


**DATABASES**

# Documentation of clinically relevant genomic biomarker allele frequencies in the next-generation FINDbase worldwide database

Fotios Kounelis<sup>1</sup> | Alexandros Kanterakis<sup>2,3</sup>  | Andreas Kanavos<sup>1,3</sup> |  
Maria-Theodora Pandi<sup>3,4</sup> | Zoe Kordou<sup>3</sup> | Olivia Manusama<sup>5</sup> |  
Gerasimos Vonitsanos<sup>1,3</sup> | Theodora Katsila<sup>3</sup> | Evangelia-Eirini Tsermpini<sup>3</sup> |  
Volker M. Lauschke<sup>6</sup>  | Maria Koromina<sup>3</sup> | Peter J. van der Spek<sup>4</sup> |  
George P. Patrinos<sup>3,4,7,8</sup> 

<sup>1</sup>Department of Computer Engineering and Informatics, Faculty of Engineering, University of Patras, Patras, Greece

<sup>2</sup>Biomedical Informatics Laboratory, Foundation of Research and Technology Hellas, Heraklion, Greece

<sup>3</sup>Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

<sup>4</sup>Bioinformatics Unit, Department of Pathology, Faculty of Medicine and Health Sciences, Medical Center, Erasmus University, Rotterdam, The Netherlands

<sup>5</sup>Department of Immunology, Faculty of Medicine and Health Sciences, Erasmus University Medical Center, Rotterdam, The Netherlands

<sup>6</sup>Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden

<sup>7</sup>Zayed Center of Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates

<sup>8</sup>Department of Pathology, College of Medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates

**Correspondence**

George P. Patrinos, Department of Pharmacy, School of Health Sciences, University of Patras, University Campus, Rion, GR-265 04 Patras, Greece.  
Email: [gpatrinos@upatras.gr](mailto:gpatrinos@upatras.gr)

**Present address**

Fotios Kounelis, Department of Computing, Imperial College London, London, London, UK  
Theodora Katsila, Institute of Chemical Biology, National Hellenic Research Foundation, Athens, Greece.

**Funding information**

General Secretariat for Research and Technology, Grant/Award Number: 5002780; European Commission, Grant/Award Numbers: 200754, 305444

**Abstract**

FINDbase (<http://www.findbase.org>) is a comprehensive data resource recording the prevalence of clinically relevant genomic variants in various populations worldwide, such as pathogenic variants underlying genetic disorders as well as pharmacogenomic biomarkers that can guide drug treatment. Here, we report significant new developments and technological advancements in the database architecture, leading to a completely revamped database structure, querying interface, accompanied with substantial extensions of data content and curation. In particular, the FINDbase upgrade further improves the user experience by introducing responsive features that support a wide variety of mobile and stationary devices, while enhancing computational runtime due to the use of a modern Javascript framework such as ReactJS. Data collection is significantly enriched, with the data records being divided in a Public and Private version, the latter being accessed on the basis of data contribution, according to the microattribution approach, while the front end was redesigned to support the new functionalities and querying tools. The above-mentioned updates further enhance the impact of FINDbase, improve the overall user experience, facilitate further data sharing by microattribution, and strengthen

the role of FINDbase as a key resource for personalized medicine applications and personalized public health.

#### KEYWORDS

allele frequencies, clinically relevant genomic variations, data visualization, genomic variation, pharmacogenomic biomarkers, population

## 1 | INTRODUCTION

In the postgenomic era, which is characterized by the exponential generation of DNA sequencing data, the need to comprehensively document and deposit these data into well-maintained and curated resources is more important than ever. There have been numerous databases developed to fulfill this need but only a fraction of those are widely recognized and utilized. Such acceptance depends on data accuracy and curation, which are crucial elements that distinguish a database that is truly useful and acknowledged by the scientific community from others that have been built by researchers at the side of their projects and tend to be abandoned, mostly due to the lack of funding and interest deflation (Patrinos & Brookes, 2005). Genomic databases with clinical relevance can be categorized into three types: (a) general (or core) genomic databases include genomic data with basic phenotypic description. Two of the most well-known are expert-curated general genomic databases with clinical relevance are ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar>; Landrum et al., 2016) and the Human Gene Mutation Database (HGMD; <http://www.hgmd.compareac.uk>; Stenson et al., 2014); (b) locus-specific databases (LSDBs), which include genotypic information accompanied with comprehensive and in-depth phenotypic description of all deposited variants, the majority of which are often contributed by researchers directly to database curators and consists of unpublished information. The HbVar database of human hemoglobin variants and thalassemia mutations (<http://globin.bx.psu.edu/hbvar>; Giardine et al., 2014) is among the most well-respected LSDBs, while the Leiden Open Variation Database (<http://www.lovd.nl>; Mitropoulou, Webb, Mitropoulos, Brookes, & Patrinos, 2010) is a collection of LSDBs which tries to rectify the extant content heterogeneity of the existing LSDBs (Fokkema et al., 2011) and, most importantly, to address technical issues that prevent the scientific community from developing and curating LSDBs, by providing an easy-to-use web interface (c) the national/ethnic genomic databases (NEGDBs), which comprehensively document the prevalence of clinically relevant alleles in different populations and ethnic groups worldwide (Patrinos, 2006).

The NEGDBs comprise a well-defined group of genomic databases that first appeared in the early 2000 (Patrinos, 2006) and are used for population genetic studies, for example, to study gene/mutation flow and admixture patterns, human demographic history, as well as in genomic medicine and public health to facilitate patient stratification and/or rationalization of drug use (Patrinos, 2006). These databases can serve as a valuable complement to the data content of the core databases and the LSDBs, in which this information is often lacking or is poorly documented (Giardine et al., 2014).

In 2006, the Frequency of INherited Disorders (FINDbase; <http://www.findbase.org>) database was established, aiming to fill the gap of a worldwide database that would comprehensively document the prevalence of clinically relevant genomic variation allele frequencies in various populations and ethnic groups worldwide (van Baal et al., 2007). From its establishment, FINDbase contains only aggregated data, that is, allele frequencies deprived from any personal information of their carriers, to ensure data anonymity. In its beginnings, FINDbase accommodated data pertaining to the prevalence of pathogenic genomic variants, leading for example, to monogenic disorders; however, as of 2010, FINDbase's data content was expanded to also include allelic frequency data for pharmacogenomic (PGx) biomarkers in distinct data modules (Georgitsi et al., 2011b). FINDbase has the most extensive content among all current NEGDBs and constitutes one of the key resources for population-specific clinically relevant genomic variation allele frequency data information deducted from the number and origin of visitors, while it also complies with the recommendations for genomic data collection from populations (Patrinos et al., 2011). To keep FINDbase up-to-date and user-friendly, its user interface and querying module have undergone major refurbishments and updates in 2010 (Georgitsi et al., 2011b), while its data content is being continuously enriched and updated, where needed, in response to user feedback (Papadopoulos et al., 2014; Viennas et al., 2017). Still, the continuous data updates and influx dictated further upgrade of the user interface so that data output is expedited and provided in a timely manner.

Here, we present the next generation of the FINDbase worldwide database, including significant technological advances in the user interface, data output, and visualization, as well as data content updates in both data modules, which substantially expands the existing functionalities and database content.

## 2 | METHODS

### 2.1 | Functionality and main user interface components

In this update, the FINDbase user interface has been redesigned from scratch, using a single-page application that was built with the use of the ReactJS framework. The interface was implemented as a responsive application to optimize the user experience and facilitate easy navigation of the user to retrieve genomic variant allele frequencies of pathogenic variants or PGx biomarkers.

The website consists of two standard components, namely the menu and the footer, as well as one variable component that depends on user navigation. The menu consists of different links, such as *Home*, *Documentation*, *Map*, *Diseases*, *Genes*, *Frequency*, and *Researchers* (Figure 1a), and by clicking on them, the user is able to navigate through the different pages, most of which comprise data filters. Some of the menu options contain dropdown submenus, which automatically display once the user hovers the cursor over it. Apart from the menu, the header component also includes the registration buttons, where new and existing users can sign up or sign in, respectively. For screens that are smaller than 700 pixels, such as tablets and mobile phones, where the menu bar would not be ideal for the user, it was replaced with a menu button, also known as a hamburger icon (Figure 1b). By tapping on this icon, the user is able to see the menu as a dropdown list (Figure 1b) and by tapping each of the options he/she can navigate through the different pages or unravel the submenus where applicable.

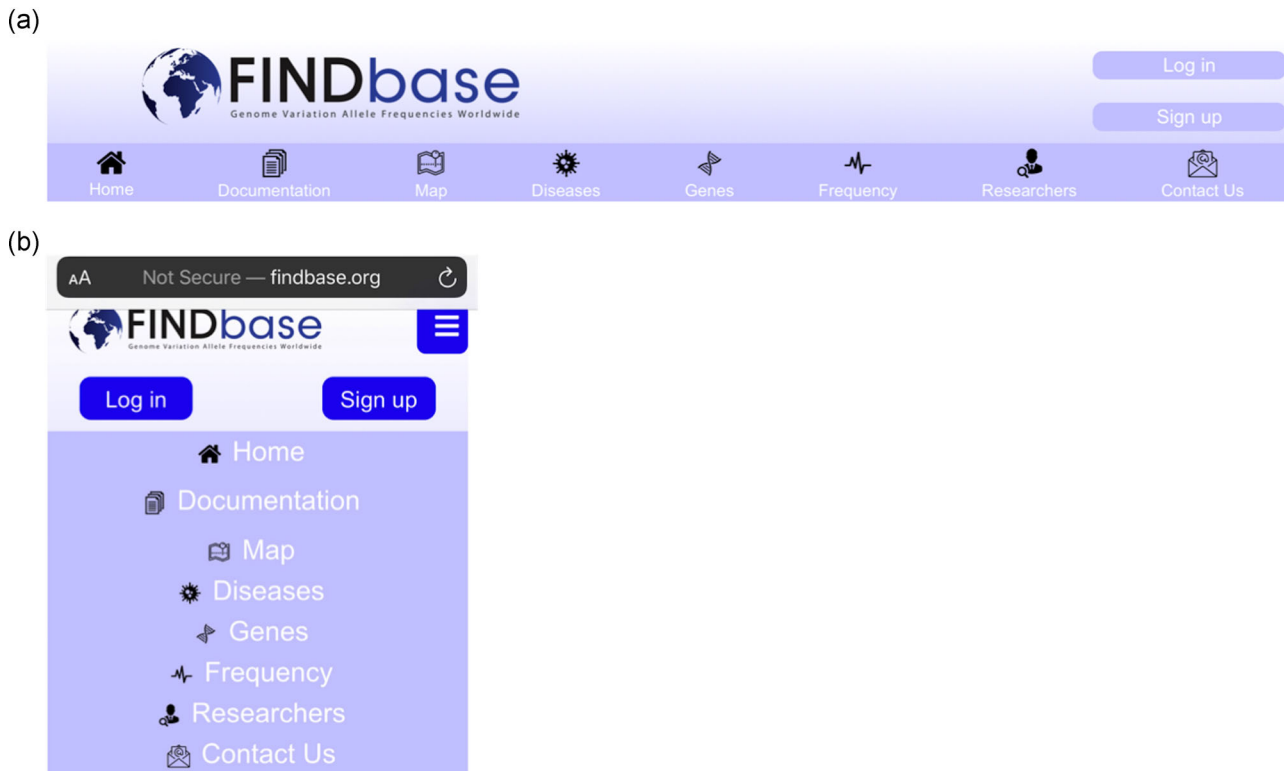
The main part between the header and the footer varies and includes different data filters. Those filters are designed to facilitate

data querying and to expedite data output, the latter delivered in a tabular format, which can be paginated for lengthy data outputs.

## 2.2 | Client side, react, and user interface

We used the ReactJS framework to implement the front-end interface. There are five different filters to help the user toward data querying, which is implemented on a different page, described as following.

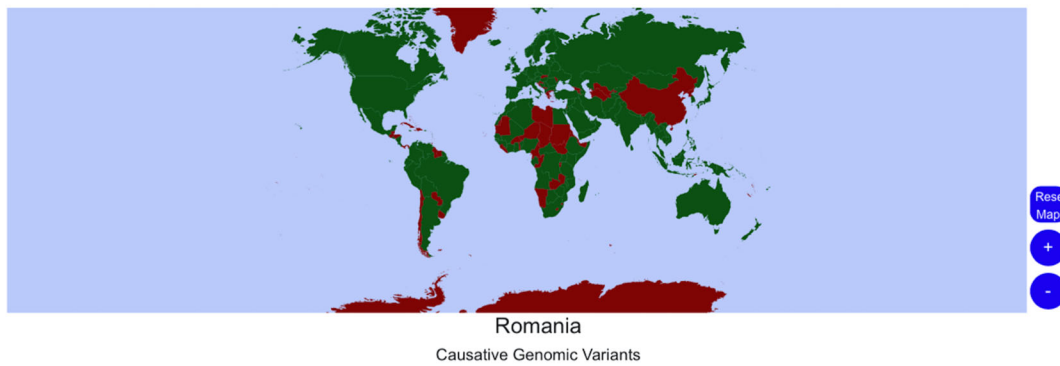
1. Data querying using the *World Map*. In this filter, as shown in Figure 2a, a world map is depicted where the user can zoom in and out using the corresponding buttons on the right side. The user is also able to slide the map to any directions just by dragging it. The countries with green color are clickable, indicating that for those populations, data are available, while no data are available for countries shown in red. Once the user clicks on a country on



FINDbase worldwide is an online resource documenting frequencies of clinically relevant genomic variants, namely **pathogenic variants leading to inherited disorders and pharmacogenomic biomarkers**, in various populations worldwide. The initial data came from previously published reports as well as from unpublished information contributed from individual researchers prior of publication.

**FIGURE 1** The new FINDbase user interface for desktops (a) and mobile devices (b). The various menu items are also graphically depicted. The Login/Sign Up options are displayed at the top right part of the menu bar (see also text for details)

Map



Showing rows 1 to 10 of 23 10 ▾

1 2 3 Next

Gene	Gene Reference	Variant	rs Number	Allele Frequency	Number of Chromosomes	Diseases	Pubmed ID	Researcher ID
CFTR	NM_000492	c.1898+1G>A		0.0051	64	Cystic Fibrosis		E-5147-2012
CFTR	NM_000492	p.R553X		0.0051	64	Cystic Fibrosis		E-5147-2012
CFTR	NM_000492	p.G542X		0.0051	64	Cystic Fibrosis		E-5147-2012
PAH	NM_000277	c.1315+1G>A		0.0674	89	Phenylketonuria		E-5147-2012
PAH	NM_000277	c.168+5G>T		0.0455	22	Phenylketonuria		E-5147-2012
CFTR	NM_000492	p.N1303K		0.0051	64	Cystic Fibrosis		E-5147-2012
CFTR	NM_000492	p.W1282X		0.0051	64	Cystic Fibrosis		E-5147-2012
CFTR	NM_000492	c.3272-26A>G		0.0051	64	Cystic Fibrosis		E-5147-2012
PAH	NM_000277	p.P225T		0.1818	22	Phenylketonuria		E-5147-2012
CFTR	NM_000492	p.I148T		0.0051	64	Cystic Fibrosis		E-5147-2012

**FIGURE 2** Data querying interface using the Population Map option. The user can zoom in and out using the options available at the bottom right corner of the map and can select a population from the countries in the world map, colored in green (see also text). In the case of this particular query, the Romanian population is selected, a query that returns 23 records in a tabular format

the world map, ReactJS will fetch all available data in the data warehouse for this population and will display them in the tables below. In the top table, allele frequency data for pathogenic variants leading to inherited diseases are provided, whereas the bottom table shows allele frequency data for PGx biomarkers.

2. Data querying per Disease. This filter is a simple input with the autocomplete feature for all the existing diseases in our database (Figure 3a). Data output is in the form of a single table that will show the allele frequencies of all variants leading to the selected genetic disease in the available populations in the data warehouse.

Diseases

Please search the disease on the following field:

Showing rows 1 to 25 of 1769 25 ▾

1 2 3 4 5 Next Last

Population	Gene	Gene Reference	Variant	rs Number	Allele Frequency	Number of Chromosomes	Pubmed ID	Researcher ID
Spanish	CFTR	NM_000492	p.R1066C		0.0124	1435		
Italian	CFTR	NM_000492	c.3849+10kbC>T		0.0003	287		
French	CFTR	NM_000492	p.D1270N/p.R74W		0.0005	6992		
French	CFTR	NM_000492	c.405+4A>G		0.0001	6992		
French	CFTR	NM_000492	p.N1303K		0.004	2641		
Belarussian	CFTR	NM_000492	p.G542X		0.012	174		
Spanish	CFTR	NM_000492	p.A561E		0.0013	1435		
Tunisian	CFTR	NM_000492	c.2766del8		0.0233	47		
Slovene	CFTR	NM_000492	p.G542X		0.0055	95		
French	CFTR	NM_000492	c.3869insG		0.0001	6992		
Belgian	CFTR	NM_000492	c.3272-1G>A		0.037	20		
French	CFTR	NM_000492	c.3849+4A>G		0.0001	6992		

**FIGURE 3** Data querying interface using the Disease option. The user can select the Disease from the specified Disease list. In the case of this particular query, query selection for Cystic Fibrosis, the output includes 1,769 records in a tabular format

3. Data querying per *Gene*: In this filter, there are two autocomplete fields, where the user can query for different variants in genes leading to genetic diseases (left-hand filter) or pharmacogenes (right-hand filter; Figure 4a). The output of this query is in the form of a table that shows the allele frequencies of this gene across all available populations. Notably, the table is a dynamic element that changes its headers depending on whether a disease-associated gene or a pharmacogene is queried.
4. Data querying per *Allele Frequency*: This filter consists of two parts, as shown in Figure 5a. In the top part, there is an autocomplete input field, where the user selects the desired country, for which data are available in the FINDbase data warehouse. Subsequently, the user can define the frequency range of the variant or allele of interest using a two-sided slider between 0.00 and 0.50. Once the frequency range is selected, the data output is given in a single table format.
5. Data querying per *Researchers*: FINDbase data contribution is based on microattribution, which encourages data sharing in the public domain (Giardine et al., 2011; Patrinos et al., 2012). Based on this concept, this filter allows querying of the data contributed to the database per contributor, based on his or her ResearcherID. Again, this filter is a single autocomplete input field, similar to the one available for Diseases, and data querying is based on the ResearchersID. For this filter, data output is provided in two different tables, one for pathogenic variants leading to genetic diseases and another table for PGx biomarkers.

## 2.3 | Registration and data entry

The updated FINDbase uses a tiered structure for data entry and modification.

1. Administrators: They have full access rights to all database contents, functionalities, and menus. This category has rights for data entry as well as modification, registration, and account activation for the second category of registered users. This category is responsible for managing the overall development and maintenance of FINDbase.
2. Registered users: This is the most important group in the next-generation FINDbase and refers to those users that tangibly contribute to the database content and maintenance. As previously mentioned, FINDbase is based on microattribution, which aims to incentivize genomic data submission to the public domain (Giardine et al., 2011). As such, registered users are individuals that have deposited clinically relevant genomic variant allele frequency data in FINDbase. These users have access to all data, that is, also to those data that are not available to unregistered users. This kind of role can be requested using the “Sign Up” button (Figure 6). The request is then automatically sent to the administrator, who will guide the user through the next steps to complete the registration and contribute data to FINDbase. This user group is crucial for the long-term sustainability of the resource.

3. Plain (nonregistered) users: This user group can only access publicly available FINDbase data.

## 3 | RESULTS

### 3.1 | Technical aspects, data querying, and database architecture

The implementation of modern development tools is among the major updates of the next-generation FINDbase. The updated back-end approach utilizing Django as well as the implementation of the front-end utilizing ReactJS, provide a notably quicker response rate, that is in less than a second. Concretely, ReactJS constitutes a synchronous, solid, and fast framework being able to quickly connect to a server and simultaneously fetch responses for multiple different queries.

The representation of the back-end unit is illustrated in the database scheme of Figure 7. The main implementation tool for FINDbase utilizes the Django REST framework and the used version constitutes the newest Django approach, which provides server link queries to receive the data. The implementation of search queries relies on knowledge of the corresponding database structure to connect the tables via join operations and finally to get the requested information that the pages provide.

Another advantage of the abovementioned frameworks concerns their widespread use among the majority of developers and specifically, which they can provide modern solutions for numerous different applications with the aim of improving and increasing these systems' stability. Furthermore, this aspect ensures that none of these frameworks will be deprecated in the next years and hence, the proposed system will maintain its performance.

### 3.2 | Data content enrichment

Contrary to the previous FINDbase data content updates in 2010 (Georgitsi et al., 2011a; 2011b), 2013 (Papadopoulos et al., 2014) and 2016 (Viennas et al., 2017), the current content update did not only include data curation, updates, and corrections, but also extensive data enrichment. In particular, FINDbase data collection was enriched mostly with PGx biomarker allele frequencies, derived from two large genotyping efforts that evaluated the prevalence of clinically actionable PGx biomarkers in more than 20, mostly European, populations (see Mizzi et al., 2016; Patrinos et al., 2012; Petrović, Pešić, & Lauschke, 2020). This data update is of relevance for geneticists, drug developers, and clinicians, as it allows to assess the prevalence of PGx biomarkers that can impact the efficacy or toxicity of a given drug in the population of interest, thus guiding the prescription of PGx testing to individualize drug treatment and assisting in finding suitable countries for clinical trials of novel drug candidates.

As stated above, FINDbase data records include two different modules, namely (a) pathogenic genomic variants and (b) PGx biomarkers. Details of the data content in both data modules are



(a)

## Genes

Please search for genes involved in Genetic Diseases:

CFTR

CFTR

Please search for pharmacogenes:

Showing rows 1 to 10 of 1746

10

1

2

3

4

5

Next

Last

Population	Gene	Gene Reference	Variant	rs Number	Allele Frequency	Number of Chromosomes	Diseases	Pubmed ID	Researcher ID
Spanish	CFTR	NM_000492	p.R1066C		0.0124	1435	Cystic Fibrosis		E-5147-2012
Italian	CFTR	NM_000492	c.3849+10kbC>T		0.0003	287	Cystic Fibrosis		E-5147-2012
French	CFTR	NM_000492	p.D1270N/p.R74W		0.0005	6992	Cystic Fibrosis		E-5147-2012
French	CFTR	NM_000492	c.405+4A>G		0.0001	6992	Cystic Fibrosis		E-5147-2012
French	CFTR	NM_000492	p.N1303K		0.004	2641	Cystic Fibrosis		E-5147-2012
Belarussian	CFTR	NM_000492	p.G542X		0.012	174	Cystic Fibrosis		E-5147-2012
Spanish	CFTR	NM_000492	p.A561E		0.0013	1435	Cystic Fibrosis		E-5147-2012
Tunisian	CFTR	NM_000492	c.2766del8		0.0233	47	Cystic Fibrosis		E-5147-2012
Slovene	CFTR	NM_000492	p.G542X		0.0055	95	Cystic Fibrosis		E-5147-2012
French	CFTR	NM_000492	c.3869insG		0.0001	6992	Cystic Fibrosis		E-5147-2012

(b)

## Genes

Please search for genes involved in Genetic Diseases:

Please search for pharmacogenes:

CYP2C9

CYP2C9

Showing rows 1 to 10 of 533

10

1

2

3

4

5

Next

Last

Populati...	Gene	Gene Reference	Variant	rs Number	Allele Frequency	Number of Chromosomes	Drug	Pubmed ID	Researcher ID	*allele
Ukrainian	CYP2C9	NM_000771	c.1075A>C	rs1057910	0.066	104	Flurbiprofen		F-6144-2016	*3
Ukrainian	CYP2C9	NM_000771	c.1075A>C	rs1057910	0.066	104	Celecoxib		F-6144-2016	*3
Ukrainian	CYP2C9	NM_000771	c.1075A>C	rs1057910	0.066	104	Warfarin		F-6144-2016	*3
South African Xhosa	CYP2C9	NM_000771	c.430C>T	rs1799853	0.095	70	Flurbiprofen		F-6144-2016	*2
South African Xhosa	CYP2C9	NM_000771	c.430C>T	rs1799853	0.095	70	Warfarin		F-6144-2016	*2
American	CYP2C9	NM_000771	c.430C>T	rs1799853	0.194	160	Flurbiprofen	18252229	F-2749-2012	*2
American	CYP2C9	NM_000771	c.430C>T	rs1799853	0.194	160	Warfarin	18252229	F-2749-2012	*2
Brazilian	CYP2C9	NM_000771	c.430C>T	rs1799853	0.124	780	Flurbiprofen	21692610	F-2749-2012	*2
Brazilian	CYP2C9	NM_000771	c.430C>T	rs1799853	0.124	780	Warfarin	21692610	F-2749-2012	*2
Cypriot	CYP2C9	NM_000771	c.430C>T	rs1799853	0.1923	80	Flurbiprofen		F-6144-2016	*2

**FIGURE 4** Data querying interface using the Gene option. The user can select between genes involved in genetic disease and pharmacogenes. In the case of the queries below, querying for G6PD (a), the query returns 16 records, while in the case of CYP3A5 (b), the query returns 142 records, both in a tabular format. In the latter case of the pharmacogene, the variant's alternative nomenclature (as a \* allele) is also displayed at the last column in the table

## Allele Frequency

First, select a country: 

Now use the slider to specify the Allele Frequency

Showing rows 1 to 10 of 10 

1

Gene	Gene Reference	Variant	rs Number	Allele Frequency	Number of Chromosomes	Pubmed ID	Researcher ID
HBB	NM_000518	c.315+1G>A		0.4706	102	17994378	I-8796-2012
PON1	NM_000446	c.163T>A	rs854560	0.41	264	19654933	B-3824-2010
PON1	NM_000446	c.575A>G	rs662	0.31	264	19654933	B-3824-2010
HBA	NM_000517	Z84721.1g.34164_37964del3801		0.4248	113	19657838	I-8796-2012
HBB	NM_000518	c.315+1G>A		0.35	120	19373586	I-8796-2012
GJB2	NM_004004	c.35delG		0.4334	30	11968091	E-5147-2012
HBB	NM_000518	c.446G>A		0.3082	2234		A-2391-2010
PON1	NM_000446	c.-108C>T	rs705379	0.473	330	18433845	B-3824-2010
NAT2	NM_000015	c.341T>C	rs1801280	0.33	176	15316701	B-5365-2010
HBB	NM_000518	c.112delT		0.3385	130	17654071	I-8796-2012

**FIGURE 5** Data querying interface using the Allele Frequency option. The user first needs to select the Population (Iran in this particular query) and then determine the allele frequency range (low- and high-frequency value) using the two-sided slider. In the case of this particular query (variants in the Iranian population with an allele frequency range between 30% and 50%), the query output includes 10 records in a tabular format., corresponding to the frequent variants documented for this population

(a)

## Sign Up

First Name

Last Name

Email Address

Password

Re-enter Password

(b)

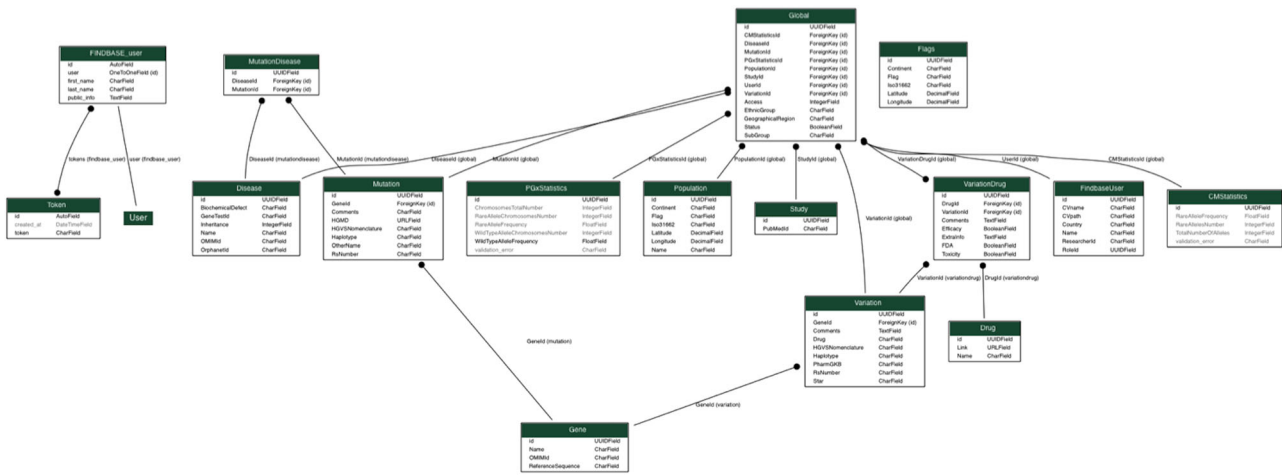
Not a member? [Sign up](#)

Sign in to continue

Email

Password

**FIGURE 6** Modules for signing up (a) and logging in to FINDbase (b)



**FIGURE 7** Database schema of the next-generation FINDbase

summarized in Table 1. An important feature of the updated FINDbase content is the split of the data content into a public section, which are open to all users, as well as a private section, which can only be accessed by registered users and administrators.

- (a) The Public section includes 8,895 publicly available data entries of which 6,035 correspond to pathogenic genomic variants distributed across 25 genes and 2,860 correspond to PGx biomarkers in 18 genes (Figure 8 and Table 1).
- (b) The Private section includes additional 21,578 data records, which can be accessed by FINDbase users that have tangibly contributed with data content. This section currently comprises only of PGx biomarkers in 233 genes (Table 1).

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

**4 | DISCUSSION**

Comprehensive documentation of the extant genetic ethnogeographic heterogeneity across different populations is key for the accurate provision of personalized medicine services. General databases such as ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar>) provide information about genotype-phenotype relationships but do not contain population-specific variant frequency data. In contrast, genomic databases, such as gnomAD (<https://gnomad.broadinstitute.org>) provide important information regarding variant frequencies. However, these data are not linked to phenotypic outcomes and frequency information is highly aggregated and not available for individual countries. In the past, several databases have tried to address this gap of providing phenotypically annotated variant allele frequencies but with limited to no success. One such resource is the Allele Frequency Net Database (<http://www.allelefreqencies.net/default.asp>), which only documents the allele frequencies of HLA alleles and other immune-related genes;

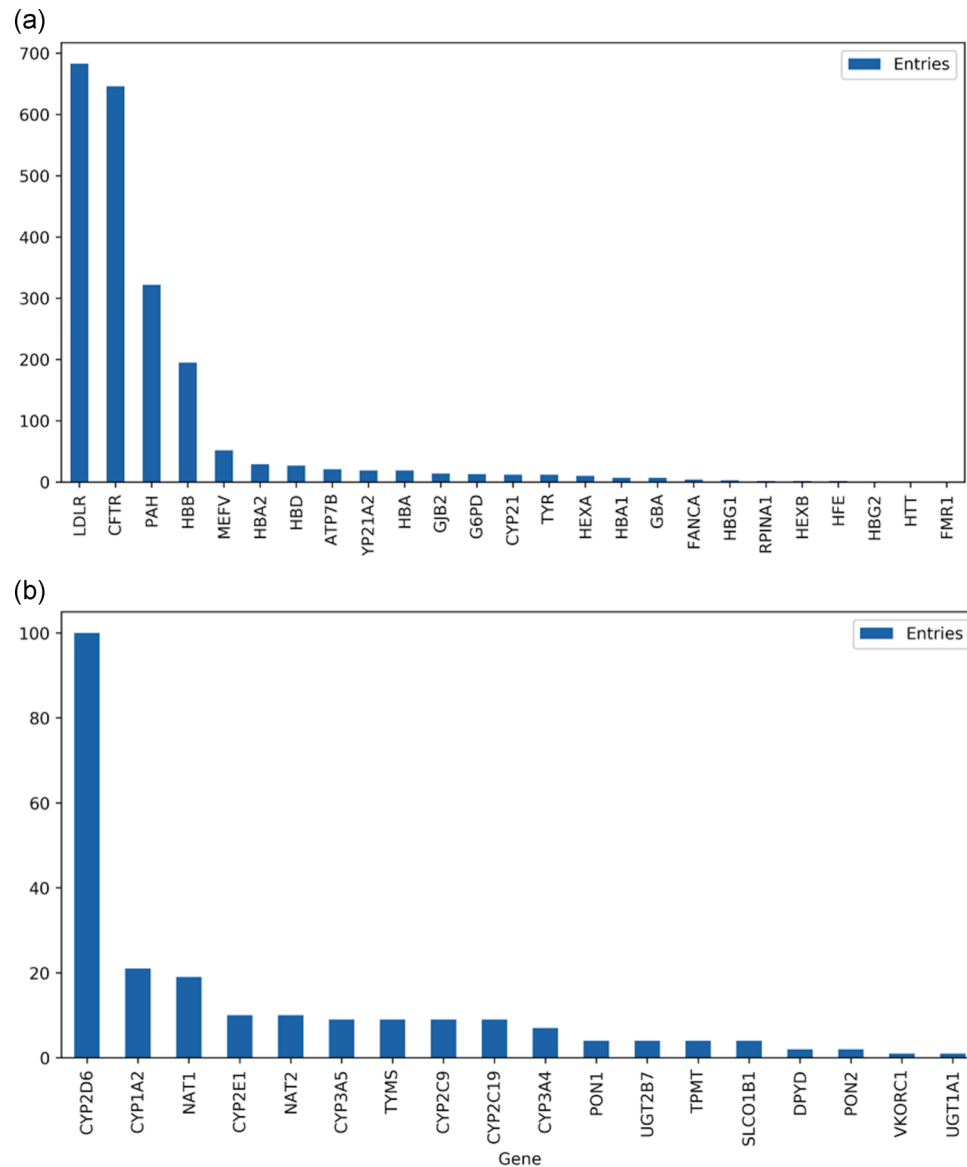
as such its content is somewhat limited. Another resource is the Allele Frequency Database (ALFRED; <https://alfred.med.yale.edu/alfred/index.asp>), which has been designed to make allele frequency data on anthropologically defined human population samples readily available to the scientific community and to link these polymorphism data to the molecular genetics-human genome databases. This resource, however, does not provide allele frequency data for clinically relevant genomic variants.

The concept of FINDbase is very different from other databases that have attempted to address this task. FINDbase focuses exclusively on clinically relevant and actionable variations for which population-specific frequency information is available and, as such, covers a niche that is not systematically covered in other genomic resources. Also, these data nicely complement with data of other resources, such as the CFRT2 ([www.cftr2.org](http://www.cftr2.org)) or the database of the International Society of Gastrointestinal Hereditary Tumors ([www.insight-group.org](http://www.insight-group.org)), which also take into account the penetrance of pathogenic variants in a more comprehensive and structured manner, compared with other LSDBs. The resulting utility for researchers and clinicians is supported by the sustainability of this resource, its

**TABLE 1** Summary of the next-generation FINDbase data content

Features	Public records	Private records	Total records
Populations	174	33	207
Genes (diseases)	25	0	25
Pharmacogenes	18	233	251
Genes (total)	43	233	276
Causative genomic variants	6,035	0	6,035
Pharmacogenomic variants	2,860	21,578	24,438
Common variants (>10%)	1,905	5,221	7,126
Rare variants (<1%)	4,568	11,895	16,463
Records of genomic variants (total)	8,895	21,578	30,473





**FIGURE 8** Histogram depicting the number of the pathogenic variants (a) and of the pharmacogenomic biomarkers (b), deposited in the public version of FINDbase, clustered per gene

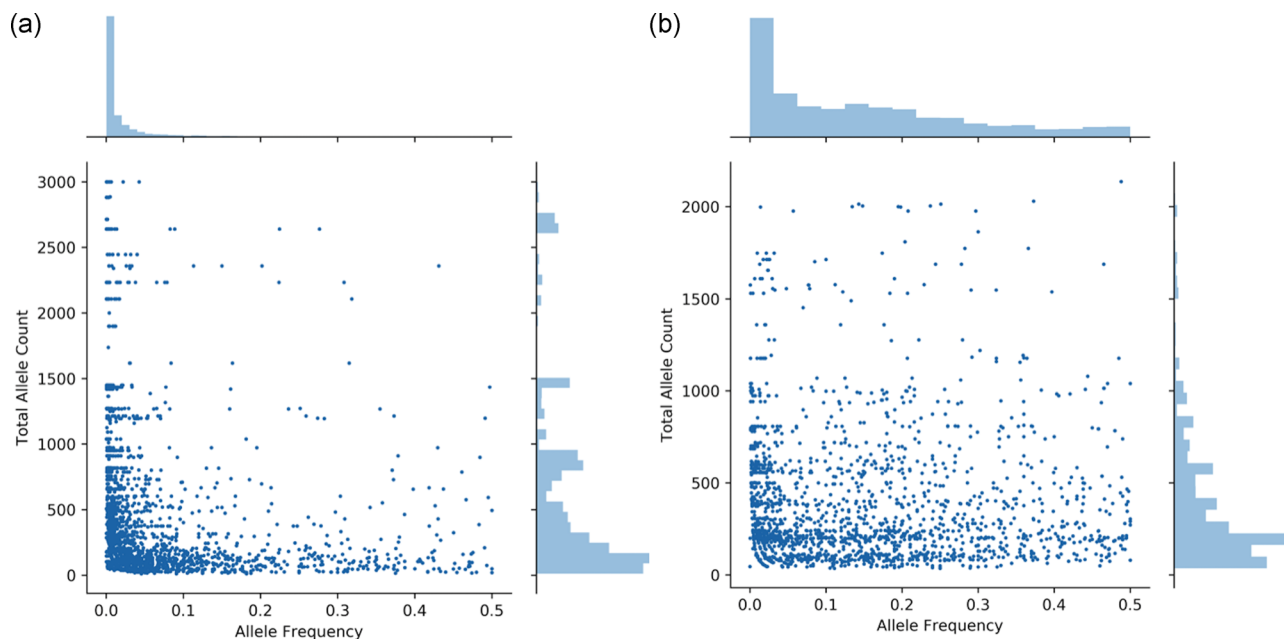
continuous update and upgrade and uninterrupted funding since its inception in 2006, one important feature of successful genomic databases and resources (Patrinos & Brookes, 2005).

The recent FINDbase update was initiated in mid-2018 and covered different aspects, aiming (a) to significantly enrich the existing data collection with PGx biomarker allele frequency data, (b) to upgrade the entire database structure and querying engine to optimize speed and efficiency of data output, (c) to refurbish the database front end to improve the end user experience, and (d) to strengthen the implementation of microattribution (Giardine et al., 2011; Patrinos et al., 2012), by rewarding genomic data sharing with extended access permissions. The latter is of utmost importance since it ensures the long-term viability and continuous update of the resource. Notably, funding for the maintenance and upgrade of FINDbase is secured for another 4-year period from European projects and private sponsors, which corroborates the value and impact of the database for the

scientific community, since opportunities for funding database projects are difficult to secure (Patrinos & Brookes, 2005).

While data updates and enrichments are crucial for genomic databases, we feel that both the redesign of the database querying engine as well as linking genomic variant allelic frequency data sharing with database access were the hallmarks of the next-generation FINDbase that would further strengthen its sustainability. In particular, the new database querying engine not only significantly expedited the data output to milliseconds, compared with the lengthy waiting times of the previous Pivot viewer-based interface but also, and most importantly, allowed the significant expansion of FINDbase data collection to more than threefold, compared to the previous version.

The updated FINDbase design does not allow yet to display autonomous NEGDBs per population, as implemented in the previous update (Viennas et al., 2017), which are either hosted separately in the FINDbase server or remotely but as part of the future design, this



**FIGURE 9** (a) Scatter plot depicting the minor allele frequency and the total allele counts for pathogenic variants. The plot depicts 5,269 records from pathogenic variants both in the public as well as the private FINDbase data collection. (b) Scatter plot depicting the minor allele frequency and the total allele counts for pharmacogenomic biomarkers, depicting 1,964 records both in the public as well as the private FINDbase data collection. In both plots, outliers have been removed

will be further implemented to allow independent management of these NEGDBs locally by experts, which would then update the central FINDbase content. The latter will be facilitated by the fact that both the NEGDBs and the central FINDbase databases will be operating under the same frame, which contributes toward NEGDB data content uniformity (Patrinos, 2006). This design will also allow local experts that will serve as NEGDB curators to maintain NEGDB-specific pages that will differ from NEGDB to NEGDB and include population-specific information, such as local human genetic societies, conferences of interest, and so forth, which may not necessarily appear in FINDbase itself.

The numerous large-scale genome projects, which are currently being implemented in several countries worldwide create unprecedented opportunities for expanding the FINDbase data collection. To accommodate the wealth of these data, FINDbase architecture will have to be further modernized and adapted not only to store but also to dynamically analyze genomic data, derived from next-generation sequencing in a continuous manner. As such, algorithms are needed that automatically calculate the allelic frequencies of genomic variants in clinically actionable genes, leading, for example, to monogenic diseases, physiological traits, and/or pharmacogenes, while guaranteeing data anonymity. These NGS data will be again contributed using the microattribution approach, flagged with a certain tag (e.g., an “mA” tag), and deposited using a DOI, hence providing credit to data contributors and curators, apart from their access to the private FINDbase data collection. Furthermore, we plan the implementation of additional forms of unambiguous contributor identifications, such as ORCID identifiers (<http://www.orcid.org>), in addition to the currently included ResearcherID. This increase in data

coverage should be accompanied by additional data visualization tools and possibly crosstalk with existing resources, such as ClinVar and gnomAD, which would increase value, particularly for the epidemiology of rare variants, which already constitute a significant proportion in the existing FINDbase data collection (Figure 9) and reveal a possible high prevalence of certain variants in specific population isolates.

In summary, the major 2019 update of FINDbase improves accessibility, functionality, and user experience and optimizes database efficiency and runtimes due to the implementation of a modern Javascript framework. Furthermore, the substantially extended content ensures that the updated FINDbase constitutes a state-of-the-art genomic database that provides important information about the prevalence of clinically relevant variants, which constitutes a cornerstone for genomic medicine and precision public health.

#### ACKNOWLEDGMENTS

This study was partly funded by a Greek General Secretariat of Research and Technology grant (Reinforcing of the Research and Innovation Infrastructure, ELIXIR-GR, Contract No: 5002780) and European Commission grants (GEN2PHEN; FP7-200754 and RD-Connect; FP7-305444) to GPP. We are indebted to Dr Aggeliki Balasopoulou, Ms Konstantina Giannakopoulou, Ms Dimitra Mandraki, Ms Aspasia Skarpathioti, Ms Maria Opoliopi, Mr Anastasio Bitsako, and Ms Panagiota Grypioti for expert data curation.

#### CONFLICT OF INTERESTS

Volker M. Lauschke is a shareholder of HepaPredict AB. Other authors declare that there are no conflict of interests.

## DATA AVAILABILITY STATEMENT

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 parties. FINDbase users have full access to the public database content, while registered users, for example, individual researchers and research groups that have deposited clinically relevant genomic variant allele frequency data in FINDbase, have access to the entire database data, that is, also to those data that are not available to unregistered users.

## ORCID

Alexandros Kanterakis  <http://orcid.org/0000-0003-4276-0115>

Volker M. Lauschke  <http://orcid.org/0000-0002-1140-6204>

George P. Patrinos  <http://orcid.org/0000-0002-0519-7776>

## REFERENCES

- van Baal, S., Kaimakis, P., Phommarinh, M., Koumbi, D., Cuppens, H., Riccardino, F., ... Patrinos, G. P. (2007). FINDbase: A relational database recording frequencies of genetic defects leading to inherited disorders worldwide. *Nucleic Acids Research*, 35, D690–D695.
- Fokkema, I. F., Taschner, P. E., Schaafsma, G. C., Celli, J., Laros, J. F., & den Dunnen, J. T. (2011). LOVD v.2.0: The next generation in gene variant databases. *Human Mutation*, 32, 557–563.
- Georgitsi, M., Viennas, E., Gkantouna, V., van Baal, S., Petricoin, E. F., Poulas, K., ... Patrinos, G. P. (2011a). FINDbase: A worldwide database for genetic variation allele frequencies updated. *Nucleic Acids Research*, 39, D926–D932.
- Georgitsi, M., Viennas, E., Gkantouna, V., Christodouloupoulou, E., Zagoriti, Z., Tafrali, C., ... Patrinos, G. P. (2011b). Population-specific documentation of pharmacogenomic markers and their allelic frequencies in FINDbase. *Pharmacogenomics*, 12, 49–58.
- Giardine, B., Borg, J., Higgs, D. R., Peterson, K. R., Philipson, S., Maglott, D., ... Patrinos, G. P. (2011). Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*, 43, 295–301.
- Giardine, B., Borg, J., Viennas, E., Pavlidis, C., Moradkhani, K., Joly, P., ... Patrinos, G. P. (2014). Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Research*, 42, D1063–D1069.
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44, D862–D868.
- Mitropoulou, C., Webb, A. J., Mitropoulos, K., Brookes, A. J., & Patrinos, G. P. (2010). Locus-specific database domain and data content analysis: Evolution and content maturation toward clinical use. *Human Mutation*, 31, 1109–1116.
- Mizzi, C., Dalabira, E., Kumuthini, J., Dzimiri, N., Balogh, I., Başak, N., ... Patrinos, G. P. (2016). A European spectrum of pharmacogenomic biomarkers: Implications for clinical pharmacogenomics. *PLoS One*, 11, e0162866.
- Papadopoulos, P., Viennas, E., Gkantouna, V., Pavlidis, C., Bartsakoulia, M., Ioannou, Z. M., ... Patrinos, G. P. (2014). Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies. *Nucleic Acids Research*, 42, D1020–D1026.
- Patrinos, G. P. (2006). National and Ethnic Mutation databases: Documenting populations' genography. *Human Mutation*, 27, 879–887.
- Patrinos, G. P., Al Aama, J., Al Aqeel, A., Al-Mulla, F., Borg, J., Devereux, A., ... Cotton, R. G. H. (2011). Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection. *Human Mutation*, 32, 2–9.
- Patrinos, G. P., & Brookes, A. J. (2005). DNA, diseases and databases: Disastrously deficient. *Trends in Genetics*, 21, 333–338.
- Patrinos, G. P., Cooper, D. N., van Mulligen, E., Gkantouna, V., Tzimas, G., Tatum, Z., ... Mons, B. (2012). Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Human Mutation*, 33, 1503–1512.
- Petrović, J., Pešić, V., & Lauschke, V. M. (2020). Frequencies of clinically important CYP2C19 and CYP2D6 alleles are graded across Europe. *European Journal of Human Genetics*, 28, 88–94.
- Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133, 1–9.
- Viennas, E., Komianou, A., Mizzi, C., Stojiljkovic, M., Mitropoulou, C., Muilu, J., ... Patrinos, G. P. (2017). Expanded national database collection and data coverage in the FINDbase worldwide database for clinically relevant genomic variation allele frequencies. *Nucleic Acids Research*, 45, D846–D853.

**How to cite this article:** Kounelis F, Kanterakis A, Kanavos A, et al. Documentation of clinically relevant genomic biomarker allele frequencies in the next-generation FINDbase worldwide database. *Human Mutation*. 2020;1–11. <https://doi.org/10.1002/humu.24018>