D926–D932 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 27 November 2010*
doi:10.1093/nar/gkq1236

FINDbase: a worldwide database for genetic variation allele frequencies updated

Marianthi Georgitsi[1], Emmanouil Viennas[2], Dimitris I. Antoniou[2], Vassiliki Gkantouna[2], Sjozef van Baal[3], Emanuel F. Petricoin III[4], Konstantinos Poulas[1], Giannis Tzimas[5] and George P. Patrinos[1,*]

[1]Department of Pharmacy, School of Health Sciences, [2]Department of Computer Engineering and Informatics, Faculty of Engineering, University of Patras, Patras, Greece, [3]Erasmus MC, Faculty of Medicine and Health Sciences, MGC-Department of Cell Biology and Genetics, Rotterdam, The Netherlands, [4]Center for Applied Proteomics and Molecular Medicine, George Mason University Manassas, VA, USA and [5]Department of Applied Informatics in Management & Finance, Faculty of Management and Economics, Technological Educational Institute of Messolonghi, Messolonghi, Greece

Received September 15, 2010; Revised November 7, 2010; Accepted November 13, 2010

ABSTRACT

Frequency of INherited Disorders database (FINDbase; http://www.findbase.org) records frequencies of causative genetic variations worldwide. Database records include the population and ethnic group or geographical region, the disorder name and the related gene, accompanied by links to any related external resources and the genetic variation together with its frequency in that population. In addition to the regular data content updates, we report the following significant advances: (i) the systematic collection and thorough documentation of population/ethnic group-specific pharmacogenomic markers allele frequencies for 144 markers in 14 genes of pharmacogenomic interest from different classes of drug-metabolizing enzymes and transporters, representing 150 populations and ethnic groups worldwide; (ii) the development of new data querying and visualization tools in the expanded FINDbase data collection, built around Microsoft's PivotViewer software (http://www.getpivot.com), based on Microsoft Silverlight technology (http://www.silverlight.net) that facilitates querying of large data sets and visualizing the results; and (iii) the establishment of the first database journal, by affiliating FINDbase with Human Genomics and Proteomics, a new open-access scientific journal, which would serve as a prime example of a non-profit model for sustainable database funding.

INTRODUCTION

National and Ethnic Mutation Databases (NEMDBs) are structured data repositories recording the various spectra of causative genetic variations for any gene or disease in different populations and ethnic groups worldwide, also between distinct ethnic groups within a geographical region (1). NEMDBs provide data that can be used to, for example, stratify national molecular diagnostic services, study human demographic history, gene/mutation flow and admixture patterns (1). Together with the central (or core) databases, such as the Online Mendelian Inheritance in Man [OMIM, http://www3.ncbi.nlm.nih.gov/omim; (2)] or the Human Gene Mutation Database [HGMD, http://www.hgmd.org; (3)] and the locus-specific databases [LSDBs; (4)], they are the main components that fall under the banner of 'genetic databases'.

We have previously described the development of FINDbase (Frequency of INherited Disorders database; http://www.findbase.org), a relational database pertaining to frequencies of causative mutations, leading to inherited disorders in various populations and ethnic groups worldwide (5). FINDbase contains only summary-level data, i.e. allele frequencies without any sensitive personal data of their carriers, in order to preserve anonymity. Content-wise, FINDbase is the richest among the NEMDB currently available and, since its establishment in August 2006, it has been broadly adopted by the scientific community, as one of the key resources to retrieve population-specific information for disease-causing mutations, as indicated by the traffic and the number of visitors.

*To whom correspondence should be addressed. Tel:/Fax: +30 2610 969 (Ext. 834); Email: gpatrinos@upatras.gr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Here, we present significant data content updates that make FINDbase more compelling to a broader user group. We also present some technological advances that facilitate data querying and visualization. Finally, we report the affiliation of FINDbase with an open-access scientific journal, inaugurating the first database journal currently available.

DATA CONTENT UPDATE

Initial FINDbase data content was acquired from the European Union FP5 Cystic Fibrosis Thematic Network Consortium data [http://www.cfnetwork.be; (6)], and population-specific mutation frequency data derived from the PAHdb phenylalanine hydroxylase locus knowledgebase [http://www.pahdb.mcgill.ca; (7)] and the HbVar database of human hemoglobin variants and thalassemia mutations [http://globin.bx.psu.edu/hbvar; (8)]. These data were curated to identify possible erroneous entries and corrected where necessary.

Apart from data curation, we also continued to enrich FINDbase data collection. To this end, we deposited in FINDbase population-specific information from the A_1ATVar database, documenting *SEPRINA1* gene variants (9), for the commonest variants leading to α_1-antitrypsin deficiency. Also, we collected information from individual NEMDBs currently available and incorporated it into the main FINDbase data collection. This was the first step toward unifying NEMDB contents into a main population-specific data repository and implementing data warehousing (currently in progress). Calculation of mutation frequencies is based on the most representative study available for each population, involving sufficient number of patients and controls, since estimation of mutation frequencies based on multiple reports has the inherent danger of including redundant cases that can alter the calculated frequencies.

All data entries were recorded against their corresponding data source. When the allele frequency data were obtained from published reports, then the corresponding unique PubMedID was recorded. If however, the data source was a LSDB, then the URL of the corresponding resource was used and/or a unique identifier, namely the ResearcherID (available from Thomson ISI; http://www.researcherid.com) of the data curator. Based on our previous experience, the latter ID not only facilitates unambiguous identification of curated data when data update or correction is needed, but also provides incentives to potential contributors of unpublished data to share their data with the broader scientific community.

FINDbase data compilation and representation are subject to copyright and usage principles to ensure that FINDbase and its contents remain freely available to all interested individuals.

NEW FEATURES

Apart from the extensive data content curation, correction and enrichment described above, several new features have been incorporated into FINDbase. These include the development of new data querying and visualization tools, a thorough documentation of population/ethnic group-specific pharmacogenomic markers allele frequencies and the establishment of the first database journal, by affiliating FINDbase with a new open-access scientific journal. These new features are described below.

Documentation of population/ethnic group-specific pharmacogenomic markers allele frequencies

Pharmacogenomics correlates an individual's genetic variation on drug response, by linking his/her gene expression profile or single nucleotide polymorphisms (SNPs) with a drug's toxicity or efficacy, hence aiming to optimize drug therapy by maximizing efficiency and/or minimizing the chances for adverse drug reactions (ADRs) (10). A plethora of SNPs have been reported to correlate with several drugs' safety or efficacy (also referred to as pharmacogenomic markers) and the incidence of these SNPs varies among different populations or ethnic groups, as expected. However, population- and ethnic group-specific allele frequencies of pharmacogenomic markers are poorly documented and not systematically collected in structured data repositories. We, therefore, developed a separate module of FINDbase, pertaining solely to the documentation of population/ethnic group-specific allele frequencies of pharmacogenomic markers in 14 genes, representing different classes of drug-metabolizing enzymes and transporters. In addition to the previous data collection, this new FINDbase module provides information on 144 pharmacogenomic markers, representing 150 populations and ethnic groups worldwide, which have been collected from the curation of 214 scientific articles, retrieved during a large-scale targeted scientific literature search (11). This effort represents the largest, so far, data collection of population/ethnic group-specific pharmacogenomic markers allelic frequencies, aiming to cover the need that is currently not fulfilled by other pharmacogenomics knowledgebases and related resources (12). Hence, this initiative would assist in the future design and development of pharmacogenomic testing, toward the advent of personalized medicine.

Development of new data querying and visualization tool

We have decided to re-develop FINDbase structure and architecture, driven not only by the recent advances in the field and the new data contents, but also from the need to facilitate data querying in the expanded FINDbase data collection. We have, therefore, chosen to build the existing (causative mutations) and new (pharmacogenomic markers) FINDbase data sets as separate modules, with the possibility of integrating them in a single module soon. The component services that comprise both modules follow the SOA (13,14) and the querying interface is built around the PivotViewer (http://www.getpivot.com), based on Microsoft Silverlight technology (http://www.silverlight.net), a recently launched development platform by Microsoft,

which offers powerful tools for dynamically querying and visualizing large data sets. The whole application provides an elegant, web-based multimedia interface for population-based genetic variation data collection and retrieval. Database records include the population, the ethnic group and/or the geographic region, the gene name and its variation parameters, the rare allele frequencies, accompanied by links to the respective OMIM and the HGMD (causative mutations module) or PharmGKB entries (pharmacogenomic markers module). The entire database schema is depicted in the Supplementary Figure S1.

The whole system architecture is based on a three-tier client-server model (15), namely the client application, the application server and the database server. The n-tier architecture is a robust model and flexible enough to aggregate multiple information sources and integrate modular developments. All the database records presented above are a collection that combines large groups of similar items. There is a set of files on a server, and a local client that has the capability to display them. The files are traditionally HTML and image files. In the collection case, the files are CXML and Deep Zoom-formatted (DZC) images (http://msdn.microsoft.com/en-us/library/cc645077%28VS.95%29.aspx).When the user browses the collection from a web page, the PivotViewer uses the Silverlight Control to display the files.

The entire FINDbase causative mutations data collection via PivotViewer is shown in Figure 1A, which enables the user to interact with large data sets at once. PivotViewer enables users to smoothly and quickly arrange FINDbase data collections according to common characteristics that can be selected from the data query menu (Figure 1B) and then zoom in for a closer look, by either filtering the collection to get a subset of information or clicking on a particular item (Figure 1C). A display item in the form of a card, with a chromosomal figure (derived from http://www.genecards.org) displaying the gene position, is provided for each genetic variation, along with a sidebar textbox with in-depth data concerning the particular genetic variation and population (Figure 1C). Hyperlinks for each gene name to OMIM database and HGMD, offer to the user the possibility of easily accessing additional information.

In particular, the new FINDbase data querying and visualization environment enables the user to visualize and sort, organize and categorize data dynamically and discover trends across all items, using different views. For example, a user can perform a query based on the frequency of the rare allele (Figure 2A). Additional filtering allows the user to see specific genetic variants and to further zoom in a particular population (Figure 2B). Compound queries can also be formulated. Such queries would be the identification of rare pathogenic mutations (frequencies 10–30%) in the Hellenic population, sorted out by gene name. The query output includes 19 alleles for 7 genes (*ATP7B*, *CYP21*, *HBA2*/*HBA1*, *HBB*, *LDLR*, *MEFV*, *PAH*; Figure 2C).

Affiliation to a new scientific journal: toward a novel publication modality

In May 2008, FINDbase officially became the first database affiliated with the open-access journal Human Genomics and Proteomics (HGP), published by SAGE-Hindawi (http://www.sage-hindawi.com/journals/hgp). HGP is a new genomics and systems biology journal that, in addition to publishing original research and review articles, includes short descriptions of genetic data sets pertaining to population/ethnic group-specific allele frequencies, namely causative mutations or biomarkers (pharmacogenomic, forensic markers and so on). These submissions are also stringently peer-reviewed and, if accepted, featured in the journal as a Mutation and Biomarker Dataset with links to the full data set in FINDbase and the PubMed literature database (16).

As the first journal with an affiliated database in this discipline, HGP offers a unique opportunity to authors to open up access to their research on the characterization of causative mutation and/or biomarker frequency spectra to the widest possible community (16). This way, HGP not only provides a forum for researchers of the post-genomic era but also increases the chances of human variation data capture and provides a centralized system for population-specific data storage and retrieval.

It is envisaged that HGP will provide the proof of principle for closely related efforts toward developing other 'database journals'. Also, the HGP–FINDbase affiliation can serve as a non-profit model for sustainable database funding, in a field that still suffers significantly from the lack of long-term funding opportunities for genetic database projects (17).

CONCLUSIONS AND FUTURE PERSPECTIVES

FINDbase is a comprehensive source of information on the extant genetic heterogeneity of different populations, documenting causative mutation and pharmacogenomic marker allele frequency data, aiming to be further established as a reference repository of such information worldwide. Database access is free of charge and there are no registration requirements for data querying.

FINDbase will continue to be updated and data collection will be further enriched. To this end, our efforts are facilitated by our participation to one of the leading database projects worldwide, namely the GEN2PHEN project (http://www.gen2phen.org) funded by the European Commission, ensuring not only interaction with the field's leading experts, but also funding that is secured for a further 3-year period. Also, FINDbase is a key component of the Human Variome Project [http://www.humanvariomeproject.org, (18)], an international initiative aiming to document genetic variation worldwide. In particular, our group has taken the lead in establishing recommendations and guidelines to develop nation-wide projects to document the genetic heterogeneity in developing countries (19), an effort that will be further facilitated by the provision of the *ETHNOS* software (20), as an aid to develop NEMDBs in these populations. In addition, as of early 2007, FINDbase is affiliated with HGMD

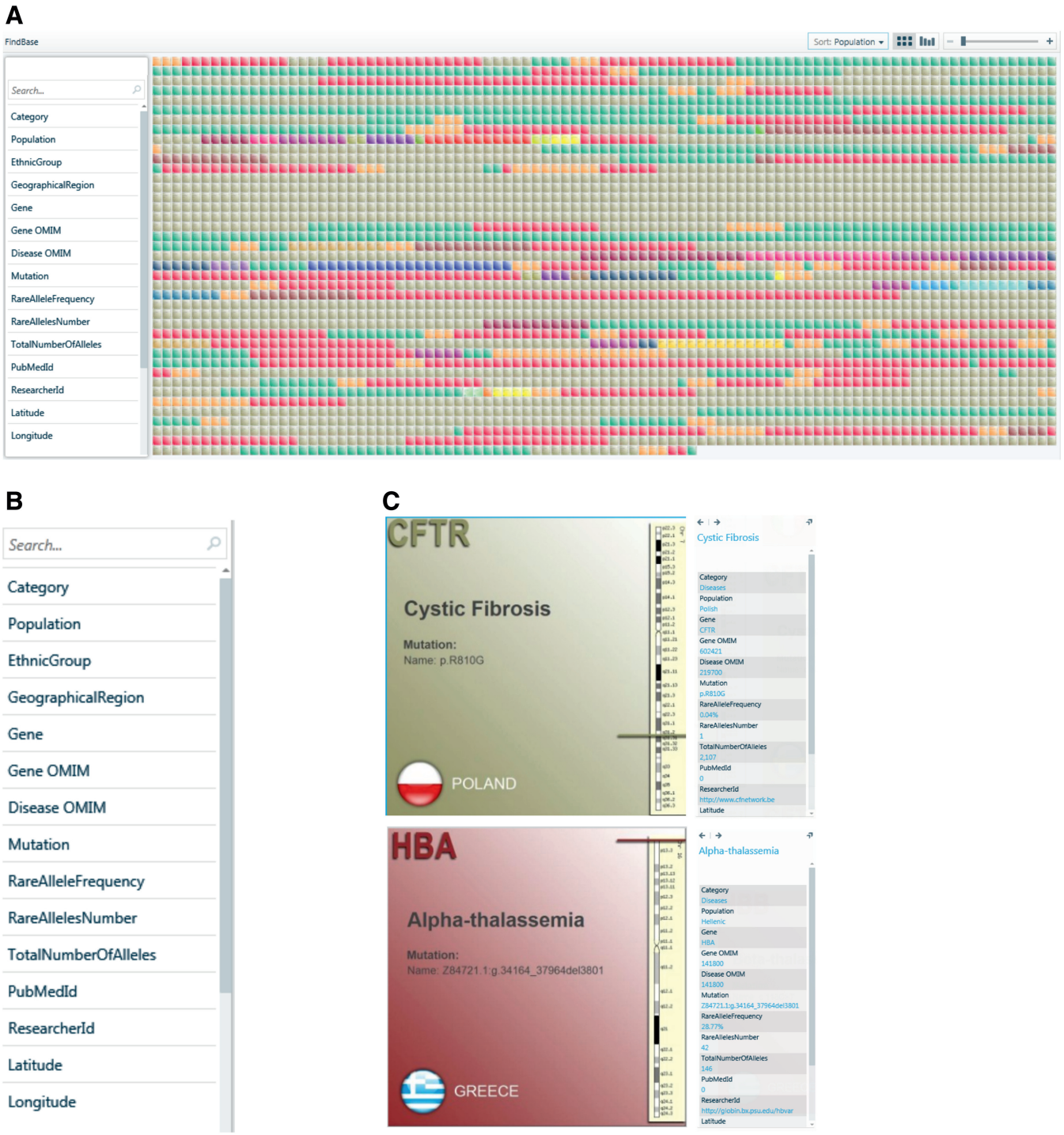

Figure 1. The updated FINDbase querying and visualization tool. (**A**) Overview of the entire FINDbase causative mutation module data collection, based on Microsoft's PivotViewer. The querying interface is shown on the left and the output option can be selected at the top-right corner of the screen. The different entries are shown as colored boxes, presented as display items (see below). The user can zoom in for a closer look or click on a particular item to get more in-depth information. (**B**) Detail of the FINDbase querying interface. (**C**) Display items provided for each causative mutation, accompanied by a sidebar textbox with in-depth data concerning the particular mutation and population. Each item includes the name of the allele in its official Human Genome Variation Society (HGVS) or other nomenclature systems, if available, the population for which this information is available (shown by the country's flag) and a chromosomal map, where the gene's position is indicated. Hyperlinks for each gene to OMIM database offer to the user the possibility of easily accessing additional information. Finally, each item displays the corresponding PubMed and Researcher IDs, if applicable.

Figure 2. FINDbase sample queries. (**A**) FINDbase causative mutation module data content sorted using the columns option (see top-right corner of the screen), according to rare allele frequency (0–100%). The user can select a specific column to retrieve more specific information. (**B**) Display of the mutant alleles in the Israeli population sorted out by gene name (indicated with an arrow). The query returns 90 alleles in equal number of display items. (**C**) Query formulation to retrieve the rare alleles in the Hellenic population with an allelic frequency between 10.08% and 30.05% (indicated with an arrow). The query returns 19 alleles in equal number of display items, sorted out in columns arranged by Gene name (arrow). The total number of alleles is also shown in the query box.

(http://www.hgmd.org) with the inclusion of bi-directional links from FINDbase records to HGMD and vice versa. Last, but not least, the affiliation with HGP is also expected to increase data influx into FINDbase, by providing incentives to potential contributors and researchers to submit their data to FINDbase, while at the same time prevents population-based allele frequency data from being lost or kept unpublished. As a result, since December 2006 when FINDbase was officially announced (5), we have recorded over 70 000 accesses to the query page from unique IP addresses spanning over 100 countries worldwide (including .com, .org, .net and .gov URLs), while many users frequently contact the administrator to report missing information for existing variants and pinpoint inconsistencies and/or erroneous entries. This is particularly important, since the user input improves data quality and accuracy.

Regarding software upgrade, we will soon dispatch a new freely available version of the *ETHNOS* software, as an off-the-shelf solution to establish new NEMDBs. We anticipate that the existing NEMDBs available at the Golden Helix Server [http://www.goldenhelix.org; (21)] as well as those that will be derived from FINDbase will migrate to the new version of the *ETHNOS* software in mid-2011. This new software version will also include some new features, namely: (i) a content management system based on the Model-View-Controller (MVC) design pattern and alternative data visualization interfaces and online data handling functionalities, in order to enhance the database's performance and user interaction features; (ii) data exposure through web services based on the oData protocol (http://www.odata.org); and (iii) expanded viewing orientation of the application's interface, providing a spatiotemporal data viewing dimension.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Elena Christodoulopoulou, Zoi Zagoriti, Christina Tafrali, Fotios Ntellos, Olga Giannakopoulou, Athanassia Boulakou, Panagiota Vlahopoulou, Eva Kyriacou and Ioanna Kohliadi for competent data mining and curation. Also, we are indebted to all FINDbase users worldwide for their valuable comments and suggestions, which helped us to keep the information as updated and complete as possible and also contributed to the continuous improvement of the database profile and contents.

FUNDING

European Commission [grants MEDGENET (FP6-31968) and GEN2PHEN (FP7-200754) to G.P.P.]; Golden Helix Institute of Biomedical Research. Funding for open access charge: FP7–GEN2PHEN Integrated Project.

Conflict of interest statement. None declared.

REFERENCES

1. Patrinos,G.P. (2006) National and Ethnic Mutation databases: documenting populations' genography. *Hum. Mutat*., **27**, 879–887.
2. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
3. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.
4. Mitropoulou,C., Webb,A.J., Mitropoulos,K., Brookes,A.J. and Patrinos,G.P. (2010) Locus-specific database domain and data content analysis: evolution and content maturation towards clinical use. *Hum. Mutat*., **31**, 1109–1116.
5. van Baal,S., Kaimakis,P., Phommarinh,M., Koumbi,D., Cuppens,H., Riccardino,F., Macek,M. Jr, Scriver,C.R. and Patrinos,G.P. (2007) FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. *Nucleic Acids Res*., **35**, D690–D695.
6. Bobadilla,J.L., Macek,M. Jr, Fine,J.P. and Farrell,P.M. (2002) Cystic fibrosis: a worldwide analysis of CFTR mutations–correlation with incidence data and application to screening. *Hum. Mutat*., **19**, 575–606.
7. Scriver,C.R., Waters,P.J., Sarkissian,C., Ryan,S., Prevost,L., Cote,D., Novak,J., Teebi,S. and Nowacki,P.M. (2000) PAHdb: a locus-specific knowledgebase. *Hum. Mutat*., **15**, 99–104.
8. Patrinos,G.P., Giardine,B., Riemer,C., Miller,W., Chui,D.H., Anagnou,N.P., Wajcman,H. and Hardison,R.C. (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.
9. Zaimidou,S., van Baal,S., Smith,T.D., Mitropoulos,K., Ljujic,M., Radojkovic,D., Cotton,R.G. and Patrinos,G.P. (2009) A1ATVar: a relational database of human SERPINA1 gene variants leading to alpha1-antitrypsin deficiency. *Hum. Mutat*., **30**, 308–313.
10. Squassina,A., Manchia,M., Manolopoulos,V.G., Artac,M., Lappa-Manakou,C., Karkabouna,S., Mitropoulos,K., Del Zompo,M. and Patrinos,G.P. (2010) The realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics*, **11**, 1149–1167.
11. Georgitsi,M., Viennas,E., Gkantouna,V., Christodoulopoulou,E., Zagoriti,Z., Tafrali,C., Ntellos,F., Giannakopoulou,O., Boulakou,A., Vlahopoulou,P. *et al*. (2011) Population-specific documentation of pharmacogenomic markers allelic frequencies in the pharmacogenomics module of the Frequency of Inherited Disorders database. *Pharmacogenomics* (in press).
12. Lagoumintzis,G., Poulas,K. and Patrinos,G.P. (2010) Genetic database and their potential in pharmacogenomics. *Curr. Pharm. Des*., **16**, 2224–2231.
13. Bell,M. (2010) *SOA Modeling patterns for service-oriented discovery and analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.
14. Valipour,M.H., AmirZafari,B., Maleki,K.N. and Daneshpour,N. (2009) A brief survey of software architecture concepts and service oriented architecture, *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT'* 09, 34–38.
15. Eckerson,W.W. (1995) Three tier client/server architecture: achieving scalability, performance, and efficiency in client server applications. *Open Inform. Syst*., **10**, 1–20.
16. Patrinos,G.P. and Petricoin,E.F. (2009) A new scientific journal linked to a genetic database: towards a novel publication modality. *Hum. Genomics Proteomics*, **1**, e597478.
17. Patrinos,G.P. and Brookes,A.J. (2005) DNA, diseases and databases: disastrously deficient. *Trends Genet*., **21**, 333–338.
18. Kaput,J., Cotton,R.G., Hardman,L., Watson,M., Al Aqeel,A.I., Al-Aama,J.Y., Al-Mulla,F., Alonso,S., Aretz,S., Auerbach,A.D. *et al*. (2009) Planning the human variome project. The Spain report. *Hum. Mutat*., **30**, 496–510.

19. Patrinos,G.P., Al Aama,J., Al Aqeel,A., Al-Mulla,F., Borg,J., Devereux,A., Felice,A.E., Macrae,F., Marafie,M.J., Petersen,M.B. *et al*. (2011) Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection. *Hum. Mutat*. (in press).
20. van Baal,S., Zlotogora,J., Lagoumintzis,G., Gkantouna,V., Tzimas,I., Poulas,K., Tsakalidis,A., Romeo,G. and Patrinos,G.P. (2010) ETHNOS: a versatile electronic tool for the development and curation of National Genetic databases. *Hum. Genomics*, **4**, 361–368.
21. Patrinos,G.P., van Baal,S., Petersen,M.B. and Papadakis,M.N. (2005) The Hellenic National Mutation database: a prototype database for inherited disorders in the Hellenic population. *Hum. Mutat*., **25**, 327–333.