Capturing all disease-causing mutations for clinical and research use: Toward an effortless system for the Human Variome Project

Richard G. H. Cotton, PhD, DSc[1,2], Aida I. Al Aqeel, BSc, MD[3], Fahd Al-Mulla, PhD[4], Paola Carrera, PhD[5], Mireille Claustres, MD, PhD[6], Rosemary Ekong, PhD[7], Valentine J. Hyland, PhD[8], Finlay A. Macrae, MBBS, MD[9], Makia J. Marafie, MD, PhD[10], Mark H. Paalman, PhD[11], George P. Patrinos, PhD[12,13], Ming Qi, PhD[14,15], Rajkumar S. Ramesar, MSc, PhD[16], Rodney J. Scott, PhD, PD[17], Rolf H. Sijmons, MD, PhD[18], María-Jesús Sobrido, MD, PhD[19,20], Mauno Vihinen, PhD[21,22]; and members of the Human Variome Project Data Collection from Clinics, Data Collection from Laboratories and Publication, Credit and Incentives Working Groups

Abstract: The collection of genetic variants that cause inherited disease (causative mutation) has occurred for decades albeit in an ad hoc way, for research and clinical purposes. More recently, the access to collections of mutations causing specific diseases has become essential for appropriate genetic health care. Because information has accumulated, it has become apparent that there are many gaps in our ability to correctly annotate all the changes that are being identified at ever increasing rates. The Human Variome Project (www.humanvariomeproject.org) was initiated to facilitate integrated and systematic collection and access to this data. This manuscript discusses how collection of such data may be facilitated through new software and strategies in the clinical genetics and diagnostic laboratory communities. ***Genet Med* 2009:11(12):843–849.**

Key Words: *mutation, LSDB, Human Variome Project, inherited disease, databases*

It is well known that research laboratories and particularly diagnostic laboratories have accumulations of described variations in patients, which cause inherited diseases that have not been made publicly available for a variety of reasons. These variants are usually referred to as "mutations" in the clinical setting but increasing numbers of "unclassified variants" are also being described as they cannot unequivocally be termed mutations. For the purposes of this article, the term "mutation" will be used throughout. If all mutations in all genes worldwide and their corresponding phenotypes were available in the one place (the main aim of the Human Variome Project [HVP]: www.humanvariomeproject.org), this would assist diagnostic laboratories, therapists, clinicians, and carers as well as researchers, not only in caring for those with inherited disease but also aid in our understanding of the pathogenetic basis of disease and ultimately save significant amounts of health-care funds.

The HUGO Mutation Database Initiative was created to address these problems, and this was subsequently formed into the Human Genome Variation Society (www.hgvs.org) and many key papers have been and continue to be published to assist in collection and storage of mutations and their effect (http://www.hgvs.org/biblio.html).

Because many excellent efforts were occurring in isolation and were, and still are, divided by gene, country, and a single expertise of those involved, and a lack of knowledge of other systems and efforts, the HVP was initiated by a group of relevant bodies and experts in 2006[1,2] (http://www.humanvariomeproject.org). This resulted in publication of 96 recommendations[3] that indicated the parlous state of the field. Consequent to this, a planning meeting was held (http://www.humanvariomeproject.org/files/Hvpprog.pdf), which outlined exact efforts required and reported efforts underway as a guide.[4] Since this meeting, dramatic progress has been made in establishing pilots for data thereby paving the way for a complete collection of mutations that are defined by their phenotypic effect, which will be an invaluable worldwide resource. The collaborators are currently focusing on collection of mutations and not variants that do not cause inherited disease.

From the [1]Genomic Disorders Research Centre, Melbourne, Australia; [2]Department of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Australia; [3]Department of Paediatrics, Riyadh Military Hospital, Riyadh, Kingdom of Saudi Arabia; [4]Molecular Pathology Unit, Kuwait University, Faculty of Medicine, Safat, Kuwait; [5]Unit of Genomics for Diagnostics of Human Disease and Laboraf, San Raffaele Scientific Institute – Dibit 2, Milano, Italy; [6]Université Montpellier 1, Faculté de Médecine et CHU, Laboratoire de Genetique Moleculaire, IURC, Montpellier, France; [7]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom; [8]Molecular Genetics Laboratory, QHPS-Central Laboratory, Royal Brisbane Hospitals Campus, Herston, Australia; [9]Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, Parkville, Australia; [10]Department of Clinical Genetics, Kuwait Medical Genetics Centre, Maternity Hospital, Sabah Medical Area, Kuwait; [11]Human Mutation, Wiley-Blackwell, Hoboken, New Jersey; [12]Department of Pharmacy, University of Patras, School of Health Sciences, Patras, Greece; [13]Department of Bioinformatics, Erasmus University Medical Center, Faculty of Medicine and Health Sciences, Rotterdam, The Netherlands; [14]ADINOVO Center for Genetic and Genomic Medicine, The First Affiliated Hospital of Zhejiang University School of Medicine, James Watson Institute of Genomic sciences, Beijing Genome Institute, Hangzhou, Zhejiang, People's Republic of China; [15]Center for Cardiovascular Research, University of Rochester Medical Center, New York; [16]Department of Human Genetics, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town; Observatory Western Cape Province, South Africa; [17]Discipline of Medical Genetics, Faculty of Health, University of Newcastle, Newcastle, Australia; [18]Department of Genetics, University Medical Center Groningen, Hanzeplein 1, Groningen, The Netherlands; [19]Fundacion Publica Galega de Medicina Xenomica, Santiago de Compostela, Spain; [20]Center for Network Biomedical Research on Rare Diseases (CIBERER), Institute of Health Carlos III, Madrid, Spain; [21]Institute of Medical Technology - Bioinformatics Group, University of Tampere, Finland; and [22]Research Unit, Tampere University Hospital, Tampere Finland.

Richard G. H. Cotton, PhD, DSc, Genomic Disorders Research Centre, Level 2, 161 Barry Street, Melbourne, VIC 3053 Australia. E-mail: cotton@unimelb.edu.au.

Disclosure: The authors declare no conflicts of interest.

Submitted for publication July 2, 2009.

Accepted for publication September 29, 2009.

DOI: 10.1097/GIM.0b013e3181c371c5

Table 1 Location of and access to published mutations

Refereed publications	PubMed/Google/bIOpORtAL indexed
	Local genetics journal not PubMed or Google indexed
Dependent repositories	Textbooks, e.g., Scriver et al.
	HGMD, OMIM, MutDB
	LSDBs, NEMDBs, disease-specific databases (e.g. BIC, InSiGHT)
	NCBI, UCSC, DECIPHER, UniProt

Table 2 Repositories of unpublished mutations

Diagnostic laboratories	Public hospital/public pathology diagnostic laboratories
	Private hospital/public pathology diagnostic laboratories
	Private hospital/private pathology diagnostic laboratories
Clinicians formal patient record	Public/private
	Hospital/clinic/clinician's formal patient record
Research laboratories	Public
	Commercial
Registries of specific hereditary diseases	E.g., Genzyme for storage diseases
Locus-Specific Databases	

This study addresses ways to ensure continued collection of data from clinics, diagnostic, and research laboratories. It has been found by surveys[5] that by far, most data reside unpublished in diagnostic laboratories and also that Locus-Specific Databases (LSDBs) contain around 50% unpublished mutations.[6]

REPOSITORIES OF MUTATIONS

Published mutations

Clearly, published mutations are scattered throughout the literature in somewhere between 20 and 30 journals at a minimum, especially as each subdiscipline has its own journal (Table 1). This makes searching for specific mutations through PubMed, Google, Google Scholar, or bIOpORtAL a tedious and time-consuming process, especially when earlier descriptions need to be corrected. Notwithstanding the publically available mutations, there are many journals containing mutations published in specific countries that are not indexed in PubMed (such as the *Journal of Medical Sciences* [Dubai]) which will be missed when undertaking a search. Furthermore, because there have been revisions of gene nomenclature, there is a plethora of data that have to be reinterpreted such that it conforms to new numbering systems that have been introduced with increased knowledge of the human genome. Considerable assistance in accessing published mutations is given by OMIM (http://www.ncbi.nlm.nih.gov/omim/), MutDB (http://mutdb.org/), HGMD (http://www.hgmd.cf.ac.uk/), and UniProt (http://www.uniprot.org/). OMIM deals with the first mutation or mutations judged by the curators as significant, MutDB at the protein level, and HGMD attempts to collect all reported mutations but its commercial status means that the public version is considerably delayed. Wildeman et al.[7] determined that only 25, 38, and 87% of variants in the PAH, BIC BRCA2, and HbVar databases were error free. This confirms earlier studies that showed literature data contained 5%,[8] 10% (D. Ravine, personal communication), and 43%[9] errors in three studies. Thus, proper curation in LSDBs or other databases is clearly essential.

Many published mutations also appear in the LSDBs and serve a convenient capture purpose (Table 1).

Unpublished mutations and data on their effects

Unpublished mutations and corresponding phenotypes are held in a finite number of key places (Table 2). Naturally, this includes both digital and paper listings in the diagnostic laboratory collections, published books, the clinicians' formal records held in the clinic or hospital, and research laboratories and registries dedicated to collecting detailed phenotypic data of patients with particular hereditary disorders.

LSDB curators often collect the unpublished mutations not only from their own laboratories but also from colleagues around the world. Often these listings are published (e.g., in a *Human Mutation* Mutation Update article) and later placed on the Internet in the form of an LSDB.

Additional LSDB portals for registering mutations have been encouraged by shared interests. Institutes with mutual interest in acquiring screening/diagnostic markers for diseases that exhibit better survival at early detection as well as genetic polymorphism (e.g., breast cancer) have prompted the initiation of disease-specific databases (e.g., Breast Cancer Mutation database; BIC). Such databases allow for refinement of submitted data, because it is constantly being reviewed/viewed by relevant health-care professionals and is being maintained and supported by federal/governmental agencies such as the NIH (http://research.nhgri.nih.gov/bic/). The International Society for Gastrointestinal Hereditary Tumors (InSiGHT; www.insight-group.org) is currently developing their database for colon cancer.

THE CHALLENGE OF COLLECTION OF ALL MUTATIONS (NOT JUST ONCE BUT FOR EVERY INSTANCE) AND THEIR EFFECT WORLDWIDE

The collection of mutations from all the sources mentioned above at first seems a massive task. However, because the data on mutations and their effects are used continuously around the world for research and particularly for clinical guidance and more recently for mutation-specific therapy, a way needs to be found. Thus, this is beyond collection for collection's sake, which of course in its own right is useful for some researchers in population movements for example. A key reason to collect all mutations is to ensure that the clinical interpretation is consistent with all users for any specific mutation. In coming years with the advent of cheaper sequencing of human genomes, there will be an extension of personalized medicine to an ever increasing number of individuals. Health-care professionals will need to have improved processes to determine how gene-specific gene variants affect human health. Increasingly, pathogenic mutations may be detected before any clinical presentation. Potentially, there may even be a demand for prenatal analyses for presymptomatic individuals who are carriers of pathogenic mutations and symptomatic phenotypically affected individuals. Further, considerable health benefits could ensue when couples plan to have a child. Thus, their genome se-

quences could be compared before conception, and there could be an alert when deleterious mutations (as indicated in a catalogue of mutations and their effects) are found in each of them in the same gene or in different genes known to contribute each to a particular shared phenotype (e.g., cases of digenic inheritance). This is an extension of the current strategy for thalassaemia in Greece where its incidence has been dramatically reduced. This would then allow the parents to make an informed decision based on this information depending on personal, cultural, and societal norms. At the same time, finding mutations will decrease the economic burden on the health-care system, the family, and the society. This is because prevention strategies are more economical than taking care of an affected patient over a lifetime. This will also overcome and will over ride the risk of stigmatization in a family once prevention measures are taken.[10]

The high prevalence of common and novel disorders in developing countries highlights the importance of research to identify their genetic basis, including mutational analysis, and should encourage an international disease testing network with developed countries. In newborn screening laboratories, identifying mutations is important for the correlation of the mutation and the phenotype and for trying to find strategies for therapy at mutational level in case of nonsense mutations.[11] Ethical issues in databasing mutation are a vexed question. Critical questions include the following: should identifying information be removed before submission to a database, should there be different levels of access to data for the public and professionals, and should consent be obtained before database submissions. These issues have been discussed in the Islamic framework[10] and recently attempts have been made to tailor guidelines from major bodies such as UNESCO, WHO, and HUGO to the needs of those curating genotype/phenotype data[12] and a manuscript with recommendation is in preparation.

The collection of mutations would not only include the novel ones but also all instances, so the frequency of the "well-known" mutations in different ethnic population can be ascertained. This is of particular importance for the clinical genetic testing in the developing countries. In China, for example, all the clinical testing is paid by the patient. Many patients cannot afford full screening of the whole coding sequence but only targeted mutation testing. If we know that there are some dominant mutations or hot spots, the testing will be much more efficient. It will also be money saving if we know a mutation is not present in a specific ethnic population.

There is clear recognition that emerging countries have an immense amount to offer, particularly in light of the known depth of genetic variation in indigenous African populations. In general, and relating to emerging countries, there are a small number of these that have dedicated clinical, diagnostic, and research facilities aimed at identifying disease-associated or-causing mutations. Nonetheless, one does need to bear in mind the wealth of information that lies there, in terms of the mission of the HVP. An attempt will be made to address some of the means for extracting data in the relevant sections hereunder.

Although at this stage just the crude collection of mutations identified worldwide is a challenging task, the issue of ensuring the quality and significance of the data collected comes immediately into play. Thus, not all the mutations published in the literature or in existing databases are pathogenic and not all the reported polymorphisms benign. Furthermore, the correlation between a given mutation and the associated clinical phenotype is often not linear. Expert curation of mutation databases is necessary to filter for errors and misclassification problems caused by a mere listing of variants. Multidisciplinary curating teams, bringing together experts in genetics, molecular pathogenesis, and clinical (phenotypic) data, would ensure the highest quality and relevance of the information available to the community.

Table 3 Mechanisms to collect and current collection of published mutations

HGMD, OMIM, DMUTDB, UCSC, dbSNP, DECIPHER, UniProt
Publishers mandate data submission on publication
LSDBs
In published listings
On the Internet

Another challenging issue is to develop appropriate mechanisms and legal context to allow mutation frequency and population distribution to be appropriately captured in databases. The current situation in LSDBs is that they often contain only one reference to the detection of a given mutation (usually the first known report). It is of great interest both for research and clinical interest to have information on the relative frequency of a given DNA variation. Currently, this knowledge must be searched through published reviews in conventional journals and may not even be available. This is because subsequent reports of the same mutation are not usually published.

MECHANISMS AND STRATEGIES FOR COLLECTING MUTATIONS AND PHENOTYPES INTO DATABASES

Published mutations

Published mutations are time consuming to access when spread across the literature and HGMD is serving an excellent purpose in collecting these but the public version is delayed (Table 3). However, if scientific journals could be persuaded to enforce data submission to databases as a condition of publication as per sequence information this would help enormously. This could be done by simply having a tick box "mutation submitted to database □," which needs to be ticked before publication.

LSDB curators may help not only by collecting mutations but also by publishing their data as well as placing them on the Internet.

Unpublished mutations

This is the area of greatest need and requires substantial and definite novel action. In simple terms, a new paradigm is needed. On the basis that the most profligate users of the data are those who also produce such data, it is logical that the producers should provide data for use by others. There are many reasons why this data have not been made available or published in the past (Table 4).

The most important of these are the lack of clear official portal or mechanism for submission and lack of incentive or time/funding to do so.

NEW DATA AND LEGACY DATA

Legacy or existing data

Many diagnostic and research laboratories will have genotype and phenotype data stretching back a decade or more. This will take a considerable time in most instances to transfer to a public database. Support for this transfer will have to be funded

Table 4 Reasons why unpublished mutations are not made available

No incentives to publish
Journals will not publish, e.g., 54th mutation in a gene
No direction to make data available
No unified, recommended method to make data available
Publication not important in developing careers of diagnostic laboratory or clinical staff
Diagnostic laboratories have no time to submit/publish mutations
Lack of clinical data to draw conclusions on pathogenicity
Commercial interest in safeguarding the information (private laboratories)
Diagnostic laboratories have no funded and trained personnel to submit/publish mutations
Lack of access to (or authorization of) a depository portal into a collecting database
No interest in reporting repeated mutations that might be of epidemiological, strategic, demographic value
Diagnostic laboratories have their own database of mutations and those of colleagues they rely on
Current consent may not be adequate

in some way either by government, patient support groups, or companies "adopting a gene" or by health professional trainees, students, or unpaid volunteers. Another source of legacy data are the registries that have collected detailed phenotypic data, and in later years also genotypic data, on particular hereditary disorders, e.g., storage disorders. Depending on the level and organization of the registry databases, extracting relevant phenotype-genotype data might be less of a technical effort. However, these registries only exist for a minority of the known hereditary disorders.

It should also be realized that, outside dedicated research settings, laboratories usually do not have good phenotype data. Medical records often do include these data, however, the percentage of data available worldwide in an electronic format, i.e., in electronic health records, is still relatively low.

Ethical issues will be different here because for full data release permission may need to be obtained, database access limited or no phenotype data put in the database.

New data

When data are produced, a system that will allow simple and seamless transmission of genotype and phenotype data to public databases from the computers of diagnostic laboratory scientists and clinicians, without extra work, the task will be simple. In the absence of this, the collection and submission will be less efficient and less convenient but systems must be developed for this interim stage.

In simple terms, if a government wants each person who is ill looked after optimally, funding should be given for collection of data from laboratories and clinics. This was achieved for Duchenne Muscular Dystrophy in a few instances in Australia (proposed government funding for a Duchenne Muscular Dystrophy Registry) and France (complete collection of data from laboratories and clinics).[13]

MECHANISMS AND STRATEGIES TO ENSURE PUBLIC CAPTURE OF MUTATIONS

Various methods to ensure capture of mutations are summarized in Table 5.

Publishing incentives and directives

Submission of data as a condition of publication is not new as this has occurred in the area of DNA sequence and three dimensional structure of macromolecules (in PDB) for many years. Because the mutations affect human health and will lead to improved diagnostics, prognostics, therapy, and care, it would seem logical that such a system for mutations would be at least as appropriate, if not more so, than sequence data. Pilot studies are being initiated.

Other activities in the publishing sphere include special conditions for articles reviewing mutations in a gene. Thus, senior authors are expected to write to all known laboratories in the world to request unpublished data in return for authorship on a paper publishing them. This policy has been initiated at *Nature Genetics* and *Human Mutation*, but need not be restricted to them. Maybe there could be two types of authorship; one for those who submit data and the other for those who analyze and write the story. However, this kind of system does not exist yet, except when describing the contributions of authors. The availability of new gene- and country-specific variant databases will result in a streamlined collection process and will allow identification of submitters to enable the database curators to prepare manuscripts for publication. This process usually leads to production of an online database if one does not exist.

A reward for publishing in databases has been proposed recently[14] and/or this has been termed "microattribution." This will involve use of the unique accession numbers (SSID) and a stable reference sequence. When data are accessed, a hit is registered and maintained for use in CV, etc. Electronic publishing is also a possibility but cost may be a disincentive to submission especially for those from developing countries.

Conditional strategies

Similarly, mutation submission could be part of the accreditation and licensing of a diagnostic laboratory. Every diagnos-

Table 5 Mechanisms and strategies to ensure mutation collection

Journals demand data submission as a condition for publication
Novel publishing strategies to ensure reward for publishing
Super reviews
Electronic publication
Microattribution
Database journal
Mutation submission a condition of grants involving mutation discovery
Mutation submission a condition of laboratory licensing
Mutation submission a condition of quality assessment of laboratories
Mutation submission part of local reporting regime
Education of relevant clinicians and scientists
Country- and state-specific collection

tic laboratory should participate in an appropriate local or international gene quality assurance program. Evidence of use of a specific local or international database for the presence of a mutation should be a standard part of any quality assurance program evaluation. Evidence of submission of all their detected mutations to a local or international database could be a requirement for registration for any external quality assurance program.

Variation submission as a condition of a grant is not new with the NIH grants for SNP discovery being conditional on submission to a dbSNP at NCBI. A sentence could easily be added to grant instructions making it necessary for EVERY variation (not only novel ones) discovered to be submitted to variation databases.

Procedural strategies

It is widely known that diagnostic laboratory staff may not have time to submit mutations and their effects as a separate exercise. In most cases, the diagnostic laboratory does not have the phenotype data. The submission should come after the physician has made the diagnosis, i.e., from the hospital. A simple strategy to achieve submission to public databases with a minimum or zero extra time would be to have an electronic form for the patients' genetic data (filled in by the diagnostic laboratory head) and clinical data (filled in by the clinician) and at the time of submission, or at a later stage, a program could be developed to extract and deidentify the relevant data and forward it to a public database.

National registries for diseases and their mutations should be encouraged and implemented by a government body, with a plan to include such data into an international disease database.

Education and training of health-care professionals working with genetic information areas could include training in the use and the importance of mutation databasing. Human genetics meetings should include sessions on databasing.

In some countries such as the United States, reporting of instances of cancer is obligatory. As such a system is already established and for inherited disease data are essential for proper genetic health care, there is a strong argument for the protocol to be extended to inherited disease.

Country-specific strategies

Reasons for needing country-specific activities in mutation documentation are shown in Table 6.

There are instances in Australia where two family members from different branches of the same family in different health systems with a specific disease and the family members with the same mutations were treated differently because of different interpretation of the mutations. This could be avoided with a nationwide or countrywide approach. Numerous examples from everyday diagnosis can be cited where previous knowledge of regional mutation distribution and the geographical area of origin of the patient can quickly (and with no cost) lead to the identification of the causal mutation. A recent example from the authors was that of a patient referred for the genetic study of familiar Tourette syndrome. However, because the patient was from a specific local area where two other instances of the same myoclonus dystonia mutation had been detected, this was the first mutation screened for and identified. Such a screen costs a few tens of dollars compared with a few thousand dollars for a complete screen. The necessary genetic information to make this connection and thus help orient the diagnosis was not available to the referring clinician, because it was only in the genetics laboratory records that the epidemiologic-genetic-phenotypic link could possibly be established. Time- and resources-consuming efforts, misdiagnosis, inappropriate genetic counseling, and other catastrophes could be avoided in this case. Similarly, the importance of ethnic-specific collection for example in the Jewish diaspora is exemplified by the Ashkenazi mutations where cost is considerably diminished by testing for these mutations dominant in a specific disease.

Table 6 Reasons for needing country-specific mutation collection of all mutations

Ensuring coherent genetic healthcare for family branches in different states to assist in classification and therapy
Documentation of disease found
Development of accurate and needed care strategies
Development of relevant and economical diagnostic strategies for the dominant ethnic group and their diaspora around the world or ethnic groups within a country, e.g. China
Spread the load of mutation collection
Ensure worldwide complete collection by redundancy
Establishment of nationwide carrier screening and prevention/ treatment strategies
Decreasing the economic burden on the health care system
Use in classifying variants as pathogenic (mutation) or not by increased numbers

Cultural and religious imposition is a significant factor in hindering efficient background data collection (and use) in some countries, specifically those with poor genetic representation in the literature. Sufficient awareness and relevant professional involvement to abolish the obscurity to the public and emphasize the importance of such data are required.

The collection by curators worldwide and country-specific collection will lead to some redundancy, but at least in the beginning this will be a good thing. However, country-specific efforts will more easily be able to document each instance of a specific mutation that is essential for connecting families and mutation screening strategies. Country-specific activities have been initiated since 2002 (Table 7; reviewed by Patrinos[15]), and database management systems for developing and ultimately curating national-/ethnic-specific genetic databases, such as ETHNOS,[16] have been made available, contributing to data uniformity. In certain instances, these efforts have been financially supported by funding agencies.

Similarly, Rare Metabolic Diseases Database (RAMEDIS; http://www.ramedis.de/) has been designed to be a hospital-based system in Germany and the diagnostic mutation database, DMuDB (http://www.ngrl.org.uk/Manchester/dmudb.html) is a centralized facility for depositing diagnostic laboratory data on individuals in the United Kingdom. The Australian system will be designed to be a pilot with a decentralized regime and cheap portability to other countries. Time will tell which will be the method of choice for other countries.

With regard to capturing data from emerging countries: The work of the HVP has already been brought to the attention of the International Federation of Human Genetic Societies (Cotton R, unpublished data). Their general support may well be leveraged for the work of HVP to gain traction with the work that should be practically done by, e.g., African Society for Human Genetics (AfSHG) and other Societies representing emerging countries. With the AfSHG, for example, a major commitment is to facilitate

Table 7 Existing country-specific activities

Finnish disease heritage	http://www.findis.org
Israeli population database	http://www.goldenhelix.org/israeli
Singapore Human Mutation and Polymorphism Database	http://shmpd.bii.a-star.edu.sg/
FINDbase	http://www.findbase.org
RAMEDIS—Germany	http://www.ramedis.de/
CETT/caBIG—USA	http://rarediseases.info.nih.gov/cettprogram/default.aspx
DMuDB—UK	https://secure.dmudb.net/ngrl-rep/Home.do
Hellenic Genetic Database Consortium	Initiated[14]
Australian node of HVP	Recently initiated
China-HVP	http://china-hvp.org/LOVD/home.php
Korean node of HVP	Recently initiated
Centre of Arab Genomics Studies (CAGS)	http://www.cags.org.ae/

education and training of African researchers. A project currently undertaken by the AfSHG has to do with developing a resources map of disease epidemiology, institutional curricula, research projects, and support structures being dedicated to African-related genetic projects. Although this is meant to cover Africa, it will of necessity aim to collect data from institutions elsewhere in the world where research is being done on African material. One may be able to gauge that an inventory of disease-related mutations will be a necessary product of this survey-based African database.

As mentioned in another section of this article, the alignment of education/training opportunities (e.g., workshops), together with the actual work being done by postgraduate students must necessarily include tasks such as gathering "mutation data," which may underpin a creditable task, e.g., a module of their Honors or Masters program. If this is set up as a task of a large enough body, e.g., AfSHG, there is little doubt that data will be gathered, but more importantly, that an effort may well be made to generate data on African populations in different countries.

Quite importantly, for Mendelian (and even non-Mendelian disorders), it is worthwhile attempting to capture as much information as exists from emerging countries. This must go hand in hand with thorough phenotyping. Many countries who do not do mutation screening, for the moment, still have excellent data on clinical phenotypes. It is important to empower these individuals (clinicians) to be part of the effort to characterize the mutations (which soon enough will happen at very low cost). It is an important means of being inclusive. The important scientific aspect for the clinical phenotypes (and research fraternity) will be the understanding of phenotypic variation related to a mutation but in relation to (i) environment and (ii) genetic background (modifying genes).

It should be noted that polymorphisms (apparently not pathogenic variants) should be collected by diagnostic laboratories also to assist in classification of the status of variants as polymorphisms are often found when sequencing disease genes. Thus all variants should be collected.

Finally, it is noteworthy that a novel publication modality, namely a database journal, has been launched to provide incentives for publishing national-/ethnic-specific genetic data. *Human Genomics and Proteomics* is a new genomics and systems biology journal that is affiliated with an international, open access database: FINDbase (http://www.findbase.org).

KEY PILOT STUDIES ACROSS DIFFERENT COUNTRIES

Many of the mechanisms, collaborations, and software to develop a seamless pipeline for mutation data and their effects are available worldwide. Because of needs, InSiGHT has a major project underway to develop this pipeline from existing resources and software. Some software will have to be improved or developed. Thus, this group will be the ultimate track-testing group for current and novel software. Considerable progress has been made in aggregating three existing databases and loading up large amounts of data from different countries.

Country-specific pilots that will be developed and watched worldwide are the UK group (DMuDB), the Indian Genome Variation Database (SNP only), the CETT group (US), the Hellenic Genetic database consortium (Greece), China-HVP, and the recently constituted Australian Node of the HVP.

FACTORS MITIGATING AGAINST CAPTURING MUTATIONS CAUSING DISEASE

The factors that need to be addressed when developing a system for routine collection of mutations and their effect are shown in Table 8.

Many of these concerns are being addressed at present, but this is where most work is needed. It could well be that different genes or countries will develop different strategies. However, it is hoped that all data will be curated and sent for permanent storage and use at NCBI, EBI, or UCSC browsers.[17] Data in LSDBs should be transferred only with the conditions indicated in this study[17] to central databases. There has been no discussion about the permanent storage between LSDBs and central databases. However, wherever data reside, curation must be by experts in each of the genes.

DISCUSSION

There is considerable expressed need and support for the collection of all variations causing human disease (mutations). In another arena, massive sums have been spent capturing common and more recently, rare variations (termed SNPs) >1% frequency that are associated with human disease, e.g., asthma, and it is hoped this will lead to the definition of variations associated with human

Table 8 Factors mitigating against mutation capture

Time available for submission
Mutations and clinical detail documented by different individuals
Ethical considerations
New paradigm needed
New software needed/electronic heath records not yet widely used
Accepted portal needed to relevant database
Accepted database needed
Fear that submission to a database would count as prior publication, precluding publication in a journal which is not true

disease. Mutations causing disease are much rarer than the SNPs associated with common diseases, for example phenylketonuria and cystic fibrosis, both rather common inherited diseases, caused by mutations being present 0.001 and 0.003% respectively. Thus, the 1000 genomes project (http://browser.1000genomes.org/index.html) will uncover few mutations causing disease when 1000 complete genomes are sequenced.

This study indicates where mutations reside after description and some of the strategies that might be used to gather them into public databases. It also outlines the problems that may be overcome with some spending, effort, and new technology.

We provide some key recommendations that will enhance the rate of collection/submission of mutations to public databases and thus available to all. These derive from some of the 96 developed of 2006 HVP meeting.[3] It is possible that some strategies will be more appropriate in some countries than others and some cheaper and more readily implemented than others.

RECOMMENDATIONS

1. The bodies responsible for accreditation of laboratories, quality assessment, and licensing add "submission of mutations and their effects to public databases" to the criteria required for assessment/practice, e.g., CLIA, EuroGentest.
2. The national- or country-wide bodies, such as the Australian node of the HVP in association with human genetics, cancer, and clinical genetics societies, should be established to ensure collection of data on all instances of each mutation and subsequent submission to public databases.
3. National registries for diseases and their mutations should be encouraged and implemented by a government body, with a plan to include such data into an international disease databases.
4. The software should be developed to trivialize the process of collection of data out of hospital laboratories and clinics. This would also contribute toward data uniformity.
5. The mutation collection/submission should be a part of routine genetic health-care practice and training (professional development).
6. An independent review committee to oversee existing data where 'inadequate' consent is preventing their release for inclusion in LSDBs.
7. When consent is deemed inadequate, the new consent forms are drawn-up seeking permission for inclusion of data in online databases.
8. The funding, research and health-care agencies encourage and support mutation database curating activities, especially those that are multidisciplinary to ensure that they are coordinated and that the needed paradigm shift to deliver accurate and economical translation of genetic discovery to health care occurs.
9. This document be forwarded and considered by bodies able to facilitate relevant recommendations including health departments, human and medical genetics societies, pathology societies, and any society interested in the genetics of their diseases.

ACKNOWLEDGMENTS

The authors thank Lauren Martin, Alex Kline, and Rania Horaitis for their assistance in the preparation of the manuscript.

The following people have expressed their support for this article: Jumana Al-Aama, Saudi Ababia; Bharati Bapat, Canada; M. Rosário N. dos Santos, Portugal; Maurizio Genuardi, Italy; Sandrine Laradi, France; and Gabriela Möslein, Germany. The members of the working groups at the time of writing this manuscript are listed below: Human Variome Project Data Collection from Clinics—Aida Al Aqeel and Jumana Al-Aama, Saudi Arabia; Fahd Al-Mulla, Kuwait; Bharati Bapat, Canada; Inge Bernstein, Denmark; Peter Byers, US; Paola Carrera, Italy; Garry Cutting, US; Maurizio Genuardi, Italy; Annika Lindblom, Sweden; Finlay Macrae, Australia; Makia Marafie, Kuwait; Gabriela Möslein, Germany; George Patrinos, The Netherlands; Ming Qi, China; David Rimoin, US; Rolf Sijmons, The Netherlands; María-Jesús Sobrido, Spain; Thomas Weber, US; Human Variome Project Data Collection from Laboratories—Fahd Al-Mulla, Kuwait; Bharati Bapat, Canada; Stacey Bleoo, Canada; Yeun-Jun Chung, Republic of Korea; Mireille Claustres, France; M. Rosário dos Santos, Portugal; Rosemary Ekong, United Kingdom; Valentine Hyland, Australia; Sandrine Laradi, France; Annika Lindblom, Sweden; Jillian Parboosingh, Canada; George Patrinos, The Netherlands; Ming Qi, China; Sue Richards, US; Rodney Scott, Australia; María-Jesús Sobrido, Spain; Graham Taylor, United Kingdom; Sylvie Tuffery Giraud, France; Mollie Ullman-Cullere, US; Human Variome Project Publication, Credit and Incentives—Myles Axton, US; Steven Brenner, US; John Burn, United Kingdom; Richard Cotton, Australia; Johan den Dunnen, The Netherlands; Robert Hoffmann, United States; Sandrine Laradi, France; Ana Maria Oller de Ramirez, Argentina; Mark Paalman, US; George Patrinos, The Netherlands; Giuditta Perozzi, Italy; Ming Qi, China; Rajkumar Ramesar, South Africa; Daniela Seminara, US; Gert-Jan van Ommen, The Netherlands.

REFERENCES

1. Ring HZ, Kwok PY, Cotton RG. Human Variome Project: an international collaboration to catalogue human genetic variation. *Pharmacogenomics* 2006;7:969–972.
2. Axton M. What is the Human Variome Project? *Nat Genet* 2007;39:423.
3. Cotton RG, Appelbe W, Auerbach AD, et al. Recommendations of the 2006 Human Variome Project meeting. *Nat Genet* 2007;39:433–436.
4. Kaput J, Cotton RG, Hardman L, et al. Planning the Human Variome Project: the Spain report. *Hum Mutat* 2009;30:496–510.
5. Cotton RG, Phillips K, Horaitis O. A survey of locus-specific database curation. *J Med Genet* 2007;44:e72.
6. Cotton RG, Horaitis O. Quality control in the discovery, reporting, and recording of genomic variation. *Hum Mutat* 2000;15:16–21.
7. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* 2008;29:6–13.
8. Gout AM, Ravine D, Harris PC, et al. Analysis of published PKD1 gene sequence variants. *Nat Genet* 2007;39:427–428.
9. Murphy JA, Barrantes-Reynolds R, Kocherlakota R, Bond JP, Greenblatt MS. The CDKN2A database: integrating allelic variants with evolution, structure, function, and disease association. *Hum Mutat* 2004;24:296–304.
10. Al Aqeel AI. Islamic ethical framework for research into and prevention of genetic diseases. *Nat Genet* 2007;39:1293–1298.
11. Welch EM, Barton ER, Zhuo J, et al. PTC124 targets genetic disorders caused by nonsense mutations. *Nature* 2007;447:87–91.
12. Cotton RG, Sallee C, Knoppers BM. Locus-specific databases: from ethical principles to practice. *Hum Mutat* 2005;26:489–493.
13. Tuffery-Giraud S, Beroud C, Leturcq F, et al. Genotype-phenotype analysis in 2,405 patients with a dystrophinopathy using the UMD-DMD database: a model of nationwide knowledgebase. *Hum Mutat* 2009;30:934–945.
14. Axton M. Human variome microattribution reviews. *Nat Genet* 2008;40:1.
15. Patrinos GP. National and ethnic mutation databases: recording populations' genography. *Hum Mutat* 2006;27:879–887.
16. Patrinos GP, van Baal S, Petersen MB, Papadakis MN. Hellenic National Mutation database: a prototype database for mutations leading to inherited disorders in the Hellenic population. *Hum Mutat* 2005;25:327–333.
17. den Dunnen JT, Sijmons RH, Andersen PS, et al. Sharing data between LSDBs and central repositories. *Hum Mutat* 2009;30:493–495.