ETHNOS: A versatile electronic tool for the development and curation of national genetic databases

*Sjozef van Baal,[1] Joël Zlotogora,[2] George Lagoumintzis,[3,4] Vassiliki Gkantouna,[5] Ioannis Tzimas,[5] Konstantinos Poulas,[3] Athanassios Tsakalidis,[5] Giovanni Romeo[6] and George P. Patrinos[3]**

[1]Erasmus MC, Faculty of Medicine and Health Sciences, MGC-Department of Cell Biology and Genetics, Rotterdam 3015CE, the Netherlands
[2]Department of Community Genetics, Public Health Services, Ministry of Health, Ramat Gan, 52621, Israel
[3]Department of Pharmacy, School of Health Sciences, University of Patras, Patras, 26504, Greece
[4]Department of Optics and Optometry, Technological Educational Institute of Patras, Patras, 26334, Greece
[5]University of Patras, Faculty of Engineering, Department of Computer Engineering and Informatics, Patras, 26504, Greece
[6]Unità di Genetica Medica, Policlinico Universitario S. Orsola-Malpighi, Bologna, 40138, Italy
**Correspondence to*: Tel/Fax: +30 2610 969834; E-mail: gpatrinos@upatras.gr

Date received (in revised form): 22 April 2010

Abstract

National and ethnic mutation databases (NEMDBs) are emerging online repositories, recording extensive information about the described genetic heterogeneity of an ethnic group or population. These resources facilitate the provision of genetic services and provide a comprehensive list of genomic variations among different populations. As such, they enhance awareness of the various genetic disorders. Here, we describe the features of the *ETHNOS* software, a simple but versatile tool based on a flat-file database that is specifically designed for the development and curation of NEMDBs. *ETHNOS* is a freely available software which runs more than half of the NEMDBs currently available. Given the emerging need for NEMDB in genetic testing services and the fact that *ETHNOS* is the only off-the-shelf software available for NEMDB development and curation, its adoption in subsequent NEMDB development would contribute towards data content uniformity, unlike the diverse contents and quality of the available gene (locus)-specific databases. Finally, we allude to the potential applications of NEMDBs, not only as worldwide central allele frequency repositories, but also, and most importantly, as data warehouses of individual-level genomic data, hence allowing for a comprehensive ethnicity-specific documentation of genomic variation.

Keywords: *genetic disorders, database, software, mutations, laboratories*

Introduction

In recent years, advances in our understanding of genotype–phenotype correlations and the evolution of genomics technology and nanotechnology have resulted in the generation of enormous amounts of genetic data. These data are usually stored in genetic databases, namely data repositories for genome variation data and their phenotypic consequences.

Genetic databases can be categorised into three types: (1) General (core) mutation databases (GMDBs); (2) locus-specific databases (LSDBs) and (3) national and ethnic mutation databases (NEMDBs).[1] GMDBs attempt to capture all

described mutations in all genes, but with each being represented in only limited detail. By contrast, LSDBs are concerned with just one or a few specific genes, usually related to a single disease entity.[2,3] NEMDBs are repositories documenting the genetic composition of an ethnic group and/or population, and the genetic defects leading to various inherited disorders and their frequencies, calculated on a population-specific basis.[4] These resources have recently emerged, mostly driven by the need to document the varying mutation spectrum observed for any gene associated with a genetic disorder, among different population and ethnic or religious groups. In general, the NEMDBs available to date can be separated into two subcategories: (1) national genetic databases, which record the existing genetic composition of a population or ethnic group but with limited or no description of mutation frequencies, and (2) national mutation frequency databases, which provide comprehensive information only on inherited disorders whose disease-causing mutation spectrum is well defined.[4]

A detailed domain analysis of the various NEMDBs presently available has been previously performed.[4] Here, we provide an overview of the different functionalities of the ETHnic and National database Operating Software (*ETHNOS*) v2.0 software and comparatively describe the various national genetic databases that have resulted from the implementation of this software — namely, the Israeli, Cypriot, Tunisian, Lebanese and Egyptian NEMDBs. These databases were some of the tangible deliverables of the European Commission-funded Euro-Mediterranean Network for Genetic Services (MEDGENET) project. Finally, we demonstrate possible outcomes from the evolution of the NEMDB field towards a truly nationwide clinical genetics database, not only for research but also for everyday clinical use.

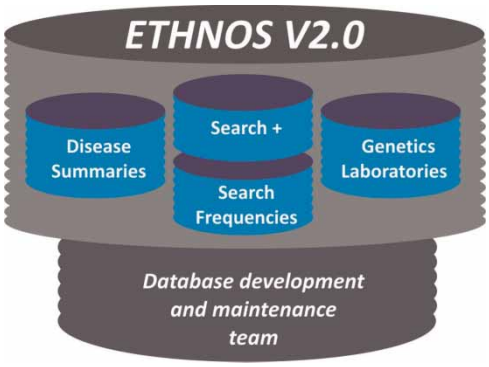

Figure 1. Outline of the Israeli national and ethnic mutation database (NEMDB). The database has three main components: disease summaries (categorised by religious groups); allele frequency search options, available separately in a public or restricted password-protected (Search +) environment; and genetic laboratories. The database is under the scientific control of a group of curators under the supervision of the Israeli NEMDB National Coordinator, while a dedicated database development and maintenance team oversees the smooth operation of the Israeli NEMDB.

Materials and methods

The *ETHNOS* v2.0 software is built around two different flat-file database techniques: single flat-file and indexed multiple flat-file databases, both of which are tab delimited — ie, fields separated with a tab — with one record per line and maintained using a custom content management system (CMS). The websites run on PHP (version 4 or higher) and are hosted by the Golden Helix Server (http://www.goldenhelix.org), which operates under Apache 2, PHP 5 with the virtual hosts feature enabled.

Database techniques

Each NEMDB that derives from this software has its own data folder in the Golden Helix Server and consists of three independent functionalities (see Figure 1):

(a) The disease summaries. This option employs the indexed multiple flat-file database technique. The records can span multiple lines and the text can be any plain text or valid HTML code. A collection of HTML files is maintained through a straightforward index file (Figure 2A). The flat-file database starts with a special line (eg ###|summaries|) that is simply skipped when reading the database, but is necessary for determining whether the correct file has been used. The second line is used to

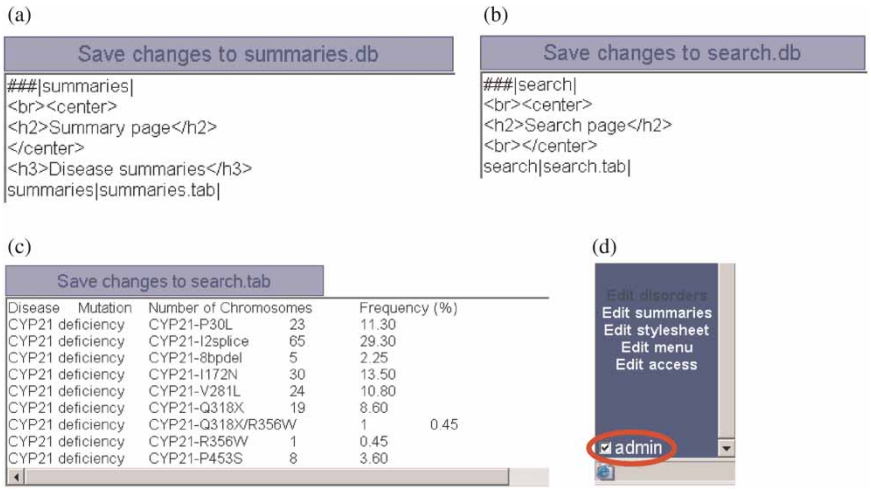

Figure 2. Basic features of the *ETHNOS* software for national and ethnic mutation database (NEMDB) development and curation. The format and structure of flat files with multiple lines per record is shown, with the 'summaries' (A) and 'search' files (B) used as examples (see text for details). (C) Structure of the flat file with one line per record, used for entering the mutation frequencies. This file consists of the data source file used in the example of Figure 1B. (D) Part of the database administration menu. The administrator is recognised by the PHP script on the basis of his/her IP address. By clicking on the 'admin' box (in the red oval), an additional screen appears at the top of each page. Again, the selected option is highlighted (disorders page for the example shown in Figure 1C), and the administrator can easily modify its contents using simple html language (see text for details).

identify the fields, and is displayed as the header of the list of disease summaries. From the third line onwards, each line represents a HTML file, the filename of which is built using the first two sections separated by the pipe '|' character. Both the index file and the individual HTML files can be edited online. The index file is edited using the basic custom-made CMS editor (see below) and the HTML files are edited using a third party (wysiwyg) editor.

(b) The allele frequency search option can be conducted in a non-restricted or secure password-protected environment. This option employs the single flat-file database technique. The text file is tab delimited and includes information on, for example, population, ethnic group, gene, Online Mendelian Inheritance in Man (OMIM) identification (ID), mutation, number (No.) of chromosomes/families, allele and carrier frequency (per cent), every time depending on the NEMDB in question (Figure 2B). Allele frequency data derived from one single study (the most representative), and not multiple studies, is mentioned in the respective disease summary. The fields 'No. of chromosomes/families' and 'allele and carrier frequency (per cent)' allow numerical values only with a period as a decimal point (not a comma) while the 'OMIM ID' field allows a URL address to be entered. The two search options allow filtering for a certain ethnic group and disorder, using drop-down menus and limiting the output to a frequency range. All data are manually entered by the curators (Figure 2C). It is estimated that, with a maximum of 3,000 records, the search options will still work relatively fast.

(c) The genetic laboratories. As with the disease summaries option, the indexed multiple flat-file database technique is also employed here, although in this case the files have a different format.

Features of the CMS

All pages, including the menu, may be edited online. The online editor is password and/or

IP-address restricted. For most pages, a basic, custom-made CMS system is used. The system runs on PHP and makes use of .db files. The key character in the file is the pipe '|' character. The presence of the pipe character defines the string in front of it as an identifier. Page content is primarily stored in a file with the '.db' extension, which is parsed by the display-file.PHP script to generate dhtml that can be read by any browser. As the file is read, the pipe '|' character is recognised and used for two purposes: (1) to identify the file's content and (2) to display special content. Another key character is the '$' character, which displays predefined content in combination with reserved words. For instance, '$last_modified_date' displays the date that the database was last modified.

Access restrictions

Being a public database, access to the *ETHNOS* v2.0 pages is unrestricted; there is no need to subscribe to query the databases. Access to the administration section of the databases is, however, restricted. This is made possible through a combined IP address-restricted and password-protected interface (Figure 2D). A special administrator's menu appears when the database is accessed from a previously specified IP address from which the administrator can access the database contents for modification or update. The current interface does not support simultaneous data entry and/or modification by more than one curator.

Online HTML editor

The built-in, custom-made wysiwygPRO online editor (http://www.wysiwygpro.com) is mainly used for basic files and the navigation bar. This editor has full HTML capabilities but requires HTML literacy. The *ETHNOS* v2.0 software makes use of this editor for full text formatting and straightforward text entry from word processors used for the disease summaries and genetic laboratories files. The online editor also allows tables and hyperlinks to be inserted and images to be included. Editing files with the wysiwygPRO online editor currently excludes the use of special characters like '|' and '$'.

Screen layout

To allow a consistent layout between various screen sizes and browsers, the database screens are resizable, which is another innovation implemented in the database design. This is accomplished using a mixture of CSS and JavaScript techniques. The mechanism used is simply having any and/or all objects on the HTML page at a relative (per cent) size to the so-called containing element. On top of the cascading hierarchy is the <html> element, almost immediately followed by the <body> element. The <html> element is set with a default text size and this is considered '100 per cent' by the CSS. Therefore, by setting the <body> tag dynamically to a percentage of the font size, the fonts on the screen automatically adjust, even when resizing the window after the document is loaded. Furthermore, images are also set to a percentage.

Results

Software implementation

The *ETHNOS* v2.0 software has been selected as the basis for the development of several NEMDBs, which was one of the thematic priorities of the MEDGENET project. We opted to use a rather simplistic implementation, based on flat-file databases, for two reasons. (1) The NEMDB field was at that time in its infancy, so we felt that adopting a simple and user-friendly platform would attract more potential curators with limited computer literacy to establish NEMDBs for their populations. (2) The resulting NEMDBs provide allele frequency information, calculated on the basis of summary- and not individual-level data. Using this software, five NEMDBs were generated. Table 1 outlines the main features of the NEMDBs that have been developed for the MEDGENET project.

The Israeli NEMDB is a freely available online prototype database, which effectively integrates information on the heterogeneity of inherited

Table 1. Features of the national and ethnic mutation database (NEMDBs) generated for the European Commission-funded MEDGENET project.

NEMDB	URL	Number of records				Search (+) option	Year of first release	References
		Disease summaries	Populations/ ethnicities	Search	Genetic laboratories			
Israeli	http://www.goldenhelix.org/israeli	476	29	945	17	Yes	2006	5,6
Cypriot*	http://www.goldenhelix.org/cypriot	28	3	70	–	Yes	2005	7
Tunisian	http://www.goldenhelix.org/tunisian	102	–	165	–	Yes	2008	This paper
Lebanese*	http://www.goldenhelix.org/lebanese	89	2	130	–	No	2007	8
Egyptian	http://www.goldenhelix.org/egyptian	38	–	94	–	No	2008	This paper

*The Cypriot and Lebanese NEMDBs were first generated as national mutation frequency databases and have been subsequently upgraded into their current form.

disorders studied for the different religious groups within the Israeli population. The Israeli population is an amalgamation of several religious groups — namely, Jewish (5.4 million) and non-Jewish, such as Muslim (1.2 million, including the Bedouin), Christian (150,000), Druze (120,000) and other, smaller groups (Table 1). Most of the Arab and Druze populations live in villages or small towns, and consanguinity is a frequent phenomenon.[9,10] The database source was previous compilations for the incidence of genetic disorders in Israel.[5] New data are contributed by database curators or are retrieved from the published literature, while database entries are periodically checked for consistency and corrected where needed.

The Israeli NEMDB is the richest in information among all NEMDBs developed for this project, with 476 disease summaries documented for the religious groups, which have been further classified into 29 ethnic subgroups in the Israeli population. This corresponds to a 28 per cent content increase since the initial publication of the database. This has led to a significant increase in user traffic in the Israeli NEMDB, while the 'Search +' feature is becoming increasingly popular among clinical geneticists in Israel (personal communication). This feature enables the acquisition of information on the genetic diseases that exist in each of the localities where Arabs and Druze live. This functionality ensures that patient privacy and anonymity are preserved, since data access and retrieval are password protected.[6] Data querying can be performed according to locality or disorder, while the query output is a list of prevalent disorders in certain localities, or different religious groups. In addition, 17 genetic laboratories are archived in the database, together with details on the genetic tests that they provide. Notably, according to a new initiative, genetic laboratories are mandated to report the details of the tests they are performing in the Israeli NEMDB in order to obtain accreditation from the Israeli Ministry of Health. Although this

feature is of clear value and has been extensively used by the Israeli genetics community, at present only the Israeli NEMDB has exploited this functionality of the *ETHNOS* sofware.

The Cypriot and Lebanese NEMDBs were initially developed as national mutation frequency databases,[7,8] based on the *ETHNOS* v1.0 software.[11] These databases have migrated into the *ETHNOS* v2.0 software by expanding the existing information by not only enriching their content, but also by introducing new software functionalities. By contrast, the Egyptian and Tunisian NEMDBs were developed from scratch. Unfortunately, genetic diseases cannot be comparatively analysed among different ethnic or religious groups in these populations, since such information has not yet been provided (Table 1).

Comparison of the previous and current versions of the flat-file *ETHNOS* software

Previously, there were two different flat-file database versions of the *ETHNOS* software. The various features of the first version of *ETHNOS* were further developed using structured query language (SQL), on which FINDbase development was based (http://www.findbase.org);[12] this will not be mentioned further here.

ETHNOS v2.0 has a more advanced search engine, with more search options for the user to query, and it extracts data from a more elaborate data file. Also, the v2.0 software includes the 'Search +' option, ensuring patient anonymity. This functionality was introduced because there is a growing need to document interesting genetic disorders that are prevalent in isolated ethnic groups without risking stigmatisation of patients and their families.[13] In the v2.0 software, disease summary data querying can also be performed not only via drop-down menus, but also by using keywords, a functionality that was not available in the v1.0 software.

The provision of the list of 'genetics laboratories' is another novelty in the v2.0 software, providing the user with a comprehensive list of genetics laboratories that perform genetic analyses in the country where the population/ethnic group(s) documented in the NEMBD reside(s). This functionality, which, unfortunately, has not yet been exploited to its full extent in some NEMDBs, offers the possibility of creating a 'one-stop shop' solution for genetics services in different countries.

Finally, the use of the custom-made online editor is another functionality that distinguishes the two flat-file versions of the *ETHNOS* software, since it facilitates data entry by curators with limited proficiency in HTML. In other words, the editor allows not only direct copying of text, tables and images from word processors, but also online editing.

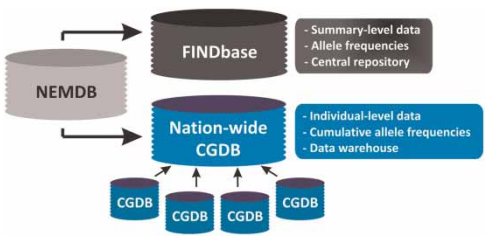

Figure 3. Schematic drawing showing the evolution of NEMDBs towards a global national mutation frequency database (eg FINDbase)[11] and a nationwide CGDB, to which data contribution occurs from different institutional CGDBs. In the former case, summary-level data are stored in a central repository, while in the latter case anonymous individual-level data are contributed unidirectionally in a data warehouse and comparatively analysed and queried.

Discussion

NEMDBs are increasingly becoming important tools for the documentation of genomic variations in various populations around the globe, and, as such, they assume an important role in the provision of genetics services.[4] Presently, the adoption of NEMDBs is not uniform and varies among different populations and national healthcare systems. Development of these NEMDBs should conform to certain guidelines and recommendations in order to assist genetic variation data capture in developing countries, ensuring a comprehensive worldwide data collection and better provision of healthcare services (Patrinos and coworkers, unpublished). Both versions of the *ETHNOS* software have contributed not only to

the establishment of similar databases for different populations, but also to database content uniformity, as more than half of the available NEMDBs are based on this software.

The NEMDB notion can be further expanded, yielding either a worldwide central repository for allele frequency data or a data warehouse of anonymous individual-level genetic data (Figure 3). One such worldwide central allele frequency repository, in which the frequency of genetic variants relating to a phenotypic alteration — namely inherited disease or variable drug response — will be deposited, is FINDbase (http://www.findbase.org).[12] This has already been established along the lines discussed here and documents a substantial amount of genetic data, allele frequencies of pathogenic mutations[12] and pharmacogenetically relevant single nucleotide polymorphisms.[1]

Apart from this application, collection of clinical genetic information on patients with particular genetic diseases, the investigation of a family's clinical history and genotype–phenotype correlations are also important.[14] Such data could be deposited in a central data warehouse derived from the various databases that exist in different healthcare institutions, namely hospitals, clinics and so on. The construction of depositories with genotype and phenotype information keyed to many individuals could be considered the ultimate database, hereafter termed 'clinical genetics databases (CGDBs)'. At present, such databases are necessary components of large population-wide epidemiological projects that have been initiated in various countries, such as Iceland, Estonia, India and the UK. Certainly, when whole-genome sequencing becomes routine, and personalised medicine is common, then we may well take these CGDBs for granted. In the meantime, the first efforts in that direction have begun to appear. In particular, a prototype software, allowing an individual's genetic profile to be stored in such a way that the information can only be retrieved by the patient and his/her physician, has been developed (Gkantouna, Tzimas and Patrinos, unpublished) (Figure 4). A convenient graphical interface is provided for the end user, who can enter data and retrieve results (through queries) from the CGDB in which the relevant information has been stored. The contents of the prototype CGDB are organised by theme, according to their semantic interpretation. At present, there are 15 topics, which generally collect information on parameters indicative of the progress of the patient's development and the results of specific genetic tests. Since the specific application is web based, end users have a number of important advantages — in particular, access from any computer or mobile device (mobile telephone or PDA), provided that there is an internet connection. In this way, the user is not limited to an application that is installed locally on his or her computer. Moreover, there is the possibility of multiple concurrent users. Such a database, currently under development, would allow patients to store all their genetic information and related phenotype securely, hence contributing decisively to customised medical treatment, better diagnosis of hereditary diseases and unambiguous personal ID. Also, with data in a structured format, the end user can easily analyse them statistically and draw useful conclusions.

Of course, CGDBs raise particularly complex ethical challenges that demand careful attention. Primarily, the inclusion of clinical and molecular data connected to specific individuals must be done in a way that ensures anonymity. How best to achieve this has not yet been established, but it is widely agreed that strict governance frameworks must be established to address any and all confidentiality concerns.[13] Other issues that need to be dealt with include copyright and intellectual property protection, the nature of informed consent, data access rights, inferential relationships and so on. There are no universally agreed solutions to these problems as yet, and, while a detailed discussion is beyond the scope of this paper, these issues must be resolved if patient databases and personalised medicine are substantially to advance. Our prototype CGDB authentication mechanism is double layered, determined by the user's password and IP address, while different user profiles are supported, thereby providing scaled access to the information. In the future, this application may be

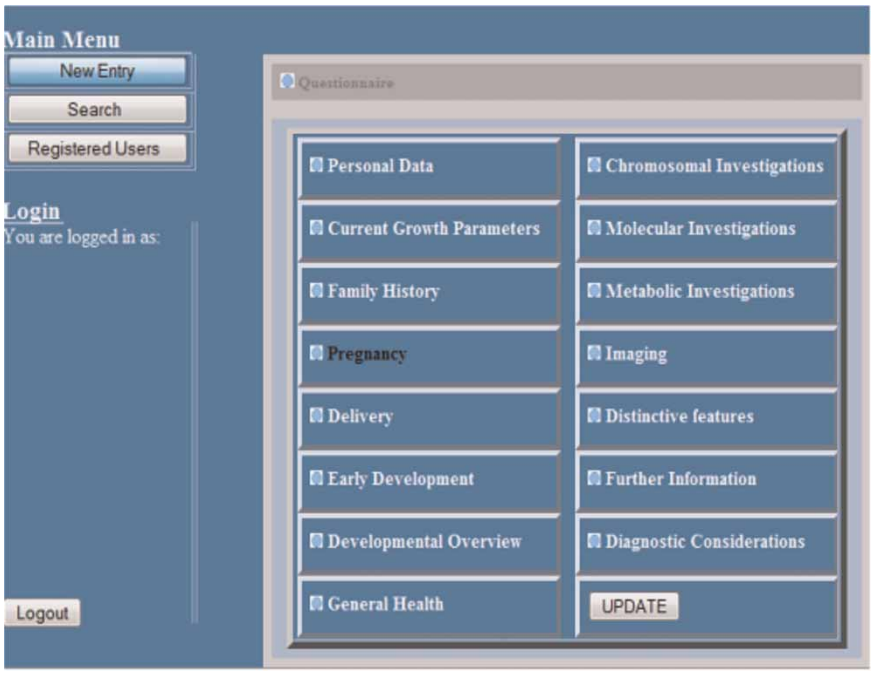

Figure 4. Secure data entry screen from a prototype CGDB, where phenotypic and genotypic data are deposited by authorised users.

extended with artificial intelligence features, in an effort to create a modern diagnostic tool, based on weighted parameters, the importance of which will be defined by the user, so as to be able to provide estimates on the patient's state of health and to facilitate the diagnosis of patients with the same or a similar genetic disease.

Acknowledgments

Part of our work has been funded by European Commission grants [MEDGENET (FP6-31968), EuroGenTest (FP6-512148) and GEN2PHEN (FP7-200574) to J.Z. and G.P.P. The authors gratefully acknowledge the contribution of Professors Andre Megarbane (Lebanon), Habiba Haabouni (Tunisia) and Marina Kleanthous (Cyprus) for their expert supervision in the development and curation of the related NEMDBs.

References

1. Lagoumintzis, G., Poulas, K. and Patrinos, G. (2010), 'Genetic databases and their potential to pharmacogenomics', *Curr. Pharm. Des.*, In press.
2. Claustres, M., Horaitis, O., Vanevski, M. and Cotton, R.G. (2002), 'Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases', *Genome Res.* Vol. 12, pp. 680–688.
3. Cotton, R.G., Phillips, K. and Horaitis, O. (2007), 'A survey of locus-specific database curation. Human Genome Variation Society', *J. Med. Genet.* Vol. 44, p. e72.
4. Patrinos, G.P. (2006), 'National and ethnic mutation databases: Recording populations' genography', *Hum. Mutat.* Vol. 27, pp. 879–887.
5. Zlotogora, J., van Baal, S. and Patrinos, G.P. (2007), 'Documentation of inherited disorders and mutation frequencies in the different religious communities in Israel in the Israeli National Genetic Database', *Hum. Mutat.* Vol. 28, pp. 944–949.
6. Zlotogora, J., van Baal, S. and Patrinos, G.P. (2009), 'The Israeli National Genetic Database', *Isr. Med. Assoc. J.* Vol. 11, pp. 373–375.
7. Kleanthous, M., Patsalis, P.C., Drousiotou, A., Motazacker, M. *et al.* (2006), 'The Cypriot and Iranian National Mutation Frequency Databases', *Hum. Mutat.* Vol. 27, pp. 598–599.
8. Megarbane, A., Chouery, E., van Baal, S. and Patrinos, G.P. (2006), 'The Lebanese National Mutation Frequency database', *Eur. J. Hum. Genet.* Vol. 14 (Suppl. 1), p. 365.
9. Cohen, T., Vardi-Saliternik, R. and Friedlender, Y. (2004), 'Consanguinity, intracommunity and intercommunity marriages in a population sample of Israeli Jews', *Ann. Hum. Biol.* Vol. 31, pp. 38–48.
10. Jaber, L., Halpern, G.J. and Shohat, T. (2000), 'Trends in the frequencies of consanguineous marriages in the Israeli Arab community', *Clin. Genet.* Vol. 58, pp. 106–110.
11. Patrinos, G.P., van Baal, S., Petersen, M.B. and Papadakis, M.N. (2005), 'Hellenic National Mutation database: A prototype database for mutations leading to inherited disorders in the Hellenic population', *Hum. Mutat.* Vol. 25, pp. 327–233.
12. van Baal, S., Kaimakis, P., Phommarinh, M., Koumbi, D. *et al.* (2007), 'FINDbase: A relational database recording frequencies of genetic defects leading to inherited disorders worldwide', *Nucleic Acids Res.* Vol. 35, pp. D690–D695.
13. Povey, S., Al Aqeel, A.I., Cambon-Thomsen, A., Dalgleish, R. *et al.* (2010), 'Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs)', *Hum. Mutat.*, In press.
14. Patrinos, G.P. and Brookes, A.J. (2005), 'DNA, diseases and databases: Disastrously deficient', *Trends Genet.* Vol. 21, pp. 333–338.