Population-specific documentation of pharmacogenomic markers and their allelic frequencies in FINDbase

Aims: Population and ethnic group-specific allele frequencies of pharmacogenomic markers are poorly documented and not systematically collected in structured data repositories. We developed the Frequency of Inherited Disorders Pharmacogenomics database (FINDbase-PGx), a separate module of the FINDbase, aiming to systematically document pharmacogenomic allele frequencies in various populations and ethnic groups worldwide. **Materials & methods:** We critically collected and curated 214 scientific articles reporting pharmacogenomic markers allele frequencies in various populations and ethnic groups worldwide. Subsequently, in order to host the curated data, support data visualization and data mining, we developed a website application, utilizing Microsoft™ PivotViewer software. **Results:** Curated allelic frequency data pertaining to 144 pharmacogenomic markers across 14 genes, representing approximately 87,000 individuals from 150 populations worldwide, are currently included in FINDbase-PGx. A user-friendly query interface allows for easy data querying, based on numerous content criteria, such as population, ethnic group, geographical region, gene, drug and rare allele frequency. **Conclusion:** FINDbase-PGx is a comprehensive database, which, unlike other pharmacogenomic knowledgebases, fulfills the much needed requirement to systematically document pharmacogenomic allelic frequencies in various populations and ethnic groups worldwide.

KEYWORDS: allelic frequencies ▪ databases ▪ data visualization ▪ ethnic groups ▪ markers ▪ pharmacogenomics ▪ populations ▪ SNPs ▪ web

Marianthi Georgitsi[1], Emmanouil Viennas[2], Vassiliki Gkantouna[2], Elena Christodoulopoulou[1], Zoi Zagoriti[1], Christina Tafrali[1], Fotios Ntellos[1], Olga Giannakopoulou[1], Athanassia Boulakou[1], Panagiota Vlahopoulou[1], Eva Kyriacou[1], John Tsaknakis[2], Athanassios Tsakalidis[2], Konstantinos Poulas[1], Giannis Tzimas[2] & George P Patrinos[†1]

[1]Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece
[2]Department of Computer Engineering & Informatics, Faculty of Engineering, University of Patras, Patras, Greece
[†]Author for correspondence:
Tel.: +30 261 096 9834
gpatrinos@upatras.gr

Pharmacogenomics studies the impact of an individual's genetic variation on drug response, by delineating his/her gene-expression profile or SNPs with a drug's toxicity or efficacy, hence optimizing drug therapy not only by maximizing efficiency, but also by minimizing the chances of adverse drug reactions [1]. In some areas of clinical research, such as cancer treatment or cardiovascular diseases, pharmacogenomics is readily applicable, while in others, such as neuropsychiatric disorders, intensive research is currently ongoing [2]. In numerous studies, a multitude of SNPs have been reported to be correlated with several drugs' safety or efficacy (also referred to as pharmacogenomic markers), often with contradicting results [3]. As expected, the incidence of these SNPs varies among different populations or ethnic groups; yet, the racial classification of populations, based on their genetic differences, for biomedical research purposes, was recently the subject of intense debate [4]. However, the observed differences among racial and ethnic groups in their clinical responses to drugs and their related toxicity, dictate a comprehensive documentation of such polymorphisms in an online repository, in order to facilitate the implementation of pharmacogenomics in different populations.

Although a limited number of pharmacogenomics-related databases are currently available, this aspect has been poorly addressed [5]. In particular, Pharmacogenomics Knowledge Base (PharmGKB) [6,7,101] sporadically records allelic frequencies of pharmacogenomic markers, but often in a racial- rather than population-specific manner, namely Caucasians, African–Americans or Asians. We have previously reported the development of the Frequency of Inherited Disorders database (FINDbase) [102], which documents the allelic frequencies of pathogenic mutations, leading to inherited disorders in various populations and ethnic groups around the world [8]. Given our previous experience, the available platform and the existing gap in the field, we have therefore decided to expand the FINDbase scope by incorporating curated allelic frequency data of pharmacogenomic markers, categorized by population and/or ethnic group.

Here, we report the development of FINDbase Pharmacogenomics (FINDbase-PGx) [103], a freely available and distinct module in FINDbase hosted at the Golden Helix Server®, which deals solely with documenting pharmacogenomic markers in various populations and ethnic groups worldwide.

Materials & methods

Inclusion criteria, literature data mining & curation

We initiated a major data mining effort, via a PubMed literature database search [104], using keywords such as the gene name and 'pharmacogenetics/pharmacogenomics' and/or 'polymorphisms' and/or 'populations', to retrieve relevant articles for 14 autosomal genes previously documented to be involved in modifying individuals' response to drug treatment and, hence, which are relevant to pharmacogenomics. These genes represent different classes of drug-metabolizing enzymes and transporters. The retrieved articles were selected for downstream data curation if they fulfilled the following criteria: the population and/or ethnicity was clearly stated and the cohort was ethnically homogeneous; the subjects were unrelated and the sample size sufficiently represented the given population, with at least 50 individuals (100 chromosomes) being analyzed in each study. Certain articles with a sample size of less than 50 individuals were exceptionally included if they reported allele frequencies of: isolated populations, such as the Nasioi (Papua New Guinea) and other Melanesian populations, the Karitiana and Surui people (Brazil), Micronesian populations, the Sandawe (Tanzania), or rare tribes, for instance the Ami and Atayal people (Taiwan), the Dogons (Mali and Burkina Faso), the Mbuti people (Congo), the Baka Pygmies (Central African Republic, Cameroon, Congo), the Cheyenne Amerindians (OK, USA), or populations rarely reported in the literature, such as the Mayas (Mexico), the Burmans (Myanmar), the Altaians and Tuvinians (Siberia), the Santomeans (Sao Tome e Principe) and the Ivorians (Côte d'Ivoire). If several articles were published on the same population and for the same pharmacogenomic marker, we would retain data from the largest available cohort, so that each population was represented once per variant and inclusion of redundant cases was avoided. If an article was studying a pharmacogenomic marker allele frequency between patients with a particular disease versus healthy controls, and if this marker was known or shown to be associated with this particular disease, we would only take into consideration the marker's allele frequencies in the healthy population.

Data collected by curators were entered in FINDbase-PGx *vis-à-vis* their published article ID in PubMed literature database and the curator's unique identifier. For the latter purpose, and based on our previous experience, we opted to use Thomson ISI ResearcherID [105], not only to clearly identify curated data when data update or correction is needed, but also to provide incentives to data contributors, particularly those submitting unpublished data.

System design & access

FINDbase-PGx is a separate module of the main FINDbase database [8,103], hosted in the Golden Helix Server [106]. The component services that comprise FINDbase-PGx follow the service oriented architectural approach [9]. Data content is freely available to the public and there are no registration requirements for data querying. The querying interface is based on a recently launched program by Microsoft (DC, USA), namely the PivotViewer [107] based on Microsoft Silverlight® technology [108], which offers useful tools for querying large datasets in multiple ways. The whole application provides an elegant, web-based multimedia interface for population-based 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 Online Mendelian Inheritance in Man (OMIM) [109] and the PharmGKB entries [101]. As previously mentioned, all entries are recorded against their unique PubMed and ResearcherIDs. The entire database schema is depicted in SUPPLEMENTARY FIGURE 1; (see www.futuremedicine.com/doi.suppl/10.27717/pgs.10.169). The system architecture and database implementation issues are detailed in the SUPPLEMENTARY TEXT (see www.futuremedicine.com/doi.suppl/10.27717/pgs.10.169).

Results

Data mining & curation

A total of 14 well-documented genes, previously implicated in the pharmacogenomics of several drug responses, were included in this first version of FINDbase-PGx, namely *CYP1A2*, *CYP2D6*, *CYP2E1*, *CYP3A4*, *CYP3A5*, *DPYD*, *NAT2*, *PON1*, *PON2*, *SLCO1B1*, *TPMT*, *TYMS*, *UGT1A1* and *UGT2B7* (TABLE 1). All data regarding *TPMT* and *UGT1A1* genes were adapted and enriched from FINDbase [8,102], after having manually adjusted all information to suit the needs of FINDbase-PGx database design and architecture.

In total, data were mined from 214 original or, rather rarely, review articles across 150 populations and ethnic groups, including North and sub-Saharan Africans, Caucasians, Northeast and Southeast Asians, populations from the South Pacific, Amerindians, Aborigineees and rare tribes. This effort resulted in the collection of data from approximately 87,000 subjects (>173,000 chromosomes; Table 1). As a result, FINDbase-PGx is a comprehensive database which, apart from consisting of a collection of important pharmacogenomic markers, currently only for 14 genes, provides additional detailed information on allele frequencies. Its content regarding the number of pharmacogenetically relevant genes will expand in the future, via continuous data curation.

As expected, several issues arose during data mining and curation. First of all, several inconsistencies were encountered during data curation, namely: inconsistencies pertaining to the calculated allele frequencies as reported by the authors in the abstract and main text or main text and tables; inconsistencies regarding the number of chromosomes analyzed for each pharmacogenomic marker; reproduction of inconsistent data in literature reviews, without prior curation, resulting in perpetuation of erroneous allele frequencies; rare allele frequency data often calculated from fewer samples than initially presented in the 'materials and methods' section of an article; difficulties in calculating rare alleles frequencies when the number of individuals with certain genotypes heterozygous for rare alleles were given as a whole [10]; uncertainty as to whether a studied haplotype was analyzed for all its component polymorphisms or solely for its most representative SNP. For instance, the **2* allele in *CYP2D6* is encountered in haplotypes **2A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *J*, *K*, *L* and *M*, yet many authors referred to **2* alone, making it unclear as to whether it was the rs16947 polymorphism studied alone or in parallel with other neighboring SNPs, such as rs61736911 and rs1135840. For such nonclarified cases, there is a note in the database, under the column 'Haplotype', reading 'Exact haplotype not defined'.

In addition, a large number of studies were conducted on cohorts whose selection was based on racial criteria (such as Caucasians or Africans) and not on ethnicity [11–15]. This raises the issue of the clinical relevance of rare allele frequency differences in pharmacogenetically relevant genes among ethnic, rather than racial groups. However, we exceptionally included in FINDbase-PGx a study on *UGT2B7* conducted on West Africans, owing to the relative genetic uniformity of West Africans and the scarcity of data reported on such populations (Côte d'Ivoire, Sierra Leone, Ghana and Senegal) [16]. Moreover, samples were often selected based on the country or area of residence [17], or from a particular medical/research center of reference [18], without clear statements regarding the ethnic background of the subjects. Interestingly, some authors stated that the cohort consisted

Table 1. Collective presentation of the datasets per pharmacogenetically relevant gene included in FINDbase-PGx.

Gene	Markers studied per gene (n)	Populations studied per gene (n)	Chromosomes analyzed per gene (n)	Ref.†
CYP1A2	17	20	12,074	[117]
CYP2D6	47	35	21,406	[118]
CYP2E1	10	45	5182	[119]
CYP3A4	6	18	9048	[120]
CYP3A5	9	51	20,320	[121]
DPYD	15	18	8652	[122]
NAT2	13	23	10,668	[123]
PON1	3	23	22,042	[124]
PON2	2	10	11,984	
SLCO1B1	4	18	11,226	[125]
TPMT	4	20	>11,776	[126]
TYMS	7	17	22,528	[127]
UGT1A1	3	23	3324	[128]
UGT2B7	4	5	3508	[129]
Total	144	‡	>173,738	

†The drugs associated with the proteins encoded by these genes are provided via their corresponding PharmGKB knowledgebase links.
‡Unlike total numbers of markers and chromosomes, a total number of populations cannot be calculated from this table, since the same population may be represented more than once in these 14 genes. In total, 150 populations and ethnic groups are represented in the database (see text).
FINDbase-PGx: Frequency of Inherited Disorders – Pharmacogenomics database; PharmGKB: Pharmacogenomics Knowledge Base.

of for example, Caucasians, of which a certain percentage was of a specific ethnic origin [19]; yet, allele frequency data was given for the whole cohort, rendering the study, according to our criteria, inappropriate for use.

Another typical issue we encountered was the inconsistent nomenclature used for the same pharmacogenomic marker across all relevant papers, either in the format of rs numbers or the * nomenclature, as well as the official Human Genome Variation Society [110] nomenclature based on genomic DNA, cDNA, or protein. For this reason, we had to manually conform all reported variations in the same three formats (rs numbers, * nomenclature and cDNA level) throughout all 14 pharmacogenetically relevant genes. In several occasions, the pharmacogenomic markers studied in certain articles and the nomenclature used by the authors was inconsistent to that used in other genomic databases, such as dbSNP [111], pharmacogenomics knowledgebases, such as PharmGKB [101], or locus-specific databases, such as the database of the Human CYP450 Nomenclature Committee [20,112].

Because of the aforementioned reasons, approximately 200 articles had to be excluded during data mining and, thus from FINDbase-PGx.

Querying FINDbase-PGx

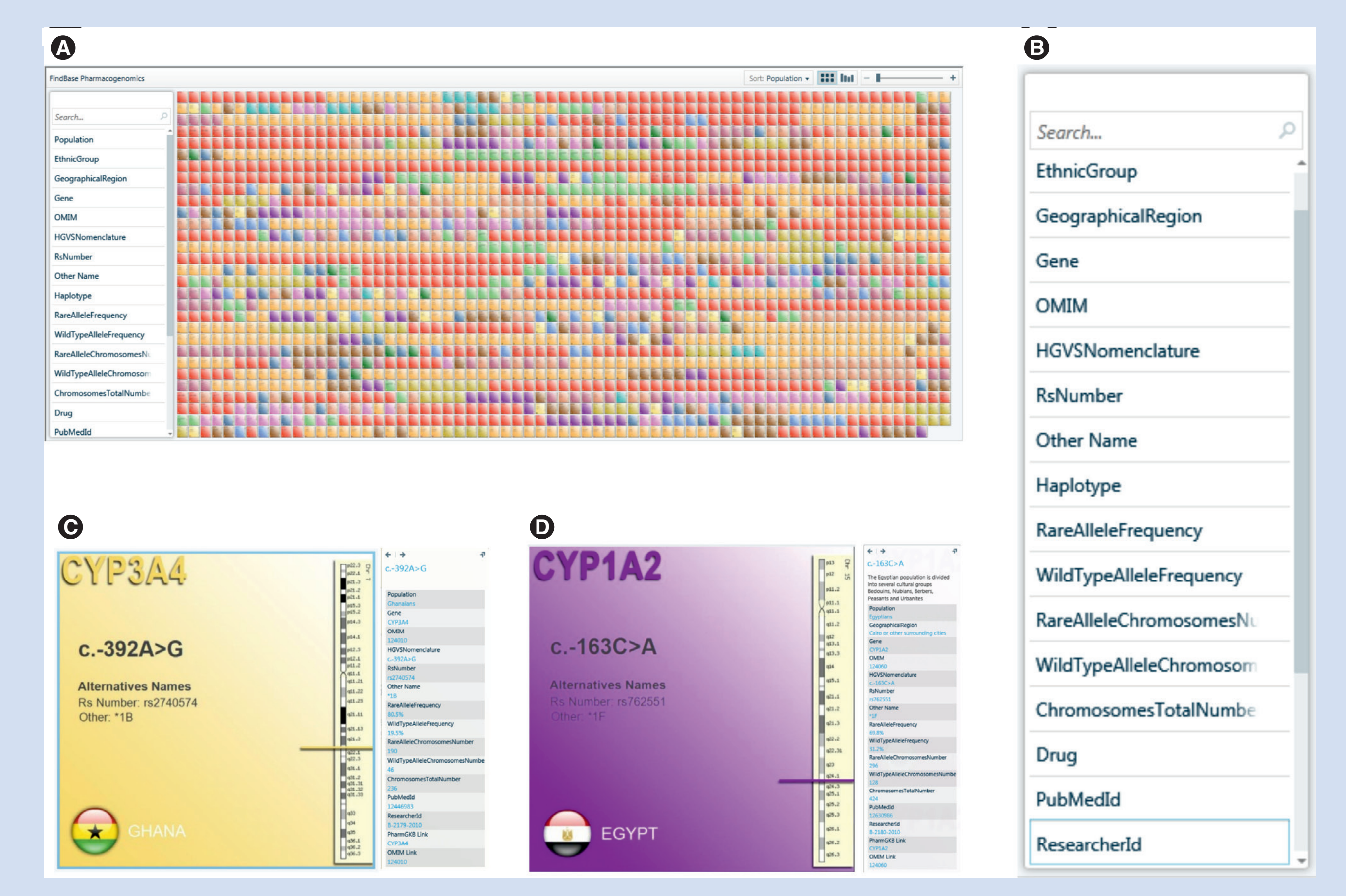
Frequency of Inherited Disorders database-PGx PivotViewer website application not only offers the user the possibility to zoom in from extensive datasets to particular gene-specific, variation-specific and/or population-specific data, but it also combines large groups of similar items, so that the user can begin viewing the relationships between individual pieces of information (Figure 1A). In summary, FINDbase-PGx provides users with a convenient environment to smoothly and quickly arrange 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 & 1D). A display item in the form of a card is provided for each marker, along with a sidebar textbox with in-depth data concerning the particular marker and population (Figure 1C & 1D). The chromosome figures on the cards are taken from the GeneCards website [113], maintained by the Weizmann Institute of Science, Israel. Hyperlinks for each gene name to OMIM database [109] and PharmGKB [101], offer the user the possibility of easily accessing additional information. Linking FINDbase-PGx to external database information enhances a growing network of genomic repositories and contributes towards data uniformity [5,8].

In particular, FINDbase-PGx enables the user to: make sense from a sheer amount of data (i.e., population-based variation data), 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 for each of the 14 pharmacogenetically relevant genes by selecting the field 'Gene' (Figure 2A). Additional filtering allows the user to observe specific allelic variants and to further zoom in on a particular population. A useful query would also lead to data arrangement based on the frequency of the rare allele (Figure 2B). Apart from these simple queries, the querying interface also allows the user to formulate compound queries. Such a query would be the identification of rare pharmacogenomic markers (frequencies 0–10%) in the Japanese population, sorted by gene name (Figure 3A). The query output includes 51 rare alleles for 11 genes (seven alleles for *CYP1A2*, 12 alleles for *CYP2D6*, three alleles for *CYP2E1*, four alleles for *CYP3A4*, one allele for *CYP3A5*, 13 alleles for *DPYD*, two alleles for *NAT2*, one allele for *PON1*, five alleles for *TPMT*, two alleles for *TYMS* and one allele for *UGT2B7*; Figure 3B).

Another example of a use of FINDbase-PGx would be the critical usage of the anticancer drug 5′-fluorouracil in relation to the *DPYD*2A* allele (also known as IVS14+1G>A). This variant has only been identified in Caucasians, rarely in Asians, but not in Africans to date (Figure 4). Yet, pretherapeutic testing for the *DPYD*2A* allele alone, in order to identify DPYD deficiency, is not considered sufficient, since 5′-fluorouracil-related toxicity cannot be attributed solely to this variant [21,22]. Contrary to the former example, a well-documented case of how the existence of an ethnic-specific marker, currently lacking from FINDbase-PGx, has led to specific US FDA-recommendations, is the *HLA-B*1502* allele in the *HLA-B* gene: a genetic test for this marker is meaningful only for patients of Asian descent [23].

Discussion

The importance of pharmacogenomic knowledgebases was highlighted in the recent review by Lagoumintzis and coworkers [5]. In this article, we describe the development of a separate module of FINDbase, namely FINDbase-PGx, including allele frequency data for 144



c.-392A>G

Population
Ghanaians
Gene
CYP3A4
OMIM
124010
HGVSNomenclature
c.-392A>G
RsNumber
rs2740574
Other Name
*1B
RareAlleleFrequency
80.5%
WildTypeAlleleFrequency
19.5%
RareAlleleChromosomesNumber
190
WildTypeAlleleChromosomesNumbe
46
ChromosomesTotalNumber
236
PubMedId
12446983
ResearcherId
B-2179-2010
PharmGKB Link
CYP3A4
OMIM Link
124010

c.-163C>A

The Egyptian population is divided into several cultural groups Bedouins, Nubians, Berbers, Peasants and Urbanites

Population
Egyptians
GeographicalRegion
Cairo or other surrounding cities
Gene
CYP1A2
OMIM
124060
HGVSNomenclature
c.-163C>A
RsNumber
rs762551
Other Name
*1F
RareAlleleFrequency
69.8%
WildTypeAlleleFrequency
31.2%
RareAlleleChromosomesNumber
296
WildTypeAlleleChromosomesNumbe
128
ChromosomesTotalNumber
424
PubMedId
12630986
ResearcherId
B-2180-2010
PharmGKB Link
CYP1A2
OMIM Link
124060

Figure 1. Detailed presentation of FINDbase-PGx features. (A) Overview of the entire FINDbase-PGx data collection, shown using Microsoft's PivotViewer. The querying interface is available at the left-hand side of the screen, while the output option can be selected at the top-right corner of the screen. The boxes represent different entries, 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)** FINDbase-PGx querying interface that provides a user-friendly tool to quickly arrange data collection according to common characteristics that can be selected from the data query menu. Display items in the form of a card are provided for each marker, along with a sidebar textbox with in-depth data concerning the particular marker and population **(C & D)**. Each item includes the name of the allele in its official Human Genome Variation Society 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 and PharmGKB offer to the user the possibility of easily accessing additional information. Finally, each item displays the corresponding PubMed and Researcher IDs.
FINDbase-PGx: Frequency of Inherited Disorders – Pharmacogenomics database; OMIM: Online Mendelian Inheritance in Man; PharmGKB: The Pharmacogenomics Knowledge Base.

pharmacogenomic markers for 14 genes of pharmacogenomic interest. Data have been collected via literature data mining of 214 scientific articles, representing approximately 87,000 individuals from 150 populations worldwide and across all continents. As a result, FINDbase-PGx represents a worldwide collection of pharmacogenomic markers, and fills in a gap left by other pharmacogenomics knowledgebases and related resources.

Pharmacogenomics Knowledge Base harbors information on the relationship between drugs, diseases, genes and their variations, including more than 200 well-documented very important pharmacogenes. It includes a wide range of genotype data, along with phenotypic data on molecular and functional assays, pharmacokinetics, pharmacodynamics and drug response, and clinical outcome [6,7]. Yet, PharmGKB focuses on variants with a well-established pharmacogenetic relationship to drug efficacy, metabolism, or toxicity, whereas population- or ethnic-specific differences are rarely documented. Rather, allelic frequencies are often referred to in terms of racial-specific differences, such as Caucasians, Asians, or African–Americans, based mainly on the HapMap groups. FINDbase-PGx presents data on well-established, as well as less established, pharmacogenomic markers, while contrary to PharmGKB, it contains data from 150 populations, including even isolated ethnic groups and rare tribes. Moreover, the locus specific

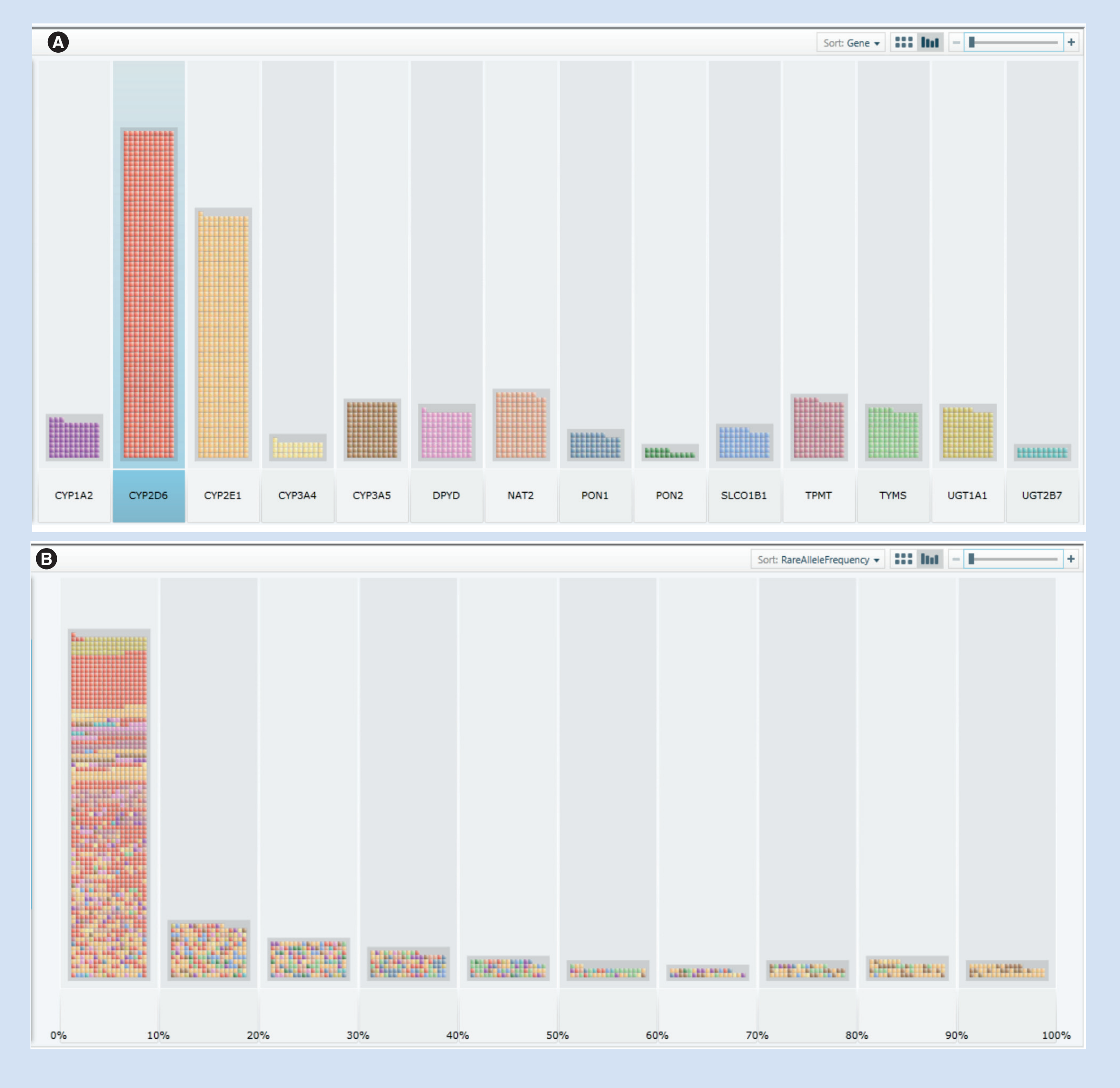


Figure 2. Sorting FINDbase-PGx data content. The database content can be sorted using the columns option (see top-right corner of the screen), according to the gene name **(A)** or rare allele frequency **(B)**. The user can select a specific column (e.g., second column in **[A]** showing variants of the *CYP2D6* gene highlighted) to retrieve more specific information.
FINDbase-PGx: Frequency of Inherited Disorders – Pharmacogenomics database.

databases of the Human *CYP* Nomenclature Committee [20,112], as well as the *UGT* Alleles Nomenclature page [114], provide a very detailed compilation of genetic variants along with information, though very limited, on their functional effects, yet they are devoid of information on allelic frequencies in different populations/ethnic groups. FINDbase-PGx adequately covers this aspect, though currently only for a subset of the *CYP* and *UGT* genes, but it aims to foster a complete list in the future, via continuous database curation and data enrichment. The latter could also be achieved by strategic partnerships with consortia currently in the process of studying the incidence of pharmacogenomic markers in different populations. The PharmacoGenomics for Every Nation Initiative [115] may well be such an example, where FINDbase-PGx could serve as the main interface for data storage and retrieval of the frequencies of pharmacogenomic

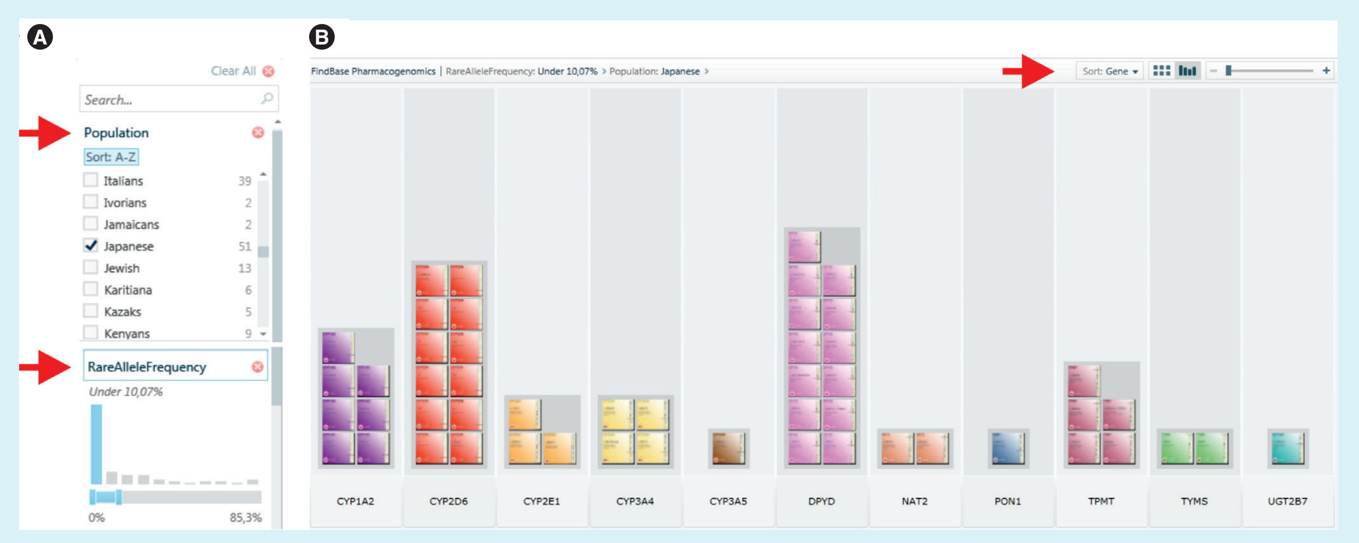


Figure 3. Example of a database query, based on a specific population. (A) Query formulation to retrieve the rare alleles in the Japanese population with an allelic frequency of 10.07% or lower (indicated with arrows). **(B)** The query returns 51 alleles in equal number of display items, sorted out in columns arranged by gene name (arrow). The total number of alleles is also shown in the query box along with the rare alleles with frequencies below 10.07% in all populations recorded in Frequency of Inherited Disorders – pharmacogenomics database. This query indicates that contrary to the Japanese and Italian populations, only two alleles in the Jamaican population have a frequency of less that 10.07%.

markers. Continuous database enrichment could be further achieved via direct data submission from individual contributors; new data submission will also be reported in the form of a manuscript in *Human Genomics and Proteomics* [116], the first database-journal that links a database (FINDbase) with a scientific journal [24]. This way, upon acceptance, pharmacogenomic marker frequency data becomes immediately part of the main FINDbase-PGx data collection and the corresponding paper becomes indexed in PubMed literature database against a PubMedID.

Apart from the examples described above, the potential impact of FINDbase-PGx in designing population-specific or ethnic group-specific strategies for personalized pharmacogenomic tests and medical services is highlighted by some other characteristic examples. It is well-established by now that the *CYP3A5*3* allele, giving rise to a nonfunctional protein product due to incorrectly spliced mRNA, shows markedly different frequencies among sub-Saharan Africans (i.e., the lowest, 6–7%) compared with Caucasian and Asian populations (i.e., the

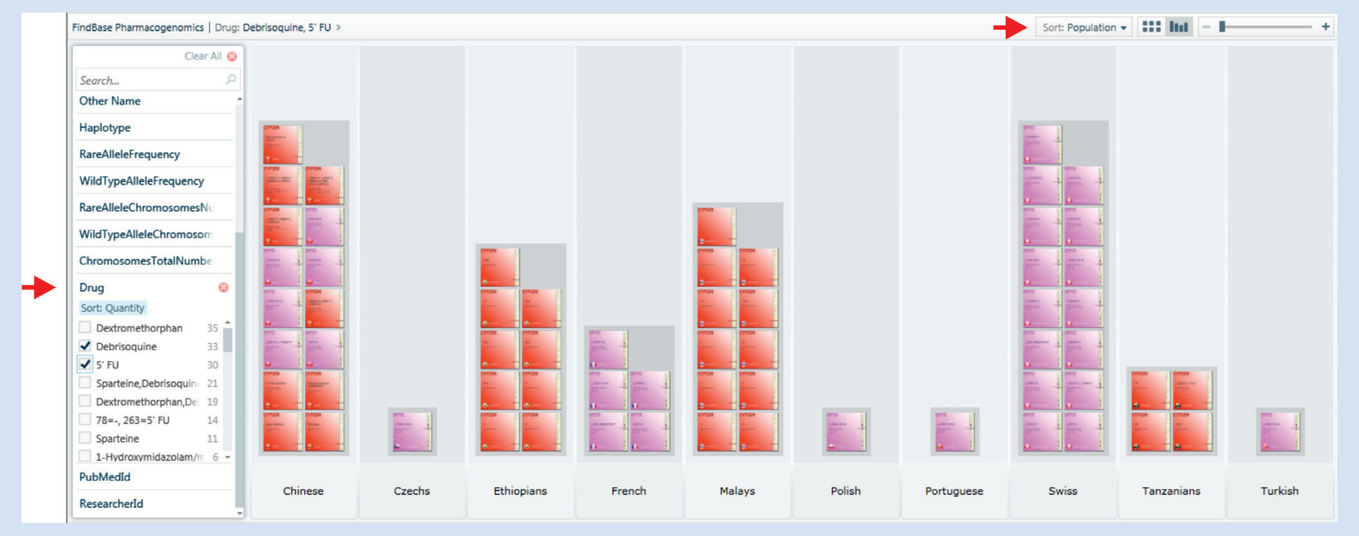


Figure 4. Query formulation to retrieve the pharmacogenomic markers for the drugs 5´-fluorouracil and debrisoquine (arrow). This query returns 63 alleles in two genes, 33 for debrisoquine (for *CYP2D6* gene) and 30 for 5´-fluorouracil (*DPYD* gene). The query output is sorted out by population (arrow), using again the columns option.

highest, 95–96%) [14], which is in agreement with our data (Supplementary Table 1; see www.futuremedicine.com/doi/suppl/10.2217/pgs.10.169). This fact has major implications in CYP3A5-mediated drug clearance and response, such as calcium channel blockers, cholesterol lowering drugs, antiretroviral drugs, cancer chemotherapeutics and immunosuppressants [14]. Interestingly, it seems that *CYP3A5*3* allele frequency increases with the geographical latitude (i.e., distance from the equator) [25].

Currently, the FINDbase-PGx user can freely access and navigate their way inside the database by performing various queries as previously described. In the future, FINDbase-PGx will adapt to the data entry and database curation scheme that has been envisaged for the renewed FINDbase (currently under construction), into which the contents of the pharmacogenomics module will be integrated later on. This FINDbase scheme relies on the use of the ResearcherID. The direct submission option provides incentives not only for expanding existing datasets, but also for incorporating new ones, by direct contribution, while strongly encouraging the concept of unique researcher identifiers (such as ResearcherID, OpenID®, Researcher Identification Primer), a concept which has been successfully implemented for microattribution [Giardine *et al.*, manuscript submitted]. Regarding the anticipated increase of data influx, the way the system is developed allows for many more data to be entered without affecting the performance of the querying and, most importantly, the visualization interface.

In addition, in certain cases where population size is relatively small, we plan to include multiple reports, if available in the literature, instead of selecting one report for these populations. Information from multiple reports will be clearly labeled and shown separately with a link to each related reference so that users have access to the most comprehensive data and can make their own judgment if they need to either select one report or combine all the available data.

An important issue that arose during literature data mining for FINDbase-PGx was the lack of nomenclature consistency pertaining to pharmacogenomic markers. This is a general problem that also involves the locus-specific databases field in general [26]. This fact highlights the necessity of establishing a uniform nomenclature system for pharmacogenomic markers, and this idea could materialize via the organization of a Pharmacogenomics Nomenclature Consortium. However, for certain genes it remains rather unclear which of all these identified variants are of proven importance for pharmacogenomics. In addition, identifying functionally important pharmacogenomic markers may be experimentally challenging and time consuming [3]. Undoubtedly, more functional studies are needed to unravel this important issue for a multitude of gene variants with ambiguous or still contradictary effects on drug metabolism, efficacy and toxicity. Last, but not least, it is worth noting that drug responses and treatment outcomes may largely rely on networks of interacting genes (i.e., pathways), rather than being attributed to monogenic traits.

Conclusion

Genetic variations in DNA may explain the spectrum of responses, including tolerability and side effects, manifested under specific

Executive summary

- Frequency of Inherited Disorders database (FINDbase) documents the frequencies of pathogenic mutations leading to inherited disorders worldwide.
- Here, we report the development of Frequency of Inherited Disorders – Pharmacogenomics database (FINDbase-PGx), a separate module of FINDbase at the Golden Helix Server®, pertaining solely to the documentation of population- and ethnic group-specific allele frequencies of pharmacogenomic markers in 14 genes, representing different classes of drug-metabolizing enzymes and transporters.
- FINDbase-PGx 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.
- The querying interface is based on a recently launched program by Microsoft (DC, USA), namely the PivotViewer based on Microsoft Silverlight™ technology, which offers useful tools for querying large datasets in multiple ways. This way, the user can quickly search from sheer amounts of data to retrieve gene-specific, variation-specific or population-specific datasets.
- This effort represents the largest collection of population- and ethnic group-specific pharmacogenomic markers allelic frequencies data to date and fills in the gap left by other pharmacogenomics knowledgebases and related resources.
- This effort aims to assist in the future design and development of pharmacogenomic testing, hence towards the advent of personalized medicine. Yet, continuous research efforts and technological improvements are needed, before we can confidently conclude on the relationship between individual genotypes and drug response/toxicity.

medication treatments. Pharmacogenomic testing is expected to assist in materializing personalized medicine and, eventually improving therapeutic modalities [27]. Registering the pharmacogenomic markers and their frequencies aims to help accustom physicians to adopt pharmacogenomic testing in different populations worldwide, by taking into consideration the individuals' ethnic, and thus, 'genographic' origin [28]. FINDbase-PGx is the first publicly available database module that will help push this idea forward. Toward this end, our group will not only continue to contribute curated data, in order to enrich FINDbase-PGx with more pharmacogenetically relevant genes and related allele frequency information, but will also simultaneously encourage the scientific community for direct data submission to keep FINDbase-PGx as comprehensive as possible.

Financial & competing interests disclosure

The study was financially supported by European Commission grants (ITHANET [FP6–026539], EuroGenTest [FP6–512148] and GEN2PHEN [FP7–200754]) to George P Patrinos and by the Golden Helix Institute of Biomedical Research. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

Papers of special note have been highlighted as:
▪ of interest
▪▪ of considerable interest

1 Patrinos GP, Innocenti F: Pharmacogenomics: paving the path to personalized medicine. *Pharmacogenomics* 11(2), 141–146 (2010).

2 Squassina A, Manchia M, Manolopoulos VG *et al.*: Realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics* 11(8), 1149–1167 (2010).

3 Evans WE, Relling MV: Moving towards individualized medicine with pharmacogenomics. *Nature* 429(6990), 464–468 (2004).

▪ **Comprehensive review with examples of how genes and their polymorphisms may affect drug response, and discussion on issues regarding the implementation of pharmacogenomics into clinical practice.**

4 Burchard EG, Ziv E, Coyle N *et al.*: The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* 348(12), 1170–1175 (2003).

5 Lagoumintzis G, Poulas K, Patrinos GP: Genetic databases and their potential in pharmacogenomics. *Curr. Pharm. Des.* 16(20), 2224–2231 (2010).

6 Klein TE, Chang JT, Cho MK *et al.*: Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge Base. *Pharmacogenomics J.* 1(3), 167–170 (2001).

7 Hewett M, Oliver DE, Rubin DL *et al.*: PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.* 30(1), 163–165 (2002).

8 van Baal S, Kaimakis P, Phommarinh M *et al.*: FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. *Nucleic Acids Res.* 35(Database issue), D690–D695 (2007).

▪▪ **Initial publication of the Frequency of Inherited Disorders database (FINDbase) recording allelic frequencies of genetic variations causing inherited disorders.**

9 Bell M: *SOA Modeling Patterns for Service-Oriented Discovery and Analy*sis. Wiley & Sons, Inc., Hoboken, NJ, USA (2010).

10 Jakubowska A, Gronwald J, Menkiszak J *et al.*: BRCA1-associated breast and ovarian cancer risks in Poland: no association with commonly studied polymorphisms. *Breast Cancer Res. Treat.* 119(1), 201–211 (2010).

11 Marsh S, Collie-Duguid ES, Li T *et al.*: Ethnic variation in the thymidylate synthase enhancer region polymorphism among Caucasian and Asian populations. *Genomics* 58(3), 310–312 (1999).

12 Lima CS, Ortega MM, Ozelo MC *et al.*: Polymorphisms of methylenetetrahydrofolate reductase (*MTHFR*), methionine synthase (*MTR*), methionine synthase reductase (*MTRR*), and thymidylate synthase (*TYMS*) in multiple myeloma risk. *Leuk. Res.* 32(3), 401–405 (2008).

13 Man M, Farmen M, Dumaual C *et al.*: Genetic variation in metabolizing enzyme and transporter genes: comprehensive assessment in 3 major East asian subpopulations with comparison to Caucasians and Africans. *J. Clin. Pharmacol.* 50(8), 929–940 (2010).

14 Kuehl P, Zhang J, Lin Y *et al.*: Sequence diversity in *CYP3A* promoters and characterization of the genetic basis of polymorphic *CYP3A5* expression. *Nat. Genet.* 27(4), 383–391 (2001).

▪▪ ***CYP3A5* polymorphisms *CYP3A5*1*, *CYP3A5*3* and *CYP3A5*6* are responsible for the varying expression levels of *CYP3A5*, and thus, for *CYP3A5*-mediated drug clearance and response, with inter-racial allele frequency differences.**

15 Ho RH, Choi L, Lee W *et al.*: Effect of drug transporter genotypes on pravastatin disposition in European– and African–American participants. *Pharmacogenet. Genomics* 17(8), 647–656 (2007).

16 Mehlotra RK, Bockarie MJ, Zimmerman PA: Prevalence of *UGT1A9* and *UGT2B7* nonsynonymous single nucleotide polymorphisms in West African, Papua New Guinean, and North American populations. *Eur. J. Clin. Pharmacol.* 63(1), 1–8 (2007).

17 Bolufer P, Collado M, Barragan E *et al.*: The potential effect of gender in combination with common genetic polymorphisms of drug-metabolizing enzymes on the risk of developing acute leukemia. *Haematologica* 92(3), 308–314 (2007).

18 Largillier R, Etienne-Grimaldi MC, Formento JL *et al.*: Pharmacogenetics of capecitabine in advanced breast cancer patients. *Clin. Cancer Res.* 12(18), 5496–5502 (2006).

19 Provenzani A, Notarbartolo M, Labbozzetta M *et al.*: The effect of *CYP3A5* and *ABCB1* single nucleotide polymorphisms on tacrolimus dose requirements in Caucasian liver transplant patients. *Ann. Transplant.* 14(1), 23–31 (2009).

20 Sim SC, Ingelman-Sundberg M: The human cytochrome P450 Allele Nomenclature Committee Website: submission criteria, procedures, and objectives. *Methods Mol. Biol.* 320, 183–191 (2006).

21 Morel A, Boisdron-Celle M, Fey L *et al.*: Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance. *Mol. Cancer. Ther.* 5(11), 2895–2904 (2006).

22 Yen JL, McLeod HL: Should DPD analysis be required prior to prescribing fluoropyrimidines? *Eur. J. Cancer* 43(6), 1011–1016 (2007).

23 Chung WH, Hung SI, Hong HS *et al.*: Medical genetics: a marker for Stevens–Johnson syndrome. *Nature* 428(6982), 486 (2004).

24 Patrinos GP, Petricoin EF: A new scientific journal linked to a genetic database: Towards a novel publication modality. *Hum. Genomics Proteomics* 1(1), e597478 (2009).

25 Thompson EE, Kuttab-Boulos H, Witonsky D *et al.*: *CYP3A* variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* 75(6), 1059–1069 (2004).

26 Mitropoulou C, Webb AJ, Mitropoulos K *et al.*: Locus-specific database domain and data content analysis: evolution and content maturation toward clinical use. *Hum. Mutat.* 31(10), 1109–1116 (2010).

27 Patrinos GP: General considerations for integrating pharmacogenomics into the mainstream medical practice. *Hum. Genomics* 4(6), 371–374 (2010).

28 Patrinos GP: National and ethnic mutation databases: recording populations' genography. *Hum. Mutat.* 27(9), 879–887 (2006).

- **Review presenting the concept of National and ethnic mutations database, along with a proposed model, and recording the various databases available worldwide.**

Websites

101 The Pharmacogenomics Knowledge base (PharmGKB)
www.pharmgkb.org

102 Frequency of Inherited Disorders database (FINDbase)
www.findbase.org

103 Frequency of Inherited Disorders – Pharmacogenomics database (FINDbase-PGx)
http://findbasepgx.goldenhelix.org

104 PubMed
www.ncbi.nlm.nih.gov/pubmed

105 Thomson ISI ResearcherID
www.researcherid.com

106 The Golden Helix Server
www.goldenhelix.org

107 PivotViewer
www.getpivot.com

108 Microsoft Silverlight
www.silverlight.net

109 Online Mendelian Inheritance in Man (OMIM)
www.ncbi.nlm.nih.gov/omim

110 Human Genome Variation Society
www.hgvs.org

111 The Single Nucleotide Polymorphism database (dbSNP)
www.ncbi.nlm.nih.gov/projects/SNP

112 The Human Cytochrome P450 (*CYP*) Allele Nomenclature Committee
www.cypalleles.ki.se

113 GeneCards
www.genecards.org

114 The UDP-glucuronosyltransferase Alleles Nomenclature page
www.pharmacogenomics.pha.ulaval.ca/sgc/ugt_alleles

115 Pharmacogenomics for Every Nation Initiative (PGENI)
www.pgeni.org

116 Human Genomics and Proteomics journal
www.sage-hindawi.com/journals/hgp

117 PharmGKB: Curated drug knowledge regarding *CYP1A2* gene
www.pharmgkb.org/do/serve?objId=PA27093&objCls=Gene#tabview=tab5

118 PharmGKB: Curated drug knowledge regarding *CYP2D6* gene
www.pharmgkb.org/do/serve?objId=PA128&objCls=Gene#tabview=tab6

119 PharmGKB: Curated drug knowledge regarding *CYP2E1* gene
www.pharmgkb.org/do/serve?objId=PA129&objCls=Gene#tabview=tab4

120 PharmGKB: Curated drug knowledge regarding *CYP3A4* gene
www.pharmgkb.org/do/serve?objId=PA130&objCls=Gene#tabview=tab6

121 PharmGKB: Curated drug knowledge regarding *CYP3A5* gene
www.pharmgkb.org/do/serve?objId=PA131&objCls=Gene#tabview=tab6

122 PharmGKB: Curated drug knowledge regarding *DPYD* gene
www.pharmgkb.org/do/serve?objId=PA145&objCls=Gene#tabview=tab6

123 PharmGKB: Curated drug knowledge regarding *NAT2* gene
www.pharmgkb.org/do/serve?objId=PA18&objCls=Gene#tabview=tab5

124 PharmGKB: Curated drug knowledge regarding *PON1* gene
www.pharmgkb.org/do/serve?objId=PA33529&objCls=Gene#tabview=tab4

125 PharmGKB: Curated drug knowledge regarding *SLCO1B1* gene
www.pharmgkb.org/do/serve?objId=PA134865839&objCls=Gene#tabview=tab6

126 PharmGKB: Curated drug knowledge regarding *TPMT* gene
www.pharmgkb.org/do/serve?objId=PA356&objCls=Gene#tabview=tab6

127 PharmGKB: Curated drug knowledge regarding *TYMS* gene
www.pharmgkb.org/do/serve?objId=PA359&objCls=Gene#tabview=tab6

128 PharmGKB: Curated drug knowledge regarding *UGT1A1* gene
www.pharmgkb.org/do/serve?objId=PA420&objCls=Gene#tabview=tab6

129 PharmGKB: Curated drug knowledge regarding *UGT2B7* gene
www.pharmgkb.org/do/serve?objId=PA361&objCls=Gene#tabview=tab5