *Bioinformatics for vaccinology. Chichester, West Sussex, England ; Hoboken, NJ*, John Wiley & Sons, Chichester, West Sussex, England ; Hoboken, NJ, 2008.

[19] Garcia-Boronat, M., C. M. Diez-Rivero, E. L. Reinherz and P. A. Reche, PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery, *Nucleic Acids Res*, 36(Web Server issue): W35–41, 2008.

[20] Gratepanche, S., B. Gamain, J. D. Smith, B. A. Robinson, A. Saul and L. H. Miller, Induction of crossreactive antibodies against the Plasmodium falciparum variant protein, *Proc Natl Acad Sci U S A*, 100(22): 13007–12, 2003.

[21] Hayes, C. N., D. Diez, N. Joannin, W. Honda, M. Kanehisa, M. Wahlgren, C. E. Wheelock and S. Goto, varDB: a pathogen-specific sequence database of protein families involved in antigenic variation, *Bioinformatics*, 24(21): 2564–5, 2008.

[22] Hurst, L. D., The Ka/Ks ratio: diagnosing the form of sequence evolution, *Trends Genet*, 18(9): 486, 2002.

[23] Ikemura, T., Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system, *J Mol*

Biol, 151(3): 389–409, 1981.

[24] Katoh, K., K. Kuma, T. Miyata and H. Toh, Improvement in the accuracy of multiple sequence alignment program MAFFT, *Genome Inform*, 16(1): 22–33, 2005.

[25] Kraemer, S. M., S. A. Kyes, G. Aggarwal, A. L. Springer, S. O. Nelson, Z. Christodoulou, L. M. Smith, W. Wang, E. Levin, C. I. Newbold, P. J. Myler and J. D. Smith, Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates, *BMC Genomics*, 8: 45, 2007.

[26] Lynch, M., *The origins of genome architecture. Sunderland, Mass.*, Sinauer Associates, Sunderland, Mass., 2007.

[27] Marcello, L. and J. D. Barry, 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*, 17(9): 1344–52, 2007.

[28] Martin, D. P., Recombination detection and analysis using RDP3, *Methods Mol Biol*, 537: 185–205, 2009.

[29] Moxon, R., C. Bayliss and D. Hood, Bacterial contingency Loci: the role of simple sequence DNA repeats in bacterial adaptation, *Annu Rev Genet*, 40: 307–33, 2006.

[30] Nei, M., Selectionism and neutralism in molecular evolution, *Mol Biol Evol*, 22(12): 2318–42, 2005.

[31] Neuberger, M. S. and C. Rada, Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase eta for A/T, *J Exp Med*, 204(1): 7–10, 2007.

[32] Pizzi, E. and C. Frontali, Low-complexity regions in Plasmodium falciparum proteins, *Genome Res*, 11(2): 218–29, 2001.

[33] Poon, A. F., S. D. Frost and S. L. Pond, Detecting signatures of selection from DNA sequences using Datamonkey, *Methods Mol Biol*, 537: 163–83, 2009.

[34] Ripley, L. S., Predictability of mutant sequences. Relationships between mutational mechanisms and mutant specificity, *Ann N Y Acad Sci*, 870: 159–72, 1999.

[35] Rogozin, I. B. and M. Diaz, Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process, *J Immunol*, 172(6): 3382–4, 2004.

[36] Rogozin, I. B. and Y. I. Pavlov, Theoretical analysis of mutation hotspots and their DNA sequence context specificity, *Mutat Res*, 544(1): 65–85, 2003.

[37] Stern, A., A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach and T. Pupko, Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach, *Nucleic Acids Res*, 35(Web Server issue): W506–11, 2007.

[38] Suyama, M., D. Torrents and P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Res*, 34(Web Server issue): W609–12, 2006.

[39] Talavera, G. and J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *Syst Biol*, 56(4): 564–77, 2007.

[40] Thorpe, I. F. and C. L. Brooks, 3rd, Molecular evolution of affinity and flexibility in the immune system, *Proc Natl Acad Sci U S A*, 104(21): 8821–6, 2007.

[41] Wernersson, R. and A. G. Pedersen, RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences, *Nucleic Acids Res*, 31(13): 3537–9, 2003.

[42] Wootton, J. C., Non-globular domains in protein sequences: automated segmentation using complexity measures, *Comput Chem*, 18(3): 269–85, 1994.