Supporting Information available from the editorial office (humu@wiley.com) upon request.

UNCORRECTED ACCEPTED ARTICLE

Special Article **Human Mutations**

DOI 10.1002/humu.20972

Planning the Human Variome Project: The Spain Report*

Jim Kaput[1], Richard G. H. Cotton[2,3], Lauren Hardman[2], Aida I. Al Aqeel[4], Jumana Y. Al-Aama[5], Fahd Al-Mulla[6], Stefan Aretz[7], Arleen D. Auerbach[8], Myles Axton[9], Bharati Bapat[10], Inge T. Bernstein[11], Jong Bhak[12], Stacey L. Bleoo[13], Helmut Blöcker[14], Steven E. Brenner[15], John Burn[16], Mariona Bustamante[17], Rita Calzone[18], Anne Cambon-Thomsen[19], Michele Cargill[20], Paola Carrera[21], Lawrence Cavedon[22], Yoon Shin Cho[23], Yeun-Jun Chung[24], Mireille Claustres[25], Garry Cutting[26], Raymond Dalgleish[27], Johan T. den Dunnen[28], Carlos Díaz[29], Steven Dobrowolski[30], M. Rosário N. dos Santos[31], Rosemary Ekong[32], Simon B. Flanagan[33], Paul Flicek[34], Yoichi Furukawa[35], Maurizio Genuardi[36], Ho Ghang[12], Maria V. Golubenko[37], Marc S. Greenblatt[38], Ada Hamosh[39], John M. Hancock[40], Ross Hardison[41], Terence M. Harrison[42], Robert Hoffmann[43], Rania Horaitis[2], Heather J. Howard[2], Carol Isaacson Barash[44], Neskuts Izagirre[45], Jongsun Jung[23], Toshio Kojima[46], Sandrine Laradi[47], Yeon-Su Lee[48], Jong-Young Lee[23], Vera L. Gil-da-Silva-Lopes[49], Finlay A. Macrae[50], Donna Maglott[51], Makia J. Marafie[52], Steven G.E. Marsh[53], Yoichi Matsubara[54], Ludwine M. Messiaen[55], Gabriela Möslein[56], Mihai G. Netea[57], Melissa L. Norton[58], Peter J. Oefner[59], William S. Oetting[60], James C. O'Leary[61], Ana Maria Oller de Ramirez[62], Mark H. Paalman[63], Jillian Parboosingh[64], George P. Patrinos[65], Giuditta Perozzi[66], Ian R. Phillips[67], Sue Povey[33], Suyash Prasad[68], Ming Qi[69], David J. Quin[70], Rajkumar S. Ramesar[71], C. Sue Richards[72], Judith Savige[73], Dagmar G.

Scheible[74], Rodney J. Scott[75], Daniela Seminara[76], Elizabeth A. Shephard[77], Rolf H. Sijmons[78], Timothy D. Smith[2], María-Jesús Sobrido[79], Toshihiro Tanaka[80], Sean V. Tavtigian[81], Graham R. Taylor[82], Jon Teague[83], Thoralf Töpel[84], Mollie Ullman-Cullere[85], Joji Utsunomiya[46], Henk J. van Kranen[86], Mauno Vihinen[87], Michael Watson[88], Elizabeth Webb[2], Thomas K. Weber[89], Meredith Yeager[90], Young I. Yeom[91], Seon-Hee Yim[92] and Hyang-Sook Yoo[93] on behalf of contributors to the Human Variome Project Planning Meeting

[1]Division of Personalised Nutrition and Medicine, FDA/National Center for Toxicological Research, Jefferson, AR USA

[2]Genomic Disorders Research Centre, Melbourne, Australia

[3]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Australia

[4]Department of Paediatrics, Riyadh Military Hospital, Kingdom of Saudi Arabia

[5]Princess Al Jawhara Center for Hereditary Disorders, King Abdulaziz University, Saudi Arabia

[6]Molecular Pathology Unit, Kuwait University, Kuwait

[7]Institute of Human Genetics, University of Bonn, Germany

[8]Laboratory of Human Genetics and Hematology, The Rockefeller University, New York, NY, USA

[9]Nature Genetics, New York, NY, USA

[10]Lab Medicine & Pathobiology, Mount Sinai Hospital, University of Toronto, Canada

[11]Surgical gastroenterology, Hvidovre Hospital, Copenhagen, Denmark

[12]Korean Bioinformation Center, KRIBB, ChungNam, South Korea

[13]Medical Genetics, University of Alberta, Edmonton, Canada

[14]Genome Analysis, HZI - Helmholtz Centre for Infection Research, Braunschweig, Germany

[15]University of California, Berkeley, USA

[16]Institute of Human Genetics, International Centre for Life, UK

[17]Centre for Genomic Regulation and Center for Network Biomedical Research on Epidemiology and Public Health, Barcelona, Spain

[18]Genetic Service ASL Napoli, Italy [19]Epidemiology and Public Health Analyses, Inserm, Toulouse, France

[20]Human Genetics, Navigenics, Redwood Shores, CA, USA

[21]Unit of Genomics for Diagnostics of Human Disease and Laboraf, San Raffaele Scientific Institute, Milano, Italy

[22]Victorian Research Laboratory, NICTA (National ICT Australia), Parkville, Australia

[23]Division of Structural and Functional Genomics, Korean National Institute of Health, Seoul, South Korea

[24]Department of Microbiology and Genomics, Catholic University of Korea, Seoul, South Korea

[25]Université Montpellier 1, Faculté de Médecine and CHU, Laboratoire de Genetique Moleculaire, Montpellier, France

[26]Johns Hopkins University School of Medicine, Institute of Genetic Medicine, Baltimore, MD, USA

[27]Department of Genetics, University of Leicester, UK

[28]Human and Clinical Genetics, Leiden University Medical Center, The Netherlands

[29]European Projects Management and Coordination Office, Fundació IMIM, Barcelona, Spain

[30]Research & Development, Idaho Technology, Inc., Salt Lake City, Utah, USA

[31]Centro de Genética Médica J., Magalhães, INSA, Portugal

[32]Department of Genetics, Evolution and Environment, University College London, UK

[33]Molecular Genetics, Pathology Queensland, Royal Brisbane and Women's Hospital, Herston, Australia

[34]EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

[35]Division of Clinical Genome Research, Institute of Medical Science, The University of Tokyo, Japan

[36]Clinical Pathophysiology, University of Florence, Italy

[37]Institute of Medical Genetics, Tomsk, Russia

[38]College of Medicine, University of Vermont, Burlington, VT, USA

[39]OMIM and Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[40]Bioinformatics Group, MRC Harwell, UK

[41]The Pennsylvania State University, University Park, PA, USA

[42]Health Sciences Library and The Victorian Mental Health Library, Royal Melbourne Hospital, Parkville, Australia

[43]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA USA

[44]Genetics, Ethics & Policy Consulting Inc, Boston, MA, USA

[45]University of the Basque Country, Spain

[46]Advanced Computational Sciences Department, RIKEN, Kanagawa, Japan

[47]Establissement Francais Du sang, Auvergne-Loire, France

[48]Functional Genomics Branch, National Cancer Center, Goyang, South Korea

[49]Department of Medical Genetics, State University of Campinas, Brazil

[50]Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, Australia

[51]OMIM, NCBI, Baltimore, MD, USA

[52]Clinical Genetics, Kuwait Medical Genetics Centre, Kuwait

[53]HLA Informatics Group, Anthony Nolan Research Institute and Department of Haematology, UCL Cancer Institute, London UK

[54]Department of Medical Genetics, Tohoku University School of Medicine, Sendai, Japan

[55]Department of Genetics, University of Alabama, Birmingham, USA

[56]St. Josefs-Hospital Bochum-Linden, Germany

[57]Department of Medicine, Radboud University Nijmegen Medical Center, The Netherlands

[58]Genome Medicine, London, UK

[59]Institute of Functional Genomics, University of Regensburg, Germany

[60]Department of Medicine, Genetics, Institute of Human Genetics, University of Minnesota, Minneapolis, USA

[61]Genetic Alliance, Washington, DC, USA

[62]Pediatric Clinic Department, School of Medicine, National University of Córdoba, Argentina

[63]Human Mutation, Hoboken, NJ, USA

[64]Medical Genetics, University of Calgary, Alberta, Canada

[65]MGC-Department of Cell Biology and Genetics, Faculty of Medicine and Health Sciences, Erasmus University Medical Center, Rotterdam and Institute of Biomedical Genomics Research, Ag., Thessaloniki, Greece

[66]INRAN - National Research Institute on Food & Nutrition, Rome, Italy

[67]School of Biological and Chemical Sciences, Queen Mary, University of London, UK

[68]Genzyme Therapeutics Ltd, Oxford, UK

[69]ADINOVO Center for Genetic & Genomic Medicine, The First Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, P.R. China and University of Rochester Medical Center, NY, USA

[70]Funding Health Information Policy, Department of Human Services, Melbourne, Australia

[71]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, South Africa

[72]OHSU DNA Diagnostic Lab, Oregon Health & Science University, Portland, OR, USA

[73] Department of Medicine, The University of Melbourne, Epping, Australia

[74]Metabolic Department, Klinik fuer Kinder-und Jugendmedizin, Reutlingen, Germany

[75]Discipline of Medical Genetics, Faculty of Health, University of Newcastle, Australia

[76]Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, MD USA

[77]Department of Structural and Molecular Biology, University College London, UK

[78]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands

[79]Fundacion Publica Galega de Medicina Xenomica, Santiago de Compostela, Spain and Center for Network Biomedical Research on Rare Diseases, Institute of Health Carlos III, Madrid, Spain.

[80]Centre for Genome Medicine, RIKEN, Tokyo, Japan

[81]International Agency for Research on Cancer (IARC), Lyon France

[82]Regional DNA Laboratory, Cancer Research, UK Mutation Detection Facility, Leeds, UK

[83]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK

[84]Technical facility, Bioinformatics Department, Bielefeld University, Germany

[85]Harvard Medical School - Partners HealthCare Center for Genetics and Genomics, Cambridge, MA, USA

[86]Nutrition & Health, National Inst. Public Health & Environment, Utrecht, The Netherlands

[87]Institute of Medical Technology - Bioinformatics Group, University of Tampere and Tampere University Hospital, Finland

[88]American College of Medical Genetics, Bethesda, MD, USA

[89]Departments of Surgery and Molecular Genetics, Albert Einstein College of Medicine, New York, NY, USA

[90]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD USA

[91]Medical Genomics Research Center, Korea Research Institute of Bioscience & Biotechnology, Daejeon, South Korea

[92]Catholic University of Korea, Seoul, South Korea

[93]Fred Hutchinson Cancer Research Center-KRIBB, collaboration center at KRIBB, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, South Korea

***More complete affiliation information for authors is available in the Supporting Information online. Members of working groups, their affiliations, institutions, and email addresses can be found at http://www.humanvariomeproject.org.**

Received 02 December 2008; accepted 22 December 2008
Published 2009 Wiley-Liss, Inc.

*** Corresponding author**

Jim Kaput

Division of Personalized Nutrition and Medicine

FDA/National Center for Toxicological Research

3900 NCTR Road

Jefferson, AR 72079, USA

James.kaput@fda.hhs.gov

+1 870 543 7997

Abstract

The remarkable progress in characterizing the human genome sequence, exemplified by the Human Genome Project and the HapMap Consortium, has led to the perception that knowledge and the tools (e.g., microarrays) are sufficient for many if not most biomedical research efforts. A large amount of data from diverse studies proves this perception inaccurate at best, and at worst, an impediment for further efforts to characterize the variation in the human genome. Since variation in genotype and environment are the fundamental basis to understand phenotypic variability and heritability at the population level, identifying the range of human genetic variation is crucial to the development of personalized nutrition and medicine. The Human Variome Project (HVP; http://www.humanvariomeproject.org/) was proposed initially to systematically collect mutations that cause human disease and create a cyber infrastructure to link locus specific databases (LSDB). We report here the discussions and recommendations from the 2008 HVP planning meeting held in San Feliu de Guixols, Spain, in May 2008.

Key Words: variome, genome, mutation, database, genetic disease

Introduction

The completion of the consensus sequence of the human genome (Lander, et al., 2001; Venter, et al., 2001) ushered in the "post-genomic era" of science – that is, experiments could be designed using the reference sequence of the genome without need for additional sequencing efforts. Subsequent publication of the human haplotype map, an analysis of sequence diversity in 270 individuals from 4 ancestral populations (International HapMap Consortium, 2003, 2004; Frazer, et al., 2007), provided knowledge for building reagents for further genetic analyses. The knowledge and sequence information provided the resources to analyze the genetic contribution to virtually all measurable phenotypes. These efforts and the resulting databases complemented the long-standing efforts by geneticists to locate, identify, and characterize mutations that cause monogenic and polygenic diseases in humans, an effort begun by McKusick and colleagues in the 1950s (reviewed in McKusick, 2006) and now catalogued in the Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim; Hamosh, et al., 2005) and the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk; Stenson, et al., 2008). NCBI's dbGaP (Genotype and Phenotype database - http://www.ncbi.nlm.nih.gov/gap; Mailman, et al., 2007), HuGEnet (Human Genome Epidemiology Network - http://www.cdc.gov/genomics/hugenet/), EBI's EGA (European Genotype Archive – http://www.ebi.ac.uk/ega/), FINDbase (Frequency of INherited Disorders database; http://www.findbase.org; van Baal, et al., 2007) and GAD (the Genetic Association Database http://geneticassociationdb.nih.gov/; Becker, et al., 2004), are repositories for data from population studies associating genetic variation with phenotypes. Most of these databases are

study-oriented and analyze existing polymorphisms rather than focusing on the discovery of new genetic variants.

A large amount of other mutation or gene variation data, however, is likely to exist on servers in laboratories scattered throughout the world. Each of these databases may contain valuable data for other studies and for the medical practitioner. The Human Variome Project (HVP; http://www.humanvariomeproject.org/) was previously proposed to systematically collect mutations that cause human disease (Cotton, et al., 2007a; Cotton, et al., 2007b; Ring, et al., 2006) and create a cyber infrastructure to link locus specific databases (LSDBs). Local experts would curate individual LSDBs but each would have similar architecture, ontologies, and data elements allowing for interoperability. Links to national and international databases such as at the National Center for Biotechnology Information (NCBI - http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (EBI - http://www.ebi.ac.uk/) would consolidate the knowledge of the curation done by local experts. We report here the discussions and recommendations from the 2008 HVP planning meeting held in San Feliu de Guixols, Spain, in May 2008, to further the development of the HVP.

The theoretical rationale for re-sequencing genes from individuals in diverse populations is that the existing databases have focused primarily on Europeans and their descendants and are therefore a relatively narrow subdivision of the entire range of human genetic diversity. Published data supporting a concerted re-sequencing effort for monogenic and complex diseases come from independent and unlinked studies:

- The molecular basis of 2393 phenotypes was known as of September 2008 (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html). Many monogenic diseases are caused by different mutations in one gene, and all monogenic diseases are known to have variable age of onset, severity and outcome (e.g., McKusick, 2007; Ropers, 2007). Differences in monogenic disease phenotype may be caused by variations in location of mutations, by modifier genes that interact with the disease causing allele (McKusick, 2007), and by gene-environment interactions (Ordovas and Corella, 2006). Characterizing causative mutations in familial and sporadic cases in diverse populations is warranted for a full understanding of each disease.

- The molecular basis of over 3700 other phenotypes are either suspected to be Mendelian disorders or are unknown (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html).

- Simple monogenic traits may be caused by different gene variants. The ability to hydrolyze lactose as an adult, which is called lactase persistence, occurs in ~30% of the world's population (Lomer, et al., 2008). Expression of the lactase gene post-weaning has been associated with a C/T variant at position -13910 from the start of the lactase (*LCT*) gene in Finnish families and 236 individuals from 4 populations (Germany, Italy, South Korea and Finland) (Enattah, et al., 2002). However, lactase persistence is associated with a different variant (G/C at -14010 from *LCT*) in Kenyans, Tanzanians, and Sudanese (Tishkoff, et al., 2007). Other populations with higher percentages of individuals with lactose tolerance have not been analyzed (Montgomery, et al., 2007). Variation in amounts of lactose required to induce intestinal bloating and diarrhea, severity, and age of onset are observed in reference

populations (lactose intolerant) and in populations where the lactase persistence variants are more common (Lomer, et al., 2008; Montgomery, et al., 2007).

- Over 290 studies associating polymorphisms in methylene tetrahyrdofolate reductase (*MTHFR*) with various disease or physiological conditions have been published (http://www.cdc.gov/genomics/search.htm -> MTHFR). The most studied variants are c.677C>T (p.A222V) and c.1298A>C (p.E429A). Marini et al (2008) recently re-sequenced 564 individuals of diverse genetic ancestry (Coriell Institute panels -- http://ccr.coriell.org/Sections/BrowseCatalog/Populations.aspx?PgId=4) and discovered 14 nonsynonymous changes including 11 alleles with frequencies <1% along with the common alleles p.A222V, p.E429A, and p.R594Q (Marini, et al., 2008). Increased levels of folate restored MTHFR activity to the normal range in 4 of the 5 variants. The sequence heterogeneity and remediation of enzyme activity by folate supports a greater emphasis on the ~600 cofactor dependent enzymes in the human proteome. Since many cofactors are derived from diet, such studies may identify individuals who require higher concentrations of vitamins for optimal health.

Analyses of populations using HapMap data or their derivative reagents also provide justification for the need to re-sequence genes in diverse populations.

- Published HapMap data analyzed by a novel algorithm identified chromosomal regions with a high Fst (Fixation index, a measure of population differentiation) between three ancestral populations (European, Chinese and African) (Myles, et al., 2008b). These regions encoded

genes involved in carbohydrate metabolism, skeletal development, and pigmentation. Such allele frequency differences may explain, for example, the differential effect in incidence of obesity and type 2 diabetes between Europeans and Pima Indians who consume similar Western diets (Schulz, et al., 2006).

- Twenty-five SNPs linked to 6 chronic diseases in genome wide association studies (GWAS) were analyzed in ~1000 individuals from 53 populations (Myles, et al., 2008a). Several risk alleles were absent from some populations and several were present at 100% frequency indicating that the allele may contribute uniquely to disease in the European population. Other polymorphisms in these genes or in other genes within the non-European populations are likely to contribute to disease incidence and severity.

- Allele frequencies of 873 tag SNPs in 82 candidate genes involved in hypertension, type 2 diabetes, obesity, dyslipidemia or metabolic syndrome varied in 54 populations (Hancock, et al., 2008). Some of these genes were linked to adaptation to cold climates and others likely to selection by other environmental factors. Exposure to infectious agents such as malaria (Kwiatkowski, 2005), high altitude (Beall, 2007) and the food environment (Perry, et al., 2007) have also been shown to select for certain genotypes.

- Culture is also known to influence allele frequencies among populations. A gradient of genotypes exists from northwest to southeast in Europe (Price, et al., 2008). At a fine mapping level, allele frequencies mirror geography, and by inference, national culture, within Europe (Lao, et al., 2008; Novembre, et al., 2008).

Variation at the genomic and gene levels demonstrates that existing data and reagents will not be sufficient to identify genes involved in maintaining health or those that contribute to the incidence and severity of disease. The newly initiated 1000 Genomes project (http://www.1000genomes.org), which is being organized by an international consortium, will employ genome-wide re-sequencing and targeted coding region sequencing in a total of approximately 1500 individuals from three human populations: Europeans, Africans, and Asians (Lang, 2008; Siva, 2008). Each of these populations will be represented by a number of sub populations consisting of approximately 100 individuals including Yoruba in Ibadan, Nigeria; Japanese in Tokyo; Chinese in Beijing; Utah residents with ancestry from northern and western Europe; Toscani in Italy; and other populations to be determined. The goal of this international effort is to characterize alleles with frequencies of approximately 1% genome-wide and less than 1% in coding regions. The phenotype of individuals sequenced in the project will not be analyzed. Hence, these data will be used for improving selection of reagents and designs for GWAS and will not be directly focused on identification of disease causing genes. Further information is available from the project website referenced above.

The Human Variome Project

The Human Variome Project differs from these other efforts in developing and fostering an international effort to systematically identify genes, their mutations, and their variants associated with phenotypic variability and indications of human disease. The HVP is an international effort linking clinical, medical, and research laboratories for developing knowledge housed within linked databases. This knowledge would be accessible to the research and medical communities to improve research strategies and clinical medical practice. The key objectives of the project are described in Box 1. An example of the need for the HVP as applied to neurological disorders has recently been published (Cotton, et al., 2008).

The HVP Planning Meeting conducted concurrent meetings that discussed (i) classifying genetic variation from unlinked clinical medicine or research laboratories, (ii) capturing data from diagnostic and service laboratories, (iii) assessment of pathogenicity, (iv) data transfer, (v) data integration access, (vi) funding and governance, (vii) emerging countries' initiative and involvement, (viii) ethics: existing and anticipated concerns, (ix) attribution and publication, and (x) pilot projects (Table 1). Reports for classifying and capturing genetic variation from laboratories (i.e., committee reports i and ii) have been combined for sake of brevity. Detailed reports for these sessions are provided in the supplementary information. Below are synopses of the main outcomes and recommendations.

Classifying Genetic Variation from Unlinked Clinical Medicine, Research, or Service Laboratories

The cyber infrastructure for biological data is extensive (Stein, 2008) but still not fully integrated or developed. The HVP is relying upon these databases for data element definition, storage, management, retrieval, and nomenclature. For example, the NCBI provides a gene-centric index for mutation nomenclature, Human Genome Organization (HUGO; http://www.hugo-international.org/) has a naming scheme for genes, and the Human Genome Variation Society (HGVS; http://www.hgvs.org/; Cotton and Horaitis, 2000) provides for naming of mutations. The Cancer Bionformatics Grid (caBIG - https://cabig.nci.nih.gov/; Fenstermacher, et al., 2005) exemplifies the need for interoperability, common languages, data standards and sharing. In addition to the disease-specific (e.g., caBIG) and international databases, LSDBs in individual laboratories and institutes exist but are not easily linked to the rest of the bioinformatic community.

While the data management infrastructure continues to expand, curated genetic data are scattered: no coordinated effort exists to harness and harmonize these efforts, data and knowledge. Gene mutation and variation data are generated from and used by diagnostic, epidemiological experiments, research laboratories, and clinicians, each of which has different missions, ability or willingness to curate information, and resources (Table 2).

Since clinical laboratories are not required or encouraged to deposit genotype or phenotype data into publicly available databases, data sharing ranges from complete to none. Efforts to encourage and develop ongoing data collection have begun and range from commercial enterprises to funded grant programs (Table 3). Each of these programs is developing unique solutions for the barriers of time, cost, concerns of patient confidentiality, IRB requirements,

maintenance of quality assurance, and difficulties in obtaining clinical information from referring centers. These challenges reinforce the concept that clinical laboratories should not be expected to develop and curate public databases. However, clinical laboratories should be expected to contribute data. Developing a standard open software suite such as the Leiden Open Variation Database (LOVD; http://www.lovd.nl; Fokkema, et al., 2005) for these initiatives will allow existing tools, for example the Universal Mutation Database (UMD; http://www.umd.be/; Beroud, et al., 2000) software, to query across the cyberspace of LSDBs to retrieve and analyze data (Table 3). Relying upon a common database design, language and interoperability will enforce quality standards across clinical and research laboratories. Nevertheless, clinical and research laboratories may have processes and quality measures which would require a "data warning" for select entries or datasets. Some of the issues and requirements to initiate adoption of these ideas are described in Box 2. Once these individual LSDBs are developed and curated locally, ethically appropriate data elements can be deposited in national or international databases (NCBI or EBI).

Pathogenicity and Clinical Utility

Understanding the consequences of genetic variation depends upon the simultaneous collection and documentation of phenotypic data for each variant (e.g., Cotton et al., 2007a; Crawford and Nickerson, 2005; Kaput, 2008; Kaput et al., 2005; Ring et al., 2006; Taylor et al., 2001). The correlation between genome and phenotype (pathogenicity) is the basis for the clinical benefit. The two broad principles for assessing pathogenicity or phenotype linked to a genetic variant are

that (i) multiple data elements must be integrated and (ii) data elements and the integration process must have standards, validation, quantification and transparency (Box 3).

The omics sciences are now capable of generating large but disparate (e.g., genomic vs metabolomic) datasets that may be used in research but also clinical applications. Although assessing pathogenicity will be an ongoing, iterative process, several specific recommendations are warranted:

Genetic and Genomic Data. Gene marker analysis is an important step in the clinical diagnosis of pediatric and adult genetic disorders. The issues associated with clinical genetic testing are well recognized for inherited cancer syndromes, where missense variants represent 10-30% of test results (e.g., (Eisinger, 2008; Metcalfe, et al., 2008; Stoffel, et al., 2008)). Many of these variants are classified as having an uncertain effect unless strong genetic epidemiologic and/or functional evidence exists associating them with the syndrome. The Breast Cancer Information Consortium (BIC - http://research.nhgri.nih.gov/bic/) classifies *BRCA1* and *BRCA2* variants as pathogenic only if the probability of pathogenicity, usually based on statistical genetic approaches is, definitely pathogenic >0.99; likely pathogenic 0.95-0.99; uncertain 0.05-0.949; likely not pathogenic/little clinical significance (LCS) 0.001-0.049; neutral or LCS <0.001 (Plon et al. 2008). For other genes, there are no standards. For each type of data, old and new, the principles of standards, validation, quantification and transparency apply. Many of the expected gene variants have not been identified or characterized in world populations. Hence, complete re-sequencing of a gene proven or suspected to be involved in monogenic and polygenic diseases will be required to determine causal linkages between genes and phenotype.

Standardizing Existing Clinical Phenotype and Pathology. A fundamental problem with assessing phenotypes is the diversity of the underlying molecular pathways that cause disease, and as a consequence the heterogeneity in clinical manifestations, age of onset, severity, complications, and age of death. Other groups (Kaput, et al., 2005; Kathiresan, et al., 2008; Makinen, et al., 2008; Rosenzweig, et al., 2002; Wong, 2006; Zaninotto, et al., 2007) have proposed using disease as a classifier (e.g., type 2 diabetes), but rely on quantitative measures of phenotype (e.g., fasting glucose, fasting insulin) as a means to reduce subjective assignments of disease (Tracy, 2008). Since the HVP seeks to collect data from laboratories and clinics, phenotype templates are needed to define ranges of (i) minimum sets of clinical data, (ii) range of subset data, and (iii) maximum datasets. Such hierarchical template structures will allow scientists in all countries to participate in data and sample collection. Developing the LOVD/UMD tools (Table 3 and below) for the HVP will also require a means to validate data and data quality prior to implementation across laboratories. Clinical and pathology data standards must be developed by experts in each genetic disorder for interpreting the effects of genetic variation.

Standards linking *in vitro* functional studies with clinical results. If a cellular function can be established that appears to correlate with the clinical syndrome, then *in vitro* assays could be used to classify whether a variant retains or loses function. Standards for performing and interpreting the assays are crucial if these methods are to be accepted as a mechanism for classifying variants clinically (Couch et al., 2008). In cancer genetics, the correlation of mismatch repair defects with Lynch syndrome (Hereditary Non-Polyposis Colorectal Cancer) is

probably the most well-established example (Ou et al., 2007); even so, the principle that multiple data elements must be integrated to achieve classification should be respected.

Computational studies. Multiple studies in recent years have confirmed the value of comparative sequence analysis in helping to predict whether a missense variant is pathogenic or not (reviewed in Tavtigian et al., 2008). However, the issues of standards, validation, and transparency also apply to computational methods. Most importantly, the quality of a multiple sequence alignment is critical to their accuracy (Ahola et al., 2006; Ahola et al., 2008). The choice of ortholog sequences that is used, the quality of cDNA or genomic data, and the methods used to construct the alignment are all important features.

Computational Studies - Predictive Algorithms. Some methods have already been validated on curated data sets of variants, establishing their Negative Predictive Value (NPV, the proportion of predictions of "neutral" that are actually neutral) and Positive Predictive Value (PPV, the proportion of predictions of "pathogenic" that are actually pathogenic) (Chan et al., 2007; Chao et al., 2008). However, several of the more commonly used algorithms have been updated. Algorithms exist for coding region variants and predictions of altered splice sites (Nalla and Rogan, 2005; Spurdle et al., 2008). New methods, including both rule-based and machine-learning approaches, are being developed (Tavtigian et al., 2008), and in the future, algorithms to assess other non-coding functions of DNA are anticipated.

Data Transfer and Databasing - Gene and Locus Specific Databases

Historically, gene variation data were first collected for specific gene(s) causing a Mendelian disorder or a change in the phenotype. These listings usually were initiated and driven by the interests of an expert using the collection for research, clinical or diagnostic applications. Currently there are over 700 such LSDBs (http://www.hgvs.org/dblist/glsdb.html; Horaitis et al., 2007), mostly web accessible. Complete collection and expert curation of gene sequence variants and their coupling to phenotypic consequences (if any), will be essential for proper future healthcare and research. Data Transfer and Databasing plans are outlined in Box 4.

Integrating Data and Providing Access

The breadth and depth of information available about human variation are rapidly expanding as new technologies (e.g., omics and imaging) analyze health and pathogenicities. Uncomplicated methods of access are needed for multiple user communities with differing expertise in genetics, clinical medicine, nutrition, physiology, and probably the public. Information generated by the HVP will have many of those dimensions, ranging from how variants are identified, the type of variant, the physiological parameters associated with the variant, and where and how records are maintained and accessed. For example, information for rare variants, SNPs, microsatellites, small insertions/deletions will range from reports of one variant in one individual, such as might be gathered from manual data entry and maintained in medical records, to large-scale screening of thousands of genomes in defined populations. How the data are represented will be a challenge since information will come from published literature or text-rich resources such as OMIM or GeneReviews (http://www.geneclinics.org/profiles/all.html; Pagon et al., 2002), include explicit records in LSDBs and/or genome-wide resources such as dbSNP or genome browsers. The scope

of the data, which includes locations of variants on reference sequences to phenotypes in humans and model organisms, increases the dimensionality. Hence, the scale of information for human genetic variation and linked phenotype can range from a single text document to petabytes of raw data derived from sequencing thousands of genomes to a high level of coverage and accuracy. Model databases capable of accessing known variants will be developed from pilot projects and provide a resource for clinicians, patient advocates, and the public (i.e., education).

Accessing known variants. Given the dimensionality issues and the challenges they represent, the HVP will help develop an infrastructure to identify variants relative to a reference standard and allow facile linking of data with appropriate tools. Notification schemes would be developed to indicate missing data for each variant. The HVP is identifying a progression of doable tasks with milestones for each. For example, the HVP will link with one such effort, the on-going collaboration among GEN2PHEN (http://www.gen2phen.org), EBI, and NCBI to develop an international set of standard Locus Specific Genomic Sequence reference (LRGS/RefSeqGene - http://www.gen2phen.org/docs/LRG_Specification_Summary_version_9.pdf). With that foundation in place, active LSDBs can report their variants in LRG/RefSeqGene coordinates to centralized databases (EBI/NCBI) to be accessioned to dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/; Sherry et al., 2001). The scope of information to be centralized requires further debate, but could include items described in Box 5. The HVP will ensure that other standards in addition to the reference sequence are developed and/or used, such as human gene nomenclature (HUGO Gene Nomenclature Committee (HGNC); http://www.genenames.org; Bruford et al., 2008), variant nomenclature (Human Genome Variation Society; http://www.hgvs.org/mutnomen; den Dunnen and Antonarakis, 2001), variant

accessions (dbSNP) and names for diseases or diagnostic tests in medical records (Logical Observation Identifier Names and Codes, LOINC; e.g., McDonald et al., 2003). Reference gene-specific standards to support LSDBs and genetic testing groups (see http://www.gen2phen.org and http://www.ncbi.nlm.nih.gov/RefSeq/RSG) are under development. In addition, assigning accessions in dbSNP to known variants, either directly to dbSNP or Ensembl (http://www.ensembl.org), or with PhenCode (http://phencode.bx.psu.edu; (Giardine et al., 2007)) serving as a data collection center are being implemented.

Public education. With the understanding that large data sets are only as useful as their ease of access, the HVP can also foster portal sites to help direct users to tools and resources of interest, and identify areas requiring additional development. Tutorials comparing and contrasting different resources might also be commissioned. Topics that must be addressed are how to access information in the published literature, the effect of variation on transcription, the significance of conserved non-coding regions on phenotypic variation, identification of mRNA splice sites, descriptions on the structure and maintenance of LSDBs, and strategies for interpreting the impact of single (e.g., SNPs) and multiple (e.g., haplotypes) variants in one allele or in multiple alleles on the phenotype of interest. The HVP intends to make greater efforts to educate the public about the project's importance and the benefits of research participation.

Attribution and Publication

A major barrier to the development of comprehensive analyses of human genetic variation, and therefore LSDBs, is the reward system for clinicians and academic researchers. Non-academic

clinicians typically receive little or no credit for contributing to the scientific literature, and academic clinicians and basic researchers cannot easily persuade journals to publish the 50th variant of a gene that has an observable effect on phenotype. These cultures may be difficult to change overnight, but specific steps could be instituted immediately to promote the submission of genotype – phenotype data to LSDBs and to "reward" contributions to team projects. Database entries could be a mandatory quality control standard for clinical laboratories and clinicians. For researchers, a publication or web-based system establishing micro-attribution and community annotation of mutations (e.g., http://www.wikigenes.org) and cited data will enable measurable contributions to the scientific knowledge base (Editorial, 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* (HGP; http://www.sage-hindawi.com/journals/hgp), which is affiliated with FINDbase, 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). 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 HVP recommends that researchers cite these attributions and citations in their *curricula vitae* to foster the transition of the academic culture. The same trend and recommendations for development of coherent tools are valid for the recognition of contribution to the setting, use and sharing of any bioresource such as biobanks (Cambon-Thomsen, 2003; Kauffmann and Cambon-Thomsen, 2008) and international efforts like P3G are being developed in the same spirit. (Knoppers, et al. 2008)

Developing / Emerging Countries – Ensuring a Worldwide Collection

Although ~90% of known SNPs are shared between Asians, Europeans and Africans, 80% of private SNPs are found within Asian and Africans (Hinds, et al., 2005; Jorde and Wooding, 2004). The recent sequencing of the Watson (Wheeler, et al., 2008), Venter (Levy, et al., 2007), Kriek (http://www.sciencedaily.com/releases/2008/05/080526155300.htm), West African (Bentley et al., 2008) and Han Chinese (Wang et al., 2008) genomes, along with gene specific re-sequencing efforts (see Introduction), suggest 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 80 to 90% of human genomic variation resides in the world's emerging countries. Any formal attempt to identify the extent of genomic variation must include geographical regions which have not been included in haplotype mapping projects. Although the Population Reference Sample (POPRES) will address some of these missing populations (Nelson et al., 2008), this effort is designed as a mapping project, not one focused on functional polymorphisms or mutations. Hence, the main focus of the HVP effort is the inclusion and analyses of clinical samples from diverse ethnic groups.

The distinct advantage of some ethnic populations is the opportunity to study genetic diseases due to consanguinity, large family size, and potential founder effects (e.g., Bittles 2001; Bittles 2002; Saadallah and Rashed, 2007). Emerging nations will be regarded as major contributors to the VARIOME project. However, biomedical research has not been the focus of resource poor

countries even though such activities are likely to produce economic and health benefits for all (Daar et al., 2002; Singer and Daar, 2001). Education of healthcare providers, the public, and government officials is needed for demonstrating the universal nature of the HVP, the need to include populations in developing countries, and the benefits from cooperating in biomedical research (Bhan et al., 2007; Cohen et al., 2008; Seguin et al., 2008; Tindana et al., 2007). Certain populations may mistrust research involving genetic analyses or fear that results can be misused to support discrimination or worse (http://www.eubios.info/ASIAE/BIAE201.htm): 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 participatory research collaborations may provide forums for addressing cultural and ethical concerns of biomedical research (McCabe-Sellers et al., 2008).

Analyzing the extent of human genomic 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 understand how general 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, we will provide evidence-based information that all populations can benefit from a global society health network. The primary objectives for including populations in emerging countries are described in Box 6.

The HVP appreciates the genomic sovereignty/equality for all countries to be involved in the Human Variome Project 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; the voluntary participation of the greatest number of countries would ensure a more general applicability and it is hoped that many countries will decide to participate.

Progress in Developing Ethical Guidelines for LSDBs: Principles to Practice to Implementation

Ethical issues remain of vital concern to the Human Variome Project. Participating researchers are committed to adhering to the highest ethical principles governing research, data sharing and ultimately enabling this new knowledge to benefit all of humanity as much as possible. Ethical guidelines specifically for LSDBs were previously published (Cotton et al., 2005).

LSDBs may contain a large amount of phenotypic data. Most LSDBs post a considerable amount of data on public websites and increasingly this information may be accessible through genome browsers. While the best intention of the HVP is to ensure that participants are acknowledged as a group, without any risk of identification, a specific challenge occurs in the case of rare mutations associated with distinctive clinical features. Since epistatic and environment interactions (reviewed in Kaput, 2008) alter age of onset, severity, complications and outcomes for monogenic and polygenic phenotypes, it may be necessary to analyze entire genomes for personalized healthcare. Such polygenic analyses generate data that could be used for re-identifying individual patients (Craig et al., 2008).

Other ethical concerns may be minimized by improving communication about the project and its goals through multiple channels such as print and broadcast media, local community outreach,

and internet sources such as the HGVS website. The HVP will develop an ethics review committee with a subcommittee focused on issues related to LSDB for (i) providing counsel when dilemmas arise, (ii) overseeing guidelines, (iii) identifying best practices, (iv) determining how best to ensure privacy in all situations, (v) formulating how to handle data for which explicit consent does not exist or is not possible to achieve, and (vi) developing a consent form that is consistent for all LSDBs but which can be adapted to the requirements of individual countries. Such consent would contain, for example, a re-contact clause. The specific recommendation and open questions are outlined in Box 7.

The HVP, through its Ethics Working Group, is committed to (i) soliciting, collecting and analyzing consent forms in order to develop a model consent form that can provide greater consistency across all LSDBs, (ii) seeking the input of relevant clinical genetic societies for comment, and (iii) using that input to develop ethical standards for LSDBs.

Funding Mechanisms and Governance

Funding for collecting data of mutations causing single gene disorders has traditionally been difficult due to the extreme fragmentation of the field even though mutations affect 60% of all individuals in a lifetime (Baird, et al., 1988). The funding possibilities are more likely if the international HVP is treated as a concerted effort. Given the limitations of existing knowledge (see Introduction and Patrinos and Brookes, 2005), this initiative will benefit research in many fields and impact prevention and clinical care of disease. Specific focus areas for developing funding streams are described in Box 8.

Governance. Those dedicated to assisting themselves and others in their clinics by collecting mutation/variants causing inherited disease have in the past acted in isolation; reinventing wheels wastes funds and time. The HGVS, formerly HUGO-MDI, was formed to alleviate this problem and accelerate the collection and management of information on mutation causing disease(s). The HVP was named and initiated to define the aims of this activity (Melbourne in 2006) by an extremely high profile group of experts in all types of genetic variation analyses. The Genomic Disorders Research Center (Melbourne, Australia; http://www.genomic.unimelb.edu.au) was voted as coordinating office, to continue its function of facilitating independent projects started at the inception of the centre in 1996. In association with world experts, Deloitte (http://www.deloitte.com) developed a business plan, which was approved by the HVP Planning Group. The HVP Planning Meeting was a designated activity of this plan and other sections of the plan will follow (e.g. board function to oversee the coordinating office's function and support it). The business plan calls for a broadly defined community of interested stakeholders to develop the HVP (Box 9).

Open questions that must be resolved in future meetings are (i) the extent of data sharing between patient records and research databases, (ii) appropriate descriptions of data elements, and (iii) data ownership and confidentiality.

Pilot Projects

The HVP has established a partnership with the International Society for Gastrointestinal Hereditary Tumors (InSiGHT; http://www.insight-group.org) to collect and classify a large set of missense variants associated with hereditary colorectal cancer. This effort will develop HVP's prototype system for interpreting variants observed in clinical genetic testing (Box 10). InSiGHT consists of a multidisciplinary scientists focused particularly on the Mendelian disorders predisposing to colorectal cancer (Familial Adenomatous Polyposis, Lynch Syndrome and MUTYH related polyposes). This effort is an ongoing project but also a model or pilot for the HVP. InSiGHT has conducted several multidisciplinary studies of Hereditary Non Polyposis Colon Cancer (HNPCC) patients that (i) require the development of a disease-specific model for integrating databases across laboratories, (ii) establish standards for data consistency for phenotypes (which include graphic pedigrees), (iii) address confidentiality, and (iv) develop a template for consent. Clinicians from multiple countries and regions are contributing and committing to the development of these systems. Some of these issues cross disciplinary boundaries and are being addressed by other committees of the HVP. The InSiGHT consortium's roadmap includes providing access to clinicians, a vital resource that will serve patients in the immediate future, and as a model for other genes and phenotypes. Among the first efforts involved uploading large datasets of mismatch repair variants generated by national consortia and laboratories into the InSiGHT MMR, a LSDB which uses the LOVD platform. Data transfer into a mirrored central database (e.g., NCBI or EBI) is also planned with an initial reciprocal agreement with the Health Data Integration project at the Australian CSIRO Centre (http://www.ict.csiro.au/HAIL/Abstracts/2004/UmaSrinivasan.htm). Other pilot projects are described in Box 11.

Discussion

The vision of the HVP, to catalogue and access all information related to human disease variation, is ambitious. One can conceptualize the challenge as a multi-dimensional, fluid matrix, with all ~20,000 genes as column headings and rows of potentially thousands of variants as descriptors. In addition, third and further dimensions would annotate other biological parameters, for example clinical and/or metabolic phenotype, microarray expression, proteomics, protein interactions, nutrient intakes, physical activity, and other functional phenotypic and epidemiological data. Separate dimensions that must be linked to these variants are the main effect of gene – environment interactions (e.g., Lim et al., 2007). These data elements relate to the cells in the initial two dimensional matrix since each may affect the genetic expression of the mutation or gene variant. These dimensions are domains of knowledge that must be integrated for understanding biological processes.

A predetermined bioinformatics structure to accommodate this matrix with forced fields for data entry is notionally appealing but practically impossible. The reality is that LSDBs, which capture the core information in any one of these domains of information, are developed by experts and curated with invaluable skill and experience. To force any change on these individual efforts would risk inestimable loss of activity by the curators and threaten data of individuals in populations. The challenge then is to integrate the existing and developing information within existing databases and public resources into a system based on this matrix of domains – the vision of the HVP.

The task of developing "super searching" software to interrogate the global information and relate it across the HVP matrix in a user-friendly fashion for enquiry represents a bioinformatics challenge already embraced by the numerous national projects (e.g., Stein, 2008). This challenge can be met with resources applied to software development or existing applications that allow searches to locate all information across all domains of the international data matrix. Hansen's SRS (sequence retrieval system) approach (http://e-hrc.net/hdi/), or the novel Genome Commons Navigator (Brenner, 2007), supported the Berkeley Computational Biology Center (http://ccb.berkeley.edu/ccb/index.html), both target this concept. The Navigator also seeks to provide algorithms for potential interpretation of pathogenicity.

The approaches presented at the HVP planning meeting contribute to this goal of integrating biological information. The HVP efforts are also consistent with the newly emerging initiative to develop standards for scientific disciplines and research strategies: the MIBBI Project (minimal reporting guidelines for biological and biomedical investigations; Taylor, 2007; Taylor et al., 2008). The challenge to catalogue and access this vast body of information relating to human biology and behavior is immense, but the HVP is leading this endeavor through international collaborations and harmonized protocols. The development of this network of LSDBs and the knowledge they generate and maintain will be beneficial not only for the genetic research community, but also for researchers in nutrition, toxicology, teratology, physiology – virtually all biological research arenas, but perhaps most importantly for the translation of basic research for improving personal and public health. The future is indeed exciting.

Disclaimer

This work includes contributions from, and was reviewed by, the FDA. This work has been approved for publication by this agency but it does not necessarily reflect official agency policy.

Acknowledgment

The authors thank Maria Mendoza and Donna Mendrick of FDA/NCTR for critically reviewing the manuscript. Mike Parker and Helen Firth are acknowledged for their contribution to the work leading up to Box 7. JMH thanks CASIMIR (funded by the European Commission under contract number LSHG-CT-2006-037811) for financial support.

References

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2006. A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 7:484.

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2008. Model-based prediction of sequence alignment quality. Bioinformatics 24(19):2165-71.

Axton M. 2008. Human Variome Microattribution Reviews. Nat Genet 40(1):1.

Baird PA, Anderson TW, Newcombe HB, Lowry RB. 1988. Genetic disorders in children and young adults: a population study. Am J Hum Genet 42(5):677-93.

Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proc Natl Acad Sci U S A 104 Suppl 1:8655-60.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet 36(5):431-2.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR and others. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53-9.

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 15(1):86-94.

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(9):e272.

Bittles A. 2001. Consanguinity and its relevance to clinical genetics. Clin Genet 60(2):89-98.

Bittles AH. 2002. Endogamy, consanguinity and community genetics. J Genet 81(3):91-8.

Brenner SE. 2007. Common sense for our genomes. Nature 449(7164):783-4.

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. 2008. The HGNC Database in 2008: a resource for the human genome. Nucleic Acids Res 36(Database issue):D445-8.

Cambon-Thomsen A. 2003. Assessing the impact of biobanks. Nat Genet. 34(1):25-6.

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ and others. 2007. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat 28(7):683-93.

Chao EC, Velasquez JL, Witherspoon MS, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennert G, Anton-Culver H and others. 2008. Accurate classification of MLH1/MSH2

missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). Hum Mutat 29(6):852-60.

Cohen ER, Masum H, Berndtson K, Saunders V, Hadfield T, Panjwani D, Persad DL, Minhas GS, Daar AS, Singh JA and others. 2008. Public engagement on global health challenges. BMC Public Health 8:168.

Consortium IH. 2004. Integrating ethics and science in the International HapMap Project. Nat Rev Genet 5(6):467-75.

Consortium IHGS. 2003. The International HapMap Project. Nature 426(6968):789-96.

Cotton RG, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A and others. 2007a. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 39(4):433-6.

Cotton RG, Auerbach AD, Brown AF, Carrera P, Christodoulou J, Claustres M, Compton J, Cox DW, De Baere E, den Dunnen JT and others. 2007b. A structured simple form for ordering genetic tests is needed to ensure coupling of clinical detail (phenotype) with DNA variants (genotype) to ensure utility in publication and databases. Hum Mutat 28(10):931-2.

Cotton RG, Horaitis O. 2000. Human Genome Variation Society. In: Cooper DN, editor. Nature Encyclopedia of the Human Genome. London: Nature Publishing Group. p 361-362.

Cotton RG, Sallee C, Knoppers BM. 2005. Locus-specific databases: from ethical principles to practice. Hum Mutat 26(5):489-93.

Cotton RGH, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting G, den Dunnen JT, Flicek P and others. 2008. GENETICS: The Human Variome Project. Science 322(5903):861-862.

Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N. 2008. Assessment of functional effects of unclassified genetic variants. Hum Mutat 29(11):1314-26.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA and others. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods.

Crawford DC, Nickerson DA. 2005. Definition and clinical importance of haplotypes. Annu Rev Med 56:303-20.

Daar AS, Thorsteinsdottir H, Martin DK, Smith AC, Nast S, Singer PA. 2002. Top ten biotechnologies for improving health in developing countries. Nat Genet 32(2):229-32.

Dantzer J, Moad C, Heiland R, Mooney S. 2005. MutDB services: interactive structural analysis of mutation data. Nucleic Acids Res 33(Web Server issue):W311-4.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. Hum Genet 109(1):121-4.

Eisinger F. 2008. Genetic testing for familial cancer. The French National Report (year 2003). Community Genet 11(1):63-7.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30(2):233-7.

Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M. 2005. The Cancer Biomedical Informatics Grid (caBIG^{TM}). Conf Proc IEEE Eng Med Biol Soc 1:743-6.

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(2):63-8.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H and others. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28(6):554-62.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database issue):D514-7.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet 4(2):e32.

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(5712):1072-9.

Hoffman, R. 2008. A wiki for the life sciences where authorship matters. Nat Genet 40(9):1047 – 1051.

Horaitis O, Talbot CC, Jr., Phommarinh M, Phillips KM, Cotton RG. 2007. A database of locus-specific databases. Nat Genet 39(4):425.

Jorde LB, Wooding SP. 2004. Genetic variation, classification and 'race'. Nat Genet 36 Suppl 1:S28-33.

Kaput J. 2008. Nutrigenomics research for personalized nutrition and medicine. Curr Opin Biotechnol 19(2):110-20.

Kaput J, Ordovas JM, Ferguson L, van Ommen B, Rodriguez RL, Allen L, Ames BN, Dawson K, German B, Krauss R and others. 2005. The case for strategic international alliances to harness nutritional genomics for public and personal health. Br J Nutr 94(5):623-32.

Kathiresan S, Musunuru K, Orho-Melander M. 2008. Defining the spectrum of alleles that contribute to blood lipid concentrations in humans. Curr Opin Lipidol 19(2):122-7.

Kauffmann F, Cambon-Thomsen A. 2008. Tracing biological collections: between books and clinical trials. JAMA. 299(19):2316-8.

Knoppers BM, Fortier I, Legault D, Burton P. 2008. The Public Population Project in Genomics (P3G): a proof of concept? Eur J Hum Genet. 16(6):664-5.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77(2):171-92.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W and others. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921. **[[COPYEDITOR: ALLOW TRUNCATED AUTHOR LIST FOR THIS REFERENCE AS PER JOURNAL STYLE]]**

Lang L. 2008. Three sequencing companies join the 1000 genomes project. Gastroenterology 135(2):336-7.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D and others. 2008. Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241-8.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G and others. 2007. The diploid genome sequence of an individual human. PLoS Biol 5(10):e254.

Lim U, Wang SS, Hartge P, Cozen W, Kelemen LE, Chanock S, Davis S, Blair A, Schenk M, Rothman N and others. 2007. Gene-nutrient interactions among determinants of folate and one-carbon metabolism on the risk of non-Hodgkin lymphoma: NCI-SEER case-control study. Blood 109(7):3050-9.

Lomer MC, Parkes GC, Sanderson JD. 2008. Review article: lactose intolerance in clinical practice--myths and realities. Aliment Pharmacol Ther 27(2):93-103.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L and others. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39(10):1181-6.

Makinen VP, Soininen P, Forsblom C, Parkkonen M, Ingman P, Kaski K, Groop PH, Ala-Korpela M. 2008. 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. Mol Syst Biol 4:167.

Marini NJ, Gin J, Ziegle J, Keho KH, Ginzinger D, Gilbert DA, Rine J. 2008. The prevalence of folate-remedial MTHFR enzyme variants in humans. Proc Natl Acad Sci U S A 105(23):8055-60.

McCabe-Sellers B, Lovera D, Nuss H, Wise C, Green B, Teitel C, Ning B, Clark B, Bogle M, Kaput J. 2008. Community Based Participatory Research and Omics Technologies. Omics, A Journal of Integrative Biology:in press.

McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J and others. 2003. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. Clin Chem 49(4):624-633.

McKusick VA. 2006. A 60-year tale of spots, maps, and genes. Annu Rev Genomics Hum Genet 7:1-27.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80(4):588-604.

Metcalfe KA, Birenbaum-Carmeli D, Lubinski J, Gronwald J, Lynch H, Moller P, Ghadirian P, Foulkes WD, Klijn J, Friedman E and others. 2008. International variation in rates of uptake of preventive options in BRCA1 and BRCA2 mutation carriers. Int J Cancer 122(9):2017-22.

Montgomery RK, Krasinski SD, Hirschhorn JN, Grand RJ. 2007. Lactose and lactase--who is lactose intolerant and why? J Pediatr Gastroenterol Nutr 45 Suppl 2:S131-7.

Myles S, Davison D, Barrett J, Stoneking M, Timpson N. 2008a. Worldwide population differentiation at disease-associated SNPs. BMC Med Genomics 1(1):22.

Myles S, Tang K, Somel M, Green RE, Kelso J, Stoneking M. 2008b. Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet 72(Pt 1):99-110.

Nalla VK, Rogan PK. 2005. Automated splicing mutation analysis by information theory. Hum Mutat 25(4):334-42.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G and others. 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83(3):347-58.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR and others. 2008. Genes mirror geography within Europe. Nature.

Ordovas JM, Corella D. 2006. Gene-Environment Interactions: Defining the Playfield. In: Kaput J, Rodriguez RL, editors. Nutritional Genomics. Discovering the Path to Personalized Nutrition. Hoboken, NJ: John Wiley and Sons. p 57 - 76.

Ou J, Niessen RC, Lutzen A, Sijmons RH, Kleibeuker JH, de Wind N, Rasmussen LJ, Hofstra RM. 2007. Functional analysis helps to clarify the clinical importance of unclassified variants in DNA mismatch repair genes. Hum Mutat 28(11):1047-54.

Pagon RA, Tarczy-Hornoch P, Baskin PK, Edwards JE, Covington ML, Espeseth M, Beahler C, Bird TD, Popovich B, Nesbitt C and others. 2002. GeneTests-GeneClinics: genetic testing information for a growing audience. Hum Mutat 19(5):501-9.

Patrinos G, Petricoin E. 2009. A new scientific journal linked to a genetic database: Towards a novel publication modality. Hum Genomics Proteomics in press.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: disastrously deficient. Trends Genet 21(6):333-8.

Peltomaki P, Vasen HF. 1997. Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer. Gastroenterology 113(4):1146-58.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R and others. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-60.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29(11):1282-1291.

Price A, Butler J, Patterson N, Capelli C, Pascali V, Scarnicci F, Ruiz-Linares A, Groop L, Saetta A, Korkolopoulou P and others. 2008. Discerning the ancestry of European Americans in genetic association studies. PLoS Genetics 4(1):e236.

Ring HZ, Kwok PY, Cotton RG. 2006. Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7(7):969-72.

Ropers HH. 2007. New perspectives for the elucidation of genetic disorders. Am J Hum Genet 81(2):199-207.

Rosenzweig JL, Weinger K, Poirier-Solomon L, Rushton M. 2002. Use of a disease severity index for evaluation of healthcare costs and management of comorbidities of patients with diabetes mellitus. Am J Manag Care 8(11):950-8.

Saadallah AA, Rashed MS. 2007. Newborn screening: experiences in the Middle East and North Africa. J Inherit Metab Dis 30(4):482-9.

Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, Valencia ME. 2006. Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the U.S. Diabetes Care 29(8):1866-71.

Seguin B, Hardy BJ, Singer PA, Daar AS. 2008. Genomic medicine and developing countries: creating a room of their own. Nat Rev Genet 9(6):487-93.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-11.

Singer PA, Daar AS. 2001. Harnessing genomics and biotechnology to improve global health equity. Science 294(5540):87-9.

Siva N. 2008. 1000 Genomes project. Nat Biotechnol 26(3):256.

Spurdle AB, Couch FJ, Hogervorst FB, Radice P, Sinilnikova OM. 2008. Prediction and assessment of splicing alterations: implications for clinical testing. Hum Mutat 29(11):1304-13.

Stein LD. 2008. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet 9(9):678-88.

Stenson, P. D., E. Ball, et al. (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet **45**(2): 124-6.

Stoffel EM, Ford B, Mercado RC, Punglia D, Kohlmann W, Conrad P, Blanco A, Shannon KM, Powell M, Gruber SB and others. 2008. Sharing genetic test results in Lynch syndrome: communication with close and distant relatives. Clin Gastroenterol Hepatol 6(3):333-8.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29(11):1327-36.

Taylor CF. 2007. Standards for reporting bioscience data: a forward look. Drug Discov Today 12(13-14):527-33.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T and others. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889-96.

Taylor JG, Choi EH, Foster CB, Chanock SJ. 2001. Using genetic variation to study human disease. Trends Mol Med 7(11):507-12.

Tindana PO, Singh JA, Tracy CS, Upshur RE, Daar AS, Singer PA, Frohlich J, Lavery JV. 2007. Grand challenges in global health: community engagement in research in developing countries. PLoS Med 4(9):e273.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M and others. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39(1):31-40.

Topel T, Hofestadt R, Scheible D, Trefz F. 2006. RAMEDIS: the rare metabolic diseases database. Appl Bioinformatics 5(2):115-8.

Tracy RP. 2008. 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. Curr Opin Lipidol 19(2):151-7.

van Baal S, Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, Macek M, Jr., Scriver CR, Patrinos GP. 2007. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 35(Database issue):D690-5.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA and others. 2001. The sequence of the human genome. Science 291(5507):1304-51. **[[COPYEDITOR: ALLOW TRUNCATED AUTHOR LIST FOR THIS REFERENCE AS PER JOURNAL STYLE]]**

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y and others. 2008. The diploid genome sequence of an Asian individual. Nature 456(7218):60-5.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT and others. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872-6.

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(1):44-54.

Wong ND. 2006. Screening and risk stratification of patients with the metabolic syndrome and diabetes. Expert Rev Cardiovasc Ther 4(2):181-90.

Zaninotto M, Mion MM, Novello E, Altinier S, Plebani M. 2007. New biochemical markers: from bench to bedside. Clin Chim Acta 381(1):14-20.

Table 1

Organization of Committee Reports [a]

Committees Meeting in Spain		Box [b]
Capturing and classifying genetic variation		
	Unlinked clinical and research laboratories	2
	diagnostic and service laboratories	2
Assessment of pathogenicity		
Data transfer		4, 5
Data Integration access		4, 5
Emerging countries		6
Ethics		7
Funding and governance		8, 9
Attribution and Publication		
Pilot Projects, including InSiGHT		10,

[a] See text for details
[b] Outlines of each committee report are available in the online Supporting Information

Box 1 describes Key Objectives of the HVP
Box 3 outlines Standards, Validation, Quantification, Transparency for data

Table 2

Laboratories Involved in Mutation/Variation Discovery and Use

Laboratory	Mission	Quality Assurance [a]	Curation [b]	Database	Other
Diagnostic or Reference	Service, no research	In place	None	Local, if any	Cost pressure
Epidemiology or Public	Associate existing variations to	Provider dependent	Population	Local & global	No new variant data generation
Research	Generate new knowledge	None	Individual	Local	Continuity risk

a. Quality Assurance is required in diagnostic or reference laboratory. QA in epidemiology studies would depend on the service provider

b. Curation would be at the population level in epidemiological studies and the individual level in research laboratories. Not all research laboratories would curate variants or mutants.

Table 3

Examples of Clinical – Research Translational Initiatives with Database Suites

Program	Name	Agency	Mission	Method	
ARUP	Clinical Laboratory	University of Utah	Galactosemia	Locus specific	h n
CETT	Collaboration, Education, Test Translation	NIH Office of Rare Diseases	Collaboration between clinical laboratory, research, and	Clinical laboratories collect phenotypic data from clinicians through online forms prior to genetic testing. Genotypic and phenotypic data to be placed in a database	h n
DMuDB	Diagnostic Mutation Database	Certus Technology Ltd, UK National Genetics Reference Laboratory	Mutation database for diagnostic genetics services	~5,000 entries including 2,900 BRCA1 and BRCA2 entries from 12 laboratories (851 unique variants, 436 of which are not in BIC), HNPCC (827 entries from 9 labs), and the following unique variants: 177 MLH1, 178 MSH2, 43 MSH6, 116 APC, 70 Lowe Syndrome, 38 Sotos Syndrome, 150 Cystic Fibrosis, 96 Neurofibromatosis, and 101 X-Linked Retinitis Pigmentosa	h e
eyeGENE	National Opthalmic Disease Genotyping Network	NIH National Eye Institute	The discovery of the genetic causes of ocular diseases: accurate diagnostic genotyping to patients with inherited eye diseases; patient information; develop populations for study; develop resources; improve	Patient meeting minimal clinical criteria sign research and a clinical testing consent forms, are provided counseling. Clinician enters the clinical information into the online data fields that populate the eyeGENE database, while the clinical laboratory enters the genotype information. The database is not public and is password protected for privacy	h /n

IDbases	Immunodeficiency Database	Tampere University Hospital	Establish database for every immunodeficiency or provide links to those maintained elsewhere.	Maintain 122 public IDbases with 3 under construction (2008). These databases contain data for 5359 patients, 27 Immunodeficiency mutation databases are maintained by others	h _
NCCRCG	National Coordinating Center for the Genetics and Newborn Screening Regional Collaborative Groups	American College of Medical Genetics	Develop standardized laboratory and clinical language for electronic medical records that will be utilized in newborn screening	Genotype and phenotype information, with long-term follow up of newborn screening patients, will be collected on the 54 conditions with over 150 genes	h
LOVD 2.0	Leiden Open Variation Database LAMP = Linux, Apache, MySQL, PHP	Leiden University Medical Centre	Software and database to establish, manage and curate a web-based gene specific DNA variant database (LSDB)	Sequence variant data and patient data stored separately - variations in several genes can be linked to one patient. The pathogenicity of each variant can be set by submitter and curator, allowing one to easily separate disease causing from non-pathogenic variants.	h
UMD	Universal Mutation Database®	INSERM - Montpellier	Creation of relational databases that run on both Macintosh and PC platforms and can create dynamic HTML pages for direct	The UMD software uses the 4th Dimension® package from 4D. The UMD central tool can query multiple Locus Specific DataBases (LSDB) developed with the Universal Mutation Database® software	h

Box 1

Key Objectives of the HVP

1. Capture and archive all human gene variation associated with human disease in a central location with mirror sites in other countries. Data governance will ensure security and integrity through the use of auditing and security technologies but nevertheless 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 gene specific (locus specific), country specific, or disease specific database perspective to improve accuracy, reduce errors and develop a comprehensive data set comprising 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 work as a dialogue 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 to ensure open access to all human variation data that are to be used for 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 knowledge to all countries of the world.
10. Continue to carry out research within the opportunities presented by investigation of human genetic variation and to present these findings to users of this information for the benefit of all.

Box 2

Developing Interoperable Clinical and Research Locus Specific Databases

1. Security, consent and ethical approval
2. Develop standards and guidelines for diagnostic labs to address database contribution
3. Develop quality standards for genotype and interpretation
4. Develop recommendations for formats for submission of data
5. Provide models for collecting clinical information
6. Provide recommendation for funding for database efforts, including bioinformatics
7. Scalability: can small databases readily handle the output from next gen sequencers?
8. Laboratory pipeline from instruments to database (LIMS : Database interface)

Box 3

Standards, Validation, Quantification and Transparency

1. **Standards -** How data (including clinical, demographic, pathologic, genetic, lifestyle and environment, computational, and laboratory) are collected, recorded, and utilized should be standardized so that results are as unambiguous as possible to users.

2. **Validation -** Before conclusions are drawn, new types of data should be compared with validated sets of variants that are known to be (or are highly likely to be) pathogenic and neutral, respectively.

3. **Quantification -** Uncertainty exists around all of the data types described above. That uncertainty should be quantified as best as possible. When properly quantified, results from all data can be combined to produce a final probability that a variant is pathogenic.

4. **Transparency -** Uncertainty can best be addressed when the source and quality of all data are known. Understanding genetic variation will be an ongoing, iterative process. Proper reporting of a conclusion regarding pathogenicity will require disclosure of all data sources, their quality, the methods of combining them, and the date of the conclusion.

Box 4

Data Transfer & Databasing Standards for LSDB

1. Develop a program to recruit curators of LSDBs who have yet to join the HVP.

2. LSDBs will be standardized and use freely available and compatible software packages such as Universal Media Disc (UMD) or Leiden Open Source Variation Database (LOVD).

3. Develop standard formats for published data by disseminating rules for LSDB submission to journals. A new standard of publication should call for as much phenotypic data as possible. Similarly, funding agencies should require submission of all gene variant data resulting from their support.

4. LSDBs should develop and follow quality standards that are measureable and transparent. The HGVS should develop a quality label for LSDBs following all standards.

5. To ensure permanent access, LSDBs should share a copy of their data to central repositories (e.g. HGVS, NCBI, EBI, Gen2Phen). In addition, to improve easy access, LSDBs should consider automated data exchanges in real time with continuously funded national databases (NCBI, EBI).

6. LSDBs can include data from expression or other *in vitro* functional studies, but these data should be clearly marked as *in vitro*.

7. New LSDBs are needed for every gene related to a Mendelian disorder. Lost or abandoned LSDBs should be revived. HGVS and the Gen2Phen consortium could play an active role here, soliciting new curators as well as finding financial support. Duplication of effort should be avoided, but if it occurs, LSDBs with common genes and variants should share data and entries.

8. An internationally acceptable database policy statement, covering ethical and privacy issues for gene variant and phenotype data collection and sharing, is urgently needed.

Box 5

Data Elements for LSDB

1. HGVS name and historical name(s) of the variant
2. Data submitter (LSDB or individual researcher)
3. URL back to data submitter per variant
4. PubMed UIDs associated with the variant
5. Number of observations of a variant
6. MIM number of the disease(s) or phenotype(s) of interest
7. MIM number of the allele of interest
8. Clinical interpretation
9. Clinical or research groups testing for this variant

Box 6

Emerging Populations: Objectives for Inclusion

1. Develop procedures to include neglected, under-resourced clinicians/researchers in emerging countries, by creating networks among those of common interest.

2. Recognize the need and challenges for recruiting participants. Considerable resources are needed to ensure an adequate consenting process as well as data collection, both of which can be cumbersome.

3. Ensure that the databases and software suite of tools that are being used or developed can be utilized by those in an under-capacitated environment, and that the software pipelines transfer genotype data easily between clinic and laboratory (e.g. Locus Specific Databases [LSDB]).

4. Researchers and their academic institutions will develop an infrastructure in anticipation of the Coordinating Centre working with international organizations (UN, UNICEF, WHO, OECD) to create a framework to facilitate interactions with national governments and regional organizations such as NEPAD (New Programme for Africa's Development) downward through national governments, their Departments of Health, Education, Science and Technology, amongst others.

5. To ensure that ethical, legal, religious and social issues of emerging countries are considered in all portals of the HVP. The HVP will enlist the regulatory/policy arm of WHO/UN to acknowledge and deal with issues pertinent to Economic, Educational, Ethical, Legal and Social Issues.

6. To foster communication, interaction and research networks between the developing and the developed countries.

7. To keep abreast of new analytical methods that may impact or influence the ability to maintain confidentiality of data.

Box 7

Ethics Sub-Group Recommendations and Remaining Issues

1. Develop a Common Ethical Framework that governs all aspects of the HVP. This guideline will ensure strict privacy protection for all participants and provide guidance for optimizing informed consent. It will strive to ensure wide access and information sharing, global benefit sharing, while preserving individual country norms, guidelines, and legal requirements. It will establish a clear basis on which to protect individuals, families, and communities from harm.

2. Curation and all communications about work involved in a particular LSDB must be transparent and clear as to the nature, purpose and limits of that particular LSDB.

3. Different types of data must be handled in ethically appropriate manners; including prospective, published and archived data. All data must, furthermore, be handled according to the conditions under which consent was originally obtained, and privacy strictly maintained (see footnote).

4. Ideally, donors should be informed when their data is being transmitted to a LSDB and to the Web. However, difficulties in achieving this goal arise. Is consent required prior to transmission? What is the ethically appropriate way to handle this requirement when consent is not possible (no re-contact clause, impossible to re-contact, or an individual

has died. If consent for transmission was not part of the original consent, must re-contact occur to obtain consent? If consent is not obtained is it unethical to transmit even anonymized data, or particular types of anonymized data such as that governing smaller more easily identified populations/individuals?).

5. What constitutes appropriate consent for prospective data collection?

6. How ought data which is not tied to consent be handled? That is, in the absence of explicit approval or refusal can data be used and if so, ought it to be de-identified or anonymized, or not used?

7. Is access to deceased individuals' information ethically appropriate? Is familial consent required if individuals did not authorize access prior to their death?

8. In fostering standardization, there is a need to consistently define 'de-identified' and 'anonymized' and ensure consistent use across the LSDBs.

9. Create an Ethics Review Committee that can ensure that the HVP adheres to the utmost ethical principles as well as serve as counsel in the event of issues in need of resolution. For example, the Committee could determine whether rare mutations ought to have more stringent protections than common mutations.

10. Ensure strict protection of privacy in cases of rare mutations in rare disease or unique combinations of clinical features, where identification of individuals is at risk.

11. Ensure against harms to vulnerable individuals or groups of individuals, including groups defined on the basis of data findings.

12. Disclosure of additional data, even for clinical reasons to medical professionals, cannot occur unless prior authorization given by donor in original consent.

13. Define conditions for removal from database.

14. Virtual or actual medical relationships with an enquirer are unwise and are discouraged (e.g., a company or person who performs the genetic test should not provide advice on the use of the results).

15. Determine the appropriate limits to access within LSDBs to prevent individual identification. (ex. Ought parts of the LSDB be password protected?)

16. Consider ethical issues arising from transferring data from LSDB to web browsers; refer to the Ethics Committee for consensus building.

Box 8

Development of Justifications for Funding

1. Focus on scientific questions and hypotheses that can be addressed using the variome database. For example, finding out what proportion of cases in each single gene disorder have unexpected phenotype(s) due to presumably modifier genes will require substantially reduced effort when collection is complete, up-to-date and easily accessible, and this information will be available at all geographical locations. Provide a flow of data from patient via clinician/laboratory and curator to central browsers/databases will be automated and seamless.

2. Both gene specific and disease specific programs can be developed when applying for funding (i.e. the InSiGHT colon cancer project).

3. Develop plans for grant applications for all venues. For examples, applications to international, national, state or province or department, country health departments; research agencies (e.g. NIH, NCI, MRC, NHMRC, EC, OECD); charitable organizations, foundations, philanthropic trusts and individuals; companies interested in data management and sequence data acquisition. Many of these efforts can be pursued in parallel. Successful applications will provide leverage for additional ones.

4. Demonstrate feasibility and productivity of approaches before application (pilot projects). The most notable at present is the InSiGHT collaboration and the “Adopt a gene” approach.

5. Lobby for general support of the HVP as well as for individual grants.

6. Proceed with systematical implementation of the Adopt a Gene approach, which has now been initiated, for funding of curators.

Box 9

HVP Governing Plan to Be Developed with Stakeholders

1. Examine overall governance strategy outside the facilitating activities of the coordinating office.

2. Examine the potential use of the term Human Variome Programme to be applied to all variation activities outside the initial remit of the Human Variome Project as initiated in Melbourne. For example, the utility of the datasets for genome wide association studies, HapMap projects, diagnostics for personalized nutrition and medicine.

3. Evaluate the need for a term or body such as the Human Variome Organisation (HUVO) and examine whether Human Proteome Organization (HUPO - http://www.hupo.org/) is a satisfactory model both for the “HVProgramme” and HVProject”.

4. Define HVP projects and HVP collaborating projects. The first example of a collaboration project will be with the InSiGHT consortium.

5. Develop draft criteria for new HVP initiatives.

6. Develop a HVP international consortium after piloting a project in Australia.

7. Develop a Governance/Strategic working group from the Planning Group and others.

8. Encourage standards in the various facets of the HVP and harmonization if this is not possible.

9. Encourage the development of miniature devices for complete genome sequencing.

Box 10

The InSiGHT Committee Description and Plans for the Prototype HVP Pilot

The problem of interpreting the pathogenicity of missense variants is most commonly encountered in genetic testing for cancer predisposition syndromes. Genetic testing is commonly performed for Hereditary Non-Polyposis Colorectal Cancer (HNPCC, Lynch Syndrome), which results from inherited pathogenic mutations in at least four genes responsible for DNA Mismatch Repair (MMR). Between 15-30% of MMR variants detected during genetic testing for Lynch syndrome are missense, and most are rare and not interpretable regarding pathogenicity (Peltomaki and Vasen, 1997). InSiGHT has established committees addressing phenotype, curation, virtual histology, and funding, which provide advice and support for the Interpretation Committee comprised of experts ranging from clinicians to basic scientists. Its two short-term goals are to (i) produce a paper discussing standards for data types, classification, and integration and (ii) convene to classify as many variants as possible from various international MMR gene mutation databases. The InSiGHT Interpretation Committee is undertaking the following actions:

1. Consolidation of multiple MMR gene databases
2. Assessment quality standards for phenotypic data, including tumor phenotype
3. Establishment of standards for clinical, epidemiologic, and pathologic data
4. Formation of standards for *in vitro* assays
5. Standardization and quality assurance methods for computational data and methods
6. Validation of methods for integrating these multiple data sources

7. Using databases to organize data and allowing access to the biomedical community

Box 11

Other Pilot and Relevant Projects

1. Kuwait (http://www.al-mulla.org/) established a Molecular Genetics Diagnostic Service Division, integrated within a research laboratory, which is focused on delivering state-of-the-art diagnostic service for the Kuwaiti population. These services provide the research laboratory with a unique opportunity to meet probands with various genetic diseases.

2. The Danish National HNPCC system has had 12 months' experience through the web-based, but closed Danish Health Data Network (www.sundhed.dk). This model is relevant for all genetic diseases. A lesson to date here is that systems are successful where the output improves the user's activity and efficiency.

3. Systems to support public and personal health in nutrition is being developed through two international nutrigenomics consortia for type 2 diabetes and micronutrient genetics (websites within http://www.nugo.org, free registration). These efforts introduce a nutritional domain in the data matrix that relate variants of the relevant genes to nutrition profiles.

4. The Rare Metabolic Disease database (RAMEDIS http://www.ramedis.de; Topel, et al., 2006) serves the needs of specialist metabolic centres, including their clinical management, and provides access for an understanding of normal biological processes

often first informed by the study of these rare disorders. These too could be linked into the broader HVP plan. The interface of microarray expression data into the HVP plan offers yet another very challenging data dense domain, which will need careful interpretation with respect to its utility in clinical medicine.

5. The UK has developed the central MutDB some time ago to collect all UK variation data and its effect from participating laboratories (http://mutdb.org/; Dantzer, et al., 2005).