Development of a universal, flexible and freely available database management system for gene-centered data collection, curation and display of DNA variation

Sophia Zaimidou, Sjozef van Baal, Timothy D. Smith, Konstantinos Mitropoulos, Mila Ljujic, Dragica Radojkovic, Richard G. Cotton, and George P. Patrinos

Abstract — **We report the development of a flexible database management system, based on a relational database format, for locus-specific database development and curation and its implementation for the development of a new locus-specific database, namely A_1ATVar, for *SERPINA1* gene variants, leading to α_1-antitrypsin deficiency (available from http://www.goldenhelix.org/a1atvar). This tool allows data entry and display of mutation summaries in a tabular format, while user-generated queries can be formulated based on fields in the database table. A separate module, linked to FINDbase database for frequencies of inherited disorders allows the user to access relevant allele frequency information in different populations worldwide. Available experimental protocols to detect variant alleles at the protein and DNA levels can also be archived in a searchable format. A visualization tool, called VariVis, is implemented to combine variant information with gene sequence and annotation data. Finally, a direct data submission tool allows registered users to submit data on novel variant alleles as well as experimental protocols via a password-protected interface. Implementation of this tool is free of charge and there are no registration requirements for data querying. This database management system can serve as an example for an all-in-one solution for locus-specific database development and curation.**

I. Introduction

RESEARCH into the genetic basis of disease has advanced in scale and sophistication, leading to very high rates of data production in many laboratories, while DNA diagnostics and electronic healthcare records are increasingly common features of modern medical practice. Therefore, it should be possible to integrate all of this information in order to establish a detailed understanding of how genome sequence differences impact human health.

Manuscript received July 4, 2008. This work was supported by the European Commission [EuroGenTest FP6 Network of Excellence (512148)], and corporate funding from Asclepion Genetics (Switzerland).

S. Z., S. v. B. and G. P. P. are with Erasmus University Medical Center, Faculty of Medicine and Health Sciences, MGC-Department of Cell Biology and Genetics, Rotterdam, The Netherlands (Phone: +31-10-704.39.49; fax: +331-10-704.47.43; e-mail: g.patrinos@erasmusmc.nl).

T. D. S. and R. G. C. are from Genomic Disorders Research Centre, Carlton South and Department of Medicine, The University of Melbourne, Parkville, Australia.

K. M. is from Asclepion Genetics, Lausanne, Switzerland.

M. L. and D. R. are from the Institute of Molecular Genetics and Genetic Engineering, Laboratory for Molecular Biology, Belgrade, Serbia.

In the last decade, major advances have been made in the cloning and characterization of genes involved in human diseases. Concurrently, advances in technology have led to the identification of numerous mutations in these genes, ranging from point mutations to large rearrangements. It rapidly has become clear that the knowledge and organization of these alterations in structured repositories will be of great importance not only for diagnosis but also for clinicians and researchers. Genetic or Mutation databases are referred to as online repositories of mutation data, described for a single (*locus-specific*) or more (*general*) genes or specifically for a population or ethnic group (*national/ethnic*). The main applications of mutation databases are: (1) To facilitate diagnosis at the DNA level and to define an optimal strategy for mutation detection, (2) To provide information about mutation-specific phenotypic patterns, and (3) To correlate locus-specific variant information with genome-wide features, such as repetitive elements, gene structures, inter-species conservation, mutation hot-spots, recombination frequencies and so on.

Locus-specific databases (LSDBs) are concerned with just one or a few specific genes [1], usually related to a single disease entity. They aim to be highly curated repositories of published and unpublished mutations within those genes, and as such they provide a much-needed complement to the core databases. Data quality and completeness is typically high, with up to 50% of stored records pertaining to otherwise unpublished mutations [1]. The data are also very rich and informative and the annotation of each mutant includes a full molecular and phenotypic description. Therefore, these databases are referred to as "inch wide and mile deep". For example, LSDBs will typically present each of the multiple discoveries of recurrent mutational events, so allowing mutation hot-spots to be identified: and when these mutations occur upon different chromosomal backgrounds (linked to other mutations) such that they result in several, or different, disease features, these correlations are also recorded. A good example of an LSDB would be HbVar database (available from http://globin.bx.psu.edu/hbvar), a relational database of hemoglobin variants and thalassemia mutations, providing information on pathology, hematology, clinical presentation and laboratory findings for a wealth of DNA alterations [2], [3]. Gene/protein variants are annotated with respect to biochemical data, analytical techniques, structure, stability, function, literature references, and qualitative and quantitative distribution in

ethnic groups and geographic locations [3]. As is common in LSDBs, entries can be accessed through summary listings or user-generated queries, which can be highly specific. A listing of the over 650 currently available LSDBs is available from http://www.hgvs.org, http://www.hgmd.org and in the literature [4].

Today, there are few database management systems (DBMSs) that can facilitate LSDB development and curation [5], [6]. These user-friendly DBMSs are designed to promote the creation of more and better LSDBs, by reducing or eliminating the requirement for substantial knowledge of computing and bio-informatics for interested parties to establish a LSDB from scratch. In addition, these off-the-shelf solutions positively impact on data uniformity, since the LSDB domain suffers from extreme structure and data content heterogeneity that impact their overall quality, while their ability to run on any platform reduces the risk of the database being "lost" if curation for some reason, e.g. lack of funding, is interrupted. In this case, data will be transferred directly between platforms or locations and they will remain accessible to all interested users.

We report here the development of a new DBMS for LSDB development and curation and its implementation for the development of a new LSDB for *SERPINA1* gene variants, namely A_1ATVar, leading to α_1-antitrypsin (AAT) deficiency.

II. ALPHA$_1$-ANTITRYPSIN DEFICIENCY AND *SERPINA1* GENE

AAT deficiency is one of the most common inherited disorders worldwide with an estimated incidence of 1 in 2,500 individuals [7]. AAT belongs to the serine proteinase inhibitor (serpin) family and is the most important protease inhibitor (Pi), significant for normal lung function. AAT deficient patients have low AAT serum, and hence alveolar levels, leading to unimpeded neutrophil elastase digestion of collagen and elastin in the alveolar walls and progressive emphysema [8]. Hence, AAT deficiency can predispose to or cause pulmonary disorders, while smoking, infections and exposure to dust and fumes can deteriorate the symptoms and/or accelerate their onset [9] AAT deficient patients can also develop liver disease, as a result of accumulation of particular AAT variants in hepatocytes [10]. AAT deficiency manifests at birth or in early childhood as a molecular dysfunction of liver and represents genetic predisposition for the subsequent development of severe liver or lung disorders later in life [11]. Early diagnosis and prevention measures are currently the best choice for the management of AAT deficiency patients. However, this disease often remains undiagnosed and less than 5% of severely AAT deficient patients are currently identified and subjected to treatment. The main reason for this is the poor genotype to phenotype correlations described for this disorder.

AAT protein is 394-amino acid long and encoded by *SERPINA1* gene (OMIM 107400). The gene is mapped on the long arm of chromosome 14 (14q32.1) and contains 7 exons and 6 introns [12]. AAT alleles are expressed in a co-dominant manner and, therefore, the combined effect of these alleles leads to certain levels of AAT protein. The normal AAT allele is M (PiM), accounting for almost 90% of all AAT alleles and with AAT serum levels of 20-53 μmol/L, while the commonest AAT deficient alleles are PiS and PiZ, the latter being the most severe AAT allele [9]. Homozygous ZZ patients present with only 3.4-7 μmol/L of AAT serum levels (approximately 10-15% of normal levels). Additionally, a relatively large number of other rare AAT variants have been described, accounting to 2-4% of the total number [9].

Although there are several patient-oriented websites to offer summary information on AAT deficiency, a locus-specific database (LSDB) that would document existing knowledge for research, genotype/phenotype correlation and disease diagnostics at the protein and/or DNA level is currently missing. LSDBs are extremely useful tools, as they can contribute towards identification of causative mutations, they can provide information about phenotypic patterns associated with a specific mutation and can enable researchers to define an optimal strategy for mutation detection [13].

III. DATA SOURCES AND DATABASE DESIGN

The primary source of information in A_1ATVar is the PubMed literature database (available from http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=PubMed) and proper variant nomenclature and annotation is done by the curators. Where DNA information was missing, the predicted DNA change is provided. A_1ATVar contains information on 91 AAT variant alleles, accounting for the vast majority of those reported in the literature and 13 experimental protocols for AAT variant and *SERPINA1* gene mutation screening protocols.

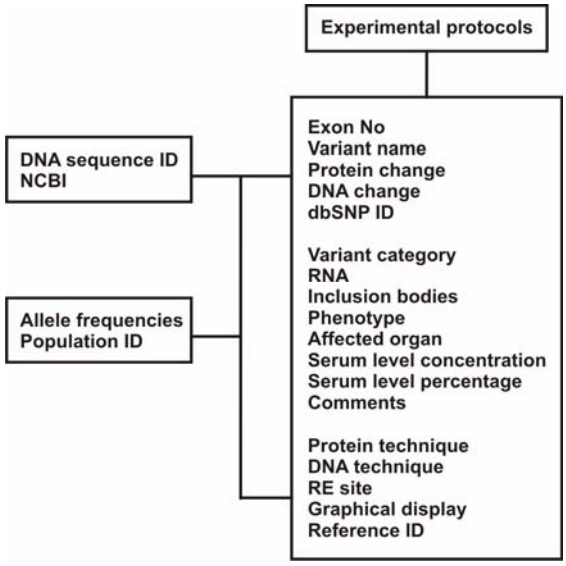

Fig. 1. Structure of the underlying database engine.

The main menu of A_1ATVar is located at the left side of each page, shown as buttons that depict the database functionalities. Each selected page is highlighted. A "*User guide*" provides simple instructions on how to navigate in and query A_1ATVar. The main database content, *i.e.*, all AAT variant alleles, can be assessed and queried upon in the "*Search*" page (see also below), sorted by exon. Automated summary listings are also provided at the end of the "*Home*" page, grouped according to exon, and allelic variant category. Frequencies of PiM, PiS and PiZ, representing the 3 most frequent AAT alleles are provided in the "*Frequencies*" page for 58 populations/ethnic groups and the VarisVis visualization tool [14] is available in the "*Graphical Display*" page, allowing for the graphical display of the A_1ATVar contents in the context of *SERPINA1* gene sequence. Links to patients' organization and societies as well as contact information are provided in the corresponding pages. Finally, the "*Data submission*" page includes an online data submission tool that allows registered users to submit novel AAT variant alleles and experimental protocols directly.

Finally, A_1ATVar fulfills all the required quality criteria, by including copyright and disclaimer notices and the date when the resource was last updated, while all mutation entries comply with the official Human Genome Variation Society (HGVS; http://www.hgvs.org) nomenclature [15].

IV. Database implementation and access

A_1ATVar is a relational database, implemented by PHP and MySQL (MySQL AB, http://www.mysql.com, Uppsala, Sweden) open source software. This choice was guided partly by the desire to facilitate easy creation and maintenance and, most importantly, to support efficient interfaces for AAT variant allele frequency data, experimental protocol archiving and sequence analysis output with other databases, such as FINDbase (http://www.findbase.org) [16]. Data are stored in a single table format (Fig. 1). Database design follows all content criteria and HGVS recommendations [17]. Finally, A_1ATVar design follows certain database guidelines in order to conform to quality.

Information for each AAT variant is summarized in three different sections in the query outcome table that are depicted in different colors (Fig. 2). The section pertaining to the structure of the AAT variant alleles (depicted in blue) include the variant's name, the genomic region that is located (e.g. exon, intron), the position and alteration in the amino acid and DNA sequences, according to the official HGVS nomenclature and the corresponding RS number from dbSNP database (http://www.ncbi.nlm.nih.gov/ projects/ SNP), if available.

The section on clinical presentation (depicted in yellow) is a summary of published results. In this section, considerable effort was made to enforce a uniform, controlled vocabulary, *i.e.*, using descriptions like "Inspiratory capacity (IC) accumulation", "IC degradation", "deficient", *etc*. Hence, data entry and queries can be performed primarily via menus and lists to enforce this controlled vocabulary. For each AAT variant, information on the RNA levels, the clinical phenotype, the AAT variant allele category and the organ affected is available, while in the "*Comments*" column, additional comments can be added, if needed. Additional information on laboratory findings, such as AAT serum levels is recorded quantitatively for some published cases.

Techniques used to identify each AAT variant are recorded, categorized as protein- and DNA-based. Where applicable, the restriction enzyme used to identify the presence of absence of an AAT variant is also provided. For some variants, a quantitative scale is made available to describe their mobility in isoelectric focusing (IEF) electrophoresis, shown in the "*Links*" page. In the "Graphical display" column, the system redirects the user to the corresponding exon, where the respective AAT variant is located. Finally, additional fields record reference, hyperlinked to the corresponding URL in the PubMed literature database. This last section is depicted in gray.

A_1ATVar is a freely available online resource that can be accessed on the World Wide Web at the Golden Helix Server at http://www.goldenhelix.org/A1ATVar. Detailed instructions for both using and querying the database are also available from the same site ("*User guide*" page). There has been no claim of ownership of the information stored in this database by anyone involved in this initiative. However, this compilation and representation of it are subject to copyright and usage principles to ensure that A_1ATVar and its contents remains freely available to all interested individuals.

V. Querying the database

For the needs of this study, we have implemented this tool to develop a new locus specific database, namely A_1ATVar, to document *SERPINA1* gene variants leading to α_1-antitrypsin deficiency. A_1ATVar data can be queried through user-generated queries. Data querying not only include the AAT allelic variants but also their frequencies in various ethnic groups and the experimental protocols used to detect them both at the protein and/or the DNA level. A_1ATVar does not yet provide a function to formulate ad hoc queries.

A_1ATVar also allows the user to submit highly specific queries. For example, one can query for all of the AAT alleles identified in exon V. This requires selecting "Exon V" as the exon respectively (Fig. 1). Sixteen AAT variants are retrieved, indicated in the respective drop-down menu. The user can select the columns to be displayed in the query output table by selecting them at the right-hand side of the respective querying options (Fig. 2). The "Exon", "Name", "Protein" and "DNA" nomenclature and restriction enzyme (RE) site are pre-selected, while the graphical display and reference columns are shown by-default. Documentation of AAT variants properties is not as comprehensive as in other LSDBs, *i.e.*, HbVar [2], [3], since

the available information for the vast majority of AAT variant alleles is limited. One can imagine a query similar to this being useful in a clinical setting as a source of data to help reach a diagnosis. For example, one such query would be to collect all the AAT variants displaying IC degradation.

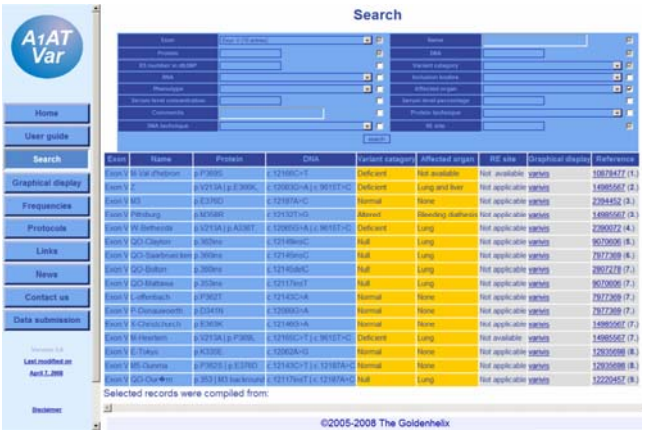

Fig. 2. Querying A_1ATVar database for all AAT variants located in exon V. Query output includes 16 records (also indicated in the drop-down menu) for which the desired information is provided (for RNA levels and phenotype) in addition to the standard columns display (selected by default and shown in gray in the selection dialog box). Mutation nomenclature follows the HGVS recommendations [15]. Based on this example, one can imagine similar queries with clinical relevance only as a means to facilitate diagnosis.

In addition, A_1ATVar can be used to query allele frequencies for the commonest AAT alleles, namely PiM, PiS and PiZ in 58 populations and ethnic groups worldwide. This information is bi-directionally linked with FINDbase database for frequencies of inherited disorders (http://www.findbase.org) [16]. The latter implies that once a FINDbase record on AAT variant allelic frequencies is updated, the corresponding record in A_1ATVar gets automatically updated. The latter is facilitated by the fact that both the AAT-specific information and FINDbase operate under the same software [18], [19], which allows for maintaining updated versions of the linked databases using a very simple server task tool [20].

TABLE I
NUMBER OF RECORDS IN A_1ATVAR

Genomic region	No of entries
Exon II	24
Exon III	9
Exon VI	3
Exon V	16
Exons II, III[a]	1
Exons II, IV[a]	1
Exons III, Vi[a]	2
Exons II-to-IV[b]	2
Intron IC	1
Intron 2	1
Intron 4	1
Not available	30

[a] These entries are AAT variants with two mutations *in cis,* [b] These entries are large deletions that remove part of exons II and IV and the entire intervening region.

A_1ATVar includes a separate archive, based on a flat-file database, to provide a succinct listing of the protocols available for AAT variants detection at the protein level and for *SERPINA1* gene mutation screening. This archive can be accessed from the button "*Protocols*" and currently contains 13 experimental protocols. These protocols are listed as protein- or DNA-based methods, named after keywords from the method itself, e.g. IEF, DGGE, *etc,* and can be retrieved by following the corresponding hyperlinks. These protocols also include hyperlink(s) to the respective citation, describing the method in question. Detailed instructions for both using and querying this archive are also available from the "*User Guide*". In this archive, all 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 archive's core engine, not only for menus and basic screens that display and parse files, but also for handling data querying. Although the structure of this database resembles XPRbase database for human globin genes experimental protocols (http://www.goldenhelix.org/xprbase) [21], a companion database to HbVar, protocol querying is done on a different and much simpler way, *i.e.*, keyword- and not gene-based querying.

Experimental protocols can be added by registered users through the "*Data submission*" page. Modifications are only possible by the administrator through a dedicated administrator module. When a new protocol is added to this file, a separate file named for the first author of the protocol is automatically created. 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. These examples show only a fraction of the querying possibilities of the A_1ATVar database.

The A_1ATVar database and associated resources are currently in use worldwide. In December 2007, A_1ATVar was accessed 624 times from 238 unique IP addresses, despite the fact that is only now being announced.

VI. GRAPHICAL DISPLAY

From the 683 LSDBs available to date, "few" possess graphical displays, especially of a dynamic nature [1]. We have therefore selected VariVis, a system designed to provide a basic set of sequence variation visualisation tools specifically for LSDBs. VariVis is designed to work in parallel with a database's existing user interface and storage and retrieval back-end. *SERPINA1* gene sequence is automatically retrieved from GenBank (NC_000014) and displayed in white (exons) and blue (introns) colours.

In A_1ATVar, VariVis can be accessed from the "Graphical display" button. Two different conceptual views based on the sequence and variation data provided. The first (Standard view) displays the gene sequence and overlays positions where variation is present (Fig. 3). Clicking the variant symbols provides the user with a brief overview of the data extracted from the database, from which the user

can link to the variant entry in the database, or perform simple PubMed and Google Scholar searches to find published articles on the variant. The second (Gel View) has the same functionality as the first viewing option, but with the sequence orientated vertically, with all 4 possible nucleotide possibilities for each nucleotide position, including deletions and insertions between positions, highlighting the nucleotides present in the reference sequence and any variations in contrasting colours. In both views, the software also displays any structural annotations, such as promoter sequences, poly-adenylation sites, and untranslated regions that are present in the sequence file being used, while the exon sequences are underlined.

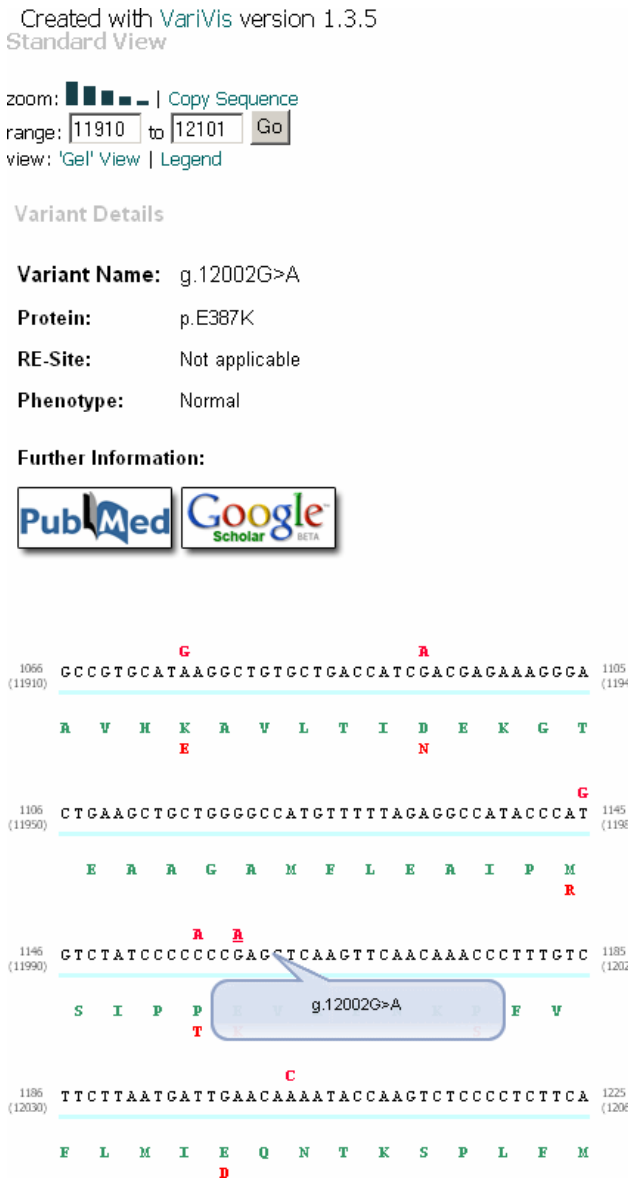

Fig. 3. VariVis visualization tool for variants located in exon V using the standard view.

The software also provides access to the raw sequence data, allowing users to copy or download the entire sequence, or specific chunks, negating the need to navigate to a dedicated sequence database [14].

VII. Data Submission

In A_1ATVar, data entry and modification is only possible for registered users. The data submission page can be accesses through the "*Data Submission*" button. Once logged in (Fig. 4A), the registered user connects with Publication data editor, for further guidance through the data entry procedure (Fig. 4B). Each entry contains empty fields in a Table format, where the registered user enters the requested information, according to the regular table output. Experimental evidence to support the data entered can be uploaded separately with the main submission. Once the data are reviewed by the database administrators and judged appropriate, they become part of the main data collection.

Furthermore, experimental protocols for AAT variant identification or *SERPINA1* gene mutation screening can be also entered through the same page, by selecting the corresponding data entry field. Data submitters may also be contacted for clarifications regarding their submission, such as provision of experimental evidence supporting the data submitted, particularly in case of unpublished mutations. Users that have submitted data in the database can be contacted through the details that they have provided when requesting for an account.

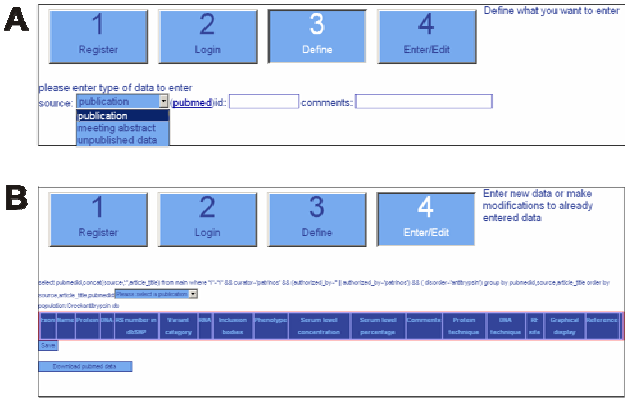

Fig. 4. A_1ATVar variant submission module, including data definition (A) and data entry (B) steps.

VIII. Conclusion

Here, we described a new universal, flexible and freely available DBMS for LSDB development and curation. We also demonstrated that this software has been successfully implemented for the development of a new LSDB for *SERPINA1* gene variants leading to AAT deficiency. This DBMS provides an all-in-one solution for database development and curation and can be easily implemented for the development of similar projects.

REFERENCES

[1] M. Claustres, O. Horaitis, M. Vanevski, R. G. Cotton, "Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases," *Genome Res.*, vol. 12, no. 5, pp. 680–688, May 2002.

[2] R. C. Hardison, D. H. Chui, B. Giardine, C. Riemer, G. P. Patrinos, N. Anagnou, W. Miller, H. Wajcman, "HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server," *Hum. Mutat.*, vol. 19, no. 3, pp. 225–233, Mar. 2002.

[3] G. P. Patrinos, B. Giardine, C. Riemer, W. Miller, D. H. Chui, N. P. Anagnou, H. Wajcman, R. C. Hardison, "Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies," *Nucleic Acids Res., vol.* 32, no. Database issue, pp. D537–D541, Jan. 2004.

[4] R. G. Cotton, K. Phillips, O. Horaitis, "A survey of locus-specific database curation. Human Genome Variation Society," *J. Med. Genet.,* vol. 44, no. 4, pp. e72, Apr. 2007.

[5] C. Beroud, G. Collod-Beroud, C. Boileau, T. Soussi, C. Junien, "UMD (Universal mutation database): a generic software to build and analyze locus-specific databases," *Hum. Mutat.,* vol. 15, no. 1, pp. 86–94, Jan. 2000.

[6] I. F. Fokkema, J. T. den Dunnen, P. E. Taschner, "LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach,". *Hum. Mutat.,* vol. 26, no. 2, pp. 63-68, Aug. 2005.

[7] M. Luisetti, N. Seersholm, "Alpha1-antitrypsin deficiency. 1: epidemiology of alpha1-antitrypsin deficiency," *Thorax,* vol. 59, no. 2, pp. 164-169, Feb. 2004.

[8] J. K. Stoller, L. S. Aboussouan, "Alpha1-antitrypsin deficiency," *Lancet,* vol. 365, no. 9478, pp. 2225-2236, Jun. 2005.

[9] D. L. DeMeo, E. K. Silverman, "Alpha1-antitrypsin deficiency. 2: genetic aspects of alpha(1)-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk," *Thorax,* vol. 59, no. 3, pp. 259-264, Mar. 2004.

[10] D. A. Lomas, R. Mahadeva, "Alpha-1-antitrypsin polymerization and the serpinopathies: pathobiology and prospects for therapy," *J. Clin. Invest.,* vol.110, no. 11, pp. 1585-1590, Dec. 2002.

[11] J. H. Teckman, "Alpha1-antitrypsin deficiency in childhood," *Semin. Liver Dis.*, vol. 27, no. 3, pp. 274-281, Aug. 2007.

[12] G. L. Long, T. Chandra, S. L. Woo, E. W. Davie, K. Kurachi, "Complete sequence of the cDNA for human alpha 1-antitrypsin and the gene for the S variant," *Biochemistry,* vol. 23, no. 21, pp. 4828-4837, Sep. 1984.

[13] G. P. Patrinos, A. J. Brookes, "DNA, diseases and databases: Disastrously deficient," *Trends Genet.,* vol. 21, no. 6, pp. 333-338, Jun. 2005.

[14] T. D. Smith, R. G. Cotton, "VariVis: A Visualisation toolkit for variation databases," *BMC Bioinformatics,* vol. 9, pp. 206, Apr. 2008.

[15] S. E. Antonarakis, and the Nomenclature Working Group, "Recommendations for a nomenclature system for human gene mutations," *Hum. Mutat.,* vol. 11, no. 1, pp. 1-3, Jan. 1998.

[16] S. van Baal, P. Kaimakis, M. Phommarinh, D. Koumbi, H. Cuppens, F. Riccardino, M. Macek Jr, C. R. Scriver, G. P. Patrinos, "FINDbase: A relational database recording frequencies of genetic defects leading to inherited disorders worldwide," *Nucleic Acids Res.,* vol. 35, no. Database issue, pp. D690-D695, Jan. 2007.

[17] R. G. Cotton, A. D. Auerbach, J. S. Beckmann, O. O. Blumenfeld, A. J. Brookes, A. F. Brown, P. Carrera, D. W. Cox, B. Gottlieb, M. S. Greenblatt, P. Hilbert, H. Lehvaslaiho, P. Liang, S. Marsh, D. W. Nebert, S. Povey, S. Rossetti, C. R. Scriver, M. Summar, D. R. Tolan, I. C. Verma, M. Vihinen, J. T. den Dunnen, "Recommendations for locus-specific databases and their curation," *Hum. Mutat.,* vol. 29, no. 1, pp. 2-5, Jan. 2008.

[18] G. P. Patrinos, "National and ethnic mutation databases: recording populations' genography". *Hum. Mutat.,* vol. 27, no. 9, pp. 879-887, Sep. 2006.

[19] G. P. Patrinos, S. van Baal, M. B. Petersen, M. N. Papadakis, "Hellenic National Mutation database: a prototype database for mutations leading to inherited disorders in the Hellenic population," *Hum. Mutat.,* vol. 25, no. 4, pp. 327-333, Apr. 2005.

[20] S. van Baal, G. P. Patrinos, "ETHNOS: A versatile database management system for population-based data collection and national/ethnic mutation database development," *Hum. Genomics Proteomics*, to be submitted.

[21] B. Giardine, S. van Baal, P. Kollia, P. Kaimakis, C. Riemer, W. Miller, M. Samara, D. H. Chui, H. Wajcman, R. C. Hardison, G. P. Patrinos, "HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update," *Hum. Mutat.,* vol. 28, no. 2, pp. 206, Feb 2007.