

Software and Data Support for the Australian Node of the Human Variome Project

D. du Sart, V. Hyland, J. MacMillan, F. Macrae, D. Thorburn and R.G.H. Cotton and many representatives of all states.

1. Overview

The collection of faults (mutations) in genes causing inherited disease and making this data available is essential for genetic health care and research in all areas of human disease. Collection of such data has been occurring for decades in approximately 10% of the 20,000 human genes, but in a manner that is incomplete and inconsistent around the globe. These collections have been compiled on either a gene by gene basis by experts into so called locus specific databases (LSDBs) (see list of LSDBs at www.hgvs.org/dblist/glsdb.html), or into large aggregated collections across all genes, e.g. OMIM (www.ncbi.nlm.nih.gov/omim/) and HGMD (www.hgmd.cf.ac.uk/). However, the presence of clinical information in these databases is often non-existent, due in large part to differences in privacy legislation across different countries. Furthermore, OMIM does not collect all possible mutations deliberately, and HGMD, whilst it is more comprehensive, is commercial and relies solely upon published data.

Diagnostic laboratories on the other hand, which usually continue on screening specific genes well after the first number of publications on patient cohorts, have a wealth of data with linked clinical information that is not in the public domain. The Human Variome Project was initiated to ensure collection of all mutations and associated clinical data in all genes worldwide and to ensure free and open access to this important data to assist in interpreting the clinical significance of these mutations in patients.

To further the aims of the Human Variome Project two HVP pilot projects have been developed:

- a) The initiation of country specific collection in individual countries. This HVP Pilot seeks to establish systems to systematically collect every mutation that is characterised by a diagnostic laboratory, clinic or research institution in an individual country. To date, this HVP Pilot includes activities in the Arab world, Kuwait, Korea, Argentina and Australia. Such initiatives will ensure complete collection of mutations and their effect worldwide. It is intended that the information from these individual databases will be shared globally through systems put in place by the Human Variome Project.
- b) The international inherited colon cancer consortium (www.insight-group.org), noting their need in treating and advising patients and at risk family members, have in collaboration with the HVP initiated an HVP Pilot to collect all data worldwide on their four genes of interest. This will be a model for the collection of mutations in all genes worldwide.

Objective:

To create a data repository called the Australian Human Variome Database to provide access to information on all genetic variants characterised by Australian laboratories and clinics in a single location. Additionally, to provide a service that streamlines the reporting of genetic variants from Australian diagnostic laboratories and integrates these reports with appropriate clinical data, to facilitate the flow of data into the Australian Human Variome Database. Together, these two services will comprise the Australian Node of the Human Variome Project, the ultimate goal of which, although it is beyond the scope of this NeAT project, is to create linkages between similar nodes in all countries and will act as a pilot system for other countries.

2. Users

Research communities

Researchers interested in the therapy of inherited disease and the functioning of the proteins/enzymes altered by mutations will be able to use the Australian Human Variome Database in modelling genotype/phenotype relationships. The database will provide access to complete data on all mutations in all

genes which will allow classes of research not available until now, due to the limitations of peer reviewed published data, its inaccuracy and its incomplete submission into public databases.

Some of the organisations and groups who will use the data made available by this project include members of the Human Genetics Society of Australia, Peter MacCallum Cancer Centre, Genetic Health Services Victoria, Familial Cancer Clinics statewide and countrywide, State and Australian Birth Defects repositories, Kathleen Cuninghame Foundation Consortium for research into Familial Breast cancer (K ConFab) and Genome-Wide Association Studies (GWAS) e.g. John Hopper.

Clinicians

The main user group for the Australian Human Variome Database will be Australian clinicians and diagnostic laboratory scientists who deal on a daily basis with results from patients suffering from inherited diseases and cancers. The data stored in the national repository will allow them to provide better quality healthcare as their diagnostic and clinical decisions will now be based on the complete body of evidence from the entire Australian population, rather than a handful of interesting cases (perhaps from a different ethnic background) that have been published in medical journals.

Resource Providers

Government funded laboratories and, increasingly, commercial entities and companies provide diagnostic tests for inherited disease and the interpretation of the results of these tests are dependent on the frequency, type and patterns of mutations in diverse ethnic populations. From a public health economics point of view, the service provision costs of the tests will decrease dramatically if results can be interpreted without having to spend hours searching local and international databases for accessible information.

Data Federating and Collaborating Research Groups

A diverse number of agencies including those such as the Birth Defects Register of Australia will utilise the data in this resource. On a much broader level, the locus specific databases that collect information on variants worldwide will find the Australian Human Variome Database an invaluable resource. This will allow the international community to use Australian data.

3. Needs

The user groups mentioned above: researchers, clinicians, diagnostic laboratories, genetic counsellors, industry, government and patients, all need access to a complete set of genetic and clinical data to provide high quality genetic healthcare. This level of data is not available at present. Specific needs are:

1. The most critical long term need is instant access to complete clinical and biochemical data on all mutations causing disease. This information will assist in interpreting the clinical significance of the mutation identified, informing risk assessment, prognosis and in some cases treatments, based on the outcomes of other cases involving that specific mutation.

Currently, investigation of the clinical significance of a gene mutation has an enormous impact on laboratory time (in both diagnostic and research), as it requires accessing all available information across the web, which, when it is available, is distributed in a vast number of databases and scientific publications. Anecdotal evidence from Dr. Cliff Meldrum, Peter MacCallum Institute, indicates that this process takes at least three hours per patient and happens many times a week.

This project is seeking to provide a service that will provide fast and efficient access to Australian data. The ultimate aim is to eventually link this service with similar projects worldwide, providing access to global data.

2. The current public databases have no or very little clinical information associated with the mutations identified. Access to Australian clinical data will add enormous power to the interpretation of clinical significance.

3. In Australia there is little to no coordination of expertise in managing the task of interpretation of genetic variations of uncertain significance. A national expert consensus on interpretation of variants of uncertain significance is required. Without such a consensus, there are already indications that different labs will provide a different interpretation of the same genetic mutation, leading to a disparity in the levels of treatment patients may receive.
4. Within a country there may be ethnic specific mutations or family specific mutations across different cities and access for prognosis is needed.
5. Information on the frequency that individual mutations occur within Australia is needed to develop cheaper and more reliable diagnostic protocols.

4. Services

1. The Australian Human Variome Database

Operator: HVP Coordinating Office/ARCS

The long-term aim of the Australian Human Variome Project is that this data repository will provide access to information on all genetic variants classified within Australia, and contain the biochemical and clinical data of every instance of a genetic variation classified within an Australian clinic or laboratory. The level of data redundancy, i.e. capturing every instance of a variant, rather than aggregating all instances into a single record, is important as it will provide a broad base of evidence from which to base future interpretations of clinical effect, as well as a mechanism for the user community to assess and improve the quality of variant interpretation.

This single data repository will not only be a useful research and diagnostic tool for Australian researchers and clinicians, but will also act as a feeder for the vast network of global variation databases that the Human Variome Project is building. In this way, the Human Variome Project can ensure that vital data generated within Australia is in a position to have a global impact.

This database has been established at <http://australianhumanvariomedatabase.arcs.org.au/> and will be operated jointly by the HVP Australian Node through the Human Variome Coordinating Office, who will provide governance and stewardship, and the Australian Research Collaboration Service (ARCS), who will provide technical and hosting support in the short term. A longer term solution for the technical operation of this service will be developed during the project. Already talks are underway with the University of Melbourne and similar institutions in regards to this. ARCS would be willing to provide long-term hosting. The Human Variome Project Coordinating Office is an entity that is funded separately from this project and will continue to be so after this project's funding cycle.

2. A system to facilitate the flow of data from Clinics and Labs into the AHV Database

Operator: HVP Coordinating Office and individual labs and clinics

The data that will be held in the Australian Human Variome Database originates in labs and clinics across the nation. The Australian Node of the Human Variome Project will develop a suite of systems to embed the timely submission of newly generated data to the national repository into the standard operating procedures of clinics and laboratories. These systems will consist of both software tools and information management systems and practices.

Currently the required data - mutation, phenotype, ethnicity, family history – is still being collected only locally in clinics and labs. The following labs and clinics have so far agreed to contribute data: - Those of Val Hyland/John MacMillan (QLD), Finlay Macrae (VIC), David Ravine (WA), Ron Trent (NSW), Ingrid Winship (VIC) and Glenice Cheetham (SA). This project aims to make data from at least these groups accessible through the AHV data repository by the end of the project. It is highly likely others will also agree and they have indicated they wish to be involved.

NeAT Characteristics

5. eResearch effect

This project will foster the adoption of eResearch practices among a group of Australian researchers who have been historical late adopters of eResearch tools and techniques. In particular, this project will:

1. Provide convenient systems for the automated submission of new data to a single national repository;
2. Enable clinicians and researchers to find valuable data in a single location;
3. Encourage greater adoption of eResearch tools and techniques; and
4. Make Australian data available to a wider audience (see Section 6 below).

6. Broader Adoption

Within Australia

1. Patients and Patient Support groups

Patients often wish to communicate and discuss their disease and progress as a group. It may be possible to have subsets of information available for general community interrogation, and establish Wiki-style platforms to enhance information on, for example, family history and phenotype – given some oversight of the information submitted.

2. Laboratories and Companies

Data from the Australian Human Variome Database will allow new diagnostic tests to be developed based on specific changes observed within the Australian population. Governance structures will be developed to protect patient privacy etc.

3. Government and Hospitals

Adoption of this system will lead to significant cost savings in two directly observable ways. First, diagnostic and clinical staff will spend less time searching the internet for the data they need. Secondly, better access to data will mean that diagnoses and possibly treatments can be delivered more successfully in a shorter time span.

4. Biochemists and protein structural scientists

A small number of computational algorithms have been developed to classify variants in a small subset of genes. This project will make these algorithms available for consideration in other genes. Already, there are interpretation algorithms based on variants occurring at evolutionarily conserved sites across a depth of species, integrated with physico-chemical predictions of effects on proteins and functional assays, usually within research laboratories, developed for some genes – particularly the tumour suppressor genes BRCA1 and 2, and the mismatch repair genes.

5. Mutation Specific Therapy Researchers

Specific mutations are now attracting interest as candidates for novel therapies, targeting the effects of germline mutations which are expressed in the phenotype. One example of this is the development of vaccines against replication errors which are a hallmark of patients with mismatch repair deficiency syndromes.

Outside Australia (other than that mentioned above)

1. The Australian Node will develop a system with multi-disciplinary input which will be readily transferable to other countries. In many respects, Australia is regarded as leader in the world in the documentation and study of human variation, lead by Australian scientists in the Human Variome Project.
2. Australian data, uploaded to international locus specific databases (LSDBs) will contribute beyond its proportional population capacity, due to our country's well integrated state genetic services and state funded DNA diagnostic laboratories.

3. Users of the central databases at University of California Santa Cruz (UCSC), the European Bioinformatics Institute (EBI) and the National Centre for Biotechnology Information (NCBI) will benefit by having the Australian mutations mounted on their genome-wide annotations which often guide judgements of pathogenesis. Processes to do this have already been tested with the InSiGHT mismatch repair gene database in Leiden.

As a consequence of these national developments, Australian experience in curation, updating, and interpreting of the national data will place Australian scientists at the forefront of variant interpretation internationally. This skill will become increasingly valuable as rapid throughput sequencing becomes a reality. This single aspect of the current project is viewed by the investigators as one of the most valuable aspects of the investment for Australia. Successful implementation of this project will be an exemplar for the coordination of genetic services and interpretation of variant information world wide.

7. Value Adding

1. Database Management System

There are a number of products available that may be suitable to run the Australian Human Variome Database. One such product is the Leiden Open Variation Database (LOVD) (<http://www.lovvd.nl>), an open-source product specifically developed to house genetic variation data. Another is the MAWSON project from the University of South Australia. An early part of the project is going to be evaluating the candidates in the field.

2. Data Model

While the DBMS products listed above are supplied with their own data models, it is anticipated that these models will need to be extended, particularly in terms of the models used to represent the phenotypic changes that genetic variants induce. Likewise, the ability to model the families of patients and their medical history is required. These extensions could be developed from existing data models, such as those employed by FamBIS, a database for colorectal cancer patient management, developed by the Victorian Department of Health. Similar software is used in NSW, with a particular focus on bowel cancer predisposition.

3. Data Display and Visualisation

To increase the usability of the services the Australian Node will provide, adequate data visualisation tools will need to be provided. VariVis (<http://www.genomic.unimelb.edu.au/varivis/>) is a software tool for genetic variation visualisation.

4. Data sharing and Access

The data sharing and access components of the Australian Node submission system could be based on the processes and technologies developed by BioGrid Australia (<http://www.biogrid.org.au/>; formerly MMIM) or similar systems. Obviously such data sharing across institutional and even state borders raises a number of ethical and legal constraints, particularly in the areas of privacy and informed consent. While not technical issues *per se*, these considerations must be given due diligence. In this area, the Australian Node can draw from the experience of the Kathleen Cuninghame Foundation Consortium for research into Familial Aspects of Breast Cancer (kConFab), an Australian consortium that provides data and biospecimens to researchers throughout Australia.

8. Standardisation

Technology Development and Standardisation Work to Be Adopted/Extended

1. Mutation Nomenclature Standards

These have been developed by the Human Genome Variation Society (HGVS) over many years and will be used (www.hgvs.org).

2. Genetic Variant Data Models

Preliminary models have been developed by the Human Genome Variation Society, but will require extension, particularly in the representation of phenotype.

3. Clinical Data Models

A number of data models for structuring clinical information currently exist, including the Phenotype and Genotype Experiment Object Model (PAGE-OM). An early part of the project will be evaluating existing standard models.

Risk Reduction by Collaboration

It has been a basic tenet of the global Human Variome Project that all possible be involved to avoid wasteful duplication of effort across different diseases, professions and countries. This policy has been highly successful in bringing together all involved e.g. NCBI, EBI, Gen2Phen, UCSC Genome Browser, Human Genetics Societies, informaticists, country specific activities, colon cancer community, Editors and more recently the Neurogenetics community.

Project Scoping

9. Key participants

1. Human Variome Project Coordinating Office (HVP)
2. Australian National Data Service (ANDS)
3. Australian Research Collaboration Service (ARCS)
4. National ICT Australia (NICTA)
5. Victorian Partnership for Advanced Computing (VPAC)
6. BioGrid Australia
7. Human Genetics Society of Australasia (HGSA)
8. Molecular Genetics Society of Australasia (MGSA)
9. Australian clinical geneticists and diagnostic lab heads

Proposed Project Committee:

- John Coghlan (Chair) - Ex Chairman, Medical Research Committee, NHMRC and Former Director, Howard Florey Institute - current University of Melbourne representative to the Human Variome Project (VIC)
- Richard Cotton – Convenor, Human Variome Project (VIC)
- Val Hyland - Diagnostic Laboratories, and Chair, Molecular Genetics Society of Australasia a special interest group of the Human Genetics Society of Australasia (QLD)
- Heather Howard (NeAT Project Manager) - Human Variome Project (VIC)
- Leon Stirling – The University of Melbourne (VIC)
- Agnes Bankier - Director, Genetic Health Services Victoria (VIC)
- Finlay Macrae – Royal Melbourne Hospital and Secretary, InSiGHT (VIC)
- John MacMillan – Genetic Health Queensland (QLD)
- David Ravine – Laboratory for Molecular Genetics, Royal Perth Hospital, University of Western Australia (WA)
- Ron Trent – Head, Department of Molecular & Clinical Genetics, Royal Prince Alfred Hospital (NSW)
- Barney Rudzki – Division of Molecular Pathology, Royal Children's Hospital (SA)
- ANDS Executive Director, Ross Wilkinson, or delegate, currently Andrew Treloar
- ARCS Executive Director, Tony Williams, or delegate, currently Paul Coddington

10. Project Scale

The scale of the project outlined in this proposal is for 2 years in duration, however we will be actively seeking alternative funding for the project that will enhance and extend the system outlined above.

Staffing requirement for the project, funded by NeAT

Project Manager	50% FTE
Project Officer	1 FTE
Bioinformatician	1 FTE

Parties willing to participate and provide in-kind effort

NICTA	1 FTE Development of software especially related to searches
VPAC	Technical support in software and system development
Human Variome Project Office	Equivalent of 1 FTE
BioGrid Australia	Assistance with system software we expect that BioGrid staff would work with this team to integrate their technology into this collection system (if it is found to be compatible).
ARCS	Hosting of website and database
Diagnostic Laboratories and Clinical support – too numerous to list here	Data submission

More details of in-kind effort will be specified in the project planning phase.

11. Major steps

Time	Milestone
Half 1	Recruitment of staff Recruitment of participating diagnostic labs and clinics Consensus reached on data requirements Ethical framework developed Existing data models reviewed
Half 2	Genetic data model developed or adapted Clinical data model developed or adapted
Half 3	Repository back-end operational Standard operating procedure for genetic test ordering developed Standard operating procedure for genetic test reporting developed Repository submission system operational
Half 4	Repository front-end operational Pilot collection with 10 labs completed

A meeting is being arranged with all collaborators to explore what information will be collected and the IT architecture that is required.

12. Dependencies

1. Collaboration of diagnostic labs
2. Collaboration of clinicians