EU Workshop Report on European Database and Analysis Resources for Research in Human Genetic Variation

Based upon a Workshop held in Brussels, Belgium on 2-3 March 2006

by Stylianos E. Antonarakis, Ewan Birney, Anthony J. Brookes, Lon Cardon, David N. Cooper, Johan T. den Dunnen, Simon Heath, Karen Kennedy, Hans Lehrach, George P. Patrinos, Stefan Schreiber, Kári Stefánsson, Johan van der Lei, Gert-Jan Van Ommen, Edgar Wingender, Richard Wooster

Editors: Frederick Marcus, Bernard Mulligan, European Commission

APRIL 2006

Executive Summary

The need for unification of databases focusing on human genetic variation and associated phenotype links has been recently highlighted in a *Nature Genetics* editorial and elsewhere. Until now, no effective strategy for achieving this has been formulated. Existing collections of genetic relationships, predominantly from Mendelian single gene variation traits, when supplemented by information from model organisms, have provided many fundamental insights into human biology, at both the body and cellular levels. Differences between healthy people, and also causes of diseases, have some genetic component. If all aspects of the genetic contribution could be identified, they would lead to advances in biomedical research, as well as furthering the cataloguing of human genotype-phenotype relationships. Nevertheless, the workshop participants strongly supported the thesis that the lack of data integration inhibits many research breakthroughs.

An integrated genetic variation catalogue would be an immense boon to bioresearch, in areas such as: general understanding of human physiological processes in both health and disease; the ability to analyse populations according to different classifications; the diagnosis and treatment of disease.

An workshop of leading European bioinformaticians, biologists, medical researchers and clinicians was held in Brussels on 2-3 March 2006 to examine if strategies could be identified. The conclusion was that an integrated database and analysis structure for much of human and model organism variation genetics should be achieved by database linking at a European level, and in the near future, by means of a pragmatic and step-by-step approach.

The organising principle of the database network would be the genotype – phenotype relationship. This combination spans the whole descriptive range of genetic variation, from single DNA base changes to highly complicated biological and clinical phenotypes and diseases. The workshop participants thought it infeasible to attempt to create a wholly new central database, with associated ontologies and standards. In fact, a wide range of databases already exists within Europe, with preliminary links to important databases elsewhere (OMIM, dbSNP in the USA). These databases and ontologies are sufficient and suitable for forming hubs to interlink in rather straightforward ways.

Database linkage could be accomplished using technologies implemented in existing EU bioinformatics grid projects and data exchange formats. This linkage should be based on a hierarchical system, with one or two major genetic sequence-based databases like ENSEMBL and its genome browser software packages acting as a hub, with links to broadly-based genome variation databases. There would be further links to the many specialized databases of four main types: locus-specific, disease-specific, population and biobank. This interlinked data should be accessed by a variety of tailored user-friendly interfaces.

Data in the public domain is required for successful and efficient access. Semi-commercial and commercial (non-public) databases could also be connected with the integrated database system. Many of these databases rely on data in the public domain, and already have arrangements for making some data available to academic researchers.

Key elements of related genetics research programs were highlighted at the workshop, such as the Wellcome Trust Case Control Consortium. The study of disease-focused association and genetic diversity, conducted at the multi-population level, would provide the type of data needed to underpin full and correct analysis of many other datasets. Unified and more complete datasets would provide improved opportunities for researchers to study association on a genome-wide level.

National and other funding agencies should promote local integration of distributed databases, support them, and encourage wider collaboration. The European Research Framework Programme should provide opportunities for collaborative projects at European and world-wide levels via topics in calls for proposals for catalysing the creation of this network of databases essential for entering the era of system biology.

For more information, **contact:** Frederick.Marcus@cec.eu.int Bernard.Mulligan@cec.eu.int **The report is available at: http://www.cordis.lu/lifescihealth/genomics/home.htm** ; http://europa.eu.int/comm/research/health/genomics/index_en.htm

TABLE OF CONTENTS

Introduction

A recent editorial [*Nature Genetics* 2005 Aug; 37(8):783 "WayStation to HUGOBase"] and a paper by Patrinos and Brookes [*Trends in Genetics* 2005, Vol 21, 333 "DNA, diseases and databases: disastrously deficient"] highlighted the need for coordination of databases focusing on human genetic variation and associated phenotype relationships. They noted that the need for coordination in this area has long been recognised, but so far no effective solutions have been found. The current database, data capture and analysis structures are highly fragmented, and a wide range of analyses are very difficult or sub-optimal. There are currently a number of worldwide efforts and discussions on moving towards a unified database, but they have all come up against the problem of combining very different data types and research fields into a single database. The problem is compounded by major difficulties with data accessibility and confidentiality. Improved access, data validation, curation and analysis would be immensely valuable and allow better utilisation of the billions of euros of expert work in human genetics research involving population and comparative genetics, biobanks, clinical trials and pharmacogenetics.

Workshop Procedure

To address these problems and to consider solutions, a workshop was hosted by the Fundamental Genomics Unit in the Health Research Directorate of DG Research of the European Commission, on 2-3 March 2006, in Brussels. Terms of reference (APPENDIX 1) and a questionnaire were circulated in advance to the participants to motivate presentations at the meeting. Eighteen top European bioinformaticians, experimental biologists, medical researchers and clinicians (APPENDIX 3) attended the workshop.

The Terms of Reference (TOR) concentrated on:

- Data access
- Genetic analysis, including human SNP (single nucleotide polymorphism), haplotype and QTL data
- National, European and world-wide collaboration possibilities.

The TOR and responses to the questionnaire were summarized by the Commission staff as an introduction to the meeting. Individuals presented papers during three sessions, which included extensive discussions. The three workshop presentation sessions were entitled:

1. Current state of the art in databases and analysis tools
2. Key biology and medical questions to be addressed with genetic variation public database capabilities and required analysis tools
3. Towards a genetic variation public database - contents, structure, standards, resources and requirements for realisation.

During the final session, other sessions were summarized by rapporteurs and further discussed, leading to the workshop conclusions. The questionnaire results (APPENDIX 5) were refined and the presentations (APPENDIX 4) were reviewed. The report was then drafted, reviewed by participants, and finalized.

The goal, which participants concluded was realistic, is to provide a database and analysis structure for much of human and model organism genetics. A means to achieve it in the near future, by a hierarchy of grid-linked databases and tools, was outlined.

New Elements Facilitating Data Integration

To facilitate data linkage, a wide range of new tools have become available, including: integrating data-grid protocols and technologies; the EMBRACE bioinformatics grid capabilities http://www.embracegrid.info, which could be implemented for genetic variation (APPENDIX 2); recent experience in integrating databases (TEMBLOR/ Integr8); integrated analysis pipelines (Biosapiens, ENFIN); and genome browsers (Ensembl), all of which are wholly or partially funded at EU-level.

New high-throughput technologies are becoming available, greatly lowering cost and allowing new data to be more complete. Major scientific support data is being provided from large-scale comparative genomics projects including non-human model organisms, for example genotype-phenotype data from EU-funded mouse projects. Relevant clinical data is increasingly computerised and publicly available.

There is a new trend, started in the UK and now also implemented by the NIH in the USA, towards public release of all data generated in large-scale genetics research projects. In some cases, there is complete

public access to control genotype data, and *bona fide* researcher access to additional data. In some projects, all raw genotype data is released, although there will be some restrictions on initial use. 'Big Pharma' is also interested in contributing to these open datasets.

In the past, genetic studies of complex diseases have not met with the anticipated success, for example in statistics from human association studies. Most researchers recognise a considerable lack of statistical analysis power and lack of genome coverage for many previous association studies. However, the current generation of association studies has reasonable power and allows genome-wide testing. Very large-scale association projects are in progress in the UK (Welcome Trust Case Control Consortium - CCC) and in the USA (the planned GAIN project).

Biological and Medical Research – Genotype to Phenotype

The guiding organising and scientific principle for linking databases is the genotype-phenotype relationship, which can involve an extremely detailed and multi-level classification.

Simple genotypes (one mutation) and phenotypes (one disease) were the key principles that Victor McKusick used to found the modern approach to genetics databases in the USA forty years ago, with OMIM http://www.ncbi.nlm.nih.gov/Omim. A genotype can be much more complicated than a single SNP in a protein coding gene [see dbSNP database http://www.ncbi.nlm.nih.gov/SNP, which is integrated with other ENTREZ databases http://www.ncbi.nih.gov/Database/datamodel]. The genotype can include SNPs, haplotypes, LSDBs (locus specific database), multiple copies, non-coding DNAs, QTLs (quantitative trait loci), epigenetic (histones and methylation) and environmental effects, full sample classification and characterisation. SNPs produce a huge number of types and "consequences", with 10 million human common variants, and additional minority variants, including non-synonymous, synonymous, UTR, regulatory, GT/AG splice changes, stop gains and frame shifts.

Genotype classification should be extended to look at the context of mutations, including the roles of:

- Abnormal copies and phenotypic differences
- Local DNA sequence context
- Mutation frequency (by type)
- Genomic loci (comparative analysis)
- Mutational spectra (design strategies and comparisons)
- Environmental and population context (providing differential effects of genetics for mutations).

Phenotype information can be used as input to data mining, genetic association, systems biology, physiology and epidemiology analyses. Phenotype classification can include normal to altered gene expression, protein-protein interaction, pathway, and the cellular, tissue and organism response. Humans have by far the most complex phenotype classification, developed in medicine in relatively modern terms over the past couple of hundred years. Even medical phenotypes can be strongly subdivided, since clinicians tend to combine and eliminate data, so as to efficiently identify treatment for a global phenotype.

To extend and to provide a firmer basis for analysis of data, further studies focusing on gene expression in relation to haplotypes and in duplicated and deleted genomic regions would provide essential data. Such data is not only available from animal models. Patient-derived cell lines provide an enormous resource for such studies (which should be collected and analyzed). Other phenotypic variability studies might include:

- Underlying genetic heterogeneity of inherited disorders in populations - Linking mutant genotypes to clinical phenotypes
- Interactions between multiple susceptibility factors and environment
- Differentiated and categorized neutral versus pathogenic variations
- Role of sequence variation and modifiers in monogenetic disease
- DNA variation in complex traits
- Health risk with variations associated with particular diseases
- Role of somatic mutations
- Careful correlations of genotype to phenotype (but not dbSNP duplicate)
- Outcomes and extended molecular phenotypes, levels of clinical sub-phenotypes, endophenotypes - e.g. osteoarthritis
- Somatic variations.

Analysis of this data presents major challenges. The highest priority issues include:

- Standardisation and database integration
- The 'phenotype data' representation challenge
- Handling association data (software tools for data generators)
- Publication bias (bring in all study findings, including negative results)
- A standard model for classifying DNA variation
- Copy-Number Variation (major genetic effects, complex informatics)
- A generic phenotype data model
- A prototype genetic association database
- Convenient database applications for genotyping labs
- Data submission tools for genotype-phenotype data.

Standards and Ontologies

Protocols for standards, ontologies, submission and data exchange are essential for successful data linking. Fortunately, these already exist in very extensive formats, such as GO http://www.geneontology.org at the genetics level and medical classifications at the physiology and pathophysiological levels. Complex XML-based submission and exchange formats have already been specified in extensive detail. Nevertheless, a significant amount of work still remains to be done in terms of fully agreed standards, especially for complex phenotypes.

Detailed standards and ontologies will require the following:

- Standard nomenclature of genetic variants
- Guidelines for contents, database structure and standards
- International collaborations
- Control population data across Europe
- Genetic epidemiology centres working on joint Standard Operating Procedures in a quality control network throughout Europe. Standards for submission and deposition are also crucial and a huge field. A major workshop in planned for these areas in October 2006.

We should first concentrate on standards for the core material. Relevant projects include: Molpage, Moltools, Wellcome Trust Conference on Biobanks, Eumorphia.

Interconnection would also be helped by developing relevant standards and tools, such as ENSEMBL Biomart http://www.biomart.org and the Polymorphism Markup Language.

Data Submission

Data may be submitted via journals or directly into databases or both. Incentives to lab personnel for database deposition should be provided by funders and by scientific journals. A full range of consultations should be initiated with journals to investigate ways of improving direct deposition and facilitating data mining via publications standards.

In the future, journals and databases may be replaced by 'database journals', wherein results are deposited directly into internet-accessible structured depositories, which are interconnected into a 'bioknowledge-web'. The genotype-phenotype challenge could encourage this practice. The web is the easiest and least expensive place to publish work, a fact which would also encourage the inclusion of negative results.

Display and Analysis Tools

To maximize utility and attractiveness of submission, display and analysis tools, a set of principles should be developed and observed, with user-friendliness at the top of the list. User-type specific front-end interfaces are essential for display, bioinformatics, association studies and systems biology analysis and simulators. At the bioinformatics level, should be developed for association studies across a wide range of genetic data to answer complex queries. Better laboratory-based capture of genotype-phenotype information is essential. Developments are required in the areas of database construction, maintenance and software packages, and a special phenotype vocabulary for locus-specific, national and linked databases.

For research geneticists, interfacing via a genome browser is the most attractive means of working, supplemented by data links. On the other hand, clinicians and medical researchers do not like to work in

genome coordinates. Using data is popular; contributing data is unpopular. This reluctance highlights the importance of having various interfaces suited to user preferences, independent of the internal links between databases.

Clinicians would follow the route LSDB → genome (Ensembl) → many links e.g. OMIM, HGMD). Their requirements include:

- reliable LSDB interfaces, with links to other (genome) information
- an up-to-date and 100% complete list of known gene variants (including rare variants)
- a Reference Sequence showing nucleotide numbering for the gene
- a field with information regarding the reported pathogenicity of that variant
- a reference to the source of the information
- any other tools / links connected to his subject of interest.

A wish list for all user communities might include:

- an open genotype-phenotype database
- initial basic functionality (tracking variants which change phenotype)
- open access (commercial data must be accessible to be integrated).

Linking to Systems Biology Analysis

Systems biology analysis requires the type of data input used by PyBioS, which simulates a wide range of metabolic and expression pathways for healthy and disease states. The data includes: substrates and products of a reaction, stoichiometry, catalyzing enzyme, kinetics, reactants and enzyme concentrations. The program also links to several databases and tools: KEGG (Kyoto Encyclopaedia of Genes and Genomes); Reactome; Transpath (Database of signal transduction pathways); SRS (Sequence Retrieval System); BioCyc; Kinetikon (Kinetics database) and "raw" experimental data (Expression data, Protein/Protein interaction). To analyze the effects of genetic variations, it is necessary to know how they activate or deactivate certain key pathways.

Appropriate links to network and pathway databases and quantitative data on the effects of genetic variations are required. A set of such databases have been developed, dealing with different aspects of gene regulation (TRANSFAC, TRANSCompel, TRANSPro, TiProD), protein-protein interactions of whole proteomes (HumanPSD) and signal transduction (TRANSPATH) for intercellular (specifically endocrine) signalling networks (EndoNet). They are complemented by databases on pathologically relevant mutations of genes encoding regulatory proteins (PathoDB, PathoSign) and disease-involvement of human proteins (HumanPSD/Disease Reports). Together they constitute an information infrastructure useful for projects that link genotype data to molecular and clinical phenotype information.

Databases to be Linked

A mixture of databases and associated analysis tools is currently available worldwide, with many in Europe. Some have been developed in publicly funded projects or with institutional budgets, are still maintained, and are fully in the public domain. Others have been transferred to commercial exploitation to generate revenue for maintenance and upgrading. Still others have been generated as purely commercial products, nevertheless providing some of their contents and services free of charge for users from non-profit entities. A range of mechanisms to balance the interests of academic researchers with those of commercial vendors has been implemented, and are often operational and sustainable. Databases can be characterized as follows:

Large public databases, partly linked by periodic data exchange and hot-links include: ENSEMBL, dbSNP, OMIM, Uniprot. ENSEMBL already handles a wide variety of variation data. Variations can be SNPs, and “reasonable” indels (insertions and deletions). It can handle multiple sources, genotypes in multiple populations, and both heterozygote (Human) and strain based (Mice and Rat) scenarios. ENSEMBL is very flexible, allowing for stable and sensible handling of variation. It has the ability to handle larger genome polymorphisms and resequencing data, and to scale to thousands of people and millions of genotypes. This allows integration with functional and comparative genomics data.

Medium size general databases include public and semi-public (charges to commercial customers)

The Human Gene Mutation Database (HGMD) http://www.hgmd.org represents a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease. Data catalogued include: single base-pair substitutions in coding, regulatory and splicing relevant regions, micro-deletions and micro-insertions, indels, triplet repeat expansions, gross deletions, insertions and duplications, and complex rearrangements. Human gene mutation is an inherently non-random process. The nature, frequency and location of the mutational lesion all are strongly influenced by the local DNA sequence context.

HGMD may be exploited to study the role of the local DNA sequence environment (e.g. repetitive sequence elements, sequence homologies and specific motifs) in mediating mutational events and to explore the nature of the underlying mechanisms. HGMD provides the only comprehensive collection of data on human gene mutations causing inherited disease and as such provides a key means of linking mutant genotypes to clinical phenotypes. Since functional SNPs with or without known disease relevance are also included, there is already a natural bridge between the pathological mutations in HGMD and the predominantly neutral SNPs catalogued in other databases.

By comparison, HGVbase-G2P http://hgvbase.cgb.ki.se will focus on all forms of variation and any association studies that connect such variants to any phenotype. This will include published and (mostly) unpublished data. In practice, it will mainly capture genetic association evidence between DNA variants and complex disease, where the genetic component alters risk but does not cause the disease (neither necessary nor sufficient to account for the observed phenotype). It will also capture environment data - the other major contributor of complex disease causation. The records will be extensive in scope, carrying detailed phenotype, sample, population, assay, genotype, allele, haplotype, marker, and sequence data, along with concluded p-values for disease associations (single point effects and synergistic interactions) plus citations, free text, and key word information. Genome annotation (e.g., exon, CDS, splice sites, repeats) will also be available for guiding database searches, along with a range of submitter information.

Predominantly commercial databases include those available through DeCode http://www.decode.com, Celera http://www.celera.com, BioBase http://www.biobase.de.

DeCode is an impressive resource, combining genealogy, phenotype and genetics. Many potential discoveries are feasible, but unlikely to be made by Decode, as they have enough "nuggets" to handle for the moment. There are also important problems with multiple mining and testing of these discovery datasets. Public release of this dataset might be possible under appropriate circumstances and funding.

Specialised locus and disease specific, population and ethnic genetic databases include COSMIC http://www.sanger.ac.uk/genetics/CGP/cosmic and p53 data. Somatic mutations are well handled by COSMIC. Many other databases exist, such as TRANSFAC http://www.gene-regulation.com for transcription factors and LOVD the Leiden Open Variation Database http://www.dmd.nl/LOVD. Different population databases to assess heterogeneity based on ethnicity, and future databases should be linked. There will be a role for many such databases to capture the full spectrum and scale of association studies being conducted globally. It is critical to capture all or most data to distinguish true from false (chance) positives, since we cannot rely on published findings alone. Hence, we rely on databases like HGVbase-G2P that gather primary data (some with specific focuses such as cancer, nations, pathways), plus interfaces allowing searches in multiple databases. Interface design and standards are essential. Central browsers such as Ensembl may only be able to (and should only) include summary level information from these many 'association databases' (i.e., markers, phenotype name, and p-values) for presentation graphically, with links back to association databases.

There is also an important requirement for genotype-phenotype raw data archives. A strategy is required to harmonize raw data collation and annotation, and for tools to store and disseminate the data. This is non-trivial, since some file sizes from single studies are tens of terabytes. Analysis tools for statistical genetics are essential, since they are not as developed as other bioinformatics areas in standardisation and capability.

In general, there are many aspects to consider in choosing databases to be linked:

- Different genetic sources in humans & model organisms
- Different types of data
- Raw vs. derived data

- Remember that all genotype-phenotype data is not equally useful
- Compilations of all associations.

Role of Model Organisms

The role and importance of model organisms is vital. All life on earth is linked by evolution. Even the most basic organisms provide relevant genotype-phenotype relationships in pathways and interactions conserved through evolution. Such data is often not available from humans, for a variety of experimental and ethical reasons. For example, lethal germ-line mutations, e.g. inherited homozygous or mutated egg or sperm heterozygous mutations, lead to death soon after fertilisation and yield no population data. Recently, phenotype categories have been vastly expanded and characterised in model organisms such as mouse and zebrafish. The best genotype-phenotype data for human genetic inferences is via mouse, fly and worm (not human!). In these model organisms, studies of large-scale variation aspects are well advanced.

Approaches to linking Databases

By concentrating on linking data which is already in the public domain, all issues concerning access, patient privacy and confidentiality are devolved to the local level. The issues are resolved at the level of each institution (e.g. hospital) or government (state or country), each with its own ethics committee, legislation, medical procedures and traditions.

Links to "restricted" databases can be developed on a case-by-case basis, keeping in mind overall goals of maximising public access. Discussions indicate that these database owners feel that there are solutions to making data publicly available, while protecting their value-added commercial viability.

Decisions as to which databases should be linked, by what means and in which order, do not have to be addressed here. In the context of EU funding schemes, consortia of organisations form themselves and make their proposals in response to EU topic- or area-based calls for proposals. The consortia would prioritise actions in their proposals, which would be competitively evaluated by external reviewers. The winning consortium could catalyze further integration, both European and world-wide. Individual databases could also provide links in the very near future by using EMBRACE grid protocols.

Researchers should consider what kinds of data, including raw, new and existing, to prioritise for capture, linking and "outreach" access**.** A pragmatic approach is required which is adjusted to availability and access types. Raw data access is critically dependent on format standardisation.

Related Genetics Research and Infrastructures – Biobanks and Testing

Key elements of a number of related and integrated genetics research programs were highlighted at the workshop, to provide the type of data needed to underpin full and correct analysis of linked database resources. These programs are being discussed in much more detail in related conferences and workshops [see *From Biobanks to Biomarkers - Translating the potential of human population genetic research to improve the quality of health of the EU citizen,* Proceedings of a conference held at the Wellcome Trust Conference Centre, Hinxton, Cambridge, 20-22 September 2005].

Activities should include very extensive studies of comparative, developmental and functional genomics in human and model organisms as appropriate, which provides both variation data and the biological knowledge underpinning analysis. All activities should be carried out in close and continuous interactive collaboration with data providers and users, including biologists, geneticists, medical researchers and clinicians.

One area particularly identified was the requirement for a control population genetics study, which could be implemented as a mixture of national and EU collaborative projects. There are also important advantages in combining data sources for disease studies, involving linkage screening, fine mapping, and whole genome association. Combined sources would enable very large-scale epidemiological studies (multi-centre studies). As an example, a study of 96 'neutral' SNPs showed surprising variation between populations, illustrating the value of combining association studies. It would be useful to have a standard set of population samples from across Europe, allowing us to identify a set of markers which can capture most of the allele frequency variation. A common set of guidelines for data release across Europe would facilitate data combination.

To further investigate genome variation of human populations, a database for all Copy Number Variation (CNV) data (in patients and healthy people) is required, coupled with studies to catalogue this variability. Currently there is no such database, and clinical diagnostic labs struggle with the information coming out of genome-wide CNV studies in relation to genetic disease. Since clinicians focus on specific pathogenic conditions, an effort to analyze a large set of controls to catalogue the non-pathogenic variations would be very worthwhile.

Discovery of new scientific knowledge is possible from large databases of measurements, observations and interpretations from population and biobank based research, by using the patient as a data source, in the sense of accidental experiments. This has the advantage of avoiding sampling bias, which can occur when specific disease or ethnic populations are chosen. Information is also obtainable about children, which is never the case in clinical trials.

Data-taking procedures require improvement, including samples prepared/stored and ready to be rapidly assessed; prioritised marker selection, data analysis and results, so as to satisfy critical statistical issues in complex disease gene-identification.

Biobanks and related genetic testing form a key part of the foundations for future health care. They allow:

- Outcome research for individuals carrying risk genotypes
- Prospective follow up of entire populations
- Detection of sub clinical manifestations
- Genotype based prevention
- Clinical trials to establish procedures
- New algorithms for genotype balanced randomisation.

To develop these possibilities, there are a number of necessary actions required:

- Establish a network of population-representative biobanks that share elements of standardisation
- Establish a network of genotyping centres that are highly standardized
- Create accessibility to these networks for clinicians and clinical expertise
- Provide background genotype frequencies to clinical projects
- Create a repository for all genotypes generated, with tags back into the originating (DNA) biobanks.

Understanding genetic etiology will further the understanding of complex diseases. With some exceptions, understanding of genetic etiology in isolation from environmental considerations will NOT lead to new therapies and will have NO direct impact on the health of European populations with chronic diseases. The coming challenge will be maintenance of health rather than cure of disease. Molecular prevention will become a major approach and most likely will involve nutrition-based strategies. These remedies will often be applicable to genetic and non-genetic pathologies. Understanding will require approaches to biobanking and data from the following sources:

- Biobanking data of the consequences of environmental triggers or markers thereof
- Open networks that allows clinicians to integrate information, and not merely to hand over samples
- Internal governance by project officers
- Decentralised systems with centralized inventories and standard.

This level of data and supporting procedures will be essential to analyze complex diseases and conditions, such as; diabetes mellitus, metabolic syndrome; arterial hypertension, arteriosclerosis, coronary heart disease; hyperlipidemia, hyperhomocysteinemia; rheumatoid arthritis/osteoarthritis; depression/bipolar disease, schizophrenia; Alzheimer's disease, dementia; multiple sclerosis; bronchial asthma, atopic eczema; sarcoidosis; psoriasis; periodontitis; malignant diseases; Crohn disease, ulcerative colitis; longevity.

The question of data access and property rights to biobanks is highly controversial. Biobank initiatives in Iceland, Estonia and the UK propose the policy of no rights of individual donors or patients to control use of their tissue, by implementing a blanket consent. This controversy is an illustration of why the workshop has chosen to concentrate on data that has already been put into the public domain, and at a local level.

National-Level Programmes

National-level funders, such as the Wellcome Trust (which although a Charitable Trust operating internationally, has the breadth of a large national program), already have wide ranging activities in data

sharing and databases. They support major initiatives to generate large-scale datasets for the research community: e.g., the genome sequencing projects, and the Structural Genomics Consortium. They also support activities specifically related to human genetic variation, e.g. the SNP Consortium, the International HapMap Project, the Case Control Consortium, the Cancer Genome Project, and research in genomic structural variation. Other related activities include collections and cohorts, e.g. ALSPAC, 1958 Birth Cohort, Biobank.

Issues for Consideration for National funders include:

- Terms of Access and use of data – 'open' or managed; there is the need to consider:
  - i Intellectual property – protection of data to exploit, e.g. HapMap click wrap;
  - ii Ethical issues – terms of consent, confidentiality & privacy;
  - iii Interests of the researchers who generate data.
- Data quality, standards and integration – use of existing 'community' standards or need for further development; ability to integrate data from different sources as well as different types of data.
- Long term preservation and sustainability of data resources – securing long-term sustainable funding for key data resources.
- Users needs – on-going evaluation & user input to ensure utility.

Specific Questions/Challenges for National Funders include:

- Promoting data sharing – how to provide incentives?
- What types of data need to be made available and/or integrated– sequence/genotype to complex phenotypes?
- The need for quality control and standards for data.
- How to make 'negative' and raw data available?
- How to control access to data – e.g. validation of bona fide researchers
- Structural Variation – what is the reference sequence? How should structural variation be represented?
- What is needed – integration and linking of existing databases or new databases?
- Co-ordination with other efforts e.g. EU-level projects; the NIH in the USA.

National-Level Support

National bodies have a key role in simultaneously supporting local resources and facilitating external collaborations. Coordinated approaches are essential for funding of databases, curators, sustainability, access, in a variety of academic/commercial environments. There are a variety of funding models, some of which do not facilitate data integration. For true, deep integration of data, one has to openly distribute all the information, and this openness is often crucially dependent on the funding model. There is often a strong relationship between database structures and funding models.

Currently, databases are usually funded locally by national funding bodies, but sometimes have great financial difficulty after the end of initial grants (depending on country and grant circumstances). As an example, HGMD is already developing interrogatory search tools, initially for distribution on a commercial basis. It is however possible to make these freely available to the academic community once a sustainable funding model is in place that guarantees a secure future for the database resource. A sustainable model for the future is public/private mixed funding, supporting free access to academic users and a subscription-based distribution for commercial users marketed by a commercial company. Moreover, the split between private and public funding can be altered over time. Thus, public funding can tip the balance toward increased free public access. Joint industry funding of open resources is an opportunity that should be explored. Industrial access to the data must be possible, as opposed to a prohibition on using data for commercial purposes such as drug development. Multinational initiatives are useful for mobilising EU-wide resources.

European-Level Support and International Collaboration

The EU-level was judged by the participants to be appropriate for providing the integrating effort. Within the EU and its Research Framework Programme, including associated states, there is a full enough range of databases (including those developed as international collaborations) to form a fully functional and valuable set of software and databases of major value. These linked resources could take the lead in

fostering world-wide collaboration. The alternative of a whole series of independent but fully comprehensive National databases would involve huge duplication. A European initiative could forge workable links with its US and other counterparts, via extended grid technology. Europe could also lead as a data provider, because of excellent annotation of data and record keeping in the health care systems.

Since Europe has many of the key resources, the EU and Member States should provide the means to link and centralize them as appropriate. There is a major role for the European Commission in leveraging and linking existing resources. Grant schemes currently provided by the EC can ideally coordinate construction and linking of databases as part of research activities and as infrastructures, also fostering development of querying tools and generic analysis tools. The workshop participants highlighted the role the European Commission could play and the essential importance of funding opportunities from the Framework Programme for Research as soon as possible, in conjunction with national funding. Financial, organisational and scientific interconnection could ideally be provided at the EU-level, for example by providing call topics for one or more Integrated projects (IP). One such IP might be for developing research tools and projects in genetics research, integrated with the database linking process. Other IPs might generate common reference genetic data.

General Conclusions of workshop –the Way Forward

The workshop participants agreed that perceived obstacles to unification of genetic variation data can be largely overcome, by means of a pragmatic and step-by-step approach. A single large database with a single interface is not feasible, the main reasons being the extreme diversity of data producers, scientific areas, funding mechanisms, and requirements of users.

Perceived obstacles included a sense that many incompatible databases could not be combined. Much data is not accessible due to “commercial” aspects, data confidentiality, ethical questions and lack of incentives to submit data. Many people consider that genetic variation is too broad a field to unify data of highly variable quality. In addition, there is a lack of long-term database funding support and financial coordination.

Rather than develop a unified database and database protocols from scratch, the participants concluded that most of the goals of a single database can be achieved in a flexible and useful way by a hierarchy of linked, existing databases. Extensive linking capabilities and grid experience have already been put in place by EU projects relying on data in the public domain. An important goal would be to achieve effective linkage between datasets, which would enable scientific results to be integrated across the entire corpus. Given the huge scale and diversity of the subject area, there will be a role for many 'data warehouses' that summarise information and discoveries, probably each with different domains of interest or focus (such as general genetic variation, cancer, cardiovascular disease, published studies, specific populations and pharmacogenomics). The future for research in these areas will depend upon data interconnection for transparent cross-database searching, aided by relevant tools and standards. The large amount of data in the public domain makes linking possible.

The unifying hub program, based on genotype identification, would probably be an existing genome browser program and database like ENSEMBL, with links to DBSNP and UNIPROT databases, using the distributed annotation system (DAS) and grid capabilities. Development of the EMBRACE life sciences grid is well advanced. The linked genetic-variation databases would be at the level of HGMD and HGVBASE. The further levels of linking would include disease, locus-specific and population genetics databases, e.g. Cosmic, deCode.

Because the databases are linked rather than unified, there can be several entry points, and several user-tailored interfaces. Even though a genome browser might act as a hub for data deposition, exchange and communications, and as an entry point for researchers in fundamental genomics, a researcher in clinical medicine could access the data via an interface specialised for locus- or disease- specific databases.

Principal Conclusions and Proposed Actions

Scientific/biomedical communities and funding organisations should attempt to implement the principal conclusions of the workshop:

1) An integrated database and analysis structure for much of human and model organism variation genetics should be achieved by database linking at a European level, and in the near future, by means of a pragmatic and step-by-step approach.
2) The organising principle of the database network would be the genotype – phenotype relationship. This combination spans the whole descriptive range of genetic variation, from single DNA base changes to highly complicated biological and clinical phenotypes and diseases.
3) Database linkage could be accomplished using technologies implemented in existing EU bioinformatics grid projects and data exchange formats. This linkage should be based on a hierarchical system, with one or two major genetic sequence-based databases like ENSEMBL and its genome browser software packages acting as a hub, with links to broadly-based genome variation databases. There would be further links to the many specialized databases of four main types: locus specific, disease specific, population and biobank. This interlinked data should be accessed by a variety of tailored user-friendly interfaces.
4) Data in the public domain is required for successful and efficient access. Semi-commercial and commercial (non-public) databases could also be connected with the integrated database system.
5) Links should be encouraged with genetics/genomics and disease-oriented research programs. Targeted research is also required. The study of disease-focused association and genetic diversity, conducted at the multi-population level, would provide the type of data needed to underpin full and correct analysis of many other datasets.
6) This workshop report should be used to provide guidance and encouragement for cooperation to scientists and institutions, international conferences, and funding agencies.
7) National and other funding agencies should promote local integration of distributed databases, support them, and encourage wider collaboration.
8) The European Research Framework Programme should provide opportunities for collaborative projects at European and world-wide levels via topics in calls for proposals for catalysing the creation of this network of databases that will be essential for entering the era of system biology.

APPENDIX 1. WORKSHOP TERMS OF REFERENCE

BACKGROUND: Two recent papers, an editorial in Nature Genetics (WayStation to HUGOBase [NG 37, Vol 37, 783]) and Patrinos and Brookes (DNA, diseases and databases: disastrously deficient [TIG 2005, Vol 21, 333]), highlighted the need for coordination of databases with a focus on human genetic variation and associated phenotype relationships. The deficit in this area has been repeatedly recognised, but so far, effective solutions have not been found. This is an obviously important area - a major motivation for funding molecular biology is its impact on human disease. Most diseases have at least some genetic component, which when identified, provides applications to biomedical research as well as furthering the cataloguing of human phenotype to genotype relationships. Undoubtedly a broad catalogue would be an immense boon to many areas of research - very directly to diagnostics and the ability to segregate disease populations into different classifications, but also into the general understanding of both human physiological ("healthy") and disease processes, which will provide important starting points for remedies. These remedies will often be equally applicable to both genetic and non-genetic forms of a disease (for example, the discovery of the molecular components responsible for genetic risk factors behind heart disease may well provide useful intervention points for all cardiovascular disease). Finally the current collection of human phenotype to genotype relationships, predominantly from Mendelian traits, have provided many fundamental insights into both mammalian and cellular biology. With the increased investment into the discovery of new genetic risk factors behind complex diseases in humans, this driving force for biology from the clinic to basic research is likely to increase. However, due to the lack of coordination of this information, many potential discoveries cannot be easily made.

Given the importance of this area, it is unsurprising that our main problem is not one of enthusiasm nor, globally, one of funding in the general area of human genetics. The papers list a number of data resources, all of which are credibly tackling parts of this problem. However, there is a massive deficit in coordination and in the focusing of resources, especially highlighted when trying to access and analyze the data for particular applications. This is best illustrated when the situation in Human is compared to other organisms; fly, worm, zebra fish and mouse all arguably have better developed phenotype to genotype datasets, which are better coordinated with other resources for that genome. Due to the large worldwide investment in experimental resources for human biology, the coordination problem is even more striking. This has been true for many other resources, from genome sequence through to cDNA clone resources and databases - human mutations are not unique in this regard - but these are only excuses for a lack of sensible coordination of this data. Like with the human genome project, we must endeavour to overcome personal, political, technical and other obstacles to coordination. We need to frame the problem of linking and unifying resources in several contexts:

1. Data Access and Analysis: The challenge posed by the Nature Genetics editorial is to construct a "HUGObase" database for human genetic variation, with universal access to all relevant data, standards for data acquisition and deposit, and powerful and productive access tools. However, there are currently a wide range of options: central databases, bioinformatics grid access, powerful data management tools, and analysis tool options including bioinformatics and systems approaches. The response to this central challenge depends crucially on data available, access possibilities, and the types of analysis and biological questions to be solved. Full association and linkage is needed with model organism data, where a much wider range of tools is available for genetic variation and phenotype linkage studies are possible.

Also in this area, care has to be taken about the ethical disclosure of personal information; **we will limit the discussion to being solely about information that can be freely shared without constraints**, and importantly this precludes releasing the coupling between an individual's phenotype and his or her genotype - even if anonymised, this data inherently ties an individual who could be identified to specific phenotypes. Instead we will focus on the common practice of releasing specific, single locus (or single base pair) genotype-phenotype correlations, in line with traditional published research. However, this information is likely, of course, to have originated from studies which have the full phenotype and genotype readings from a set of people - we wish to focus our coordination efforts on the summaries of this research which is commonly made public via publication. These considerations lead to a number of guiding principles for the scope of this workshop:

a) We will consider linking or unifying only those data which are already in the public domain, either in databases or publications

b) The content, structure and analysis possibilities of central and linked databases should be designed and analyzed against a number of defined research areas, which should also be specified at the workshop, as constituting a minimum specification and requirement for content and format.
c) We acknowledge that there is a wide range of data, which from a combination of medical, ethical, personal, legal and commercial considerations, which should not be included in this integrated range of data for a range of highly justifiable reasons, and we will not attempt to do so.
d) Having recognized the necessity to leave certain databases out of the public database, where possible we should take account of the existence of these other databases, and provide access and structure within the public database to facilitate the possibility of these private databases and activities making optimum use of the public databases for their own research and medical application purposes. For example, a medical patient database should be able to link to the public genetic variation database(s).
e) Standards for deposition in the database should be developed to encourage the maximum use and content of this public and open database, with all contributions based on use without restrictions.

2. Human SNP, Haplotype and QTL analysis: There has been incredible technological progress in genome-wide marker determination, in particular by SNPs. This is making fully powered genome-wide genetic association studies economically feasible. The recent discovery of genetic markers associated with a 6 fold increase of risk of age-related macular degeneration in the homozygote is an example of how fully powered association studies have progressed. Though it would be naive to think that all genetic associated diseases will be trivially discovered, it is clear that genetic association studies will become an important if not dominant discovery route for phenotype to genotype relationships (especially in humans, where knock-out research is unthinkable). Of course, although these SNP based assays are currently the most feasible approach for genome-wide association, this does not mean that we should only consider SNPs as the causative mutations. The whole spectrum of genetic mutations may be the causative genotype change; however we expect many of these will be found by virtue of them arising on specific haplotype backgrounds, and so amenable to genome wide association studies.

3. World wide Collaboration: Any work in this area must be seen in an international context, which means collaborations with both US and Japan but also with emerging scientifically rich countries such as China, Brazil and India. This international context requires sensible European coordination to provide a clear way to coordinate European resources with the worldwide community.

4. Action at European level: The contribution of EU member states in this area is already impressive and potentially can be world class. Many of the Locus specific databases that capture genotype-phenotype relationships for individual genes are maintained by European researchers. Perhaps more importantly some of the largest and most effective cohort based genetic studies are European member state supported, including Biobank (UK, MRC and Wellcome), Finish, Estonian and Icelandic cohorts, some of which have already yielded exciting new discoveries [Nature Genetics 2005 Feb:37(2):129-37], the Case Control Consortium (UK, Wellcome Trust), and major EU funded projects such as Genome EU Twin (www.genomeutwin.org). Europe also has strong bioinformatics capabilities in both genomic databases and analysis of genotype information. Finally, Europe has world class depth in mammalian research and in particular the new EU project, the European Conditional Mouse Mutagenesis Programme, which will provide easy shuttling between human phenotype-genotype relationships and a premiere mammalian experimental model, mouse. Europe therefore is already playing a strong scientific role, and there is clearly the ability to become a world leader in this area by a mixture of more focused resourcing, coupled with effective coordination of existing projects.

With this background, the EU plans to hold a workshop on 2-3 March 2006 with the overall goal of discussing the future roadmap for this area in Europe. This workshop will be an EU-focused workshop which can then give rise to more international workshops for coordination world-wide. There will be explicit recommendations made both for short and longer term plans in this area. To ensure this workshop is effective, it is important to have a well defined scientific scope. We have decided to focus on genotype-phenotype relationships, targeted at human genetic variation data, and strictly the result of biomedical research led genetics, as opposed to the broader remit of patient cohort management or clinical data. Obviously there is a complete continuum from clinical studies and cohort management through to focused molecular research on a particular gene; we will be focusing on the capture of data linking particular phenotypes to localised genotypes. Although many of the genetic association studies will be via SNP based markers, it is important to realise that the causative mutation need not be a SNP, and could indeed be rather complex genomic scenarios.

APPENDIX 2. BIOINFORMATICS GRID TECHNOLOGIES

(Note: Although not directly discussed, this appendix is provided as essential reference material)

Linking of databases can efficiently be accomplished by providing Bioinformatics Grid capabilities for data and computing, using established European and National Grid hardware and middleware networks (e.g. The GÉANT2 http://www.geant2.net and EGEE http://public.eu-egee.org grid projects funded by European Commission DG Information Society). This is the main objective of the EMBRACE project http://www.embracegrid.info , as the name suggests, is to embrace the European scientific community, to enable biomedical research in the 'omics' era. The EMBRACE final goal is to allow biomedical researchers to get answers to their questions using interfaces, graphics, and terminology with which they are familiar. The objective is to build programmatic interfaces to a range of biological databases. In Europe, we have different families of projects with different focuses. Grid infrastructures are developing low-level middleware and deploying resources. Grid middleware projects are developing high-level services. User oriented projects aim at using the grid for improved performances, i.e. Embrace. On top of the existing e-infrastructures in Europe, the plan is to deploy high-level services to serve the life sciences research community. Embrace aims at building a 'knowledge grid' allowing integrated exploitation of data collection, curation and provision of biomolecular information, and providing availability of most of the popular databases and software products, and tools and programming interfaces to exploit that information by taking away the need for maintaining local copies of databases and software.

The method adopted to solve these problems is to use standard grid technology to present each database through an applications programming interface (API), which can be used by any software to exploit it. The goal will be to exploit the developments of grid computing in the technical methods used. This will enable any group to develop software which accesses all of the databases in all of their richness. Many of the databases created, extended via TEMBLOR http://www.ebi.ac.uk/Information/funding/temblor.html and the FELICS Integrated Infrastructure Initiative, funded by the EU Research Infrastructures programme, will be further linked, greatly increasing their capabilities and utility for analysis.

However, aside from the benefits to relatively independent software developers like this, the APIs developed will also improve the efficiency and effectiveness of core service centres such as the EBI by allowing rapid development of systems of greater richness. While it is hoped that most typical user requirements will thus be satisfied, the open APIs to databases mean that, where there is a specialist need for custom software, it will be easy to produce. As Europe's largest server of biological databases, the EBI forms a natural focus for this content integration work package, however, a very explicit goal is to build a system which in no way depends on centralisation. A range of data resources from other partners in the project will also be integrated, and, aside from the valuable scientific content which that will bring, it will also ensure that the technology adopted can function in a distributed way.

The next step is interpreting the wide range of genomics data. A major project in this field is the EU network BioSapiens http://www.biosapiens.info, in which 25 prominent European bioinformatics institutes work together under Commission funding to combine data to answer questions from the biomedical community. The new EU systems biology network of Excellence called ENFIN http://www.ebi.ac.uk/Information/funding/temblor.html will further extend these capabilities, and provide toolboxes for lab experimentalists to integrate into the network data capabilities.

APPENDIX 3. WORKSHOP PARTICIPANTS

Stylianos E Antonarakis MD, DSc; Professor and Chairman,; Department of Genetic Medicine and Development; University of Geneva Medical School,; and University Hospitals of Geneva; 1 rue Michel-Servet; 1211 Geneva, Switzerland; Stylianos.Antonarakis@medecine.unige.ch; http://medgen.unige.ch/; http://www.frontiers-in-genetics.org

Dr. Ewan Birney; ENSEMBL Group Leader, EMBL Outstation - Hinxton, European Bioinformatics Institute,; Wellcome Trust Genome Campus; Hinxton, Cambridge, CB10 1SD; United Kingdom; birney@ebi.ac.uk

Professor Anthony J Brookes; Department of Genetics; University of Leicester; University Road; Leicester, LE1 7RH, UK; ajb97@leicester.ac.uk

Professor Lon Cardon; Wellcome Trust Centre for Human Genetics; University of Oxford; Oxford OX3 7BN; United Kingdom; lon.cardon@well.ox.ac.uk

Professor David N. Cooper; Curator, HGMD; Professor of Human Molecular Genetics; Institute of Medical Genetics; Cardiff University; Heath Park; Cardiff, CF14 4XN; United Kingdom; cooperdn@cardiff.ac.uk

Dr Johan T. den Dunnen; Human and Clinical Genetics; Leiden University Medical Center; Einthovenweg 20, 2333 LEIDEN; Nederland; ddunnen@HumGen.nl

Dr. Gudmundur Einarsson (observer), DeCode Genetics; Sturlugata 8; IS-101 Reykjavik; Iceland

Dr. Hakon Gudbjartsson (observer); DeCode Genetics; Sturlugata 8; IS-101 Reykjavik; Iceland

Dr. Simon Heath; Head of statistical genetics; Centre National de Génotypage (CNG); 2 rue Gaston Crémieux - CP 5721; 91057 Evry Cedex; France; heath@cng.fr

Dr. Karen Kennedy; Science Programme Manager (Molecules, Genes and Cells); Wellcome Trust; Gibbs Building, 215 Euston Road; London NW1 2BE; United Kingdom; k.kennedy@wellcome.ac.uk

Prof. Dr. Hans Lehrach, Director; Department of Vertebrate Genomics; Max-Planck-Institute for Molecular Genetics; Ihnestraße 63-73; 14195 Berlin; Germany; lehrach@molgen.mpg.de

Dr. George P. Patrinos (observer); Erasmus University Medical Center Rotterdam; Faculty of Medicine and Health Sciences; MGC-Department of Cell Biology and Genetics; P.O. Box 1738, 3000 DR, Rotterdam; The Netherlands; g.patrinos@erasmusmc.nl

Prof. Dr. Stefan Schreiber, Director Institute for Clinical Molecular Biology and Department of General Internal Medicine; Christian-Albrechts-University/University Hospital Schleswig-Holstein; Schittenhelmstrasse 12; 24105 Kiel; Germany; s.schreiber@mucosa.de

Dr. Kári Stefánsson, CEO; DeCode Genetics; Sturlugata 8; IS-101 Reykjavik; Iceland; kari.stefansson@decode.is

Professor Johan van der Lei; Erasmus MC, University Medical Center Rotterdam; Department of Medical Informatics; Dr. Molewaterplein 50; 3015 GE Rotterdam; The Netherlands; j.vanderlei@erasmusmc.nl; http://www.eur.nl/fgg/mi

Prof dr G.J.B. van Ommen; Head, Dept. of Human Genetics http://www.humgen.nl,; Center of Human and Clinical Genetics,; Center for Medical Systems Biology http://www.cmsb.nl; Leiden University Medical Center; Netherlands; gjvo@lumc.nl

Prof. Dr. Edgar Wingender; Director; Department of Bioinformatics; Center of Informatics, Statistics and Epidemiology; Universität Göttingen - Bereich Humanmedizin; Robert-Koch-Str. 40; 37075 Göttingen; Germany; edgar.wingender@.bioinf.med.uni-goettingen.de

Dr. Richard Wooster; Curator COSMIC Cancer Database; Co-leader Cancer Genome Project; Wellcome Trust Sanger Institute; Wellcome Trust Genome Campus; Hinxton, Cambridge, CB10 1SA; United Kingdom; rw1@sanger.ac.uk

EUROPEAN COMMISSION ORGANIZERS

Frederick Marcus, D.Phil., Principal Scientific Officer; European Commission - Research Directorate General, Directorate F: Health Research, Unit F.4: Fundamental Genomics, Bioinformatics; E-mail: frederick.marcus@cec.eu.int; Unit Website: http://www.cordis.lu/lifescihealth/genomics/home.htm; Mailing Address: European Commission (Office - CDMA 2/157), B-1049 Brussels, Belgium; Physical Address: 21 Rue du Champs de Mars, Brussels, Belgium

Bernard Mulligan, Ph.D., Acting Head of unit F.4, Fundamental Genomics; European Commission - Research Directorate General, Directorate F: Health Research, Unit; E-mail: bernard.mulligan@cec.eu.int Unit Website: http://www.cordis.lu/lifescihealth/genomics/home.htm; Mailing Address: European Commission (Office -CDMA 2/152), B-1049 Brussels, Belgium; Physical Address: 21 Rue du Champs de Mars, Brussels, Belgium.

APPENDIX 4. PRESENTATIONS

Session 1 - CURRENT STATE OF THE ART IN DATABASES AND ANALYSIS TOOLS

Dr. Ewan Birney, EMBL Outstation - Hinxton, European Bioinformatics Institute - RAPPORTEUR

SESSION 1 SUMMARY

Session 1 showed the current situation of databases and analysis tools. The main points were as follows:

Ewan Birney (Ensembl) - Dr Birney expressed frustration that Human had the least complete and integrated genotype-phenotype resource compared to other well studied organisms such as Fruitfly and Worm. He expressed optimism however that this need not continue. He present the Ensembl Variation resource which handle variation data for vertebrate genomes, including Human. This resource scales well, can handle large-scale resequencing information and is presented intuitively to the user through the Ensembl web site. Ensembl can also handle structural polymorphism in particular as it alters the reference genome, and indeed human has a less challenging scenario than some other organisms, e.g., *Anopheles gambiae*.

Lon Cardon (Oxford) - Dr Cardon presented an overview of how large-scale association studies are maturing in both the UK and the US, with other studies expected to come on-line. These studies are reasonably well powered in both case/control numbers (around 1,000 cases and pooled set of 3,000 controls) and will be genome-wide (currently using the 500k Affymetrix SNP panel). Control data will be made public very shortly for the UK Case Control Consortium (CCC). Case data will be made available to bona-fide researchers, but this is likely to be even more permissive in the US. This means that there will be large amounts of publicly available association study data at the raw level.

There was lively support for the openness of both control and case data. One clearly useful idea would be to extend open control population studies across Europe to provide reassurance that using "European" controls for association would be valid (or if not, to obtain appropriate controls allele frequencies). Simon Heath expressed the desire to have a European wide coordination on data controls to encourage appropriate data sharing (mindful of ethical concerns).

Dr Johan T. den Dunnen - Dr den Dunnen outlined his experiences in both running and providing software for Locus Specific Databases (LSDBs). These databases, often focused on Mendelian diseases are sometimes forgotten about but they are a rich source of genotype-phenotype information. Dr den Dunnen showed how there was considerable interest in the use of LSDB information, but far less interest in populating LSDB information. In the ensuing discussion this imbalance was discussed; clearly one would like to take advantage of the intense interest some individuals have but it was also important to have consistency of information.

Dr Kári Stefánsson - Dr Stefansson outlined the impressive DeCode genetics resource in Iceland. This combines phenotype, genotype and pedigree information to provide a rich source of hypothesis generating information. DeCode genetics treats this database as a source for hypotheses but then confirms them in other populations. It already has enough different disease areas which it is looking at to keep DeCode fully occupied, but there is potentially many more interesting discoveries to make. Dr Stefansson was very interested in finding a way to put this information in the public domain whilst making this a sensible proposition for DeCode. He was also worried about the potential for excessive data mining on one resource and noted that this multiple testing problem was complex when a large database is repeatedly used for hypothesis generation. The discussion focused on the details of DeCode processes, which potentially many other groups could learn from and the issues around multiple testing.

Dr Richard Wooster - Dr Wooster presented COSMIC, a genotype-phenotype database focused on somatic mutations found in cancers. This database was driven by an in-house need, but has become a useful resource in it's own right. At first the COSMIC curators focused on capturing a broad range of information from the literature but this contracted to only a few key items (the mutation and the cancer type being the two main features). COSMIC integrates well with Ensembl, and is becoming more "large-scale" (one can't just download as an excel spreadsheet any more). It is clear, due to the increased Cancer genome projects both in the UK and the US, that this information will continue to be generated and will be a useful source of genotype to phenotype.

Dr. Ewan Birney, EMBL Outstation - Hinxton, European Bioinformatics Institute

Ensembl and Variation

ENSEMBL http://www.ensembl.org is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. It already handles a wide variety of variation data. One of the most frustrating aspects however is the lack of good genotype-phenotype information in Human compared to other model organisms, such as Drosophila and C.elegans.

The main points in Ensembl variation are as follows:

- Variations can be SNPs, and "reasonable" indels
- Can handle multiple sources
- Can handle genotypes in multiple populations
- Can handle both hetreozygote (Human), Strain based (Mice and Rat) scenarios
- Way of handling structurally different parts of the genome
  - MHC, CYPD26
  - PAR (Pseudoautomsomal) regions also handled this way
- Human genome not the worst case
  - Anopheles and Ciona far, far worse

There is a wide range of SNP "consequences" on Human Variations:

- 10 million variants
  - 88,676 Non synonymous (0.8%)
  - 76,021 Synonymous (0.7%)
  - 124,289 UTR (1.1%)
  - 16,588 Regulatory (0.1%)
  - 1599 GT/AG splice changes (0.01%)
  - 1729 Stop gains (0.01%)
  - ... other classes (frame shift)

Strain data in is available in TranscriptSNPView and BioMart .

ENSEMBL is very flexible, allowing for:

- Stable and sensible handling of variation
- Ability to handle larger genome polymorphisms
- Ability to handle resequencing data
- Ability to scale to 1,000 people, millions of genotypes
- Integration with functional data
- Future integration with comparative genomics
  - Close primates
  - Distant "conserved regions"
  - Protein mappings

However, ENSEMBL does not have good genotype-phenotype relationships:

- OMIM is v. difficult to handle
- COSMIC is great, but only covers Cancer
- Uniprot focus on protein coding cases
- HGMD do not allow integration and then redistribution of their data

A wish list would include:

- An open genotype-phenotype database
  - Starts with good simple goals - tracking variants which change phenotype
  - Is open (does not rule out industry funding, but must be open to be integrated)
  - Could extend beyond the simple things

Prof. Lon Cardon, Wellcome Trust Centre for Human Genetics, University of Oxford

Complex Disease Genetics, Analysis Strategies & Tools (or lack thereof)

In the past, genetic studies of complex diseases not met anticipated success.

Some Human Association Studies statistics:

Pubmed: 12 Feb 2006. "Genetic association" gives 36,561 hits ; ~1% confirmed (unique) loci for complex traits; Claims of "replicated genetic association" → 377 (1%); Claims of "validated genetic association" → 151 (0.4%) ; However, at present there is a renewed promise for disease gene finding.

Whole Genome Association

- Objective: screen entire human genome at high density in population samples (e.g case/control)
- Currently main focus area of human complex trait genetics research

Upcoming studies:; 100,000's genetic markers (high-throughput); 1000s individuals (but not 10,000s) ; High cost (€1,000,000s); Whole Genome Association

Wellcome Trust Case-Control Consortium, Data Release

- All RAW genotype data to be put in public domain
  - Includes raw intensity files and called genotypes (CHP, CEL, DAT)
- 1958 BC data completed April 2006
  - Accessible by scientific community immediately

Disease data completed by end of 2006; Genotypes & phenotypes (disease +/- only); Released asap, ≤ 6 months following QC/analysis; Future (Need): Powerful, population-based replication studies

PRESENT

- WGA will yield novel genes, and they will need to be validated
- Validation will involve 10-100s genetic markers, but very large samples (1000s → 10,000s)

PROBLEM

- Long time-lag from initial finding to replication
  - But value of replication >> initial study

FUTURE NEEDS

- Need samples prepared/stored and ready to rapidly assess
- Need tools to prioritise marker selection, data analysis, results

Critical Statistical Issues in Complex Disease Gene-Identification

- Raw Genotype/Phenotype Data ; Prediction: V. large sets of raw genotypes and (simple) phenotypes will be publicly available over next 2-3 years ; WTCCC, NIH, Pharma already pushing ; Journals have problems addressing 'real' associations – beginning to see need for raw data (cf expression array studies) ; Databases for this type of information do not exist ; Immediate need at least as great for 'raw' data as for 'derived' dbs of genotype-phenotype correlations; 'derived' information currently very poor quality – will soon get worse ; Db of raw genotypes/phenotypes complements multi-disciplinary dbs (sequence, mouse, expression and pathways) ; Enables testing new hypotheses, dev methods.

Raw genotype/phenotype data: what to do today?

- Develop strategy to harmonize raw data collation/annotation
- Build tools to store, disseminate the data
  - Non-trivial, some files for single study are 10s Tb
- Encourage development of analysis tools
  - Statistical genetics **far** behind bioinfo in standardized tools

Overall genotype/phenotype data: what to do today?

Many aspects to consider

- Different genetic sources in humans & model organisms
- Different types of data
- Raw vs derived data
- Need to remember that genotype-phenotype data is not equally useful
  - Need compilations of all associations (high information, low value)
  - Also need compilations of 'real' or 'validated' associations (cf dbSNP) (low info, high value)

What questions are we trying to answer?

- Prioritisation? Interpretation? Hypothesis generation v confirmation?
- Immediate questions not necessarily same as those in even 1-2 yrs.
- No single solution fits all needs; Requires > 1 country; Requires > 1 funding source

Dr Johan T. den Dunnen, Human and Clinical Genetics, Leiden University Medical Center, Einthovenweg 20, 2333 LEIDEN, Nederland http://www.HumGen.nl

Leiden Muscular Dystrophy pages - Human & Clinical Genetics

I will focus on the most urgent needs from the viewpoint of a scientist close to the patient, in relation with DNA based clinical diagnosis. The most likely starting point for this person is then a Locus-Specific DataBase (when available) or OMIM (using the disease as entry). Second choice start points, when no LSDB exists, will be OMIM or HGMD. Using a genome-browser, this person will get lost immediately.
The needs then are;

* a reliable LSDB, with links to other (genome) information
* an up-to-date and 100% complete list of gene variants
* a Reference Sequence showing nucleotide numbering for the gene
* a field with information regarding the reported pathogenicity of that variant
* a reference to the source of the information (to enable contacting the person/lab for more information)
* any other tools / links connected to his subject of interest

So the route to follow would be LSDB > genome (Ensembl) > many links (a.o. OMIM, HGMD). As discussed, OMIM and HGMD are important but neither collect ALL mutations in the disease gene of interest. What should be promoted in this respect is;
1) a centralised starting point for LSDBs
2) one LSDB for every disease gene (e.g. using LOVD); tight links to HGMD, WayStation, OMIM and genome viewers (e.g. Ensembl) to ensure timely updates of all information, to reduce undesired copying (double work) and to improve links to all other information. The LSDB should be a submission point for new entries. Ways should be found to promote that information is submitted (and this should somehow be rewarded), incl. as much phenotypic information as possible. Ultimately a connection to Quality Assurance schemes can be envisaged.
3) easy to work with software to determine whether a variant found is "pathogenic or not" to help the scientist to draw the right conclusion based on all available knowledge and prediction tools.

Genome variation of the human population. Promote the generation of a database where all information regarding Copy Number Variation is stored (in patients and normal). Promote studies to catalogue all this variability. Currently there is no such database and clinical diagnostic labs struggle with all information coming out of genome-wide CNV studies in relation to genetic disease; which changes to follow up ?. Regarding 'non-pathogenic', since clinicians are paid to focus on pathogenic, an effort to analyze a large set of controls to catalogue the non-pathogenic variation would be very worthwhile.

Genotype to phenotype

I support the idea of S. Antonarakis to go step by step and promote studies that go from DNA to RNA and not directly from DNA to phenotype. Such studies should focus on gene expression in relation to haplotypes and/or in relation to duplicating and/or deleting genomic regions. For this it is not always necessary to go to animal models. In the format of patient-derived cell lines the world has collected an enormous resource for such studies (these should be collected and analyzed in a European project).

* Data quality - often discussed but I think with too much emphasis. When more and more data come in it will be easy to separate trustworthy from junk data (NB the first few years the sequences in GenBank contained many mistakes, ...but were very helpful).
* Publication - discussed extensively. Connected to e.g. the LSDB promote the electronic publication of all work linked to the gene of interest. In this electronic age the web is the easiest place to publish work, incl. negative studies (and nearly for free). Using search engines like Google the work will be found, a scientific journal is not necessary for this. Placing it all on a centralized place will of course help to find relevant information.

Dr. Kári Stefánsson, CEO, DeCode Genetics

Phenome/Genome Association Databases - How and Why

http://www.decode.com

Population genetic research, Information Systems & Data Protection - deCODE's informatics systems

- Electronic Data Capture – Questor
- Patient privacy - IPS
- Workflow systems
  - LIMS
  - Automatic allele caller (DAC) for microsatelites
- Downstream analysis
  - DiseaseMiner
    - Phenotype modelling
    - Familial clustering
    - Genetic analysis – Allegro - NEMO
      - Multipoint linkage analysis
      - Association and haplotype analysis
      - Linkage disequilibrium
  - SequenceMiner
    - Mutation analysis
    - SNP & InDel detection

From Genes to Drugs and Population genetic research

MI: The Leukotriene Pathway

MI: HapK risk in African Americans compared to other risk factors for heart attack

Dr. Richard Wooster, Curator COSMIC Cancer Database, Co-leader Cancer Genome Project, Wellcome Trust Sanger Institute

The COSMIC database and web site

Catalogue Of Somatic Mutations In Cancer
Known cancer genes in the human genome
www.sanger.ac.uk/genetics/CGP/Census

The literature contains small intragenic somatic mutation data on over 200,000 tumours.

Example was shown of the response to gefitinib in a patient with refractory non–small-cell lung cancer and a somatic mutation in EGFR.

==

WHY DEVELOP COSMIC

- There was no other similar resource
- To preserve somatic mutation data
- To share somatic mutation data
- To standardise genotype and phenotype information
- To integrate published data with the output of the Cancer Genome Project

Data types captured in COSMIC

- Individual, Tumour, Gene, Location

The data on the web site is available in multiple formats;

- Export function from the web site for individual genes (HTML, csv, Excel)
- ftp.sanger.ac.uk csv and Excel for individual genes
- Oracle export of the whole database

Integration with Ensembl

- COSMIC DAS track for Ensembl

WHAT IS AVAILABLE IN COSMIC?
Statistics for Feb 2006 release

Experiments	228,669
Tumours	142,569
Mutant samples	25,176
Mutations	26,194
Papers curated	3,013
Genes	1,035

SUMMARY

- COSMIC holds data for small intragenic somatic mutations in cancer
- Approximately half of the available published data has been curated (excluding TP53) – there is plenty left to do
- Unpublished data is being submitted by the Cancer Genome Project at the Sanger Institute with the possibility of data from the NCI
- The COSMIC web site gives access to the data

Prof. Stylianos E. Antonarakis, Chairman, Department of Genetic Medicine and Development, University of Geneva Medical School - RAPPORTEUR

SESSION 2 SUMMARY

This session took place in the afternoon of March 2, 2006. The main points of the speakers' presentations were:

1. S. E. Antonarakis emphasized the need to consider several levels of phenotypic variability. This ranges from the disease outcome at one extreme, to gene expression variation at the other. Many different intermediate phenotypes could be considered. Thus, the databases that provide risk links between genetic variation and phenotypic variation, require many different liability classes of phenotypic expression.
2. H. Lehrach argued that the organism "computes" its phenotype from its genotype given a specific environment. Thus all data resources from the genotype to the gene expression level and control, protein quantities, metabolic pathways, networks of metabolic communications are important parameters in the prediction and/or manifestation of a phenotype.
3. S. Schreiber used the example of inflammatory bowel disease to illustrate the issues of genetic predisposition and risk genotype factors. A particularly interesting point of his presentation was the notion that different variants of a given gene or other functional elements are associated with different phenotypes. He emphasized the needs for biobanks, standardisation of experimental approaches and phenotypic assessment.
4. G.-J. van Ommen elaborated on biobanks and problems of ownership. He explained the issues and controversies regarding property rights of tissues in biobanks, definition of interests and consequences of "commons" and "anticommons", the gene patent proliferation, international developments and examples from existing biobanks.
5. S. Heath presented the centre of large-scale genotyping in Paris, and the issues related to different allele frequencies of SNPs in European population. He suggested to establish a standard set of population samples from across Europe, to identify a set of markers which can capture most of the allele frequency variation, and to combine datasets for large association studies. He also emphasised the issues of combining data from different platforms, and data release.

S. E. Antonarakis comment: I strongly support the creation and maintenance of an international database that provides a comprehensive link between genotypic and phenotypic variation. Primary data, published and unpublished, as well as meta-analysis and summary statements, need to be included in this database.

Prof. Stylianos E. Antonarakis, Chairman, Department of Genetic Medicine and Development, University of Geneva Medical School
Stylianos.Antonarakis@medecine.unige.ch

Which phenotypes ?

Gene expression variation and genomic variation

- Is there variation of gene expression ?
- Is gene expression variation genetically determined ?
- What is the genomic variation that controls gene expression variation ?

Gene expression variation examples:

40 CEPH grandparents, LBCL, ReTi-qPCR
41 test genes, 7 normalisation genes,
6 RT/RNA, 6 PCR replicates
10 CEPH families = **135 individuals**
RNA from **lymphoblastoid cell lines**
eQTLs using
1700 **Microsatellites**
2600 **SNPs**
genomewide – publicly available

Heritability Results

19/25 genes showed significant heritability of expression level

eQTL mapping results

Significant eQTL for 9/19 genes selected because of significant heritability
The majority of eQTLs are in trans

Models are presented for the Pathogenesis of Down Syndrome

Gene Expression threshold hypothesis for the Pathogenesis of Down Syndrome

eQTL mapping results : Conclusions

Gene expression variation is a **common phenomenon.**
We could find an **eQTL for 9/19** genes with significant heritability .
The **majority of eQTLs are trans** !

Prof. Dr. Hans Lehrach, Director, Department of Vertebrate Genomics, Max-Planck-Institute for Molecular Genetics

Genome to Phenotype

Life is the translation of the information in the genome into the organism:
The organism 'computes' its phenotype from its genotype, given a specific environment

PyBioS – a modelling and simulation platform for cellular systems

Modelling and Data Resources
Data required for modelling

- Information about the substrates and products of a reaction and its stoichiometry
- Information about the catalysing enzyme and its kinetics
- Information about the reactants and enzyme concentrations

DATABASES

- KEGG (Kyoto Encyclopaedia of Genes and Genomes)
- Reactome
- Transpath (Database of signal transduction pathways)
- SRS (Sequence Retrieval System)
- BioCyc
- Kinetikon (Kinetics database)
- Database of experimental data (Expression data, Protein/Protein interaction, ...)

Automatically generated graph of the reaction network

Concentrations and fluxes of simulation results are visualized by the node color or inserted graphs

APPLICATIONS

Simulation studies of large metabolic networks of
-Trisomy 21 (Down syndrome)
-Skin Aging
-Listeria metabolism
Modelling and simulations of somitogenesis in mouse
Simulation of gene regulatory motifs for the development of reverse engineering strategies
Simulation procedure for large metabolic networks with unknown kinetic parameters

ESBIC-D (an EU Coordination Action)

ANNOTATION AND MODELLING OF PATHWAYS RELEVANT FOR CANCER, APOPTOSIS, CELL CYCLE, EGF RECEPTOR SIGNALING PATHWAY, TOOLS, REACTOME DATABASE AND ITS CURATOR TOOL, PyBioS

Prof. Dr. Stefan Schreiber, Director, Institute for Clinical Molecular Biology and Department of General Internal Medicine, Christian-Albrechts-Universität zu Kiel

Biobanks, Genetic Testing and Maintenance of Health

Inflammatory Bowel Disease

- No Mendelian Trait
- Polygenic Etiology
- Incomplete Penetrance due to Epigenetic Factors
- Crohn‘s disease, 15-50

Characteristics of Complex Disorders

- Syndromes rather than single diseases
- Definition of disease subgroups not possible with techniques of clinical phenotyping
- Diverse Pathophysiology
- Polygenic Etiology
- Differential response to targeted therapies

Complex Diseases and Conditions

- diabetes mellitus, metabolic syndrome; arterial hypertension, arteriosclerosis, coronary heart disease; hyperlipidemia, hyperhomocysteinemia; rheumatoid arthritis/osteoarthritis; depression/bipolar disease, schizophrenia; Alzheimer disease, dementia; multiple sclerosis; bronchial asthma, atopic eczema; sarcoidosis; psoriasis; periodontitis; malignant diseases; Crohn disease, ulcerative colitis ; longevity

Polygenic Aetiology

- Genetic Composition is Associated with Distinct Subphenotypes?
- NOD2 insertion mutation: Case-control
- Clinical Implications: Genotype and Phenotype

Necessities for Future Health research

- Outcome Research for individuals carrying risk genotypes
- Prospective follow up of entire populations
- Detection of sub clinical manifestations
- Genotype based prevention
- Clinical trials to establish procedures
- New algorithms for genotype balanced randomisation

Necessities for EU Developments

- Establish a network of population-representative biobanks that share elements of standardisation
- Establish a network of genotyping centres that are highly standardized
- Create accessibility for clinicians and clinical expertise
- Provide background genotype frequencies to clinical projects
- Create a repository for ALL genotypes generated with tags back into the originating (DNA) biobanks

From Genetic Etiology to Pathophysiology, Environment

- Understanding genetic etiology will further the understanding of complex diseases
- With some exceptions understanding of genetic etiology will NOT lead to new therapies and will have NO direct impact on the health of European populations with chronic diseases
- The coming challenge will be maintenance of health rather than cure of disease
- Molecular prevention will be the game and this most likely will involve nutrition based strategies
- Targeted prevention will be a new industry, without regulations but with many pre-requisites.
- Present scientific „players“ can not self organize to solve these questions

Necessities for EU Developments II

- Biobanking of consequences of environmental triggers or markers thereof
- Open network that allows clinicians to integrate not to merely hand over samples
- Internal governance by project officers
- Decentralized systems with centralized inventories and standard

Prof. Gert-Jan Van Ommen, Trustee, HUGO; Head, Dept. of Human Genetics, Leiden University Medical Center

Towards a Dutch Biobank

Biobanks and Ownership

Reference: Centre for Medical Systems BiologyStart Symposium, Leiden. September 3, 2004

by Jasper A. Bovenberg, Leiden University, Faculty of Law

Do individual donors have property rights in their tissue as such?

Do individual donors have property rights in their tissue collected in a Biobank?

What is property?

International applications

Unesco and HUGO declarations:
human genome is heritage of mankind
Biobanks are a global public good
Concept of "common good":
argument against commodification of human genome by the commercial sector
However, also an argument against commodification of tissue in Biobanks by individuals supplying the 'raw material'

National applications

Biobank initiatives in Iceland, Estonia and the UK propose following policy:
No individual property rights
No right to individual benefits
No individual right to control use of tissue: blanket consent

International developments

Trend towards blanket consent for research on Biobanks
Recent OECD workshop on large-scale human genetic databases strong voices:
Genetic information not necessarily different from other medical information
Caution against overregulation
Only context specific safeguards
Emphasis on benefits >> threats

Conclusion

There have been calls from some people for the creation of property rights in human tissue.
However, property rights in human tissue are problematic
Property rights in Biobanks (de novo and existing repositories) may lead to an anticommons:
- Suboptimal research use
- Suboptimal valorisation

(Inter)national policies are needed for tissue commons.

Dr. Simon Heath, Team Leader, Statistical Genetics, Centre National de Genotypage (CNG)

Issues relating to combining studies

Mainly concerned with disease-genotype studies

High throughput genotyping

Different platforms

Varying populations and species

Internal and (mainly) external projects

Internal and external funding

Disease Studies

Linkage screening (6000 SNP panels)

Fine mapping (custom 1536 SNP panels)

Whole genome association (10k -> 500k SNP panels)

Often have multiple platforms used for the same project

Various issues with combining data

Combining sources

Potential for very large-scale epidemiological studies (multi-centre studies)

Unidentified population level variation (or other forms of batch level variation) could lead to biased analyses

Ethnic/batch information not always that useful (missing/unreliable/insufficient)

Allele frequency variation in Europe

Most European populations commonly considered to be homogenous

Study of 96 'neutral' SNPs showed surprising variation between populations

Important implications for combining association studies

Diabetic Nephropathy

Large case control study

~3000 individuals

French, Danish and Finnish samples

Candidate gene study

Concerns about population differences(Large number of differences between populations)

Handling frequency variation

Genotype samples at a number of independent markers

Assign samples to groups using genotype information

Correct for bias due to stratification (genomic controls)

Standard sample and marker sets?

Useful to have a standard set of population samples from across Europe

Identify a set of markers which can capture most of the allele frequency variation

Great help in combining datasets for very large-scale association studies

Combining data from different platforms

CNG produces SNP genotypes using:; Affymetrix; Illumina (Infinium/Golden Gate); SNPlex; TaqMan; Mass Spectrometry, Sequencing

Problems with cross-platform data

Different file formats; Matching alleles/markers; Pooled/individual DNA; Platform specific biases; Project specific reliability; Unbalanced patterns of data availability

Data Release

Production of genotype data from many sources at CNG (internal & external)

Considering a policy of data release for genotypes generated at CNG

i.e., data generated from publicly funded studies must be released

Questions of timing/ethics

Data Release

Useful to have a common set of guidelines for data release across Europe

Possibly this will be driven by events in US

Greatly increase the possibilities for combining studies - maybe even have enough data for reliable association studies.

Session 3 - TOWARDS A GENETIC VARIATION PUBLIC DATABASE - CONTENTS, STRUCTURE, STANDARDS, RESOURCES AND REQUIREMENTS FOR REALISATION

Prof. David Cooper, Human Molecular Genetics, Institute of Medical Genetics, Cardiff University - RAPPORTEUR

SESSION 3 SUMMARY

Key Biology and Medical Questions

- Mutant genotype → clinical phenotype → 'evidence-based genetic medicine'
- Clinical phenotype is different from laboratory phenotype.
- Human bearing inherited mutations can be viewed as 'transgenic' in some sense. It is our task to interpret 'nature's experiments' for the benefit of future patients with these conditions.
- Mutations may be considered part of a spectrum of genetic variation from (i) differences between orthologous genes to (ii) neutral polymorphisms to (iii) functional and disease/associated polymorphisms to (iv) pathological lesions (inherited) and (v) pathological lesions (somatic). Underlying causes of mutagenesis similar.
- Studies of mutational spectra important for optimisation of mutation screening techniques and for the identification of the specific mutation-inducing characteristics of exogenous mutagens. An appreciation of the background endogenous mutational spectrum is essential here.

Databases/data ready to integrate

- Identification of new required databases important.
- Must find way to incorporate databases from public and private sources without undermining the sustainability of the latter.

Contents / Structure / Standards

- Serious attention should be paid to data quality issues, particularly for unpublished data.
- Funding should be made available specifically for database initiative in terms of the commissioning of new databases, the support of existing databases, and the linking together of those databases deemed to be both useful and important.

Prof. David Cooper, Human Molecular Genetics, Institute of Medical Genetics, Cardiff University

The Human Gene Mutation Database (HGMD)

Human Gene Mutation Database http://www.hgmd.org represents a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease. Data catalogued include single base-pair substitutions in coding, regulatory and splicing relevant regions, micro-deletions and micro-insertions, indels, triplet repeat expansions, gross deletions, insertions and duplications, and complex rearrangements. Human gene mutation is an inherently non-random process, with the nature of the mutational lesion, its frequency and location all being strongly influenced by the local DNA sequence context. HGMD may be exploited to study the role of the local DNA sequence environment (e.g. repetitive sequence elements, sequence homologies and specific motifs) in mediating mutational events and to explore the nature of the underlying mechanisms. HGMD provides the only comprehensive collection of data on human gene mutations causing inherited disease and as such provides a key means of linking mutant genotypes to clinical phenotypes. Since functional SNPs with or without known disease relevance are also included, there is already a natural bridge between the pathological mutations in HGMD and the predominantly neutral SNPs catalogued in other databases. HGMD is already developing interrogatory search initially for distribution on a commercial basis. It is however possible to make these freely available to the academic community once a sustainable funding model is in place which guarantees the secure future for the database resource. Ideal sustainable model for the future that we are working toward: public/private mixed funding potentiating free access to academic users and subscription-based distribution for commercial users marketed by a commercial company.

Data Submission

- Provided in collaboration with the journal Human Genetics and its publishers, Springer-Verlag (Heidelberg & Berlin)
- Mutation data submitted online via the Springer website (http://www.springer.com/HGMD)
- Checked for originality, accuracy and uniformity
- After publication, the mutation data are transmitted to Cardiff for inclusion in HGMD

Future Developments

- Hypertext links to papers describing functional analysis of mutations, prevalence, origin and penetrance
- Recording of different clinical phenotypes associated with the same mutation
- Inclusion of study data (for disease-associated SNPs) including cohort size, statistical tests and level of significance obtained
- Inclusion of haplotype data from disease association studies

Future Developments

- Links to genomic reference sequences
- Improved mutation maps
- Links to available protein structures and homology models
- Links to orthologous gene sequences
- Development of computational tools to analyze mutations
- Development of advanced database search software

Financial Sustainability

- HGMD has never received any public funding
- Commercial partner: BIOBASE GmbH
- Subscription model: exclusivity period/search tool provision
- HGMD & BIOBASE are committed to moving toward free access to the academic community, potentiated by a ‘mixed economy’ of public/private funding

Prof. Dr. Edgar Wingender, Director, Dept. of Bioinformatics, University of Göttingen / Medical School, Goldschmidtstr. 1, D-37077 Göttingen, Germany.
BIOBASE GmbH, Halchtersche Str. 33, D-37077 Wolfenbüttel, Germany

A set of databases for systems biology

A set of databases have been developed, ranging from different aspects of gene regulation (TRANSFAC, TRANSCompel, TRANSPro, TiProD) through protein-protein interactions of whole proteomes (e.g., HumanPSD) and signal transduction (TRANSPATH) up to intercellular, specifically endocrine, signalling networks (EndoNet). They are complemented by databases on pathologically relevant mutations of genes encoding regulatory proteins (PathoDB, PathoSign) and disease-involvement of human proteins in general (HumanPSD/Disease Reports). Altogether they build up an information infrastructure that will prove useful for projects that aim at linking genotype data with molecular and clinical phenotype information from a systems biology perspective.

Some of these databases have been developed as research tools in publicly funded projects or with institutional budgets, where some are still maintained whereas others have been transferred for commercial exploitation. Other resources have been generated as purely commercial products, nevertheless providing some of their contents and services free of charge for users from non-profit entities. A range of mechanisms to balance the interests of academic researchers with those of commercial vendors has been implemented and proven operational.

To render data on human genetic variations instrumental for systems biological approaches, appropriate links to network/pathway databases and quantitative data on the effects of these variations would be required.

Integrative vs. Systems Biology

REQUIREMENTS FOR INTEGRATIVE BIOLOGY

- Mapping variant genes / gene products onto the relevant pathways and networks
- requires: (links to) network database(s)
- Explaining known and infer potential effects from network topology
- requires: (links to) network database(s)
- Simulate the dynamics
- requires: quantitative data

INTEGRATIVE DATA MODELS

- Specific for each range of objects / relations
- Unifying view on the distinct objects / relations
- Standards
- Cooperation models between proprietary and public databases

Prof. Anthony J Brookes, Department of Genetics, University of Leicester

Genotype-Phenotype Databases: Challenges and Solutions - HGVbase

Given the infinite scale and diversity of the subject area (genotype-phenotype connections) there will be a need for many 'data warehouses' that summary information and discoveries, probably each with different domains of interest or focus (e.g., cancer, cardiovascular disease, published studies, specific populations, pharmacogenomics,...)

These many needed/emerging 'association databases' will range from large to small enterprises, with academic through to commercial models, but a glorious future will depend upon these systems being interconnected for transparent cross-database searching. That will be helped greatly by today emphasising standard and tools to that end (e.g., the Ensembl Biomart tool, and the PML data model). Issues explicitly elaborated upon as most urgent included:
* Standardisation / DB integration *
* The 'phenotype data' challenge * (how to represent phenotypes)
* Handling association data * (software tools for data generators)
* Publication bias * (need to bring in all study findings!)
To address the key issues as we see them, we are working on:
* PML * (a standard model for DNA variation (& soon phenotypes))
* Copy-Number Variation * (major genetic effects, complex informatics)
* a generic phenotype data model * (EAV based, emphasising 'method')
* HGVbase-G2P * (as a prototype genetic association database)
* GenoScore * (convenient database application for genotyping labs)
* Phenobase * (data submission tool for genotype-phenotype data)

I proposed that journals and databases will merge in the future, to be replaced totally by 'database journals', wherein all science gets immediately deposited into internet accessible structured depositories, all of which will be interconnected into a 'web-web'. We could lead the way on this with the genotype-phenotype challenge, and on the way, thereby, produce a non-profit model for sustainable funding for such an entity.

HGVbase-G2P - will focus on all forms of variation and any association studies that connect such variants to any phenotype. This will include published and (mostly!) unpublished data! In practice, we will mainly capture genetic association evidence between DNA variants and complex disease, where the genetic component alters risk but does not cause the disease (i.e., is neither necessary nor sufficient to account for the observed phenotype). We will also capture environment data - the other major contributor of complex disease causation. Our records will be extensive in scope, carrying detailed phenotype, sample, population, assay, genotype, allele, haplotype, marker, and sequence data, along with concluded p-values for disease associations (single point effects and synergistic interactions) plus citations, free text, and key word information. Genome annotation (e.g., exon, CDS, splice sites, repeats) will also be available for guiding database searches, along with a range of submitter information.

For Comparison, HGMD - summarizes *published* mutations that *cause* disease, and only collects the first report of each change. Records are very limited in content - typically just DNA sequence and name of disease, plus a citation! So they really are quite different project - even though they both focus upon genotype-phenotype relationships.

There will be a need for many such databases to capture the full spectrum and scale of association studies being conducted globally. It is critical that we do capture most/all of this otherwise we cannot distinguish true form false (chance) positives, but this will never happen if we rely on published findings alone! Hence, we need databases like HGVbase-G2P that gather primary data (some with specific focuses such as cancer, nations, pathways), plus a network of interactions between these databases so they can all be searched transparently from any one of many database interfaces. For this we need to harmonise their designs and provide standards as soon as possible in the evolutions of this field. Central browsers such as Ensembl will only be able to (and should only) include summary level information from these many 'association databases' (i.e., markers, phenotype name, and p-values)- for presentation graphically with links back to these association databases.

Prof. Johan van der Lei, Erasmus MC, University Medical Center Rotterdam, Department of Medical Informatics

Databases for Knowledge Discovery

Biomedicine and Health Sciences

Basic Research - experiments
Clinical care - patients
Health Research - populations

Discovery of new scientific knowledge from large databases of measurements, observations and interpretations

Population-based research:
Prospective

Prospective longitudinal database
10,000 persons > 55 years of age
relationships between risks and diseases
cardiovascular and vessel-wall diseases, glaucoma
neurologic diseases (Alzheimer), osteoporosis

- Expensive
- Requires freezers
- Fight over time and blood
- "Easy" ICT structure: patient centered
- "Easy" to combine
- Etiology (and therefore therapy) of common diseases
- If not: out of luck

For time's sake …We'll ignore RCTs

phenotype is intentionality-driven (consequences)
incomplete/inaccurate date
hypothesis-specific manual curation
patient privacy and data security
source population
not database but communication network

Next:

- Randomised naturalistic (pragmatic) trials
  - EPR does inclusion/randomisation
  - EPR used for naturalistic follow-up
  - Patient as source of data
- "patient as biobank"
  - Nonresponders (overresponders)
- Patient as "accidental experiment"
  - Children
  - Off-label use
  - Pathway-driven

Requirements??

- "the" phenotype does not exist
- Clinical data is process / context specific
- "Quality of data" is also context specific
- Hypothesis generation and verification
- Significance and consequence
- "Divide and conquer"

Dr. Karen Kennedy, Science Programme Manager (Molecules, Genes and Cells), Wellcome Trust

Wellcome Trust Activities in Data Sharing, Databases & Human Genetic Variation Research

Wellcome Trust Activities in Data Sharing & Databases

- Support for major initiatives to generate large-scale datasets for the research community. E.g., the genome sequencing projects, the SNP Consortium, the International HapMap Project and the Structural Genomics Consortium.
- Funding to support the development and maintenance of a number of major biological databases for the research community including funding via the Functional Genomics Development Initiative and funding for the WTSI. E.g, Ensembl, the European Macromolecular Structures Database, GeneDB, FlyMine.
- Worked with other funders and the research community to advance the policy debate on data sharing and facilitate the development of best practice standards. E.g, co-sponsored a workshop in Fort Lauderdale in January 2003 that developed a statement on sharing of pre-publication sharing of data.

Wellcome Trust Activities in Human Genetic Variation Research

PROJECTS

Human Genome Project; The SNP Consortium – characterisation of ~1.8M SNPs; HapMap Project – genotyping of ~4M common SNPs in 269 samples; WTSI resequencing projects ; WT Case Control Consortium – genotyping of ~ 700K SNPs in 2K cases for 8 diseases & 3K common controls ; Cancer Genome Project – cataloguing somatic mutations in human cancers ; Genomic Structural Variation – characterising CNVs in HapMap & disease samples.

ASSOCIATED DATABASES

Trace Archive, Ensembl & Vega http://www.ensembl.org http://vega.sanger.ac.uk ; dbSNP http://www.ncbi.nlm.nih.gov/ ; projects/SNP/ ; http://www.hapmap.org/ ; GLOVAR http://www.glovar.org/ ; WTCCC dedicated project database http://www.wtccc.org.uk/ ; COSMIC http://www.sanger.ac.uk/genetics/ ; CGP/cosmic/ ; DECIPHER http://www.sanger.ac.uk/ ; PostGenomics/decipher/

OTHER ACTIVITIES: collections & cohorts, e.g. ALSPAC, 1958 BC, Biobank

Issues for Consideration

- Terms of Access – 'open' or managed:
  - Intellectual property – need for protection of data to protect/exploit, e.g. HapMap click wrap;
  - Ethical issues – terms of consent, confidentiality & privacy;
  - Use of data – limits of consent, interests of the researchers who generate data.
- Data quality, standards and integration – use of existing 'community' standards or need for further development; ability to integrate data from different sources as well as different types of data.
- Long term preservation and sustainability of data resources – securing long-term sustainable funding for key data resources.
- Users needs – on-going evaluation & user input to ensure utility.

Specific Questions/Challenges

- Promoting data sharing – how to provide incentives?
- What types of data – sequence/genotype to complex phenotypes?
- Quality control and standards for data?
- Making raw data available?
- Making 'negative' data available?
- How to control access – e.g. validation of bona fide researchers
- Structural Variation – reference sequence? How to represent?
- Integration/linking of existing databases or new database?
- Co-ordination with other efforts e.g. NIH

Dr. George P. Patrinos,
Erasmus MC, Faculty of Medicine, Department of Cell Biology and Genetics, Rotterdam, The Netherlands; Asclepion Genetics, Lausanne, Switzerland

Development of National Mutation databases and related tools

Rationale: The mutation spectrum observed for any gene (or multiple genes), associated with a genetic disorder, varies between different population groups across the globe.

Background: Contrary to the more than 370 locus-specific databases, recording detailed information for every published and unpublished mutation identified in various genes, by 2004 there were only 2 true databases dedicated to record the genetic heterogeneity population-wise (Finnish, Arab) and 2 websites (Turkish, Cypriot,) containing related information on a textual format, hence not allowing data querying.

Aim: To develop a comprehensive online depository to record the different mutation frequencies observed for various genetic disorders in different population and ethnic groups and a generic software, allowing for easy creation of similar projects. The latter also contributes towards database uniformity.

Benefits to society:
1. To help optimizing national molecular diagnostic services, not only by providing essential reference information for the design and implementation of mutation detection efforts, but also by enhancing awareness among clinicians and scientists involved in disease research and patient care and the general public about the range of most-common genetic disorders suffered by certain populations and/or ethnic groups.
2. To provide the platform for comparative studies
3. To interpret diagnostic test results in countries with heterogeneous populations, particularly where interpretation of test results in minority ethnic groups may be ambiguous or problematic
4. To provide insights into the demographic history of human populations
5. To serve societies, via ethnic identification, in their rapid transition through migration to multi-ethnicity

Present status:
1. Development of the *ETHNOS* (Ethnic and National database Operation Software) V1.0, supporting flat-file database creation.
2. Creation of 4 fully functional National Mutation frequency databases (Greek, Cypriot, Iranian, Lebanese), while 5 other databases are currently under development (Serbian, Tunisian, Moroccan, Egyptian, Czech) using the same software. These databases can be found at http://www.goldenhelix.org.
3. Development of the *ETHNOS* V2.0 based on MySQL, supporting relational database creation.
4. Establishment of *FINDbase* (Frequencies of Inherited Disorders) Worldwide (http://www.findbase.org), as a main source of information on frequencies of genetic disorders worldwide.

Membership: Human Variome Project and EuroGenTest FP6 NoE
Sources of funding: Corporate funding and ITHANET EC Collaboration action.

Future needs:
• To enhance coordination between researchers active in various countries in the field of human molecular genetics and diagnostics for enriching the database contents with more and up-to-date information
• To contribute data, related to mutation frequencies in European populations, to a Central European Genetic Variation database
• To provide researchers with incentives (carrots) to contribute their data in the database, e.g. through publication to a related scientific journal

APPENDIX 5. CONSOLIDATED QUESTIONNAIRE FINDINGS

Following are the integrated results of the questionnaire replies, which were iterated by all workshop participants, in light of general workshop conclusions and discussions:

Key biology and medical questions that should be addressable with a Central / Linked Genetic Variation Database (s) –

A number of key areas were identified. These seem to have some of the fundamental requirements for central / linked database information content. In particular, for genotype-phenotype data, we need to carefully consider the role of model organism data, e.g. and especially mouse. The questions and research imperatives may be categorized as follows:

- GENERAL GENETICS

- We need to make good general use of genomic variation information
- Research should be based on improved genotype-phenotype data. Both genotype and phenotype actually represent various information levels.

- BEYOND SNPS – CONTEXT OF MUTATIONS

- Role of abnormal copies and phenotypic differences
- Role of local DNA sequence context
- Mutation frequency by type
- Comparative analysis of genomic loci
- Use of mutational spectra –design strategies
- Comparison of mutational spectra
- Environmental and population context, providing differential effects of genetics for mutations

Key biology and medical issues to be considered with Central / Linked Genetic Variation Database (s) and analysis tools to be taken account of:

Given the range of data from "one SNP-one OMIM disease entry" to whole databases on a disease, we need to think carefully about how to appropriately incorporate access to data.

- DISEASE SPECIFIC

- There are huge numbers of disease specific and locus specific databases. Since much of the relevant data will be restricted, we will need to think how the public databases can provide links for "restricted' analysis.

- GENERAL phenotypic variability approaches

- Underlying genetic heterogeneity of inherited disorders in populations - Linking mutant genotypes to clinical phenotypes
- Interaction between multiple susceptibility factors and environment
- Differentiated and categorized neutral versus pathogenic variations

- SPECIFIC phenotypic variability, including DISEASE APPROACHES

- Role of sequence variation and modifiers in monogenetic diseases
- DNA variation in complex traits
- Health risk with variations associated with particular diseases
- Role of somatic mutations

Towards Genetic Variation Database/tools integrated views–What should be included eventually? – access, interrogate, query, analysis, present: basic requirements

- Careful correlations of genotype-phenotype (but not dbSNP duplicate)
- Outcomes and extended molecular phenotypes, levels of clinical sub-phenotypes, endophenotype, e.g. osteoarthritis
- Somatic variations
- Raw data archive

Databases to be Linked (consider "one-stop-shop" database and analysis integrated view):

- We need to consider what kind of data to prioritise for capture, hierarchy of centralized, linked, and "outreach" access.
- Key types of data, exemplified by the databases such as, but are not limited to: dbSNP, Ensembl Variation database, HGMD, OMIM, TRANSFAC, HGVbase, LOVD, FINDbase, Different population databases to assess heterogeneity, COSMIC, p53 databases, uniprot, and future databases.

- Existing and new types of data capture need to be considered

Once 'CORE' database contents are decided, what should be centralized vs. linked?:

- A practical approach is needed. We cannot initially include everything.
- Need pragmatic approach , depending on availability and access types. Special attention is needed for raw data. Standardisation is very important

Towards a Genetic Variation Integration environment :

- Act at European, local, or global level?

- Europe could be a leader, while also acting at a global level, because of excellent annotation of data and health care.
- Europe has many of the key resources, and the EU and Member States could help to provide some of the means to link / centralize them as appropriate.

- Structures and standards

- Standard nomenclature of genetic variants
- Guidelines for contents, database structure and standards, EuroGenTest, European GT.
- International Collaboration
- Series of genetic epidemiology centres working on joint Standard Operating Procedures in a quality control driven network throughout Europe
- Need for Control normal population data generation across Europe
- Standards for submission and deposition are key, but this is a huge field. We should first concentrate on standards for the core material. Many projects already trying: Molpage, Moltools, Wellcome Trust Conf.

Analysis tools:

This points to the need for looking at the various stages for interfaces: data capture, data storage, data analysis and interpretation, linking all information needed to answer complex queries

- Better capture of known genotype-phenotype information and better links to analysis tools

Several database construction and maintenance software packages for locus-specific and national databases.

- User-friendly data entry and access tools querying and downloading resources and a special phenotypic vocabulary should be developed

Resources and Requirements for realisation - cooperation, organisation, financial:

- How can the EU help?

- Major role for EU in leveraging existing resources
- Grant schemes currently provided by the EC can ideally coordinate construction of databases as infrastructures, development of querying tools and generic codes.

- How can National Agencies work together better?

- Increase awareness to national governing bodies and facilitating coordination

- Approaches to Database funding, curators, sustainability; access, academic/commercial - there are a variety of funding models required.

- For true, deep integration of data, one has to openly distribute all the information.
- Fundamental resources are best exploited when they are fully disclosed
- Joint industry funding of open resources is an opportunity.
- Currently, databases can be funded locally by national funding bodies, but often have great difficulty.
- Multinational initiatives are useful for mobilising EU-wide resources.
- Curators are needed, as is encouragement of lab personnel for data submissions.
- Incentives for data depositing can be provided by scientific journals
- Industrial access and commercial exploitation of data must be possible
- There is a strong relation between database structures and funding models.

Summary

Although the field is highly diverse and complex, there does seem to be enough common ground to consider "unified" database (s) and integration activities here at this workshop. There is widespread support for pursuing this initiative at a European level.