HAL
open science

# Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection

George P. Patrinos, Jumana Al Aama, Aida I Al Aqeel, Fahd Al-Mulla, Joseph Borg, Andrew Devereux, Alex E. Felice, Finlay Macrae, Makia J Marafie, Michael B Petersen, et al.

## To cite this version:

George P. Patrinos, Jumana Al Aama, Aida I Al Aqeel, Fahd Al-Mulla, Joseph Borg, et al.. Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection. Human Mutation, 2010, 32 (1), pp.2. 10.1002/humu.21397 . hal-00599480

**HAL Id: hal-00599480**

**https://hal.science/hal-00599480v1**

Submitted on 10 Jun 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Human Mutation

# Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection

| | |
|---|---|
| Journal: | *Human Mutation* |
| Manuscript ID: | humu-2010-0429.R1 |
| Wiley - Manuscript type: | Special Article |
| Date Submitted by the Author: | 15-Oct-2010 |
| Complete List of Authors: | Patrinos, George; University of Patras, School of Health Sciences, Dept of Pharmacy; Erasmus MC, Bioinformatics<br>Al Aama, Jumana; King Abdulaziz University, Department of Genetic Medicine, Faculty of Medicine and Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders<br>AL Aqeel, Aida; Riyadh Military Hospital and Stem Cell therapy program, King Faisal Specialist Hospital and Research Center, Department of Pediatrics<br>Al-Mulla, Fahd; Kuwait University, Molecular Pathology Unit,<br>Borg, Joseph; University of Malta, Department of Physiology and Biochemistry, Laboratory of Molecular Genetics<br>Devereux, Andrew; NGRL Manchester, St Mary's Hospital, Department of Medical Genetics<br>Felice, Alex; University of Malta, Department of Physiology and Biochemistry, Laboratory of Molecular Genetics<br>Macrae, Finlay; University of Melbourne, The Royal Melbourne Hospital,, Department of Colorectal Medicine and Genetics<br>Marafie, Makia; Kuwait Medical Genetics Centre, Maternity Hospital, Sabah Medical Area, Clinical Genetics,<br>Petersen, Michael; Institute of Child Health, Department of Genetics<br>Qi, Ming; Center for Genetic and Genomic Medicine, The First Affiliated Hospital of Zhejiang University School of Medicine, James Watson Institute of Genomic Sciences<br>Ramesar, Raj; University of Cape Town Medical school, Department of Human genetics<br>Zlotogora, Joel; Ministry of Health, Community genetics<br>Cotton, Richard; Genomic Disorders Research Centre, Howard Florey Institute |
| Key Words: | Developing countries, National/ethnic mutation databases, Populations, Genetic variation |

SCHOLARONE™
Manuscripts

For Peer Review

# Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection

George P. Patrinos [1,*], Jumana Al Aama [2], Aida Al Aqeel [3], Fahd Al-Mulla [4], Joseph Borg [5], Andrew Devereux [6], Alex E. Felice [7], Finlay Macrae [8], Makia J. Marafie [9], Michael B. Petersen [10], Ming Qi [11], Rajkumar S. Ramesar [12], Joel Zlotogora [13], and Richard G. Cotton [14]

[1] University of Patras, School of Health Sciences, Department of Pharmacy, Patras, Greece;

[2] Department of Genetic Medicine, Faculty of Medicine and Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia;

[3] Department of Pediatrics, Riyadh Military Hospital and Stem Cell therapy program, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia;

[4] Molecular Pathology Unit, Kuwait University, Kuwait;

[5] University of Malta, Faculty of Health Sciences, Department of Applied Biomedical Sciences, Msida, Malta;

[6] NGRL Manchester, Department of Medical Genetics, St Mary's Hospital, Manchester, United Kingdom;

[7] University of Malta, Department of Physiology and Biochemistry, Laboratory of Molecular Genetics, Msida, Malta;

[8] Department of Colorectal Medicine and Genetics and University of Melbourne, The Royal Melbourne Hospital, Australia;

For Peer Review

[9] Clinical Genetics, Kuwait Medical Genetics Centre, Maternity Hospital, Sabah Medical Area, Kuwait;

[10] Institute of Child Health, Department of Genetics, Athens, Greece;

[11] Center for Genetic and Genomic Medicine, The First Affiliated Hospital of Zhejiang University School of Medicine and James Watson Institute of Genomic Sciences, P. R. China;

[12] MRC Human Genetics Research Unit, Division of Human Genetics, Institute for Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa;

[13] Department of Community Genetics, Public Health Services, Ministry of Health, Jerusalem, Israel;

[14] Genomic Disorders Research Centre, Howard Florey Institute, Melbourne, Australia & Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Australia

For Peer Review

Running head: Recommendations for genetic variation data capture in developing countries

* Corresponding author at:

Department of Pharmacy, School of Health Sciences, University of Patras,

University Campus, Rion, GR-265 04, Patras,

Greece

Telephone/Fax: +30-2610-969.834,

E-mail: gpatrinos@upatras.gr

## Abstract

Developing countries have significantly contributed to the elucidation of the genetic basis of both common and rare disorders, providing an invaluable resource of cases due to large family sizes, consanguinity and potential founder effects. Moreover, the recognised depth of genomic variation in indigenous African populations, reflecting the ancient origins of humanity on the African continent, and the effect of selection pressures on the genome, will be valuable in understanding the range of both pathological and non-pathological variations. The involvement of these populations in accurately documenting the extant genetic heterogeneity is more than essential. Developing nations are regarded as key contributors to the Human Variome Project (HVP; http://www.humanvariomeproject.org), a major effort to systematically collect mutations that contribute to or cause human disease and create a cyber infrastructure to tie databases together. However, biomedical research has not been the primary focus in these countries even though such activities are likely to produce economic and health benefits for all. Here, we propose several recommendations and guidelines to facilitate participation of developing countries in genetic variation data documentation, ensuring an accurate and comprehensive worldwide data collection. We also summarize a few well-coordinated genetic data collection initiatives that would serve as paradigms for similar projects.


**Key words:** Developing countries, national/ethnic mutation databases, populations, genetic variation

For Peer Review

## Introduction

The recent elucidation of the genome sequence of several individuals, such as Causacians (Jim Watson [Wheeler et al., 2008] and Craig Venter [Levy et al., 2007]), West African [Bentley et al., 2008], and Han Chinese [Wang et al., 2008], suggests that a large number of SNPs and other sequence variation exists in the human population. Estimates from African genetic diversity and the Pan Asian SNP initiative suggest that approximately 80–90% of human genomic variation resides in the world's developing countries. It is therefore obvious that there is a huge need to comprehensively document the extant genetic variation in various ethnic groups and populations around the globe to ensure maximal data coverage.

Since 2002, there have been several initiatives to thoroughly document the genetic heterogeneity in various populations worldwide in National/Ethnic Mutation databases (NEMDBs), a large number of which reside in developing countries. Despite the fact that data content is rather homogeneous, facilitated by the existence of an off-the-shelf database management system [van Baal et al., 2010], these databases have been developed without appropriate guidelines and recommendations to ensure not only a thorough description of common and rare genetic diseases in these populations but also maximal visibility and usefulness of these resources in the respective countries while maintaining the individual local ethical and cultural values in them. Furthermore, developing countries are often a valuable resource of clinical samples that derive from consanguineous families with very unusual phenotypes. These genotype and phenotype data are worth being recorded in NEMDBs not only to make them readily available to interested parties but also to develop and implement genetic epidemiological approaches for estimating the relative frequency of rare autosomal recessive disorders and an estimate of their prevalence in these populations.

For Peer Review

Here, we propose a number of guidelines and recommendations that would assist in genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection, one of the aims and goals of the Human Variome Project [HVP; Cotton et al 2007; Kaput et al., 2009; www.humanvariomeproject.org]. These guidelines are largely the result of discussions among the participants of the developing countries working group during the International HVP Planning and Implementation meetings in May 2008 and 2010, respectively, coming from a wide range of developed and developing countries and extended, based on ongoing data documentation efforts in these countries.

## Background

Any formal attempt to identify the extent of genetic variation must include geographical regions that have not been included in previous haplotype mapping projects, e.g., the HapMap project, particularly since the latter project included racial and not population groups. Although the Population Reference Sample (POPRES) will address some of these missing populations [Nelson et al., 2008], this effort is again designed as a mapping project that is not expected to focus on functional polymorphisms or mutations. Also, the Arab region has long played a key role in discovering new genes and it is really awkward why this region has not been so actively involved in today's genomics research and genome-wide association studies.

Hence, one of the main activities of the HVP is the analysis and subsequent inclusion of genetic variation data from clinical samples originating from diverse populations or ethnic groups [Kaput et al., 2009]. The distinct advantage of some ethnic populations is the opportunity to study genetic diseases due to: (a) large family size, (b) consanguinity, and (c) potential founder effects [Saadallah and Rashed, 2007]. In fact, several disorders have been described for the first time in certain populations residing in developing countries e.g Sanjad-Sakatti

For Peer Review

syndrome, Teebi-Shaltoot syndrome, Teebi Hypertelorism and others. Based on the abovementioned advantages, developing nations are regarded as major contributors to the HVP even though biomedical research has not been the focus of these countries, despite such activities being likely to produce economic and health benefits for all [Singer and Daar, 2001; Kaput et al., 2009].

The NEMDBs are one of the main components of the HVP that could accommodate the extant genetic heterogeneity in the developing countries and, for that matter, all countries. In particular, the NEMDBs are mutation depositories needing continuous updating, recording extensive information over the described genetic heterogeneity of an ethnic group or population [reviewed in Patrinos, 2006]. NEMDBs can help optimize national molecular diagnostic services, by stratifying the corresponding mutation spectra in various inherited diseases, and assist in interpreting diagnostic test results in countries with heterogeneous populations, particularly in minority ethnic groups. NEMDBs can also contribute toward the elucidation of the origin and migration of populations, acting as platforms for comparative genomic studies that can reciprocally provide insights into, e.g., the demographic history of human populations, patterns of their migration and admixture, gene/mutation flow, *etc* [Patrinos, 2006]. Such comprehensive databases would include information on genotype and reasonable phenotype, in a level of detail that should be agreed by the community, which would also enable those interested in these diseases to investigate the possible influence of geographical environment or genetic modifiers on expression of the disease/phenotype. Examples include the genetics of obesity, hypertension or type II diabetes. Relevant genotype/phenotype collection projects, such as GEN2PHEN (http://www.gen2phen.org) or P3G (http://www.p3g.org) could serve as models to establish the necessary requirements for the depth of the recorded genotype and phenotype information.

For Peer Review

## Recommendations

The recommendations and guidelines that follow (summarized in Box 1) are intended to facilitate genetic variation data capture in developing countries, either by suggesting specific measures aiming to enhance awareness among clinicians, bioscientists, and the general public about the range of commonest genetic disorders suffered by certain populations and/or ethnic groups, or by sharing the experience from ongoing projects to document genetic heterogeneity in different countries.

For Peer Review

**1. Develop procedures to include clinicians and researchers in developing nations, mostly by creating networks among those of common interest.**

This would be particularly important for clinicians/researchers working in under-resourced institutes and research environments. Where construction of a NEMDB is decided, selected representatives from participating laboratories, mainly established scientists and principal investigators involved in human genetics should take the lead in this effort. These individuals would form the advisory committee of the NEMDB with one of them assigned as national coordinator with a defined (e.g. yearly) term. These members will then assign eligible members from their groups that would serve as data curators responsible for data entry/modification.

Such efforts would flourish if placed under the umbrella of specific scientific initiatives or consortia (see also below), such as the Human Genome Variation Society (HGVS; www.hgvs.org), a Society which generates recommendations and guidelines towards human genome variation capture, the HVP and ORPHANET (http://www.orpha.net), a database of information on rare diseases and orphan drugs, with presence in 37 mostly European and Arab countries, that aims to contribute to the improvement of the diagnosis, care and treatment of

patients with rare diseases. This approach is already adopted in the Israeli and Greek genetics communities, mostly to share genetic laboratory information, but also to accelerate the data collection process. Partnering with local human genetic societies is also liable to have the same result.

Moreover, in order to increase awareness among interested parties, the resulting NEMDBs should be made known to their potential users using various communication channels. These do not necessarily include international scientific journals to which local institutes in these countries might not have access, but local scientific journals, news bulletins, scientific events (symposia, workshops), one-to-one communication, and so on. Such an approach was followed by the Israeli NEMDB, by producing a review article in the Israeli Medical Association Journal, a leading peer-reviewed scientific medical journal in Israel [Zlotogora et al., 2009].

**2. Create international networks from closely related populations to enhance communication, interaction, and initiate research networks between them.**

In this case, it would be mutually beneficial for closely related countries such as in Latin America, the Middle East or regions in Africa, to initiate new or strengthen existing collaborative ties towards comprehensive genetic variation data capture. Similarly, specific resources would be committed in different countries so that there will be no redundancy in their research efforts. In this regard, the Catalogue for Transmission Genetics in Arabs (CTGA; Tadmouri et al., 2006; see also below) initiated from the Centre for Arab Genomic Studies (http://www.cags.org.ae) in Dubai, United Arab Emirates, is a pilot project to document genetic disorders in Arab populations from the published literature. The current version of CTGA database contains nearly 760 full-text records for nearly 1,290 Mendelian disorders and related genes, including extensive data from the United Arab Emirates, Bahrain, Oman, and Qatar.

For Peer Review

**3. Foster communication, interaction, and research networks between developing and developed countries.**

This would be particularly beneficial in both data collection from the literature or generation of new genetic variation data and would include:

(a) Education of healthcare providers, the public, and government officials from researchers and health care providers from developed countries, regarding data collection, data generation and interpretation, and so on. To this end, the Pan-Arab Human Genetics conferences, with international participation from developed countries, could very well be an example of such interactions. The African Society for Human Genetics is also spearheading efforts to create a database of genetic variation in indigenous African populations (Raj Ramesar; personal communication; Al Aama and co-workers, submitted)

(b) Assistance in data handling and storage in large databases and servers in developed countries, particularly where massive storage units (e.g. GRIDs) are needed to accommodate large datasets. Such possibilities are currently offered by LSDB and NEMDB software developers (LOVD [Fokkema et al., 2005] and ETHNOS [van Baal et al., 2010], respectively) in the Netherlands and Greece, respectively, to either download for use or to store in their local servers, data from interested parties and curators that do not have either the knowledge or the resources needed, although ultimately the aim would be to provide the infrastructure locally.

(c) Consultancy in the design and coordination of large-scale genome projects (see also below) if such need arises, including patients recruitment, experimental design, software/database development (see also Point 5),

(d) Outsourcing of high-throughput experiments, such as microarray or next-generation sequencing experiments, to laboratories in the developed countries, where such

For Peer Review

infrastructure and expertise is more likely to be available, but also provide expert external training to those countries where the facilities are available but outstrip the local expertise.

(e) To work interactively, leading from all of the above, to train postgraduate students and other researchers towards capacity development in developing countries.

**4. Ensure that the database management systems that are being used or developed can be utilized by those in a limited resource environment.**

Nowadays, LSDBs and NEMDBs, which capture the core information in any one of these domains of information, are developed by experts and curated with invaluable skill and experience. Often, developing countries suffer from a limited resource research environment; hence this reality should be carefully considered when database management systems are developed for this purpose.

Therefore, the database management system should be designed and built such that the resulting database would:

(a) Record the specific variants of inherited disorders found in a population or ethnic group in a comprehensive and user-friendly manner, i.e., by providing consistent and simplified information accompanied by numerous links to other databases and related resources [Patrinos, 2006],

(b) Provide the end user with an interface for easy data query and retrieval of information pertaining to the mutation spectrum of each disorder together with the corresponding mutation frequencies and references,

(c) Allow permanent data storage and secure data entry and modification. In this case, software pipelines should also allow automated genotype data transfer between the clinic and/or laboratory to the database, and ideally from the genotyping equipment (sequencer, scanner, *etc*) directly to the database,

For Peer Review

(d) Operate under strict quality assurance guidelines, such as intellectual property rights, the professional code of ethics (http://ihealthcoalition.org/ethics/ehcode.html) or specific ethical guidelines [Povey et al., 2010], laws on computing and liberties (e.g., www.ama-assn.org/ama/pub/category/1905.html), electronic data protection, specific guidelines and any law or regulation applicable in addition to the local ethical issues that may be specific to a given country.

Such databases should be able to operate remotely, should the required support not be available locally (see also Point 3)

**5. Provide support to developing countries to build capacity and to fully participate in the collection, analysis and sharing of genetic variation information.**

Comprehensive genetic variation data capture usually entails highly specialized researchers or diagnostic laboratory staff and state-of-the-art infrastructure, particularly where generation of new genomic data comes to play. Developing countries usually suffer from brain drain and lack of highly sophisticated equipment to perform cutting–edge research. To this end, developed countries have the obligation to alleviate this burden and to assist developing nations by every means possible [Kaput et al., 2009]. It should be noted that developing countries have a valuable resource of clinical samples, often with very unusual phenotypes, particularly when consanguineous families are involved since in this case, the elucidation of the genetic basis of the underlying phenotype becomes much easier. Developing countries should therefore:

(a) Educate healthcare providers and government officials to demonstrate the need to initiate documentation of the genetic variation in their countries and to underline the benefits of cooperation in biomedical research [Bhan et al., 2007],

For Peer Review

(b) Increase awareness in the general population about the need to comprehensively record the underlying genetic heterogeneity in their country and informing them objectively about the benefits and risks of such projects,

(c) Ascertain all parties involved that the research outcome will not be misused. Certain populations or ethnic groups in developing countries usually mistrust research involving genetic analyses or fear that results can stigmatize them or worse (http://www.eubios.info/ASIAE/BIAE201.htm). For example, Malay-Muslims, Chinese, and Indians in Singapore expressed anxiety about breach of confidentiality, the misuse of their DNA for cloning, and possibilities of being diagnosed with disease [Wong et al., 2004]. Community-based research collaborations may provide fora for addressing cultural and ethical concerns of biomedical research [Povey et al., 2010]. Analyzing the extent of human genome variation creates an ideal opportunity for the developed and the developing nations of the world to forge meaningful partnerships and to work together in an unprecedented way, initially to identify variation causing disease, and then to simplify mutation detection in specific ethnic groups, and then to understand how genetic variation contributes to human phenotypic diversity. By ensuring that all nations and ethnic groups have an equal and fair opportunity to share data and technology, evidence-based information will be provided such that all populations can benefit from a global society health network.

(d) Emphasize in every stage of the collaborative project that external input is needed from the design and coordination of NEMDB development and curation to database hosting and technical support, patient recruitment, and experimental design to performing high-throughput experiments in the case of large-scale genome projects (see also Point 3). To this end, funding agencies (e.g. UKIERI; UK-India Education and Research Initiative) could assist by funding innovative projects so that developing countries become independent and self-sustained.

For Peer Review

(e) Develop partnerships for the equitable sharing of knowledge and technology between developed and developing countries. These partnerships should be ideally formed under the umbrella of international organizations, such as the HGVS or the Human Genome Organization (HUGO; http://www.hugo-international.org) that would also organize inter-institutional specialized educational programmes and workshops. The biennial mutation detection workshop would be a paradigm for this initiative and may sometimes be held in developing countries.

**6. Develop a framework to facilitate interactions between the coordinating centre and national, regional and international agencies.**

Researchers and their academic institutions that participate in national data collection projects and/or NEMDBs development should jointly develop infrastructures that would ensure smooth communication between them and national and international agencies. These include: (a) national governments and their Departments of Health, Science and Technology, Development, Education and so on, local human genetic societies, (b) regional organizations, such as the NEPAD (New Programme for Africa's Development) and (c) international organizations, such as the European Commission (EC), United Nations (UN), United Nations Children's Fund (UNICEF), World's Health Organization (WHO), Organization for Economic Cooperation and Development (OECD), *etc*.

This in turn would: (a) create new collaboration ties with neighbouring countries interested in initiating similar projects, (b) increase the chances of receiving funding for these activities (see also below), (c) increase visibility of the progress to the international scientific community.

For Peer Review

**7. Ensure that all ethical, legal, religious, and social issues are thoroughly considered when NEMDBs and/or data capture projects are launched in developing countries.**

The goal of establishing LSDBs and NEMDBs in developing countries is the sharing of genetic variation information for the benefit of the respective populations. This includes the protection of privacy, which in this context is the right of the individual and members of their family to be protected against intrusion into their personal information and further intrusions ensuing from access to this, by publication of information. Of course, there is no such issue when summary level data are reported for a population, as it is the case in some NEMDBs. Therefore, there should be a balance between the public's interest in the value of the shared information and its interest in the strict protection of privacy. This balance would be viewed differently in different cultures and ethnicities [Al-Aqeel, 2007].

For many of these NEMDBs, it seems reasonable that an independent group of qualified scientists oversees these resources and related activities not only at their initiation but on an ongoing basis. Where needed, the main purpose of the particular NEMDB should be clarified, recognising that this may change with time. This will allow evaluation of the exact requirements for data deposition in the NEMDBs, and whether compliance with the remaining guidelines will be possible with the database content freely available to the public or partly restricted to registered persons, namely curators, physicians, and so on. This measure was implemented in the Israeli NEMDB, where not all information is made publicly available [Zlogotora et al., 2007)], since access to specific cases is password-protected [Zlotogora et al., 2009]. A similar measure would entail anonymizing genetic data, either by permanently removing or coding the personal identifiers [Povey et al., 2010]. Of interest, the newest version of the LOVD software [Fokkema et al., 2005] has the option to store data that are not public but that can be queried. The result of a query hitting non-public data is a notification that there is such information in the database but that the curator needs to be contacted to get more information.

For Peer Review

The depth of the ethnic and geographic origin data should be determined in advance, particularly in the case of small villages where carriers of rare variants could be identified even if personal information were not provided. In this case, data donors should be well informed and provided with an explanation as to why their data documentation is useful and how their privacy is secured. These individuals should consent to the inclusion of their genetic information in the database and reserve the right to withdraw their data from the database whenever they desire. This is particularly important for persons unable to consent because of disability or young age. Standardized consent forms need to be translated to different local languages and/or dialects.

Finally, these databases would also benefit from some independent/external ethical review, perhaps from an international group (HGVS or HVP) who could advise them accordingly.

## Nation-wide genomic projects that would serve as paradigm for genetic variation data collection

By definition, a nation-wide genomic project is a national multicentre initiative to determine the underlying genetic heterogeneity of a population with the goal to ascertain the genetic variation leading to common or rare inherited diseases. Such initiatives consist of three parts: (a) documenting the genetic heterogeneity that has already been reported in the published literature, (b) generating new genetic data by screening a large proportion of the population using high throughput genotyping approaches, and (c) accumulating data causing inherited disease from hospitals, laboratories and other healthcare facilities. In every case, except perhaps (c), a large national genetic database/portal is being developed to accommodate the sheer amount of data gathered. To date, there are few nation-wide genomic projects in various developing countries that are currently in progress. These projects (summarized in Table 1) implement some



John Wiley & Sons, Inc.

For Peer Review

of the guidelines described previously. Also, there are several other NEMDB projects, all of which more or less follow these guidelines, that have already been reviewed [Patrinos, 2006; van Baal et al., 2007] and will, therefore, not be further discussed since they lie outside the scope of the present article.

The Estonian Genome Project (EGP; http://www.geenivaramu.ee/index.php?lang=eng) is a research venture of the University of Tartu initiated in 2001 [Metspalu, 2002] and aims to create a database of health, genealogical and genetic data representing approximately 10% of Estonia's population. The database will make it possible for researchers both in Estonia and outside to look for links between genes, environmental factors and common diseases (such as cancer, diabetes, depression, cardio-vascular diseases, *etc*). Currently, EGP gene bank contains data contributed by over 20,000 donors with an expected increase to 100,000 within the next three or four years.

The Indian Genome Variation database (IGVdb; http://www.igvdb.res.in) is a consortium activity established in 2003 to document the inherent genetic variability of the Indian subpopulations as a first step towards identifying genetic biomarkers to understand susceptibility for any genetic disease and population migration, pharmacogenomic markers to understand variable drug response. The Indian Genome Variation (IGV) consortium aims to provide data on validated SNPs and repeats, both novel and reported, along with gene duplications, in over 1,000 genes (selected on the basis of their relevance as functional and positional candidates in many common diseases including genes relevant to pharmacogenomics), in 15,000 individuals drawn from Indian subpopulations [Indian Genome Variation Consortium, 2005].

Kuwait established within the Faculty of Medicine, Department of Pathology a Molecular Genetics Diagnostic Service Division (http://www.al-mulla.org), integrated within a research laboratory, which is focused on delivering state-of-the-art diagnostic service for the Kuwaiti population. These services, which focus on autozygosity mapping in families with

For Peer Review

consanguineous marriages, provide the research laboratory with a unique opportunity to identify novel genes in autosomal recessive disorders. Moreover, various unique cases suffering from cancer are currently being analysed using the Affymetrix Array 6.0 microarray technology to identify copy number changes and homozygosity stretches. The data generated from these studies are deposited in CAGS database (see above), and LOVD, while the Kuwaiti NEMDB has been recently established databases to encourage data storage for genetic diseases in different subpopulations in Kuwait.

The United Kingdom has developed a central database to collect all UK genome variation data, and their effects, from participating diagnostic laboratories. The diagnostic mutation database (DMuDB; http://ngrl.org.uk/Manchester/dmudb.html) provides a route for sharing mutation data within and between diagnostic laboratories in the United Kingdom to support genetic testing services for patients. Currently DMuDB has over 4,884 records for 33 genes, namely Breast Cancer (*BRCA1, BRCA2*), Lowe Syndrome (*OCRL*), Sotos Syndrome (*NSD1*), HNPCC (*MLH1, MSH2*), Cystic Fibrosis (*CFTR*), X-Linked Retinitis Pigmentosa (*XLRP*), Neurofibromatosis, Types I and II (*NF1, NF2*), Alstrom Syndrome (*ALMS1*), CADASIL (*NOTCH3*), Familial adenomatous polyposis (*FAP*) and Muscular Dystrophy (*DMD*), containing over 10,985 individual variants.

The Hellenic Genome Variation database aims to document the genetic heterogeneity in the Hellenic population. The first version of the Hellenic NEMDB was created in 2004 [Patrinos et al., 2005]. In 2007, research institutes involved in hemoglobinopathies research initiated a pilot multicenter project to determine the mutation spectrum of the thalassemia syndromes in various geographical regions in the country with the goal to document the findings in the Hellenic NEMDB. These data showed that although Greece is a rather homogeneous country, there are big qualitative (type of mutations) and quantitative (allelic frequencies) differences in various parts of the country, namely central Greece [Samara et al., 2007] and South-Western

For Peer Review

Greece [Papachatzopoulou et al., 2010]. These data compared to findings from North-Western Greece [Georgiou et al., 2003] and the rest of the country [Patrinos et al., 2005] reveal valuable information for region-specific population carrier screening and for efficient provision of prenatal diagnosis of β-thalassemia in Greece.

Finally, there are four international consortia with activities related to the documentation of causative mutations and allelic frequencies in developing countries:

(a) The International Society for Gastrointestinal Hereditary Tumours (InSiGHT; http://www.insight-group.org) is an international multidisciplinary, scientific organisation with a mission "…to improve the quality of care of patients and their families with any condition resulting in hereditary gastrointestinal tumours." The InSiGHT consortium maintains a LOVD-based database for genetic variations for the *EPCAM*, *MLH1*, *MLH3*, *MSH2*, *MSH6*, *MUTYH*, *PMS1*, and *PMS2* genes leading to gastrointestinal tumours and encourages the establishment of national repositories for patients suffering from gastrointestinal tumours, particularly in developing countries.

(b) The International Thalassemia Network (ITHANET; http://www.ithanet.eu) is a European Commission-funded project that comprises a network of all major European research institutions active in the field of molecular and clinical research of thalassemia and related hemoglobinopathies [Lederer et al., 2009]. Among its activities, ITHANET hosts a database (ITHANETbase) that includes mutation frequencies in various populations at-risk for thalassemia, particularly from developing countries. In many instances though, this repository is largely redundant to HbVar database of hemoglobin variants and thalassemia mutations [http://globin/bx.psu.edu/hbvar; Patrinos et al., 2004].

(c) The MEDGENET project (http://medgenet.tredueuno.it) is a European Consortium, funded by the European Commission, aiming to expand the human expertise in clinical genetics in



For Peer Review

Mediterranean Partner Countries (MPC) through the transfer of knowledge and technology between the two rims of the Mediterranean which share a common burden of genetic diseases. Among the actions foreseen by MEDGENET is the building of Medical Genetics databases for the various partner countries. Several NEMDBs are built from scratch, such as the Tunisian and Egyptian NEMDBs or expanded from the existing NEMDBs, such as the Israeli [Zlotogora et al., 2007], Cypriot [Kleanthous et al., 2006], and Lebanese NEMDBs [Megarbane et al., 2007], using the ETHNOS database management system for NEMDB development and curation [van Baal et al., 2010].

(d) The Pharmacogenomics for Every Nation Initiative (PGENI; www.pgeni.org) that aims to determine the frequencies of the commonest pharmacogenomic markers in various populations, particularly from developing countries worldwide.

Finally, there are a few nation-wide DNA biobanks, such as the National Laboratory for the Genetics of Israeli Populations, representing a national repository for human cell-lines and DNA from the unique and large ethnic variation of the Israeli populations (Tel Aviv University; available at http://www.tau.ac.il/medicine/NLGIP) or the Biomaterial Banking Medical Research Initative (BBMRI) in Malta, which of course lie at the borderline of the nation-wide genomic projects described above.

## Funding and incentives

There are currently two main obstacles that hinder further efforts to expand the NEMDB field, namely funding and providing incentives to data submitters.

Funding for collecting genetic variation data has traditionally been difficult due to the extreme fragmentation of the field [Patrinos and Brookes, 2005]. Therefore, it would be expected that efforts to receive funding for genetic variation data capture are more likely to flourish if



For Peer Review

treated as concerted efforts. GEN2PHEN (http://www.gen2phen.org) represents such initiative that have already received funding from the EC and aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype data, and to link this system into other biomedical knowledge sources via genome browser functionality. Linking population and ethnic-specific information into LSDBs and diagnostic databases, creating NEMDBs and developing genomic databases, similar to the Indian Genome Variation database are key activities for the main GEN2PHEN work-packages. Obviously, other strategically minded funding agencies (e.g. NIH), charities, and foundations would also launch similar calls. Recently the Australian Node of the HVP has been funded to pilot software for data flow from hospitals/clinics via country specific databases to LSDBs and NCBI/EBI/UCSC browsers (unpublished). Involving universal organizations such as the WHO and UNESCO in the fund-raising project will clearly have a strong impact.

On the other hand, the development of comprehensive analyses of human genetic variation often reaches a barrier in the form of the reward system for clinicians and academic researchers. First of all, database entries could be made a mandatory quality control standard for clinical laboratories and clinicians. Notably, according to a new initiative of the Israeli government, 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 [van Baal et al., 2010].

For researchers, a publication or Web-based system establishing microattribution and community annotation of mutations and cited data will enable measurable contributions to the scientific knowledge base [Anonymous, 2008; Hoffmann, 2008]. Similarly, database-journals would also serve this task by providing a forum for publishing gene variation data that would be eventually deposited in the PubMed literature database. One such journal, the open access Human Genomics and Proteomics (http://www.sage-hindawi.com/journals/hgp), which is

For Peer Review

affiliated with FINDbase [http://www.findbase.org; van Baal et al., 2007], will focus on studies characterizing causative mutation and/or biomarker frequency spectra. Accepted contributions including datasets will be linked in FINDbase and deposited in PubMed [Patrinos and Petricoin, 2009]. Also, the journal *Human Genetics* has been accepting mutation data and this is linked to HGMD (http://www.hgmd.org).

Journals, tenure, and promotion committees, and funding agencies would be encouraged to cite these contributions to and citations of LSDB and international national databases [Patrinos and Brookes, 2005]. The rapid progress in genetic research has led to a declining interest of scientific clinical genetic journals in case reports with single or few novel mutations. This may deter clinicians from reporting these new mutations, which collectively become important, unless another venue becomes visible.

## Concluding remarks

As stated above, documenting population/ethnic-specific data in developing countries are among the primary objectives of the HVP that appreciates the genomic sovereignty/equality for all countries to be involved and acknowledges the value of ‘‘human capital’’ within all populations. Real and tangible benefits of the HVP to improve health will be generated for participating populations. Hence, the voluntary participation of the greatest number of countries would ensure a more general applicability, and it is hoped that many countries will eventually decide to participate.

## Acknowledgements

For Peer Review

We thank the members of the developing countries working group at the HVP planning and implementation meetings for lively and helpful discussions. Part of this work has received funding from the European Commission [ITHANET (FP6-026539), EuroGenTest (FP6-512148), MEDGENET (FP6-31968) and GEN2PHEN (FP7-200754) projects], and the Kuwait Foundation for the advancement of Sciences (KFAS) grant number 2006-1302-07.

## References

Al Aqeel AI. 2007. Islamic ethical framework for research into and prevention of genetic diseases. Nat Genet 39:1293-1298.

Anonymous. 2008. Human variome microattribution reviews. Nat Genet 40:1.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Bhan A, Singh JA, Upshur RE, Singer PA, Daar AS. 2007. Grand challenges in global health: engaging civil society organizations in biomedical research in developing countries. PLoS Med 4:e272.

Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Marafie MJ, Paalman MH, Patrinos GP, Qi M, Ramesar RS, Scott RJ, Sijmons RH, Sobrido MJ, Vihinen M. (2009). Capturing All Disease Causing Mutations For Clinical And Research Use: Towards An Effortless System For The Human Variome Project. Genet Med, 11:843-849.

Cotton RG, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, Cantu JM, Cassiman JJ, Claustres M, Concannon P, Cotton RG, den Dunnen JT, Flicek P, Gibbs R, Hall J, Hasler J, Katz M, Kwok PY, Laradi S, Lindblom A, Maglott D, Marsh S, Masimirembwa CM, Minoshima S,



John Wiley & Sons, Inc.

For Peer Review

de Ramirez HZ, Scriver CR, Sherry S, Shimizu N, Shimizu N, Stein L, Tadmouri GO, Taylor G, Watson M. 2007. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 39:433–436.

Fokkema IF, den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63-68.

Georgiou I, Makis A, Chaidos A, Bouba I, Hatzi E, Kranas V, Zilidis C, Bourantas KL. 2003. Distribution and frequency of beta-thalassemia mutations in northwestern and central Greece. Eur J Haematol 70:75-78.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

Hoffman R. 2008. A wiki for the life sciences where authorship matters. Nat Genet 40:1047–1051.

Indian Genome Variation Consortium. 2005. The Indian Genome Variation database (IGVdb): a project overview. Hum Genet 118:1-11.

Kaput J, Cotton RG, Hardman L, Watson M, Al Aqeel AI, Al-Aama JY, Al-Mulla F, Alonso S, Aretz S, Auerbach AD, Bapat B, Bernstein IT, Bhak J, Bleoo SL, Blöcker H, Brenner SE, Burn J, Bustamante M, Calzone R, Cambon-Thomsen A, Cargill M, Carrera P, Cavedon L, Cho YS, Chung YJ, Claustres M, Cutting G, Dalgleish R, den Dunnen JT, Díaz C, Dobrowolski S, Dos Santos MR, Ekong R, Flanagan SB, Flicek P, Furukawa Y, Genuardi M, Ghang H, Golubenko MV, Greenblatt MS, Hamosh A, Hancock JM, Hardison R, Harrison TM, Hoffmann R, Horaitis R, Howard HJ, Barash CI, Izagirre N, Jung J, Kojima T, Laradi S, Lee YS, Lee JY, Gil-da-Silva-Lopes VL, Macrae FA, Maglott D, Marafie MJ, Marsh SG, Matsubara Y, Messiaen LM, Möslein G, Netea MG, Norton ML, Oefner PJ, Oetting WS, O'Leary JC, de Ramirez AM, Paalman MH, Parboosingh J, Patrinos GP,

For Peer Review

Perozzi G, Phillips IR, Povey S, Prasad S, Qi M, Quin DJ, Ramesar RS, Richards CS, Savige J, Scheible DG, Scott RJ, Seminara D, Shephard EA, Sijmons RH, Smith TD, Sobrido MJ, Tanaka T, Tavtigian SV, Taylor GR, Teague J, Töpel T, Ullman-Cullere M, Utsunomiya J, van Kranen HJ, Vihinen M, Webb E, Weber TK, Yeager M, Yeom YI, Yim SH, Yoo HS. 2009. Planning the Human Variome Project. The Spain report. Hum Mutat 30:496-510.

Kleanthous M, Patsalis PC, Drousiotou A, Motazacker M, Christodoulou K, Cariolou M, Baysal E, Khrizi K, Pourfarzad F, Moghimi B, van Baal S, Deltas CC, Najmabadi S, Patrinos GP. 2006. The Cypriot and Iranian National Mutation databases. Hum Mutat 27:598-599.

Lederer CW, Basak AN, Aydinok Y, Christou S, El-Beshlawy A, Eleftheriou A, Fattoum S, Felice AE, Fibach E, Galanello R, Gambari R, Gavrila L, Giordano PC, Grosveld F, Hassapopoulou H, Hladka E, Kanavakis E, Locatelli F, Old J, Patrinos GP, Romeo G, Taher A, Traeger-Synodinos J, Vassiliou P, Villegas A, Voskaridou E, Wajcman H, Zafeiropoulos A, Kleanthous M. 2009. An electronic infrastructure for research and treatment of the thalassemias and other hemoglobinopathies: the Euro-mediterranean ITHANET project. Hemoglobin 33:163-176.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Berman J, Rogers YH, Frazier ME, Scherer SW, Strausberg FL, Venter JC. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Mégarbané A, Chouery E, van Baal S, Patrinos GP. 2006. The Lebanese National Mutation frequency databases. Eur J Hum Genet 14(Suppl 1):365.

Metspalu A. 2002. Estonian Genome Project--before the take-off and take-off. Bioinformatics 18 Suppl 2:S152.

For Peer Review

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Sitrnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamanta CD, Roses AD, Burns DK, Ehm MG, Lia EH. 2008. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347–358.

Papachatzopoulou A, Vantarakis A, Kourakli A, Stavrou E, Frangou E, Patrinos GP, Athanassiadou A. 2010. Documentation of the incidence of beta-thalassemia and *HBB* mutation frequencies in South-Western Greece. Hemoglobin 34:333-342.

Patrinos GP. 2006. National/Ethnic mutation databases: Recording populations' genography. Hum Mutat 27:879-887.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: Disastrously deficient. Trends Genet, 21:333-338.

Patrinos G, Petricoin E. 2009. A new scientific journal linked to a genetic database: towards a novel publication modality. Hum Genomics Proteomics 1:e597478.

Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, Anagnou NP, Wajcman H, Hardison RC. 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 32:D537-D541.

Patrinos GP, van Baal S, Petersen MB, Papadakis MN. 2005. The Hellenic National Mutation database: A prototype database for inherited disorders in the Hellenic population. Hum Mutat, 25:327-333.

Povey S, Al Aqeel AI, Barash CI, Cambon-Thomsen A, Dalgleish R, den Dunnen JT, Firth HV, Greenblatt M, Parker M, Patrinos GP, Savige J, Sobrido MJ, Cotton RG. 2010. Practical Guidelines addressing ethical issues pertaining to the curation of human Locus specific variation databases (LSDBs). Hum Mutat (in press).

For Peer Review

Saadallah AA, Rashed MS. 2007. Newborn screening: experiences in the Middle East and North Africa. J Inherit Metab Dis 30:482–489.

Samara M, Chiotoglou I, Kalamaras A, Likousi S, Chassanidis C, Vagena A, Vagenas C, Eftichiadis E, Vamvakopoulos N, Patrinos GP, Kollia P. 2007. Large-scale population genetic analysis for hemoglobinopathies reveals different mutation spectra in central Greece compared to the rest of the country. Am J Hematol 82:634-636.

Singer PA, Daar AS. 2001. Harnessing genomics and biotechnology to improve global health equity. Science 294:87–89.

Tadmouri GO, Al Ali MT, Al-Haj Ali S, Al Khaja N. 2006. CTGA: the database for genetic disorders in Arab populations. Nucleic Acids Res. 34:D602-D606.

van Baal S, Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, Macek Jr M, Scriver CR, Patrinos GP. 2007. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 35:D690–D695.

van Baal S, Zlotogora J, Lagoumintzis G, Gkantouna V, Tzimas I, Poulas K, Tsakalidis A, Romeo G, Patrinos GP. 2010. ETHNOS: A versatile electronic tool for the development and curation of National Genetic databases. Hum Genomics 4:361-368.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, FanW, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Feng X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang M, Wang J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60–65.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR,

For Peer Review

Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Marqulies M, Weinstock GM, Gibbs RA, Rothberg GM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.

Wong ML, Chia KS, Wee S, Chia SE, Lee J, Koh WP, Shen HM, Thumboo J, Sofjan D. 2004. Concerns over participation in genetic research among Malay-Muslims, Chinese and Indians in Singapore: a focus group study. Community Genet 7:44–54.

Zlotogora J, van Baal S, Patrinos GP. 2007. Documentation of inherited disorders and mutation frequencies in the different religious communities in Israel in the Israeli National Genetic Database. Hum Mutat 28:944-949.

Zlotogora J, van Baal S, Patrinos GP. 2009. The Israeli National Genetic Database. IMAJ 6:373-375.

For Peer Review

**Box 1**: Guidelines and recommendations for data capture in developing countries

1. Develop procedures to include clinicians and researchers in developing nations, mostly by creating networks among those of common interest.
2. Create international networks from closely related populations to enhance communication, interaction, and initiate research networks between them.
3. Foster communication, interaction, and research networks between developing and developed countries.
4. Ensure that the database management systems that are being used or developed can be utilized by those in a limited resource environment.
5. Provide support to developing countries to build capacity and to fully participate in the collection, analysis and sharing of genetic variation information.
6. Develop a framework to facilitate interactions between the coordinating centre and national, regional and international agencies.
7. Ensure that all ethical, legal, religious, and social issues are thoroughly considered when NEMDBs and/or data capture projects are launched in developing countries.

For Peer Review

**Table 1**. Nation-wide genome and related projects (N/A: Not available).

| Project | URL | Aims and scope | References |
|---|---|---|---|
| **Nation-wide genome projects** | | | |
| The Estonian Genome Project | http://www.geenivaramu.ee/index.php?lang=eng | To create a database of health, genealogical and genome data representing approximately 10% of Estonia's population | Metspalu, 2002 |
| The MEDGENET Project | http://medgenet.tredueuno.it | To expand the human expertise in clinical genetics in Mediterranean Partner Countries (MPC) that share a common burden of genetic diseases. | van Baal et al., 2010 |
| The Indian Genome Variation database | http://www.igvdb.res.in | To create a DNA variation database of the people of India and make it available to researchers for understanding human biology with respect to disease predisposition, adverse drug reaction, population migration | Indian Genome Variation Consortium, 2005 |
| The Catalogue of Transmission Genetics in Arabs | http://www.cags.org.ae/ctga_search.html | To provide the user with clinical and molecular genetic descriptions of genetic disorders, epidemiological details of reports of the disorders in the Arab population and clinical case details wherever available | Tadmouri et al., 2006 |
| **Other related projects** | | | |
| Kuwaiti Molecular Genetics Diagnostic Service | http://www.al-mulla.org | To provide autozygosity mapping services, Information on Arab genome and molecular karyotying of cancers | N/A |
| UK diagnostic mutation database | http://ngrl.org.uk/Manchester/dmudb.html | To provide a route for sharing mutation data within and between diagnostic laboratories in the United Kingdom to support genetic testing services for patients. | N/A |
| Hellenic Genome Variation database | http://www.goldenhelix.org/hellenic | To create a compendium for genetic diseases and to encourage genetic studies to elucidate the genetic basis of inherited diseases in Greece | Patrinos et al., 2005 |
| The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) | http://www.insight-group.org | To improve the quality of care of patients and their families with any condition resulting in hereditary gastrointestinal tumours | N/A |
| International Thalassemia Network (ITHANET) | http://www.ithanet.eu | To strengthen research and treatment of hemoglobinopathies | Lederer et al., 2009 |
| Pharmacogenomics for Every Nation Initiative | http://www.pgeni.org | To determine the frequencies of the commonest pharmacogenomic markers in various populations, particularly from developing countries worldwide | N/A |

For Peer Review

# Recommendations for genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection

George P. Patrinos [1,*], Jumana Al Aama [2], Aida Al Aqeel [3], Fahd Al-Mulla [4], Joseph Borg [5], Andrew Devereux [6], Alex E. Felice [7], Finlay Macrae [8], Makia J. Marafie [9], Michael B. Petersen [10], Ming Qi [11], Rajkumar S. Ramesar [12], Joel Zlotogora [13], and Richard G. Cotton [14]

[1] University of Patras, School of Health Sciences, Department of Pharmacy, Patras, Greece;

[2] Department of Genetic Medicine, Faculty of Medicine and Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia;

[3] Department of Pediatrics, Riyadh Military Hospital and Stem Cell therapy program, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia;

[4] Molecular Pathology Unit, Kuwait University, Kuwait;

[5] University of Malta, Faculty of Health Sciences, Department of Applied Biomedical Sciences, Msida, Malta;

[6] NGRL Manchester, Department of Medical Genetics, St Mary's Hospital, Manchester, United Kingdom;

[7] University of Malta, Department of Physiology and Biochemistry, Laboratory of Molecular Genetics, Msida, Malta;

[8] Department of Colorectal Medicine and Genetics and University of Melbourne, The Royal Melbourne Hospital, Australia;

Deleted: [5]
Deleted: [7]
Deleted: [8]
Deleted: [9]
Deleted: [0]
Deleted: [1]
Deleted: [2]
Deleted: [3]

Deleted: University of Malta, Department of Physiology and Biochemistry, Laboratory of Molecular Genetics, Msida, Malta;

Deleted: [7]

For Peer Review

[9] Clinical Genetics, Kuwait Medical Genetics Centre, Maternity Hospital, Sabah Medical Area, Kuwait;

[10] Institute of Child Health, Department of Genetics, Athens, Greece;

[11] Center for Genetic and Genomic Medicine, The First Affiliated Hospital of Zhejiang University School of Medicine and James Watson Institute of Genomic Sciences, P. R. China;

[12] MRC Human Genetics Research Unit, Division of Human Genetics, Institute for Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa;

[13] Department of Community Genetics, Public Health Services, Ministry of Health, Jerusalem, Israel;

[14] Genomic Disorders Research Centre, Howard Florey Institute, Melbourne, Australia & Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Australia

Deleted: [8]

Deleted: [9]

Deleted: [0]

Deleted: [1]

Deleted: [2]

Deleted: [3]

For Peer Review

Running head: Recommendations for genetic variation data capture in developing countries

* Corresponding author at:

Department of Pharmacy, School of Health Sciences, University of Patras,

University Campus, Rion, GR-265 04, Patras,

Greece

Telephone/Fax: +30-2610-969.834,

E-mail: gpatrinos@upatras.gr

## Abstract

Developing countries have significantly contributed to the elucidation of the genetic basis of both common and rare disorders, providing an invaluable resource of cases due to large family sizes, consanguinity and potential founder effects. Moreover, the recognised depth of genomic variation in indigenous African populations, reflecting the ancient origins of humanity on the African continent, and the effect of selection pressures on the genome, will be valuable in understanding the range of both pathological and non-pathological variations. The involvement of these populations in accurately documenting the extant genetic heterogeneity is more than essential. Developing nations are regarded as key contributors to the Human Variome Project (HVP; http://www.humanvariomeproject.org), a major effort to systematically collect mutations that contribute to or cause human disease and create a cyber infrastructure to tie databases together. However, biomedical research has not been the primary focus in these countries even though such activities are likely to produce economic and health benefits for all. Here, we propose several recommendations and guidelines to facilitate participation of developing countries in genetic variation data documentation, ensuring an accurate and comprehensive worldwide data collection. We also summarize a few well-coordinated genetic data collection initiatives that would serve as paradigms for similar projects.


**Key words:** Developing countries, national/ethnic mutation databases, populations, genetic variation



For Peer Review

## Introduction

The recent elucidation of the genome sequence of several individuals, such as Causacians (Jim Watson [Wheeler et al., 2008] and Craig Venter [Levy et al., 2007]), West African [Bentley et al., 2008], and Han Chinese [Wang et al., 2008], suggests that a large number of SNPs and other sequence variation exists in the human population. Estimates from African genetic diversity and the Pan Asian SNP initiative suggest that approximately 80–90% of human genomic variation resides in the world's developing countries. It is therefore obvious that there is a huge need to comprehensively document the extant genetic variation in various ethnic groups and populations around the globe to ensure maximal data coverage.

Since 2002, there have been several initiatives to thoroughly document the genetic heterogeneity in various populations worldwide in National/Ethnic Mutation databases (NEMDBs), a large number of which reside in developing countries. Despite the fact that data content is rather homogeneous, facilitated by the existence of an off-the-shelf database management system [van Baal et al., 2010], these databases have been developed without appropriate guidelines and recommendations to ensure not only a thorough description of common and rare genetic diseases in these populations but also maximal visibility and usefulness of these resources in the respective countries while maintaining the individual local ethical and cultural values in them. Furthermore, developing countries are often a valuable resource of clinical samples that derive from consanguineous families with very unusual phenotypes. These genotype and phenotype data are worth being recorded in NEMDBs not only to make them readily available to interested parties but also to develop and implement genetic epidemiological approaches for estimating the relative frequency of rare autosomal recessive disorders and an estimate of their prevalence in these populations.



For Peer Review

Here, we propose a number of guidelines and recommendations that would assist in genetic variation data capture in developing countries to ensure a comprehensive worldwide data collection, one of the aims and goals of the Human Variome Project [HVP; Cotton et al 2007; Kaput et al., 2009; www.humanvariomeproject.org]. These guidelines are largely the result of discussions among the participants of the developing countries working group during the International HVP Planning and Implementation meetings in May 2008 and 2010, respectively, coming from a wide range of developed and developing countries and extended, based on ongoing data documentation efforts in these countries.

Deleted: respective

## Background

Any formal attempt to identify the extent of genetic variation must include geographical regions that have not been included in previous haplotype mapping projects, e.g., the HapMap project, particularly since the latter project included racial and not population groups. Although the Population Reference Sample (POPRES) will address some of these missing populations [Nelson et al., 2008], this effort is again designed as a mapping project that is not expected to focus on functional polymorphisms or mutations. Also, the Arab region has long played a key role in discovering new genes and it is really awkward why this region has not been so actively involved in today's genomics research and genome-wide association studies.

Deleted: i.

Hence, one of the main activities of the HVP is the analysis and subsequent inclusion of genetic variation data from clinical samples originating from diverse populations or ethnic groups [Kaput et al., 2009]. The distinct advantage of some ethnic populations is the opportunity to study genetic diseases due to: (a) large family size, (b) consanguinity, and (c) potential founder effects [Saadallah and Rashed, 2007]. In fact, several disorders have been described for the first time in



For Peer Review

certain populations residing in developing countries e.g Sanjad-Sakatti syndrome, Teebi-Shaltoot syndrome, Teebi Hypertelorism and others. Based on the abovementioned advantages, developing nations are regarded as major contributors to the HVP even though biomedical research has not been the focus of these countries, despite such activities being likely to produce economic and health benefits for all [Singer and Daar, 2001; Kaput et al., 2009].

The NEMDBs are one of the main components of the HVP that could accommodate the extant genetic heterogeneity in the developing countries and, for that matter, all countries. In particular, the NEMDBs are mutation depositories needing continuous updating, recording extensive information over the described genetic heterogeneity of an ethnic group or population [reviewed in Patrinos, 2006]. NEMDBs can help optimize national molecular diagnostic services, by stratifying the corresponding mutation spectra in various inherited diseases, and assist in interpreting diagnostic test results in countries with heterogeneous populations, particularly in minority ethnic groups. NEMDBs can also contribute toward the elucidation of the origin and migration of populations, acting as platforms for comparative genomic studies that can reciprocally provide insights into, e.g., the demographic history of human populations, patterns of their migration and admixture, gene/mutation flow, *etc* [Patrinos, 2006]. Such comprehensive databases would include information on genotype and reasonable phenotype, in a level of detail that should be agreed by the community, which would also enable those interested in these diseases to investigate the possible influence of geographical environment or genetic modifiers on expression of the disease/phenotype. Examples include the genetics of obesity, hypertension or type II diabetes. Relevant genotype/phenotype collection projects, such as GEN2PHEN (http://www.gen2phen.org) or P3G (http://www.p3g.org) could serve as models to establish the necessary requirements for the depth of the recorded genotype and phenotype information.

**Deleted:** detail of

**Formatted:** Indent: First line: 0 pt



For Peer Review

## Recommendations

Formatted: Font: 14 pt, Bold
Formatted: Font: 14 pt
Formatted: Font: 14 pt, Bold
Formatted: Font: 14 pt, Bold

The recommendations and guidelines that follow (summarized in Box 1) are intended to facilitate genetic variation data capture in developing countries, either by suggesting specific measures aiming to enhance awareness among clinicians, bioscientists, and the general public about the range of commonest genetic disorders suffered by certain populations and/or ethnic groups, or by sharing the experience from ongoing projects to document genetic heterogeneity in different countries.

For Peer Review

1. **Develop procedures to include clinicians and researchers in developing nations, mostly by creating networks among those of common interest.**

This would be particularly important for clinicians/researchers working in under-resourced institutes and research environments. Where construction of a NEMDB is decided, selected representatives from participating laboratories, mainly established scientists and principal investigators involved in human genetics should take the lead in this effort. These individuals would form the advisory committee of the NEMDB with one of them assigned as national coordinator with a defined (e.g. yearly) term. These members will then assign eligible members from their groups that would serve as data curators responsible for data entry/modification.

Such efforts would flourish if placed under the umbrella of specific scientific initiatives or consortia (see also below), such as the Human Genome Variation Society (HGVS; www.hgvs.org), a Society which generates recommendations and guidelines towards human genome variation capture, the HVP and ORPHANET (http://www.orpha.net), a database of information on rare diseases and orphan drugs, with presence in 37 mostly European and Arab

countries, that aims to contribute to the improvement of the diagnosis, care and treatment of patients with rare diseases. This approach is already adopted in the Israeli and Greek genetics communities, mostly to share genetic laboratory information, but also to accelerate the data collection process. Partnering with local human genetic societies is also liable to have the same result.

Moreover, in order to increase awareness among interested parties, the resulting NEMDBs should be made known to their potential users using various communication channels. These do not necessarily include international scientific journals to which local institutes in these countries might not have access, but local scientific journals, news bulletins, scientific events (symposia, workshops), one-to-one communication, and so on. Such an approach was followed by the Israeli NEMDB, by producing a review article in the Israeli Medical Association Journal, a leading peer-reviewed scientific medical journal in Israel [Zlotogora et al., 2009].

**2. Create international networks from closely related populations to enhance communication, interaction, and initiate research networks between them.**

In this case, it would be mutually beneficial for closely related countries such as in Latin America, the Middle East or regions in Africa, to initiate new or strengthen existing collaborative ties towards comprehensive genetic variation data capture. Similarly, specific resources would be committed in different countries so that there will be no redundancy in their research efforts. In this regard, the Catalogue for Transmission Genetics in Arabs (CTGA; Tadmouri et al., 2006; see also below) initiated from the Centre for Arab Genomic Studies (http://www.cags.org.ae) in Dubai, United Arab Emirates, is a pilot project to document genetic disorders in Arab populations from the published literature. The current version of CTGA database contains nearly 760 full-text

Deleted: the



For Peer Review

records for nearly 1,290 Mendelian disorders and related genes, including extensive data from the United Arab Emirates, Bahrain, Oman, and Qatar.

**3. Foster communication, interaction, and research networks between developing and developed countries.**

This would be particularly beneficial in both data collection from the literature or generation of new genetic variation data and would include:

(a) Education of healthcare providers, the public, and government officials from researchers and health care providers from developed countries, regarding data collection, data generation and interpretation, and so on. To this end, the Pan-Arab Human Genetics conferences, with international participation from developed countries, could very well be an example of such interactions. The African Society for Human Genetics is also spearheading efforts to create a database of genetic variation in indigenous African populations (Raj Ramesar; personal communication; Al Aama and co-workers, submitted)

(b) Assistance in data handling and storage in large databases and servers in developed countries, particularly where massive storage units (e.g. GRIDs) are needed to accommodate large datasets. Such possibilities are currently offered by LSDB and NEMDB software developers (LOVD [Fokkema et al., 2005] and ETHNOS [van Baal et al., 2010], respectively) in the Netherlands and Greece, respectively, to either download for use or to store in their local servers, data from interested parties and curators that do not have either the knowledge or the resources needed, although ultimately the aim would be to provide the infrastructure locally.

(c) Consultancy in the design and coordination of large-scale genome projects (see also below) if such need arises, including patients recruitment, experimental design, software/database development (see also Point 5),



For Peer Review

(d) Outsourcing of high-throughput experiments, such as microarray or next-generation sequencing experiments, to laboratories in the developed countries, where such infrastructure and expertise is more likely to be available, but also provide expert external training to those countries where the facilities are available but outstrip the local expertise.

(e) To work interactively, leading from all of the above, to train postgraduate students and other researchers towards capacity development in developing countries.

**4. Ensure that the database management systems that are being used or developed can be utilized by those in a limited resource environment.**

Deleted: s

Nowadays, LSDBs and NEMDBs, which capture the core information in any one of these domains of information, are developed by experts and curated with invaluable skill and experience. Often, developing countries suffer from a limited resource research environment; hence this reality should be carefully considered when database management systems are developed for this purpose.

Therefore, the database management system should be designed and built such that the resulting database would:

(a) Record the specific variants of inherited disorders found in a population or ethnic group in a comprehensive and user-friendly manner, i.e., by providing consistent and simplified information accompanied by numerous links to other databases and related resources [Patrinos, 2006],

(b) Provide the end user with an interface for easy data query and retrieval of information pertaining to the mutation spectrum of each disorder together with the corresponding mutation frequencies and references,



For Peer Review

(c) Allow permanent data storage and secure data entry and modification. In this case, software pipelines should also allow automated genotype data transfer between the clinic and/or laboratory to the database, and ideally from the genotyping equipment (sequencer, scanner, *etc*) directly to the database,

(d) Operate under strict quality assurance guidelines, such as intellectual property rights, the professional code of ethics (http://ihealthcoalition.org/ethics/ehcode.html) or specific ethical guidelines [Povey et al., 2010], laws on computing and liberties (e.g., www.ama-assn.org/ama/pub/category/1905.html), electronic data protection, specific guidelines and any law or regulation applicable in addition to the local ethical issues that may be specific to a given country.

Such databases should be able to operate remotely, should the required support not be available locally (see also Point 3)

**5. Provide support to developing countries to build capacity and to fully participate in the collection, analysis and sharing of genetic variation information.**

Comprehensive genetic variation data capture usually entails highly specialized researchers or diagnostic laboratory staff and state-of-the-art infrastructure, particularly where generation of new genomic data comes to play. Developing countries usually suffer from brain drain and lack of highly sophisticated equipment to perform cutting–edge research. To this end, developed countries have the obligation to alleviate this burden and to assist developing nations by every means possible [Kaput et al., 2009]. It should be noted that developing countries have a valuable resource of clinical samples, often with very unusual phenotypes, particularly when consanguineous families are involved since in this case, the elucidation of the genetic basis of the underlying phenotype becomes much easier. Developing countries should therefore:



For Peer Review

(a) Educate healthcare providers and government officials to demonstrate the need to initiate documentation of the genetic variation in their countries and to underline the benefits of cooperation in biomedical research [Bhan et al., 2007],

(b) Increase awareness in the general population about the need to comprehensively record the underlying genetic heterogeneity in their country and informing them objectively about the benefits and risks of such projects,

(c) Ascertain all parties involved that the research outcome will not be misused. Certain populations or ethnic groups in developing countries usually mistrust research involving genetic analyses or fear that results can stigmatize them or worse (http://www.eubios.info/ASIAE/BIAE201.htm). For example, Malay-Muslims, Chinese, and Indians in Singapore expressed anxiety about breach of confidentiality, the misuse of their DNA for cloning, and possibilities of being diagnosed with disease [Wong et al., 2004]. Community-based research collaborations may provide fora for addressing cultural and ethical concerns of biomedical research [Povey et al., 2010]. Analyzing the extent of human genome variation creates an ideal opportunity for the developed and the developing nations of the world to forge meaningful partnerships and to work together in an unprecedented way, initially to identify variation causing disease, and then to simplify mutation detection in specific ethnic groups, and then to understand how genetic variation contributes to human phenotypic diversity. By ensuring that all nations and ethnic groups have an equal and fair opportunity to share data and technology, evidence-based information will be provided such that all populations can benefit from a global society health network.

(d) Emphasize in every stage of the collaborative project that external input is needed from the design and coordination of NEMDB development and curation to database hosting and technical support, patient recruitment, and experimental design to performing high-



For Peer Review

throughput experiments in the case of large-scale genome projects (see also Point 3). To this end, funding agencies (e.g. UKIERI; UK-India Education and Research Initiative) could assist by funding innovative projects so that developing countries become independent and self-sustained.

(e) Develop partnerships for the equitable sharing of knowledge and technology between developed and developing countries. These partnerships should be ideally formed under the umbrella of international organizations, such as the HGVS or the Human Genome Organization (HUGO; http://www.hugo-international.org) that would also organize inter-institutional specialized educational programmes and workshops. The biennial mutation detection workshop would be a paradigm for this initiative and may sometimes be held in developing countries.

Deleted: -annual

**6. Develop a framework to facilitate interactions between the coordinating centre and national, regional and international agencies.**

Researchers and their academic institutions that participate in national data collection projects and/or NEMDBs development should jointly develop infrastructures that would ensure smooth communication between them and national and international agencies. These include: (a) national governments and their Departments of Health, Science and Technology, Development, Education and so on, local human genetic societies, (b) regional organizations, such as the NEPAD (New Programme for Africa's Development) and (c) international organizations, such as the European Commission (EC), United Nations (UN), United Nations Children's Fund (UNICEF), World's Health Organization (WHO), Organization for Economic Cooperation and Development (OECD), *etc*.



For Peer Review

This in turn would: (a) create new collaboration ties with neighbouring countries interested in initiating similar projects, (b) increase the chances of receiving funding for these activities (see also below), (c) increase visibility of the progress to the international scientific community.

**7. Ensure that all ethical, legal, religious, and social issues are thoroughly considered when NEMDBs and/or data capture projects are launched in developing countries.**

The goal of establishing LSDBs and NEMDBs in developing countries is the sharing of genetic variation information for the benefit of the respective populations. This includes the protection of privacy, which in this context is the right of the individual and members of their family to be protected against intrusion into their personal information and further intrusions ensuing from access to this, by publication of information. Of course, there is no such issue when summary level data are reported for a population, as it is the case in some NEMDBs. Therefore, there should be a balance between the public's interest in the value of the shared information and its interest in the strict protection of privacy. This balance would be viewed differently in different cultures and ethnicities [Al-Aqeel, 2007].

For many of these NEMDBs, it seems reasonable that an independent group of qualified scientists oversees these resources and related activities not only at their initiation but on an ongoing basis. Where needed, the main purpose of the particular NEMDB should be clarified, recognising that this may change with time. This will allow evaluation of the exact requirements for data deposition in the NEMDBs, and whether compliance with the remaining guidelines will be possible with the database content freely available to the public or partly restricted to registered persons, namely curators, physicians, and so on. This measure was implemented in the Israeli NEMDB, where not all information is made publicly available [Zlogotora et al., 2007)],



For Peer Review

since access to specific cases is password-protected [Zlotogora et al., 2009]. A similar measure would entail anonymizing genetic data, either by permanently removing or coding the personal identifiers [Povey et al., 2010]. Of interest, the newest version of the LOVD software [Fokkema et al., 2005] has the option to store data that are not public but that can be queried. The result of a query hitting non-public data is a notification that there is such information in the database but that the curator needs to be contacted to get more information.

The depth of the ethnic and geographic origin data should be determined in advance, particularly in the case of small villages where carriers of rare variants could be identified even if personal information were not provided. In this case, data donors should be well informed and provided with an explanation as to why their data documentation is useful and how their privacy is secured. These individuals should consent to the inclusion of their genetic information in the database and reserve the right to withdraw their data from the database whenever they desire. This is particularly important for persons unable to consent because of disability or young age. Standardized consent forms need to be translated to different local languages and/or dialects.

Finally, these databases would also benefit from some independent/external ethical review, perhaps from an international group (HGVS or HVP) who could advise them accordingly.

## Nation-wide genomic projects that would serve as paradigm for genetic variation data collection

By definition, a nation-wide genomic project is a national multicentre initiative to determine the underlying genetic heterogeneity of a population with the goal to ascertain the genetic variation leading to common or rare inherited diseases. Such initiatives consist of three



For Peer Review

parts: (a) documenting the genetic heterogeneity that has already been reported in the published literature, (b) generating new genetic data by screening a large proportion of the population using high throughput genotyping approaches, and (c) accumulating data causing inherited disease from hospitals, laboratories and other healthcare facilities. In every case, except perhaps (c), a large national genetic database/portal is being developed to accommodate the sheer amount of data gathered. To date, there are few nation-wide genomic projects in various developing countries that are currently in progress. These projects (summarized in Table 1) implement some of the guidelines described previously. Also, there are several other NEMDB projects, all of which more or less follow these guidelines, that have already been reviewed [Patrinos, 2006; van Baal et al., 2007] and will, therefore, not be further discussed since they lie outside the scope of the present article.

The Estonian Genome Project (EGP; http://www.geenivaramu.ee/index.php?lang=eng) is a research venture of the University of Tartu initiated in 2001 [Metspalu, 2002] and aims to create a database of health, genealogical and genetic data representing approximately 10% of Estonia's population. The database will make it possible for researchers both in Estonia and outside to look for links between genes, environmental factors and common diseases (such as cancer, diabetes, depression, cardio-vascular diseases, *etc*). Currently, EGP gene bank contains data contributed by over 20,000 donors with an expected increase to 100,000 within the next three or four years.

The Indian Genome Variation database (IGVdb; http://www.igvdb.res.in) is a consortium activity established in 2003 to document the inherent genetic variability of the Indian subpopulations as a first step towards identifying genetic biomarkers to understand susceptibility for any genetic disease and population migration, pharmacogenomic markers to understand variable drug response. The Indian Genome Variation (IGV) consortium aims to provide data on



For Peer Review

validated SNPs and repeats, both novel and reported, along with gene duplications, in over 1,000 genes (selected on the basis of their relevance as functional and positional candidates in many common diseases including genes relevant to pharmacogenomics), in 15,000 individuals drawn from Indian subpopulations [Indian Genome Variation Consortium, 2005].

Kuwait established within the Faculty of Medicine, Department of Pathology a Molecular Genetics Diagnostic Service Division (http://www.al-mulla.org), integrated within a research laboratory, which is focused on delivering state-of-the-art diagnostic service for the Kuwaiti population. These services, which focus on autozygosity mapping in families with consanguineous marriages, provide the research laboratory with a unique opportunity to identify novel genes in autosomal recessive disorders. Moreover, various unique cases suffering from cancer are currently being analysed using the Affymetrix Array 6.0 microarray technology to identify copy number changes and homozygosity stretches. The data generated from these studies are deposited in CAGS database (see above), and LOVD, while the Kuwaiti NEMDB has been recently established databases to encourage data storage for genetic diseases in different subpopulations in Kuwait.

The United Kingdom has developed a central database to collect all UK genome variation data, and their effects, from participating diagnostic laboratories. The diagnostic mutation database (DMuDB; http://ngrl.org.uk/Manchester/dmudb.html) provides a route for sharing mutation data within and between diagnostic laboratories in the United Kingdom to support genetic testing services for patients. Currently DMuDB has over 4,884 records for 33 genes, namely Breast Cancer (*BRCA1, BRCA2*), Lowe Syndrome (*OCRL*), Sotos Syndrome (*NSD1*), HNPCC (*MLH1, MSH2*), Cystic Fibrosis (*CFTR*), X-Linked Retinitis Pigmentosa (*XLRP*), Neurofibromatosis, Types I and II (*NF1, NF2*), Alstrom Syndrome (*ALMS1*), CADASIL

**Deleted:** the

**Deleted:** some time ago

**Deleted:** diagnostic and its effect from



For Peer Review

(*NOTCH3*), Familial adenomatous polyposis (*FAP*) and Muscular Dystrophy (*DMD*), containing over 10,985 individual variants.

The Hellenic Genome Variation database aims to document the genetic heterogeneity in the Hellenic population. The first version of the Hellenic NEMDB was created in 2004 [Patrinos et al., 2005]. In 2007, research institutes involved in hemoglobinopathies research initiated a pilot multicenter project to determine the mutation spectrum of the thalassemia syndromes in various geographical regions in the country with the goal to document the findings in the Hellenic NEMDB. These data showed that although Greece is a rather homogeneous country, there are big qualitative (type of mutations) and quantitative (allelic frequencies) differences in various parts of the country, namely central Greece [Samara et al., 2007] and South-Western Greece [Papachatzopoulou et al., 2010]. These data compared to findings from North-Western Greece [Georgiou et al., 2003] and the rest of the country [Patrinos et al., 2005] reveal valuable information for region-specific population carrier screening and for efficient provision of prenatal diagnosis of β-thalassemia in Greece.

Finally, there are four international consortia with activities related to the documentation of causative mutations and allelic frequencies in developing countries:

(a) The International Society for Gastrointestinal Hereditary Tumours (InSiGHT; http://www.insight-group.org) is an international multidisciplinary, scientific organisation with a mission "…to improve the quality of care of patients and their families with any condition resulting in hereditary gastrointestinal tumours." The InSiGHT consortium maintains a LOVD-based database for genetic variations for the *EPCAM*, *MLH1*, *MLH3*, *MSH2*, *MSH6*, *MUTYH*, *PMS1*, and *PMS2* genes leading to gastrointestinal tumours and encourages the establishment of national repositories for patients suffering from gastrointestinal tumours, particularly in developing countries.



For Peer Review

(b) The International Thalassemia Network (ITHANET; http://www.ithanet.eu) is a European Commission-funded project that comprises a network of all major European research institutions active in the field of molecular and clinical research of thalassemia and related hemoglobinopathies [Lederer et al., 2009]. Among its activities, ITHANET hosts a database (ITHANETbase) that includes mutation frequencies in various populations at-risk for thalassemia, particularly from developing countries. In many instances though, this repository is largely redundant to HbVar database of hemoglobin variants and thalassemia mutations [http://globin/bx.psu.edu/hbvar; Patrinos et al., 2004].

(c) The MEDGENET project (http://medgenet.tredueuno.it) is a European Consortium, funded by the European Commission, aiming to expand the human expertise in clinical genetics in Mediterranean Partner Countries (MPC) through the transfer of knowledge and technology between the two rims of the Mediterranean which share a common burden of genetic diseases. Among the actions foreseen by MEDGENET is the building of Medical Genetics databases for the various partner countries. Several NEMDBs are built from scratch, such as the Tunisian and Egyptian NEMDBs or expanded from the existing NEMDBs, such as the Israeli [Zlotogora et al., 2007], Cypriot [Kleanthous et al., 2006], and Lebanese NEMDBs [Megarbane et al., 2007], using the ETHNOS database management system for NEMDB development and curation [van Baal et al., 2010].

(d) The Pharmacogenomics for Every Nation Initiative (PGENI; www.pgeni.org) that aims to determine the frequencies of the commonest pharmacogenomic markers in various populations, particularly from developing countries worldwide.

Finally, there are a few nation-wide DNA biobanks, such as the National Laboratory for the Genetics of Israeli Populations, representing a national repository for human cell-lines and



For Peer Review

DNA from the unique and large ethnic variation of the Israeli populations (Tel Aviv University; available at http://www.tau.ac.il/medicine/NLGIP) or the Biomaterial Banking Medical Research Initative (BBMRI) in Malta, which of course lie at the borderline of the nation-wide genomic projects described above.

## Funding and incentives

Deleted: ation

There are currently two main obstacles that hinder further efforts to expand the NEMDB field, namely funding and providing incentives to data submitters.

Deleted: ation

Deleted:

Funding for collecting genetic variation data has traditionally been difficult due to the extreme fragmentation of the field [Patrinos and Brookes, 2005]. Therefore, it would be expected that efforts to receive funding for genetic variation data capture are more likely to flourish if treated as concerted efforts. GEN2PHEN (http://www.gen2phen.org) represents such initiative that have already received funding from the EC and aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype data, and to link this system into other biomedical knowledge sources via genome browser functionality. Linking population and ethnic-specific information into LSDBs and diagnostic databases, creating NEMDBs and developing genomic databases, similar to the Indian Genome Variation database are key activities for the main GEN2PHEN work-packages. Obviously, other strategically minded funding agencies (e.g. NIH), charities, and foundations would also launch similar calls. Recently the Australian Node of the HVP has been funded to pilot software for data flow from hospitals/clinics via country specific databases to LSDBs and NCBI/EBI/UCSC browsers (unpublished). Involving universal organizations such as the WHO and UNESCO in the fund-raising project will clearly have a strong impact.



For Peer Review

On the other hand, the development of comprehensive analyses of human genetic variation often reaches a barrier in the form of the reward system for clinicians and academic researchers. First of all, database entries could be made a mandatory quality control standard for clinical laboratories and clinicians. Notably, according to a new initiative of the Israeli government, 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 [van Baal et al., 2010].

For researchers, a publication or Web-based system establishing microattribution and community annotation of mutations and cited data will enable measurable contributions to the scientific knowledge base [Anonymous, 2008; Hoffmann, 2008]. Similarly, database-journals would also serve this task by providing a forum for publishing gene variation data that would be eventually deposited in the PubMed literature database. One such journal, the open access Human Genomics and Proteomics (http://www.sage-hindawi.com/journals/hgp), which is affiliated with FINDbase [http://www.findbase.org; van Baal et al., 2007], will focus on studies characterizing causative mutation and/or biomarker frequency spectra. Accepted contributions including datasets will be linked in FINDbase and deposited in PubMed [Patrinos and Petricoin, 2009]. Also, the journal *Human Genetics* has been accepting mutation data and this is linked to HGMD (http://www.hgmd.org).

Journals, tenure, and promotion committees, and funding agencies would be encouraged to cite these contributions to and citations of LSDB and international national databases [Patrinos and Brookes, 2005]. The rapid progress in genetic research has led to a declining interest of scientific clinical genetic journals in case reports with single or few novel mutations. This may deter clinicians from reporting these new mutations, which collectively become important, unless another venue becomes visible.



For Peer Review

## Concluding remarks

As stated above, documenting population/ethnic-specific data in developing countries are among the primary objectives of the HVP that appreciates the genomic sovereignty/equality for all countries to be involved and acknowledges the value of ‘‘human capital’’ within all populations. Real and tangible benefits of the HVP to improve health will be generated for participating populations. Hence, the voluntary participation of the greatest number of countries would ensure a more general applicability, and it is hoped that many countries will eventually decide to participate.

## Acknowledgements

We thank the members of the developing countries working group at the HVP planning and implementation meetings for lively and helpful discussions. Part of this work has received funding from the European Commission [ITHANET (FP6-026539), EuroGenTest (FP6-512148), MEDGENET (FP6-31968) and GEN2PHEN (FP7-200754) projects], and the Kuwait Foundation for the advancement of Sciences (KFAS) grant number 2006-1302-07.

## References

Al Aqeel AI. 2007. Islamic ethical framework for research into and prevention of genetic diseases. Nat Genet 39:1293-1298.



For Peer Review

Anonymous. 2008. Human variome microattribution reviews. Nat Genet 40:1.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Bhan A, Singh JA, Upshur RE, Singer PA, Daar AS. 2007. Grand challenges in global health: engaging civil society organizations in biomedical research in developing countries. PLoS Med 4:e272.

Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Marafie MJ, Paalman MH, Patrinos GP, Qi M, Ramesar RS, Scott RJ, Sijmons RH, Sobrido MJ, Vihinen M. (2009). Capturing All Disease Causing Mutations For Clinical And Research Use: Towards An Effortless System For The Human Variome Project. Genet Med, 11:843-849.

Cotton RG, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, Cantu JM, Cassiman JJ, Claustres M, Concannon P, Cotton RG, den Dunnen JT, Flicek P, Gibbs R, Hall J, Hasler J, Katz M, Kwok PY, Laradi S, Lindblom A, Maglott D, Marsh S, Masimirembwa CM, Minoshima S, de Ramirez HZ, Scriver CR, Sherry S, Shimizu N, Shimizu N, Stein L, Tadmouri GO, Taylor G, Watson M. 2007. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 39:433–436.

Fokkema IF, den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63-68.

Georgiou I, Makis A, Chaidos A, Bouba I, Hatzi E, Kranas V, Zilidis C, Bourantas KL. 2003. Distribution and frequency of beta-thalassemia mutations in northwestern and central Greece. Eur J Haematol 70:75-78.



For Peer Review

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

Hoffman R. 2008. A wiki for the life sciences where authorship matters. Nat Genet 40:1047–1051.

Indian Genome Variation Consortium. 2005. The Indian Genome Variation database (IGVdb): a project overview. Hum Genet 118:1-11.

Kaput J, Cotton RG, Hardman L, Watson M, Al Aqeel AI, Al-Aama JY, Al-Mulla F, Alonso S, Aretz S, Auerbach AD, Bapat B, Bernstein IT, Bhak J, Bleoo SL, Blöcker H, Brenner SE, Burn J, Bustamante M, Calzone R, Cambon-Thomsen A, Cargill M, Carrera P, Cavedon L, Cho YS, Chung YJ, Claustres M, Cutting G, Dalgleish R, den Dunnen JT, Díaz C, Dobrowolski S, Dos Santos MR, Ekong R, Flanagan SB, Flicek P, Furukawa Y, Genuardi M, Ghang H, Golubenko MV, Greenblatt MS, Hamosh A, Hancock JM, Hardison R, Harrison TM, Hoffmann R, Horaitis R, Howard HJ, Barash CI, Izagirre N, Jung J, Kojima T, Laradi S, Lee YS, Lee JY, Gil-da-Silva-Lopes VL, Macrae FA, Maglott D, Marafie MJ, Marsh SG, Matsubara Y, Messiaen LM, Möslein G, Netea MG, Norton ML, Oefner PJ, Oetting WS, O'Leary JC, de Ramirez AM, Paalman MH, Parboosingh J, Patrinos GP, Perozzi G, Phillips IR, Povey S, Prasad S, Qi M, Quin DJ, Ramesar RS, Richards CS, Savige J, Scheible DG, Scott RJ, Seminara D, Shephard EA, Sijmons RH, Smith TD, Sobrido MJ, Tanaka T, Tavtigian SV, Taylor GR, Teague J, Töpel T, Ullman-Cullere M, Utsunomiya J, van Kranen HJ, Vihinen M, Webb E, Weber TK, Yeager M, Yeom YI, Yim SH, Yoo HS. 2009. Planning the Human Variome Project. The Spain report. Hum Mutat 30:496-510.



For Peer Review

Kleanthous M, Patsalis PC, Drousiotou A, Motazacker M, Christodoulou K, Cariolou M, Baysal E, Khrizi K, Pourfarzad F, Moghimi B, van Baal S, Deltas CC, Najmabadi S, Patrinos GP. 2006. The Cypriot and Iranian National Mutation databases. Hum Mutat 27:598-599.

Lederer CW, Basak AN, Aydinok Y, Christou S, El-Beshlawy A, Eleftheriou A, Fattoum S, Felice AE, Fibach E, Galanello R, Gambari R, Gavrila L, Giordano PC, Grosveld F, Hassapopoulou H, Hladka E, Kanavakis E, Locatelli F, Old J, Patrinos GP, Romeo G, Taher A, Traeger-Synodinos J, Vassiliou P, Villegas A, Voskaridou E, Wajcman H, Zafeiropoulos A, Kleanthous M. 2009. An electronic infrastructure for research and treatment of the thalassemias and other hemoglobinopathies: the Euro-mediterranean ITHANET project. Hemoglobin 33:163-176.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Berman J, Rogers YH, Frazier ME, Scherer SW, Strausberg FL, Venter JC. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Mégarbané A, Chouery E, van Baal S, Patrinos GP. 2006. The Lebanese National Mutation frequency databases. Eur J Hum Genet 14(Suppl 1):365.

Metspalu A. 2002. Estonian Genome Project--before the take-off and take-off. Bioinformatics 18 Suppl 2:S152.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Sitrnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamanta CD, Roses AD, Burns DK, Ehm MG, Lia EH. 2008. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347–358.



For Peer Review

Papachatzopoulou A, Vantarakis A, Kourakli A, Stavrou E, Frangou E, Patrinos GP, Athanassiadou A. 2010. Documentation of the incidence of beta-thalassemia and *HBB* mutation frequencies in South-Western Greece. Hemoglobin 34:333-342.

Patrinos GP. 2006. National/Ethnic mutation databases: Recording populations' genography. Hum Mutat 27:879-887.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: Disastrously deficient. Trends Genet, 21:333-338.

Patrinos G, Petricoin E. 2009. A new scientific journal linked to a genetic database: towards a novel publication modality. Hum Genomics Proteomics 1:e597478.

Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, Anagnou NP, Wajcman H, Hardison RC. 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 32:D537-D541.

Patrinos GP, van Baal S, Petersen MB, Papadakis MN. 2005. The Hellenic National Mutation database: A prototype database for inherited disorders in the Hellenic population. Hum Mutat, 25:327-333.

Povey S, Al Aqeel AI, Barash CI, Cambon-Thomsen A, Dalgleish R, den Dunnen JT, Firth HV, Greenblatt M, Parker M, Patrinos GP, Savige J, Sobrido MJ, Cotton RG. 2010. Practical Guidelines addressing ethical issues pertaining to the curation of human Locus specific variation databases (LSDBs). Hum Mutat (in press).

Saadallah AA, Rashed MS. 2007. Newborn screening: experiences in the Middle East and North Africa. J Inherit Metab Dis 30:482–489.

Samara M, Chiotoglou I, Kalamaras A, Likousi S, Chassanidis C, Vagena A, Vagenas C, Eftichiadis E, Vamvakopoulos N, Patrinos GP, Kollia P. 2007. Large-scale population genetic



For Peer Review

analysis for hemoglobinopathies reveals different mutation spectra in central Greece compared to the rest of the country. Am J Hematol 82:634-636.

Singer PA, Daar AS. 2001. Harnessing genomics and biotechnology to improve global health equity. Science 294:87–89.

Tadmouri GO, Al Ali MT, Al-Haj Ali S, Al Khaja N. 2006. CTGA: the database for genetic disorders in Arab populations. Nucleic Acids Res. 34:D602-D606.

van Baal S, Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, Macek Jr M, Scriver CR, Patrinos GP. 2007. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 35:D690–D695.

van Baal S, Zlotogora J, Lagoumintzis G, Gkantouna V, Tzimas I, Poulas K, Tsakalidis A, Romeo G, Patrinos GP. 2010. ETHNOS: A versatile electronic tool for the development and curation of National Genetic databases. Hum Genomics 4:361-368.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, FanW, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Feng X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang M, Wang J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60–65.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Marqulies M, Weinstock GM, Gibbs RA, Rothberg GM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.



For Peer Review

Wong ML, Chia KS, Wee S, Chia SE, Lee J, Koh WP, Shen HM, Thumboo J, Sofjan D. 2004. Concerns over participation in genetic research among Malay-Muslims, Chinese and Indians in Singapore: a focus group study. Community Genet 7:44–54.

Zlotogora J, van Baal S, Patrinos GP. 2007. Documentation of inherited disorders and mutation frequencies in the different religious communities in Israel in the Israeli National Genetic Database. Hum Mutat 28:944-949.

Zlotogora J, van Baal S, Patrinos GP. 2009. The Israeli National Genetic Database. IMAJ 6:373-375.

For Peer Review

**Box 1**: Guidelines and recommendations for data capture in developing countries

1. Develop procedures to include clinicians and researchers in developing nations, mostly by creating networks among those of common interest.
2. Create international networks from closely related populations to enhance communication, interaction, and initiate research networks between them.
3. Foster communication, interaction, and research networks between developing and developed countries.
4. Ensure that the database management systems that are being used or developed can be utilized by those in a limited resource environment.
5. Provide support to developing countries to build capacity and to fully participate in the collection, analysis and sharing of genetic variation information.
6. Develop a framework to facilitate interactions between the coordinating centre and national, regional and international agencies.
7. Ensure that all ethical, legal, religious, and social issues are thoroughly considered when NEMDBs and/or data capture projects are launched in developing countries.

Deleted: s

For Peer Review

**Deleted:** **Table 1**. Nation-wide genome and related projects (N/A: Not available).¶
¶
**Project** ... [1]

For Peer Review

Page 30: [1] Deleted ADMIN 10/15/2010 11:15:00 AM

**Table 1**. Nation-wide genome and related projects (N/A: Not available).

| Project | URL | Aims and scope |
|---|---|---|
| **Nation-wide genome projects** | | |
| The Estonian Genome Project | http://www.geenivaramu.ee/index.php?lang=eng | To create a database of health, genealogical genome data representing approximately 10 Estonia's population |
| The MEDGENET Project | http://medgenet.tredueuno.it | To expand the human expertise in clinical g Mediterranean Partner Countries (MPC) tha common burden of genetic diseases. |
| The Indian Genome Variation database | http://www.igvdb.res.in | To create a DNA variation database of the p India and make it available to researchers fo understanding human biology with respect t predisposition, adverse drug reaction, popu migration |
| The Catalogue of Transmission Genetics in Arabs | http://www.cags.org.ae/ctga_search.html | To provide the user with clinical and molec descriptions of genetic disorders, epidemiol details of reports of the disorders in the Ara and clinical case details wherever available |
| **Other related projects** | | |
| Kuwaiti Molecular Genetics Diagnostic Service | http://www.al-mulla.org | To provide autozygosity mapping services, on Arab genome and molecular karyotying |
| UK diagnostic mutation database | http://ngrl.org.uk/Manchester/dmudb.html | To provide a route for sharing mutation dat between diagnostic laboratories in the Unite to support genetic testing services for patier |
| Hellenic Genome Variation database | http://www.goldenhelix.org/hellenic | To create a compendium for genetic disease encourage genetic studies to elucidate the g of inherited diseases in Greece |
| The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) | http://www.insight-group.org | To improve the quality of care of patients a families with any condition resulting in her gastrointestinal tumours |
| International Thalassemia Network (ITHANET) | http://www.ithanet.eu | To strengthen research and treatment of hemoglobinopathies |
| Pharmacogenomics for Every Nation Initiative | http://www.pgeni.org | To determine the frequencies of the commo pharmacogenomic markers in various popu particularly from developing countries worl |

For Peer Review

**We thank both referees and the communicating and managing editors for their very positive and constructive criticism. Please find below our reply to the comments in boldface.**

Referee: 1
Comments to the Author
This paper represents an interesting extension of previous work carried out by the authors and is very much in keeping with the federation of network of LSDBs and CGDBs recently proposed by Patrinos et al. (Locus-specific database domain and data content analysis: Evolution and content maturation towards clinical use. Human Mutation, Vol. 31, Issue 10, 2010).
However in my opinion one important aspect is completely missing among the recommendations for genetic variation data capture in developing countries. Although the authors acknowledge "that developing countries have a valuable resource of clinical samples, often with very unusual phenotypes, particularly when consanguineous families are involved ...." they do not stress enough among their recommendations the importance of collecting information about the possible consanguinity of parents of each individual patient and of recording the causative mutation(s) the patient carries, when this information is available.
As Dr. Patrinos knows very well, today it is possible to develop an approach for estimating the relative frequency of autosomal recessive disorders using mutation data originated by the molecular genetic centers which have diagnosed the affected offspring of consanguineous couples. The rationale of this new approach is based on the possibility that the child born of consanguineous parents carries the same mutation in double copy (true homozygosity) or alternatively carries two different mutations in the same gene (compound heterozygosity). In the latter case the two mutations must have been inherited through two different ancestors of the consanguineous parents (in such case the two mutated alleles are not "identical by descent"). In a very recent paper, Leo ten Kate et al. showed a positive correlation between compound heterozygosity and gene frequency (L.P. Ten Kate et al.: Autosomal recessive disease in children of consanguineous parents: inferences from the proportion of compound heterozygous. J Community Genet. Jan 2010).
The interesting aspect of this new genetic epidemiological approach is that it makes use of data already existing in molecular diagnostic laboratories. I believe it would be very valuable to find this information also in mutation databases. The practical benefit for developing countries will be that the genetic epidemiological approach just mentioned will yield not only a ranking order of the frequencies of rare monogenic disorders in each of them but also an estimate of prevalence.
I strongly encourage the authors to add this recommendation to their paper and to promote a comprehensive worldwide data collection also of this information.

**We thank Referee 1 for his comments. Based also on the Communicating editor's comments, we have chosen to add a couple of sentences in the Introduction to emphasize the Referee 1 comments on consanguinity.**

Referee: 2
Comments to the Author
This manuscript is a laudable and timely effort to provide guidelines on the collection of human genetic variation data in the developing world, and to offer advice on how these data may be gathered in a fashion and format which maximizes their usefulness

For Peer Review

to the local (national) and international communities. This paper will be of interest to readers of Human Mutation and I recommend it for publication, pending one change and a few minor points detailed below.

Major point:

p6 l29: "...reasonable detail of phenotype..." is a rather vague description. Whilst the authors need not explain in detail what "reasonable detail" means, they should perhaps rephrase with something like "...appropriate detail..." or "...level of detail agreed by the community...". Could some relevant references to other phenotype collection projects be included e.g. P3G, GEN2PHEN, Biobanks, etc.?

**We have rephrased this sentence and add some references from relevant projects.**

Minor corrections:

Abstract, last sentence: insert "a" before "few" - "We also summarize a few well-coordinated genetic data..."

p4 l58: suggest naming the "...respective working group..." (respective to what?).

p5 l19/20: "...mapping projects, i.e., the HapMap..." - should this be "e.g." rather than "i.e."?

p8 l17: Remove "The" preceding "Latin America"

p9 l50: Replace "...the databases management systems..." with "...the database management systems..."

p17 l12/13: Replace "The United Kingdom has developed the central database some time ago to collect all UK genome variation data from diagnostic and its effect from participating laboratories." with "The United Kingdom has developed a central database to collect all UK genome variation data, and their effects, from participating diagnostic laboratories."

p19 l27/28: Replace "...there are few nation-wide DNA biobanks..." with "...there are a few nation-wide DNA biobanks..."

p19 l44/45/ and 51/52: "incentivation" is a little (if ever) used portmanteau word and should be avoided. Perhaps replace with "motivation", "encouragment", or "incentive".

**We thank Referee 2 for thoroughly going through the text and providing the places that corrections are required. We have fully corrected all minor points indicated.**

EDITORIAL BOARD'S COMMENTS
This is a very useful manuscript which summarizes a small set of important recommendations to be considered regarding the collection of genetic variation data

For Peer Review

in developing countries. This paper should be published with top priority, pending some revision.

The authors are asked to attend to the points raised by Reviewer 2, all of which are minor.

**We have fully addressed Referee's 2 comments**

The comments made by Reviewer 1 require some consideration, and some input/response/discussion from the authors would be welcome. Reviewer 1's comments seem perfectly reasonable, but care must be taken not to unbalance the manuscript. The authors might like to consider these options (or suggest another):

1) add a sentence or two, emphasizing Reviewer 1's point, in the Introduction where consanguinity is mentioned;
2) make an additional recommendation along the lines of "the minimum set of data recommended by HUGO-MDI/HGVS/HVP be collected, but that more data should be gathered if possible", then provide examples of additional data, e.g. consanguinity, and why it is important?

**We have chosen to add a couple of sentences in the Introduction to emphasize the Referee 1 comments on consanguinity.**

Please make edits to the manuscript using "track changes" if possible, or by highlighting added text.

**Changes are made using the "track changes" option and two file versions are provided, one with the changes indicated and another one with the changes accepted, the latter file to be forwarded for typesetting.**

MANAGING EDITOR COMMENTS:

Please respond to these points under "Response to Managing Editor", otherwise the final decision could be delayed.

1) Paper organization:
a-Please ensure that the subheadings are presented in the proper hierarchy to convey your meaning. For example, While Introduction is a top-level heading and Background is subordinate to Intro, the numbered recommendations actually appear as subordinate to Background. A new top-level heading called "Recommendations" might be necessary before the paragraph starting "The recommendations and guidelines…"

**The new top-level heading has been added.**

b-Please cite "Box 1" in the text.

**Box 1 has been cited in the text.**

2) On resubmission:

For Peer Review

a-Please double check the author names and affiliations carefully. These are often a source of typographical errors.

**Names were double-checked.**

b-Title Page, Abstract (180-200 words max), Key Words, Main Text, and References should be combined into one file for the manuscript and submitted as a *.doc file.

**Title Page, Abstract (180-200 words max), Key Words, Main Text, and References were combined and submitted as one main manuscript (.doc format).**

c-Tables must be submitted individually as separate *.doc files (with their titles and legends included). Please use the MS Word table format if possible. Excel (*.xls) files should not be submitted.
--Do not use custom paper sizes - only use Letter or A4.

**Table 1 has been saved and submitted separately to the main text**

3) Please check to see that the references follow the journal's standard format and are cited properly. See our online Author Instructions
http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291098-1004/homepage/ForAuthors.html

**Reference format is according to the journal's style.**

4) Human Mutation abides by the NIH Mandate. If your work was funded by the NIH, be sure to include a grant number in an "Acknowledgments" section of the manuscript right before the References. Visit this site for more information:
http://www.wiley.com/go/nihmandate

**Not applicable**

5) IMPORTANT INFORMATION REGARDING PREPRINTS
a-Human Mutation is now publishing online preprints of accepted manuscripts prior to typesetting and page proof corrections. It is therefore crucial that you revise your manuscript carefully so that errors (typographical and grammatical) are corrected BEFORE the final accepted manuscript is posted online. The accepted preprint version will remain online until the corrected proofs are received and the typeset manuscript is finalized. At that time, the preprint version will be replaced with the final, typeset version online, in Early View.

b-It is essential that you submit a copyright transfer agreement (CTA) upon submission of your revised manuscript. This will avoid delay in publication of your article upon acceptance. If possible, the CTA must be signed by the corresponding author and should be signed by all contributing authors if practical. All authors must be made aware of the CTA and the rights it conveys to them.
The CTA can be found here:
www.wiley.com/go/ctaaus
and must be filled out completely, including the article title and manuscript number. Please fax it to this number: (201) 748-6091.

For Peer Review

**The CTA has been filled in and faxed to the above number.**

6) N.B.: Referee #1 offered to forgo anonymity and discuss the issue raised with you if necessary. However, the editorial board placed considerably less emphasis on Reviewer 1's comments than Reviewer 2's and suggested a means of dealing with that. It it up to you whether you wish to communicate with Referee 1 and I'd be happy to put you in contact if you wish. Please write me as needed.
**We feel that there is no need to communicate with Referee 1 since the issue was addressed based on the Communicating editor's suggestions.**

For Peer Review