

Recommendations of the 2006 Human Variome Project meeting

Richard G H Cotton & participants of the 2006 Human Variome Project meeting

Lists of variations in genomic DNA and their effects have been kept for some time and have been used in diagnostics and research. Although these lists have been carefully gathered and curated, there has been little standardization and coordination, complicating their use. Given the myriad possible variations in the estimated 24,000 genes in the human genome, it would be useful to have standard criteria for databases of variation. Incomplete collection and ascertainment of variants demonstrates a need for a universally accessible system. These and other problems led to the World Health Organization–cosponsored meeting on June 20–23, 2006 in Melbourne, Australia, which launched the Human Variome Project. This meeting addressed all areas of human genetics relevant to collection of information on variation and its effects. Members of each of eight sessions (the clinic and phenotype, the diagnostic laboratory, the research laboratory, curation and collection, informatics, relevance to the emerging world, integration and federation and funding and sustainability) developed a number of recommendations that were then organized into a total of 96 recommendations to act as a foundation for future work worldwide. Here we summarize the background of the project, the meeting and its recommendations.

Disease-causing gene mutations (or variations) in humans were first clearly described at the molecular level in 1949, although it was not until the 1970s that the collection of human mutations was started by Victor McKusick. Today, his OMIM is an excellent resource for human genetics, but it is unable to collect all mutations. The Human Gene Mutation Database (HGMD), which has collated the most mutations to date, collects from the literature. The HGMD is not complete, as many genes do not have up-to-date lists of mutations, the data are fragmentary and public access is considerably delayed because it is a commercial entity. Publishing mutations is becoming increasingly difficult, and service laboratories do not consider reporting to be an essential part of their work, partly because they do not

have an efficient way to report the variants they encounter during molecular diagnostic testing.

The lack of systematic collection of mutations led to a meeting in Montreal in 1994 (chaired by Richard Cotton) of some of the world's leading geneticists, who concluded that as experts in genes are the best curators, the collation of mutations is most efficiently arranged by a federation of locus-specific database curators. This would lead to curation meeting a standard not possible in central databases (such as OMIM or the HGMD). This 1994 meeting led to the foundation of the Human Genome Organization (HUGO)-sponsored Mutation Database Initiative¹, which recently became a society, the Human Genome Variation Society (HGVS, <http://www.hgvs.org>). The society aims to promote the collection and display of mutations causing single-gene disorders and related polymorphisms. Its members have produced various recommendations and software to promote collection and display of mutations.

More recently, the importance of genetic variation possibly causing common diseases and adverse drug reactions has led to an enormous investment in catalogs of polymorphisms

(commonly known as SNPs) used to map genes and in association studies. These sequence variations have been collected mainly in the US National Institutes of Health (NIH)-sponsored dbSNP and HGVbase and the industrially financed SNP Consortium (TSC) database². After collecting SNPs, researchers turned their efforts toward determining which SNPs are real and also defining an appropriate set of mapping and association studies. Another aspect requiring curation is the accumulation of reports of association studies, both positive and negative. For example, what association studies have been done with what SNPs in asthma? Or, for what diseases have association studies been done with specific SNPs at a specific gene? This unsatisfactory situation calls for a well-organized international consortium to avoid the current inevitably fragmented and disconnected efforts that have resulted in systems that may be less than ideal in the collection and navigation of data.

Mutation collection has been very much on an *ad hoc* basis, with major initiatives led by Victor McKusick (OMIM)³ and David Cooper (HGMD)⁴, which is now commercial but has plans for future free, noncommercial access)

Richard G.H. Cotton is at the Genomic Disorders Research Centre, St. Vincent's Hospital Melbourne, 35 Victoria Parade, Melbourne, Victoria 3065, Australia. The complete list of the authors appears at the end of the paper.
e-mail: cotton@unimelb.edu.au

Published online 28 March 2007; doi:10.1038/ng2024

and with mutations in specific genes being collected by curators on the basis of need (for example, see PAHdb⁵ at <http://www.pahdb.mcgill.ca>). There are currently some 672 such genes with databases. MutationView, an aggregation of locus-specific databases (LSDBs), is an intermediate between these classes (<http://mutview.dmb.med.keio.ac.jp>). Certain databases allow direct submission of mutations.

Various journals publish mutations, but it has become more difficult to publish, say, the 25th mutation in gene 'X'. The journal *Human Mutation* attempts to trigger the creation of databases by inviting and promoting summaries of mutations ('mutation updates').

In the SNP area, the NIH and drug companies have provided large sums of money to find SNPs and collect them in their databases. More recently, a listing of somatic mutations in genes in cancer (COSMIC) has been promoted by the Human Cancer Project (<http://www.sanger.ac.uk/genetics/CGP/cosmic/>).

The need for international cooperation in these efforts is underscored by the fact that migration around the world is massive, particularly from low-resource countries to high-resource countries. Thus, the latter need to know the mutations specific to particular ethnic groups to deliver the most efficient healthcare. From the perspective of the HGVS, we have encouraged ethnicity- and country-specific databases of mutations not only for this reason but also as a further net with which one day to collect all mutations. From an individual country's perspective, such databases assist in the delivery of genetic healthcare.

Obviously, there is no common focus for these activities besides that of harmful variation itself. Inasmuch as some 60% of all humans will be affected by mutations in their lifetime⁶, there is a clear need for a common base of variations.

This situation prompted a meeting in Melbourne in June 2006 (<http://www.humanvariomeproject.org>) that addressed these problems and provided recommendations (listed in the **Supplementary Note**) on how to achieve the global objective of efficient, complete collection and accurate curation. Currently, collection and curation efforts are fragmented, and some databases do not relate well to others both in a practical and an informatics sense. The field is diverse, ranging from gene-specific databases to ethnicity-specific databases (for example, a database containing information on Arab populations) to cancer databases and OMIM and HGMD, to dbSNP and TSC databases. The collection of SNP data, although massive, is not coordinated, except by dbSNP, the HapMap databases and some commercial companies. There is need for consensus

and unification and for improvement in cross-talk, collection and curation. Feasibility to search a complete database is crucial to all researchers, regardless of their mission. Only joint action will allow economization and result in better information.

Objectives of the meeting

The aims of the meeting were as follows.

1. To gather representatives of all sections of the human gene variation industry and community, including clinicians, diagnosticians and informatics experts and general, locus-specific and ethnicity-specific or national database curators.

2. To define the current situation in various key areas of content, software, collection, completeness, compatibility, availability, ability to search, access, suitability, etc.

3. To identify problems and deficiencies.

4. To identify areas of need and areas with the potential for collaboration.

5. To develop guidelines for the collection and display of variations.

6. To develop guidelines to capture all variations and their phenotypes.

Meeting outcome

At the meeting, delegates from various organizations related to genetic health and funding agencies agreed to launch the Human Variome Project (HVP; see <http://www.humanvariomeproject.org>)⁷. In simple terms, the project aims to systematically collect human gene variation information with associated phenotype information and make it generally available to those who need it. This will involve global collaboration, with a number of major interacting projects developed, funded and carried out by working groups. Although the exact details of the project business plan will be developed and finalized by the working groups, the project is based on the 96 recommendations coming from the meeting (**Supplementary Note**) and the considerable work done worldwide by many people.

The HVP will collect, curate and electronically record DNA variations (alleles), in the form of either mutation or polymorphism, in the canonical human genomic nucleotide sequence, with emphasis on disease and other phenotype relationships. This process is feasible because the Human Genome Project has generated a reference nucleotide sequence for the human species and because there is a systematic vocabulary for genes, exons, mutations, etc. LSDBs provide 'inch-wide, mile-deep' views, whereas genomic repositories give complementary 'mile-wide, inch-deep' perspectives. LSDB data sets typically concentrate on variations that have a major or 'causative'

direct influence on one or more disease-related phenotypes. Genome-wide variation databases tend to contain neutral variation and variants that modify the resulting protein only slightly or are associated only indirectly with disease. In other cases, genome-wide databases represent mutations in genes but fail to distinguish benign mutations from those that cause disease, or they have little detail or do not contain all mutations and are often delayed before being published.

Mission

The Human Variome Project aims to improve health outcomes by facilitating the unification of data on human genetic variation and its impact on human health. It will support the use of human variation information in clinical environments across the world by developing the resources required to undertake the following key objectives.

1. Capture and archive all human gene variation associated with human disease, via gene-specific curation in a central location. Provide mirror sites in other countries to maximize data security and integrity and allow searching across all genes using a common interface.

2. Provide a standardized system of gene variation nomenclature, reference sequences and support systems that will enable diagnostic laboratories to use and contribute to total human variation knowledge.

3. Establish systems that ensure adequate curation of human variation knowledge from a gene-specific (locus-specific), country-specific or disease-specific database perspective to improve accuracy, reduce errors and develop a comprehensive data set covering all human genes.

4. Facilitate the development of software to collect and exchange human variation data in a federation of gene-specific (locus-specific), country-specific, disease-specific and general databases.

5. Establish a structured and tiered mechanism that clinicians can use to determine the health outcomes associated with genetic variation. This will allow communication between those who use human variation data and those who provide them. Clinicians will be encouraged to provide data and will have open access to complete variation data.

6. Create a support system for research laboratories that provides for the collection of genotypic and phenotypic data together using the defined reference sequence in a free, unrestricted and open access system, and create a simple mechanism for logging discoveries.

7. Develop ethical standards that ensure open access to all human variation data that are to be used for the global public good and address

the needs of 'indigenous' communities under threat of dilution in emerging countries.

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

9. Establish a communication and education program to collect and spread knowledge related to human variation to all countries of the world.

10. Continue to carry out research investigating human genetic variation, and present findings to users of this information for the benefit of all.

The mission will be pursued and the objectives will be realized by implementing at least the 96 recommendations of the June 2006 meeting, which were finalized October 6th, 2006 (**Supplementary Note**). Below, we describe part of the plan developed from some of these recommendations.

The working groups

The 96 recommendations (**Supplementary Note**) from the delegates at the 2006 Human Variome Project meeting fall under 12 headings, each of which can be a major project:

- The clinic and phenotype
- Diagnostic laboratories
- Variation/linkage of common diseases/research laboratory
- Informatics and central databases
- Curation, collection and locus-specific databases
- Developing countries' international liaison
- Funding and sustainability
- Nomenclature and standards
- Ethics and education
- Publication and scientific journals
- Translation
- Coordinating center

Each topic will have a chair and co-chair, and they will form a project working group. Because of the detailed nature and number of the 96 recommendations and because of the status of those making the recommendations, a large amount of planning has already been done. It now remains to obtain funding to implement the recommendations. The first step in this process was the development of a concept plan in late October 2006.

Progress so far

Major progress of the consortium so far includes standards for database content, variation nomenclature and entry forms, the WayStation (an input portal; see <http://www.centralmutations.org>), tailored software and a coordination office. Other items can be seen on the HGVS website (<http://www.hgvs.org>).

Summary of recommendations

The full recommendations are shown in the **Supplementary Note**. Here we summarize the key points in each of the areas.

1. Coordinating office

The office has essentially been coordinating activities since 1996 and has been involved in almost all activities. It was voted to be located at the Genomic Disorders Research Centre in Melbourne under the direction of Richard Cotton, the Human Variome Project convenor. The office is to develop an oversight committee with key individuals representing key organizations. These will then develop into a steering committee, board and working groups. The recommended business plan will be developed for fundraising purposes. Governments are to be lobbied to facilitate data collection. Participants suggested that a World Health Organization-accredited list of databases should be developed. Stakeholders must be informed by newsletters, meetings and a website.

2. The clinic and phenotype

The clinic and the symptoms of patients are fundamental to the project. The recommendations center on the importance of accurate and complete documentation of phenotype with simplification through electronic means. Bioinformaticians and clinicians should work together to create systems capable of evolution and accredit a range of data types and users. Guidance and incentives should be given to all those involved to collect, report and interpret data. Standards should be developed for machine-readable data and structured descriptions of disease-specific phenotypes, by way of data sheets.

3. Diagnostic laboratories

Diagnostic laboratories should be encouraged to participate. This would be made easier by encouraging that consent forms and archiving and databasing should be part of quality control or licensing. Design of databases should allow for inclusion of less-wealthy societies. Prediction of the effect of variants is crucial in this area.

4. Variation/linkage of common diseases/research laboratory

Disease-specific databases are to be supported, as well as a network of networks in collaboration with HuGENet and the Human Genome Variation Society. Reporting of haplotype associations with normal and mutant alleles as well as low-frequency, putatively functional variants should be generated from medical resequencing projects.

5. Informatics and central databases

Thus far, the focus of the massive central databases such as the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) in

the area of variation has been high-throughput SNP data. These recommendations, which focused on systems needed for clinical variation, included (i) the use of standard data models (ii) the synchronization of sequence with build (iii) the placement of variation-on-assemblies links to LSDBs (iv) the design of LSDBs as a unified set but with individually curated development of user-friendly interfaces (v) maximal automation and (vi) maximal linkages between databases.

6. Curation, collection and locus-specific databases

LSDBs should be accredited by the Human Variome Project and use an agreed reference sequence, nomenclature and database standard, and they should accommodate high-quality clinical and phenotype information, state each instance of a variant, flag data quality and encourage material collection and sequence submissions. A federation of curators is to be formed with informatics support and state-of-the-art open-source software. Any sequence possible should be submitted to the public databases to ensure searchability. An expert group of curators for each gene should be organized.

7. Developing countries' international liaison

The Human Variome Project should be regarded as working toward the global public good and thus should be inclusive, develop capacity and skill in the emerging world, enhance collaboration at national and international levels, encourage sample preservation, involve emerging countries in the generation of molecular data, seek pharmacogenetics knowledge and encourage and fund training programs and HVP reference centers within the emerging country groupings.

8. Funding and sustainability

The Human Variome Project should assure support for curation and development of databases, and to ensure their continuity, the World Health Organization, the United Nations Educational, Scientific and Cultural Organization (UNESCO) and other similar organizations should have critical roles in encouraging government support. A consortium of funding bodies should be formed along with a practical funding model, and an appropriate not-for-profit body should be established, political visibility should be promoted, a business plan with a working group should be developed and family and patient groups should be engaged in the organizations. Scientists in the emerging world should be supported.

9. Nomenclature and standards

Guidelines should be developed for sharing genetic and genomic information and promoting

the use of standard reference sequences from the current build that relate directly to BACs. Systems to translate readily from older to newer nomenclature should be developed, data quality will be paramount, submission should include flanking sequences as part of a submission standard, the use of the word 'variants' should be encouraged and systematic allele nomenclature and standardized data formats should be used.

10. Ethics and education

The highest ethical standards for sharing intellectual property should be the basis for the activity, an educational program should be developed to gain traction, and ethical and legal issues should be followed and identified.

11. Publications and scientific journals

Reporting criteria for phenotypic data with use of online data for large clinical tables should be developed. Published data, including positive and negative association data (with some criteria for quality control), should be deposited in databases, and an open-access electronic database journal is suggested.

12. Translation

Collaboration should occur between research clinicians, clinical laboratories and patient advocates to assist in translating discoveries of new tests to the clinic and to promote genetic research as an indispensable tool for the progress of healthcare in developing countries, where prenatal diagnosis should be encouraged. An international rare-disease testing network should be formed to offer quality service, and genetic variants must be linked to clinical outcome.

Current situation

Systematic and targeted spreading of the details of the promise of the Human Variome Project has elicited enthusiastic responses from major funding and health-related organizations as well as governments and pharmas. Funding sources are being systematically targeted for grant applications. We are fortunate that two major genetics meetings have allowed sessions on this topic in 2007 and that key publishers have become interested. However, we await substantial funding until the massive effort required can begin. The business plan should be available by April 2007.

Conclusion

With the Human Genome Project all but complete and with massive medical resequencing and genome-wide association studies about to begin, the time has come for the complete

human gene sequence to be annotated from one end to the other with each variation found and a standard link to all the available information on it.

Gathering high-throughput data is relatively easy, but gathering of information on related phenotypes is not so simple. The gathering of data on single-gene disorders and their phenotypes worldwide is low-throughput and challenging and will need the involvement of perhaps thousands of individuals. Not only will the developing world contribute, but it will also benefit. This, then, is the challenge of the Human Variome Project, which could be regarded as the 'globalization' of human and clinical genetics.

ACKNOWLEDGMENTS

Collection of fragments of the total picture was begun by V. McKusick, D. Cooper, A. Brookes, NCBI, EBI and others, and the vision of complete and coordinated collection was kept alive by a dedicated group of core members of the HUGO-Mutation Database Initiative/Human Genome Variation Society and their funders.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

1. Cotton, R.G. *Hum. Mutat.* **15**, 4–6 (2000).
2. Porter, C.J., Talbot, C.C. Jr. & Cuticchia, A.J. *Hum. Mut.* **15**, 1236–1244 (2000).
3. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. *Nucleic Acids Res.* **33**, D514–D517 (2005).
4. Stenson, P.D. *et al.* *Hum. Mutat.* **21**, 577–581 (2003).
5. Scriver, C.R. *et al.* *Hum. Mutat.* **21**, 333–344 (2003).
6. Baird, P.A., Anderson, T.W., Newcombe, H.B. & Lowry, R.B. *Am. J. Hum. Genet.* **42**, 677–693 (1988).
7. Ring, H.Z., Kwok, P.Y. & Cotton, R.G. *Pharmacogenomics* **7**, 969–972 (2006).

The complete list of authors is as follows:

William Appelbe¹, Arleen D Auerbach², Kevin Becker³, Walter Bodmer⁴, D Joe Boone⁵, Victor Boulyjenkov⁶, Samir Brahmachari⁷, Lawrence Brody⁸, Anthony Brookes⁹, Alastair F Brown¹⁰, Peter Byers¹¹, Jose Maria Cantu¹², Jean-Jacques Cassiman¹³, Mireille Claustres¹⁴, Patrick Concannon¹⁵, Richard G H Cotton¹⁶, Johan T den Dunnen¹⁷, Paul Flicek¹⁸, Richard Gibbs¹⁹, Judith Hall²⁰, Julia Hasler²¹, Michael Katz²², Pui-Yan Kwok²³, Sandrine Laradi²⁴, Annika Lindblom²⁵, Donna Maglott²⁶, Steven Marsh²⁷, Collen Muto Masimirembwa²⁸, Shinsei Minoshima²⁹, Ana Maria Oller de Ramirez³⁰, Roberta Pagon³¹, Raj Ramesar³², David Ravine³³, Sue Richards³⁴, David Rimoin³⁵, Huijun Z Ring³⁶, Charles R Scriver³⁷, Stephen Sherry³⁸, Nobuyoshi Shimizu³⁸, Lincoln Stein³⁹, Ghazi Omar Tadmouri⁴⁰, Graham Taylor⁴¹ & Michael Watson⁴²

¹Victorian Partnership for Advanced Computing, Melbourne, Victoria, Australia. ²The Rockefeller

University, New York, New York, USA. ³National Institute on Aging, Bethesda, Maryland, USA. ⁴Weatherall Institute of Molecular Medicine, Oxford, UK. ⁵National Center for Health Marketing, US Centers for Disease Control and Prevention, Atlanta, Georgia, USA. ⁶World Health Organization, Geneva, Switzerland. ⁷Institute of Genomics and Integrative Biology, Delhi, India. ⁸National Human Genome Research Institute, Bethesda, Maryland, USA. ⁹University of Leicester, Leicester, UK. ¹⁰Medical Research Council, Human Genetics Unit, Edinburgh, Scotland, UK. ¹¹University of Washington, Seattle, Washington, USA. ¹²Universidad de Guadalajara, Guadalajara, Mexico. ¹³Center for Human Genetics, University of Leuven, Leuven, Belgium. ¹⁴CHU de Montpellier, Montpellier, France. ¹⁵Benaroya Research Institute at Virginia Mason, Seattle, Washington, USA. ¹⁶Genomic Disorders Research Centre, Melbourne, Victoria, Australia. ¹⁷Leiden University Medical Center, Leiden, The Netherlands. ¹⁸European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. ¹⁹Baylor College of Medicine, Baylor, Texas, USA. ²⁰University of British Columbia, Vancouver, British Columbia, Canada. ²¹Natural Sciences Sector, United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France. ²²March of Dimes Birth Defects Foundation, Mamaroneck, New York, USA. ²³University of California San Francisco, San Francisco, California, USA. ²⁴Faculté de Pharmacie, Monastir, Tunisia. ²⁵Karolinska Institutet, Stockholm, Sweden. ²⁶National Center for Biotechnology Information (NCBI), US National Institutes of Health (NIH), Bethesda, Maryland, USA. ²⁷Anthony Nolan Research Institute, UK. ²⁸African Institute of Biomedical Science and Technology, Harare, Zimbabwe. ²⁹Hamamatsu University School of Medicine, Hamamatsu, Japan. ³⁰Centro de Estudio de las Metabolopatías Congénitas (CEMECO), National University of Córdoba and Santísima Trinidad Children's Hospital, Córdoba, Argentina. ³¹University of Washington, Seattle, Washington, USA. ³²University of Cape Town, Cape Town, South Africa. ³³Royal Perth Hospital, Perth, Western Australia, Australia. ³⁴Oregon Health & Science University, Portland, Oregon, USA. ³⁵Cedars-Sinai Medical Center, Los Angeles, California, USA. ³⁶NIH Pharmacogenetics Research Network, Bethesda, Maryland, USA. ³⁷Montreal Children's Hospital Research Institute, Montreal, Quebec, Canada. ³⁸Keio University School of Medicine, Tokyo, Japan. ³⁹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. ⁴⁰Centre for Arab Genomic Studies, Dubai, United Arab Emirates. ⁴¹St. James' University Hospital, Leeds, UK. ⁴²American College of Medical Genetics, Bethesda, Maryland, USA.