DATABASE IN BRIEF

HbVar Database of Human Hemoglobin Variants and Thalassemia Mutations: 2007 Update

Belinda Giardine [1], Sjozef van Baal [2], Polynikis Kaimakis [2], Cathy Riemer [1], Webb Miller [1], Maria Samara [3], Panagoula Kollia [3], Nicholas P. Anagnou [4], David H. K. Chui [5], Henri Wajcman [6], Ross C. Hardison [1,7], and George P. Patrinos [2,*]

[1] The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, Pennsylvania; [2] Erasmus MC, Faculty of Medicine and Health Sciences, MGC-Department of Cell Biology and Genetics, Rotterdam, The Netherlands; [3] Department of Biology, University of Thessaly School of Medicine, Larissa, Greece; [4]Department of General Biology, University of Athens, School of Medicine, Athens, Greece; [5]Departments of Medicine and Pathology, Boston University School of Medicine, Boston, Massachusetts; [6] INSERM-U654 Bases Moléculaires et Cellulaires des Maladies Genetiques, Hôpital Henri Mondor, Créteil, France; [7]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania

*Correspondence to: George P. Patrinos, MGC-Department of Cell Biology and Genetics, Erasmus MC, Faculty of Medicine and Health Sciences, PO Box 2040, 3000 CA, Rotterdam, The Netherlands. Tel.: +31-10-408.7454; Fax: +31-10-408.9468; E-mail: g.patrinos@erasmusmc.nl

Communicated by Stylianos E. Antonarakis

HbVar (http://globin.bx.psu.edu/hbvar) is a locus-specific database (LSDB) developed in 2001 by a multi-center academic effort to provide timely information on the genomic sequence changes leading to hemoglobin variants and all types of thalassemia and hemoglobinopathies. Database records include extensive phenotypic descriptions, biochemical and hematological effects, associated pathology, and ethnic occurrence, accompanied by mutation frequencies and references. In addition to the regular updates to entries, we report significant advances and updates, which can be useful not only for HbVar users but also for other LSDB development and curation in general. The query page provides more functionality but in a simpler, more user-friendly format and known single nucleotide polymorphisms in the human α- and β-globin loci are provided automatically. Population-specific β-thalassemia mutation frequencies for 31 population groups have been added and/or modified and the previously reported δ- and α-thalassemia mutation frequency data from 10 population groups have also been incorporated. In addition, an independent flat-file database, named XPRbase (http://www.goldenhelix.org/xprbase), has been developed and linked to the main HbVar web page to provide a succinct listing of 51 experimental protocols available for globin gene mutation screening. These updates significantly augment the database profile and quality of information provided, which should increase the already high impact of the HbVar database, while its combination with the UCSC powerful genome browser and the ITHANET web portal paves the way for drawing connections of clinical importance, that is from genome to function to phenotype. © 2007 Wiley-Liss, Inc.

KEY WORDS: LSDB; globin genes; thalassemia; variants; mutation screening; software

INTRODUCTION

Hemoglobinopathies are the most common inherited disorders in humans, resulting from mutations in the α- and β-globin gene clusters (reviewed in Forget et al., 2001). The α-globin gene cluster is composed of the genes *HBZ* (MIM# 142310), *HBA2* (MIM# 141850), *HBA1* (MIM# 141800), *HBM* (MIM# 609639) and *HBQ1* (MIM#

Received 10 August 2006; accepted revised manuscript 31 October 2006.

DOI: 10.1002/humu.9479

142240), which encode the ζ-, α2-, α1-, and possibly μ- and θ-globin polypeptides, respectively. The β-globin gene cluster is composed of the genes *HBE1* (MIM# 142100), *HBG2* (MIM# 142250), *HBG1* (MIM# 142200), *HBD* (MIM# 142000) and *HBB* (MIM# 141900), which encode the ε-, $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin polypeptides, respectively. Single nucleotide substitutions can lead to hemoglobin (Hb) variants, due to amino acid replacements, while molecular defects in either regulatory or coding regions of the human α-, β- or δ-globin genes can minimally or drastically reduce their expression, leading to α-, β- or δ- thalassemia respectively.

In 2001, we developed HbVar, derived from previous compilations (Huisman et al., 1997, 1998), as a publicly available locus-specific database (LSDB) to provide timely information to interested users, e.g. the globin research community, providers of genetic services and counseling, patients and their parents, pharmaceutical industries, *etc*. New hemoglobin variants and thalassemias continue to be discovered, and thus HbVar was designed for regular entry updates and corrections. The query interface provides easy access to this information for the research community and for physicians as an aid in diagnosis. HbVar has rapidly become an important aid in the globin research community and is considered to be one of the premier LSDBs available to date (Claustres et al., 2002, Beroud, 2005). The major advantages of HbVar compared to other LSDBs, are the mutation information quality and depth of coverage, its regular updates and upgrades, and, most importantly, its interrelation with other databases, such as GALA (Giardine et al., 2003), GenPhen and PhenCode (Giardine et al., submitted). Minor disadvantages were the lack of information on the available experimental protocols to screen for globin gene mutations and the database's lengthy query page.

We report here several new updates in HbVar's structure and contents, aiming at increasing the quality of the database and its impact to society. Also, the possible implications of these updates in the curation of other LSDBs are discussed.

UPDATES TO EXISTING DATA

HbVar's information has been expanded by more than 400 additional entries and corrections, made continually by the database curators during the two recent upgrades (Table 1). In order to identify new Hb variants and thalassemia mutations not previously documented in the database, HbVar records were compared against Online Mendelian Inheritance in Man (MIM) records and external links were placed in each HbVar entry, redirecting the user to the respective MIM entry. Also, articles from the specialized journal *Hemoglobin*, which frequently publishes new Hb variants and thalassemia mutations, were manually scanned and where applicable, previously undocumented mutations have been included into HbVar. Although empirical nomenclature is widely used in HbVar, proper [Human Genome Variation Society (HGVS); http://www.hgvs.org] mutation nomenclature is also automatically provided next to each entry, based on the reference sequence.

As far as population-specific data are concerned, β-thalassemia mutation frequencies for 31 population groups have been added and/or modified where needed, while the previously reported δ- and α-thalassemia mutation frequency data from 10 population groups have also been extracted from the published literature and made available for the HbVar users..

Finally, we have recently introduced a unique identifier for each HbVar entry. This feature will be particularly useful for curators to avoid duplicate entries and also for users when referring to a particular mutant, especially those lying within mutational hotspots.

QUERY PAGE UPDATES AND NEW FUNCTIONALITIES

The previous query page for HbVar was lengthy, so in order to streamline it without losing any functionality, we have grouped all previous options by type and included them in 13 different hyperlinked sections (Fig. 1).

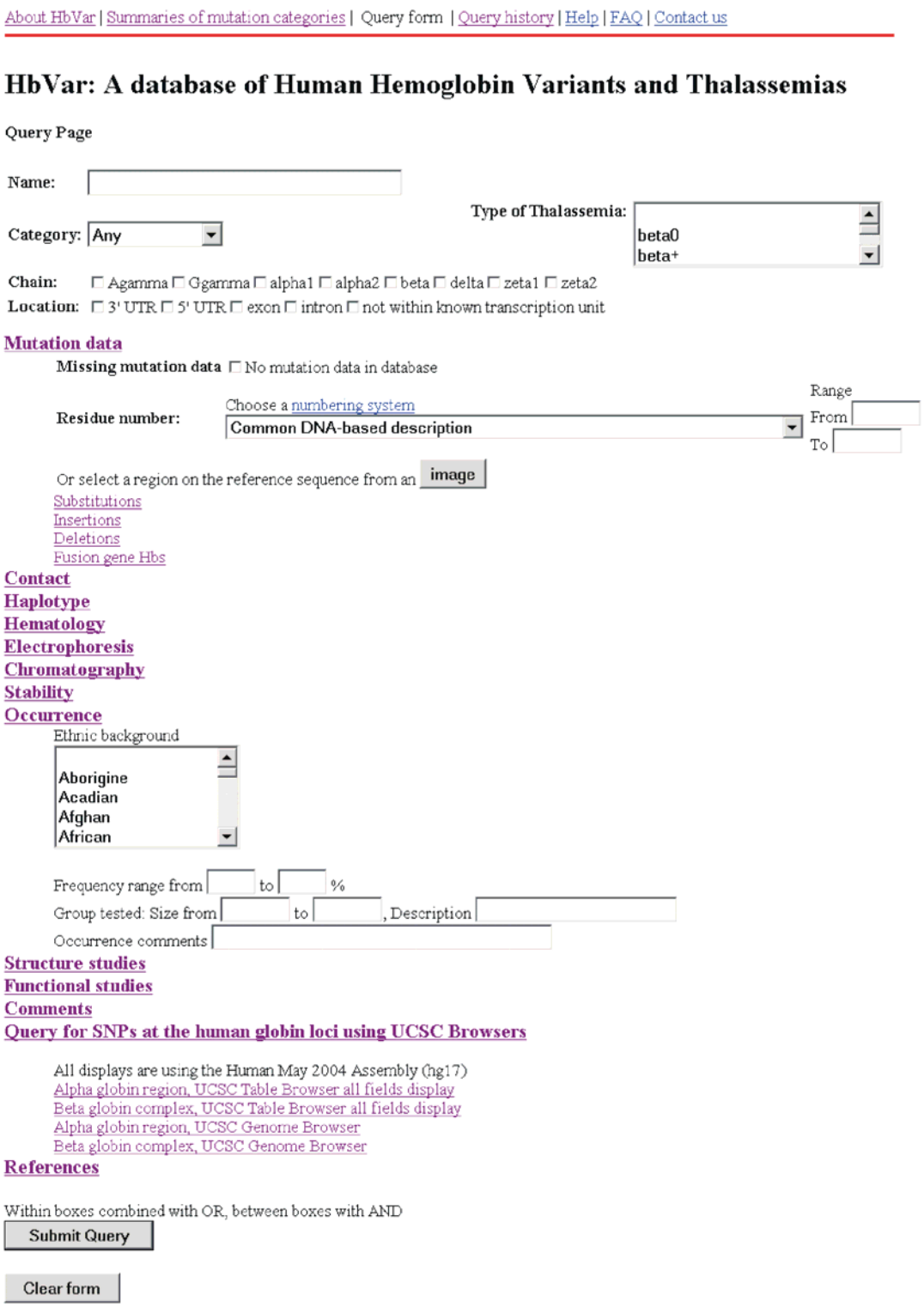

Figure 1. The new HbVar query page. All of the features of the previously lengthy query page (Hardison et al., 2002) have been incorporated in the various hyperlinks. Also, new functionalities have been added for screening for SNPs and for graphically displaying the reported mutants per region for every gene (see text for details).

Therefore, if a user wishes to query the database based on a variant's occurrence, he/she only needs to open the "Occurrence" section and narrow the search by selecting the relevant criteria included therein. This minimize/maximize option has improved the clarity of display of the query page. Also, a "query history" has been added, which displays a list of the queries a user has previously run. Tools on this page not only enable the user to retrieve previous query results but also to perform operations on the results, such as unions, intersections and subtractions, to refine them.

Table 1. Summaries of Mutation Categories in the HbVar Database for Hemoglobin Variants and Thalassemia Mutations (Note That Some Entries Appear in More Than One Category)

Entries	Count of results	
	October 2001 [a]	October 2006 [b]
Entries involving the *HBA1* gene (OMIM# 141800)	212	254
Entries involving the *HBA2* gene (OMIM# 141850)	250	295
Entries involving the *HBB* gene (OMIM# 141900)	635	711
Entries involving the *HBD* gene (OMIM# 142000)	57	72
Entries involving the *HBG2* gene (OMIM# 142200)	44	49
Entries involving the *HBG1* gene (OMIM# 142250)	52	55
Entries with a fusion gene mutation	8	8
Entries with a substitution mutation	899	1,031
Entries with an insertion mutation	46	51
Entries with a deletion mutation	116	150
Hemoglobins with high oxygen affinity	79	90
Unstable hemoglobins	121	134
Methemoglobins	9	9
Total hemoglobin variant entries	832	938
Total thalassemia entries	336	383
Total entries in both variant and thalassemia categories	43	48
Total entries in database	**1,125**	**1,273**

[a]: Data from Hardison et al., 2002, [b]: Present report.

In addition to these changes in the operation of the query and history pages, there are several new features that have been added. Users can now query HbVar by gene location as well as by chain. For example, by selecting "Beta" and "Exon" the user will only get the mutations that are in the exons of the β-globin gene. Also under the "Mutation data" section there is a button to select regions by clicking on an image. There, a user can select a gene or a region between genes on the first image (Fig. 2A). When the user zooms in on a particular gene in order to select an exon or intron, the mutations documented in HbVar are displayed in a dense mode below the genes, grouped by category (substitutions, deletions, insertions, *etc.*) to give an impression of how much data is included in HbVar for each region (Fig. 2B).

In the "Occurrence" section, additional fields have been included to refine the query based on the population sample size tested. This is an important parameter, which is indicative of how representative the mutation frequencies are.

Automatic routines have been implemented to query for single nucleotide polymorphisms (SNPs) in the human α- and β-globin loci via the University of California at Santa Cruz (UCSC) Genome Browser (Hinrichs et al., 2006). The user can choose to view the results either in a table format or as graphical display.

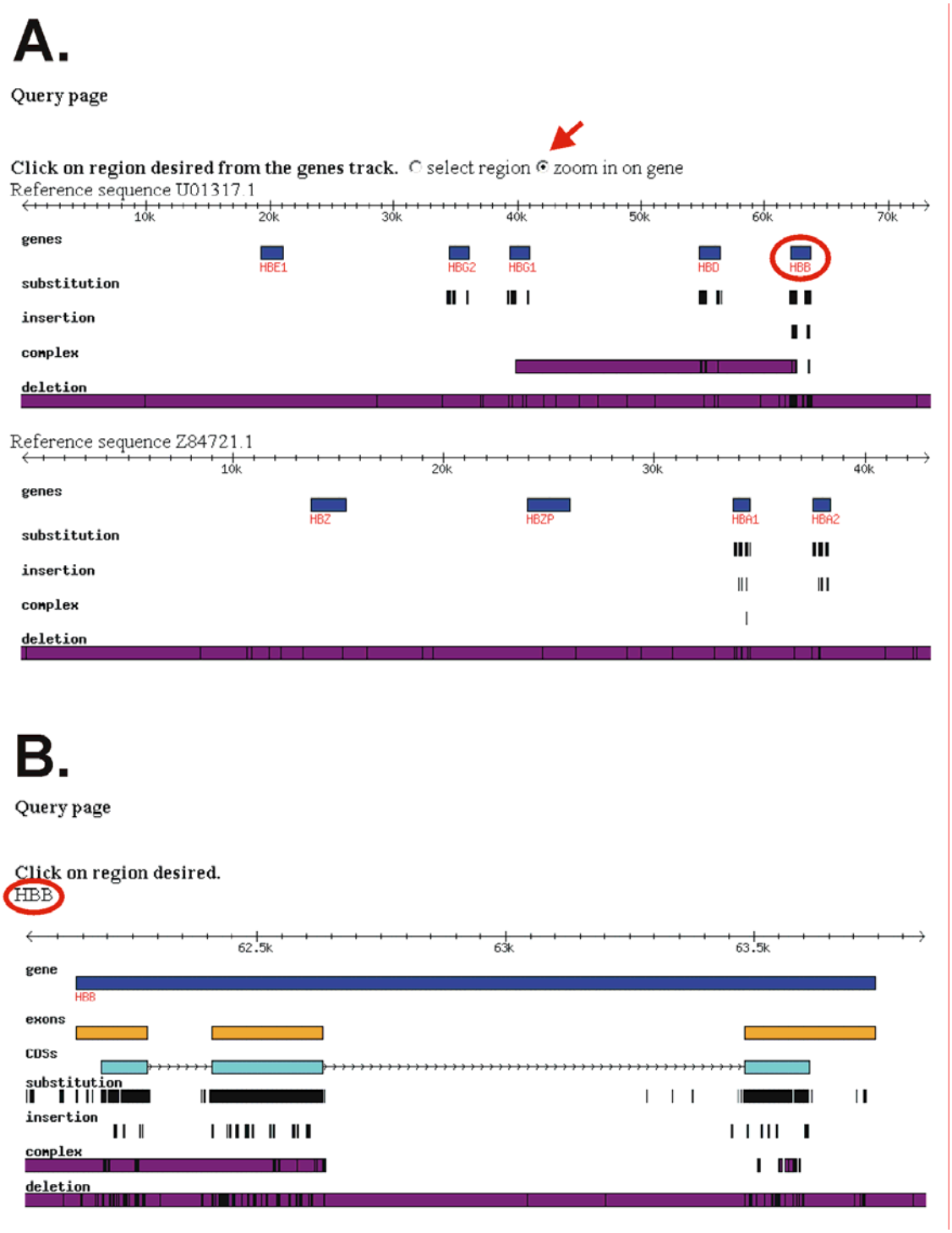

Figure 2. Overview of the new HbVar querying feature to select regions from an image. **A:** A user can select a particular region between genes or zoom in on a gene (indicated by the red arrow) to select a particular exon or intron. **B:** Outcome from selecting the human β-globin gene (*HBB*) region (circled in red). The various types of mutations included in HbVar (substitutions, insertions, deletions, and complex rearrangements) are displayed in a dense mode below the gene to give an indication of the amount of data currently available for each region. Mutant entries in these images are not "clickable."

DATABASE IMPLEMENTATION AND ACCESS

In its underlying implementation, HbVar was initially designed and built as a relational database, using Oracle as the database management system. We have recently switched from Oracle to PostgreSQL (http://www.postgresql.org), an object-oriented relational database management system (sometimes referred to as an object-relational database), which is available in several free and commercial versions. This choice was guided partly by the high costs of the Oracle license, compared to the free PostgreSQL system.

The HbVar database and associated resources at the Globin Gene Server (http://globin.bx.psu.edu, Hardison et al., 1994), such as the online *Syllabi*, are currently in use worldwide. Since January 2000, we recorded 15,859 accesses to the HbVar query page and 87,328 accesses to the online *Syllabi*, a significant increase compared to the previous years (Hardison et al., 2002, Patrinos et al., 2004). Users frequently contact the curators and the rest of the HbVar team members in order to submit new hemoglobin variants and/or thalassemia mutations, report missing information for existing mutants and pinpoint inconsistencies and/or erroneous entries. This is particularly important, since user input is vital in our effort to improve the data quality and accuracy. Therefore we encourage HbVar users to notify the curators regarding such issues (detailed contact information is available at the Globin Gene Server site).

XPRBASE, A HUMAN GLOBIN GENE MUTATION SCREENING PROTOCOLS DATABASE

We have constructed a separate database, named XPRbase (Fig. 3A), to provide a succinct summary listing of the protocols available for human globin gene mutation screening. The database currently contains 51 protocols for whole-gene, mutation-screening strategies (see also Patrinos et al., 2005a). These protocols are summarized in Table 2. XPRbase can be accessed on the World Wide Web at: http://www.goldenhelix.org/xprbase, and a link is provided from the HbVar homepage. Detailed instructions for both using and querying the database are also available from the same site. There has been no claim of ownership of the information stored in this database by anyone involved in this initiative. However, this compilation and representations of it are subject to copyright and usage principles to ensure that this resource remains freely available to all interested individuals.

Table 2. Summary of the Experimental Protocols Available in XPRbase to Screen for Sequence Alterations in the Human Globin Loci, Categorized by Gene and Mutation Detection Strategy

Type	Protocols	Globin genes					Total
		α	β	γ	δ	ε	
Known mutations	Restriction analysis	1	1	1	1	0	**4**
	Reverse dot-blot	1	3	0	0	0	**4**
	PCR-ASO	2	1	0	0	0	**3**
	PCR-ARMS	1	3	0	0	0	**4**
	Real-time PCR	0	2	0	0	0	**2**
	GAP-PCR	8	0	1	1	0	**10**
	Microarrays	2	4	0	0	0	**6**
Unknown mutations	SSCP	3	2	0	0	1	**6**
	DGGE	1	3	2	1	0	**7**
	DG-DGGE	1	0	0	0	0	**1**
	DHPLC	1	3	0	0	0	**4**
Total		**21**	**22**	**4**	**3**	**1**	**51**

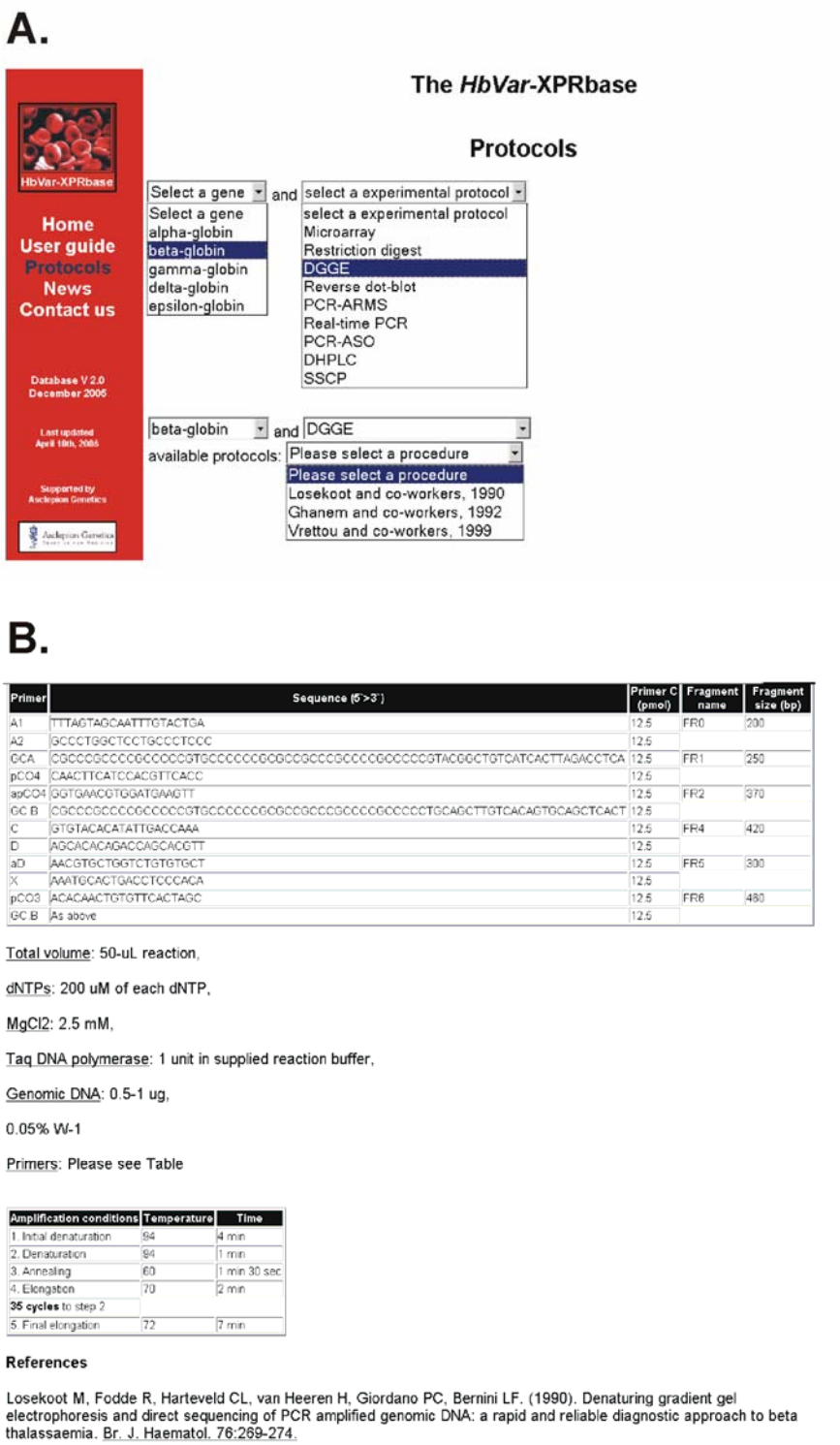

Figure 3. A: The protocols page from XPRbase. A user can select from two drop-down menus, which give the protocols that are available for mutation screening in every globin gene. When running the query "Find all DGGE protocols available for β-globin gene mutation screening," three protocols are displayed. **B:** Outcome of the query as described in A, consisting of the primer sequence, amplification conditions, reaction ingredients and the source reference(s), hyperlinked to PubMed, from where the publication(s) can be retrieved.

All database screens are based on the HTML language with some JavaScript and rely on Cascading Style Sheet (CSS) support. They are built using a custom-made PHP script that comprises the database's core engine, not only for menus and basic screens that display and parse files, but also for handling data querying. XPRbase is a flat-file database, derived from the *ETHNOS* V1.0 software, which facilitates the establishment of National Mutation Frequency databases (Patrinos et al., 2005b). A user guide provides some brief information on the operation and querying principles of XPRbase.

Database protocols can be added and/or modified only by the administrator through a dedicated administrator module (Sjozef van Baal and George Patrinos, unpublished). They are contained within an index file, namely "protocols.tab". When a new protocol is added to this file, a separate file named for the first author of the protocol is automatically created. This enables the administrator to enter the new protocol details by selecting this file from the protocol drop-down menu. To modify an existing protocol, the administrator only needs to select the desired protocol from the list and modify its contents in the designated area. All XPRbase data are freely available for academic research purposes, and data and accompanying descriptions are available for this website.

FUTURE PROSPECTS

During the 5 years that HbVar has been fully functional, it has become a key resource for information about sequence variation leading to hemoglobinopathies. One important feature is its constant update and improvement, mostly driven by the devotion of the researchers involved in this project and the valuable input from the end users, reporting erroneous entries and omissions and suggesting new additions. Given the increasing impact of HbVar on society and its growing content, we plan to expand the curator team by adding more expert advisors, and further to perform a user satisfaction survey to be able to identify weaknesses in the database contents and structure. The positive impact HbVar has on the globin research community is also illustrated by the fact that funding, whether dedicated or related to other projects, has always been available for keeping this resource alive, in an environment where funding opportunities for database development and curation are often limited and very hard to find (Patrinos and Brookes, 2005), frequently resulting in the discontinuation of many useful databases.

In order to ensure continuous data influx into HbVar, we plan to implement a broader search strategy that combines manual and electronic search procedures. This involves an expanded computerized search of all articles listed in the PubMed literature database (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=PubMed), in addition to the manual screening of a selection of core hematology journals. Also, in order to prevent globin gene mutation data from being overlooked, it is anticipated that HbVar will be more tightly linked to the journal *Hemoglobin*, following the successful recipe of the Waystation project (http://www.centralmutations.org) linked to the journal *Human Mutation*, and of the Human Gene Mutation Database (HGMD; http://www.hgmd.org, Stenson et al., 2003) currently linked to the journal *Human Genetics*. In addition, HbVar will be used as the main globin gene mutation data source for the portal currently under construction for the ITHANET project (http://www.ithanet.eu).

We also plan to enrich XPRbase, a vital component of the ITHANET portal, with additional experimental protocols and to create multiple links between individual HbVar entries and the protocols by which those mutations can be detected. Also, a graphical display of all oligonucleotide primers documented in XPRbase in relation to their respective genomic regions can improve the service provided by this database. Although this will require a more advanced technological platform, this feature would help in designing gap-PCR protocols to detect the over 40 known δβ-thalassemia and Hereditary Persistence of Fetal Hb (HPFH) deletional mutants, provided that their deletion breakpoints are precisely mapped within HbVar. That is an ongoing effort, facilitated by the fact that the DNA sequence for the human β-globin gene cluster and surrounding regions is now complete. This will allow the precise identification of all deletion junctions and annotation of the DNA features affected by the deletions. Indeed, this information will be critical for interpreting all deletional mutants. Not only will the junction sequences allow better analysis and interpretation of the mutations, but they will also allow specialized screening strategies to be designed and implemented for each mutation.

The link between the HbVar and GALA databases (Giardine et al., 2003, Patrinos et al., 2004), coupled with the UCSC Genome Browser (Hinrichs et al., 2006), was the first step towards integrating the available resources at the Globin Gene Server (Hardison et al., 1994). A step further is PhenCode (Giardine et al., submitted), which connects the phenotype and clinical data in a variety of LSDBs, including HbVar and GenPhen, to the data on genome sequences, evolutionary history, and function available at the Genome Browser. Displaying locus-specific mutation data with clinical relevance on the Genome Browser makes it readily available to a wide audience, and facilitates the viewing of data from many sources in one context. On the other hand, links back to the original databases allow detailed queries within individual loci, which can then be further analyzed accordingly. This

combination of LSDBs and a powerful genome browser paves the way for drawing connections from genome to function to phenotype. Also, expansion of these connections to additional loci of clinical importance, perhaps building on the current WayStation project, will help in fulfilling the promise of the Human Genome Project to improve human health.

ACKNOWLEDGMENTS

We thank all the HbVar users worldwide for their valuable comments and suggestions, which help us to keep the information as updated and complete as possible and also contribute to the continuous improvement of the database and its contents. We will always be indebted to the late Prof. Titus H.J. Huisman and his colleagues for their detailed compilations of hemoglobin variants and thalassemia mutations. This work was supported by United States Public Health Service grants HG02238 to WM and DK065806 to RCH, by a European Commission grant RI026539 to GPP, and by financial support from Tobacco Settlement Funds of the Commonwealth of Pennsylvania and the Huck Institutes of the Life Sciences at Penn State University and Asclepion Genetics (Lausanne, Switzerland).

REFERENCES

Beroud C. 2005. The use of mutation databases in molecular diagnostics. In: Patrinos GP, Ansorge W, editors. Molecular Diagnostics. San Diego, Academic Press/Elsevier. p 319-325.

Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12:680-688.

Forget BG, Higgs DR, Steinberg M, Nagel RL. 2001. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management. Cambridge University Press, Cambridge, UK.

Giardine B, Elnitski L, Riemer C, Makalowska I, Schwartz S, Miller W, Hardison RC. 2003. GALA, a database for genomic sequence alignments and annotations. Genome Res. 13:732-741.

Hardison RC, Chao KM, Schwartz S, Stojanovic N, Ganetsky M, Miller W. 1994. Globin gene server: A prototype E-mail database server featuring extensive multiple alignments and data compilation. Genomics 21:344-353.

Hardison RC, Chui DH, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W, Wajcman H. 2002. HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum Mutat 19:225-233.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. 2006. The UCSC Genome Browser Database: update 2006. Nucleid Acids Res 34:D590-D598.

Huisman TH, Carver MF, Baysal E. 1997. A Syllabus of Thalassemia Mutations. The Sickle Cell Anemia Foundation, Augusta, GA.

Huisman TH, Carver MF, Efremov GD. 1998. A Syllabus of Human Hemoglobin Variants (2nd Edition). The Sickle Cell Anemia Foundation, Augusta, GA.

Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, Anagnou NP, Wajcman H, Hardison RC. 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 32:D537-D541.

Patrinos GP, Kollia P, Papadakis MN. 2005a. Molecular diagnosis of inherited disorders: Lessons from hemoglobinopathies. Hum Mutat 26:399-412.

Patrinos GP, van Baal S, Petersen MB, Papadakis MN. 2005b. Hellenic National Mutation database: A prototype database for mutations leading to inherited disorders in the Hellenic population. Hum Mutat 25:327-333.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: Disastrously deficient. Trends Genet 21:333-338.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577-581.