INFORMATICS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

An Informatics Project and Online "Knowledge Centre" Supporting Modern Genotype-to-Phenotype Research

Adam J. Webb,* Gudmundur A. Thorisson, and Anthony J. Brookes, on behalf of the GEN2PHEN Consortium

Department of Genetics, University of Leicester, University Road, Leicester, United Kingdom

For the HVP Bioinformatics Special Issue

Received 20 January 2011; accepted revised manuscript 28 January 2011.
Published online 22 March 2011 in Wiley Online Library (www.wiley.com/humanmutation). DOI 10.1002/humu.21469

ABSTRACT: Explosive growth in the generation of genotype-to-phenotype (G2P) data necessitates a concerted effort to tackle the logistical and informatics challenges this presents. The GEN2PHEN Project represents one such effort, with a broad strategy of uniting disparate G2P resources into a hybrid centralized-federated network. This is achieved through a holistic strategy focussed on three overlapping areas: data input standards and pipelines through which to submit and collect data (data in); federated, independent, extendable, yet interoperable database platforms on which to store and curate widely diverse datasets (data storage); and data formats and mechanisms with which to exchange, combine, and extract data (data exchange and output). To fully leverage this data network, we have constructed the "G2P Knowledge Centre" (http://www.gen2phen.org). This central platform provides holistic searching of the G2P data domain allied with facilities for data annotation and user feedback, access to extensive G2P and informatics resources, and tools for constructing online working communities centered on the G2P domain. Through the efforts of GEN2PHEN, and through combining data with broader community-derived knowledge, the Knowledge Centre opens up exciting possibilities for organizing, integrating, sharing, and interpreting new waves of G2P data in a collaborative fashion. Hum Mutat 32:543–550, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: genotype–phenotype; association; GWAS; informatics; database; integration; Web services; variation

The Problem Domain

Life sciences are being transformed by a tremendous growth in the scale and complexity of new data and knowledge, reflecting an era of unprecedented technology development that is enabling increasingly high-throughput and low-cost experimentation. This is all part of a "multiomics" approach to research and the veritable information bonanza it brings, but this, however, is a double-edged sword. Certainly, the resulting genomics, transcriptomics, proteomics, metabolomics, and other large datasets promise greatly improved understanding of biological processes, and translational application thereof. But such progress will depend upon scientists being able to organize, integrate, share, and interpret this wealth of new information in sophisticated and effective ways. Meeting this challenge is far from trivial, and to the extent we fail in this endeavor we risk missing potential discoveries and, even more critically, missing the truth and drawing false conclusions. Obvious examples of such problems from the field of genetic association analysis would include: not being able to account for publication bias when aggregating datasets; employing too little phenotype data to distinguish between similar phenotypes with differing etiologies; and performing meta-analysis without being able to incorporate information about differences in population environments or haplotype structures. New systems biology studies and other projects that consider data across multiple omics disciplines are even more vulnerable to such confounding influences.

The myriad problems involved in properly managing and exploiting today's and tomorrow's life science data relate to things such as the fragmentation of data across hundreds of heterogeneous databases, the lack of standardization, the inconsistent identification of biological objects and concepts [Goble and Stevens, 2008], poor enabling of resource discovery [Cannata et al., 2005], difficulties in facilitating data quality assurance and curation [Howe et al., 2008], and approaches to promoting extensive and yet ethically and culturally acceptable data sharing [Walport and Brest, 2011; Wellcome Trust, 2011]. There is also a need for more effective representation of scientific knowledge distilled from research data, and for linking data and other research objects into future modalities for semantic publishing [Bourne, 2005, 2010; Neylon, 2009; Shotton et al., 2009]. Furthermore, as the real and virtual worlds of science increasingly merge so that research is "done" not just "reported" online, there is a need to come up with completely new paradigms for socioscientific interaction in the digital age [Stafford, 2010], to promote highly collaborative and interactive modes of Internet-based scholarly debate and communication.

In this present communication, to explore and illustrate the challenges and current progress in some of the areas listed above, we will concentrate our focus upon the science of genotype-to-phenotype (G2P) relationships in human and model organisms. Even within this one domain there are many multidimensional challenges to be tackled. On a very basic level, the massive data volumes generated by next-generation sequencing instruments present major informatics challenges for smaller laboratories that utilize these devices (either locally or via external service providers), and this considerably curtails the scientific impact that these new technologies are having [Editors, 2008]. More generally, scientists are facing the herculean task of reporting, cataloging, and managing the seemingly limitless number of G2P interactions being identified by research and diagnostic laboratories

*Correspondence to: Adam J. Webb, Department of Genetics, University of Leicester, University Road, Leicester, LE1 7RH, UK. E-mail: ajw51@leicester.ac.uk

Contract grant sponsor: European Community's Seventh Framework Programme (FP7/2007–2013); Contract grant number: 200754—The GEN2PHEN Project.

on a daily basis. For example, according to the NHGRI GWAS Catalog [Hindorff et al., 2009] (http://www.genome.gov/gwastudies/), genome-wide association studies (GWAS) have been published at a rate of approximately five research articles per week during the past 2 years—but because partial or negative studies are generally not reported, then even this number is a substantial underestimate of the true frequency at which genetic association findings are being produced. Similarly, diagnostic labs routinely perform many DNA mutation scans on patients with traits that have a heritable component, but very little of this information ever gets to be released and utilized by others. Furthermore, there are many other pressing challenges relating to G2P data, not least ethical, legal, and social issues relevant to promoting and achieving the sharing of potentially identifiable data from human subjects [Kaye et al., 2009, 2010; Povey et al., 2010].

Tackling G2P Data Challenges

Traditional approaches to biological databasing have been mostly based on the "centralized" model, characterized by gathering data into a large central hub for storage, integration, and display. Historically, this strategy has proved highly successful. Examples include the global collaboration of nucleotide sequence archives (http://www.insdc.org) established in the 1980s, and sophisticated resources for data analysis and visualization provided by bioinformatics centers such as NCBI (http://www.ncbi.nlm.nih.gov), UCSC (http://genome.ucsc.edu), and EBI/EMBL (http://www.ebi.ac.uk). However, as we argued elsewhere [Thorisson et al., 2008b], centralization alone is insufficient for dealing with the full quantitative breadth and qualitative depth of contemporary G2P data, and so hybrid models combining centralized databases with "federated" networks of distributed data and analytical resources are required to tackle the new challenges facing the G2P data field.

Federation of data storage, provision, and analysis across sites is well established in some other scientific disciplines that have a longer history of dealing with "big data," such as astronomy and particle physics, and it is also a cornerstone of data-intensive scientific research, or e-Science [Buetow, 2005; Hey and Trefethen, 2005]. Increasingly, such projects employ Web service-based grid computing to enable automated resource discovery and data analysis. A prominent example is the multi-institutional caBIG project (https://cabig.nci.nih.gov), which has constructed a centrally managed and tightly integrated network designed to seamlessly link dozens of cancer research institutions in the United States and internationally [Buetow, 2009; Saltz et al., 2006]. Another example based on many of the same technologies but with a contrasting, decentralized style is the UK-based myGrid family of tools [Bhagat et al., 2010; Goble et al., 2010; Hull et al., 2006; Oinn et al., 2004] (http://www.mygrid.org.uk).

Unfortunately, the majority projects and institutions that produce and analyze G2P data do not participate in these federated and open grid initiatives. Hence, there is a real problem in ensuring that all their valuable data and discoveries become shared and merged into the online universe of G2P information. To help enable this, and to promote and support blended federated-centralized approaches to G2P data exploitation in general, a 5-year Genotype to Phenotype databasing (GEN2PHEN) project was launched at the start of 2008, via a €12 M award under the European Community's Seventh Framework Programme (FP7: http://cordis.europa.eu/fp7/). GEN2PHEN specifically aims to help establish holistic access to G2P information, through modular tool and data standards developments toward a federated network of online G2P resources, and simultaneously to facilitate the bidirectional flow of knowledge between public G2P databases and G2P researchers. Below, we provide a broad overview of the GEN2PHEN project and provide details of one of its main deliverables: the "G2P Knowledge Centre"—an integrated G2P community Website, information resource, tool repository, and comprehensive data access portal.

The GEN2PHEN Project

The GEN2PHEN consortium is made up of representatives from 25 research organizations and companies based in 10 countries in Europe, in Saudi Arabia, and in India (see full list of partners at http://www.gen2phen.org/about/partners), providing exceptional competence and broad expertise in various aspects of G2P data management and exploitation. Their common goal is to improve the effectiveness of G2P databasing, described as "disastrously deficient" in a review written shortly before the project's conception [Patrinos and Brookes, 2005]. In practice, this means enabling heterogeneous and largely unconnected G2P data resources to evolve toward a comprehensive "G2P biomedical knowledge environment."

The Strategy

Most GEN2PHEN activities are assembled around three core, practical aspects of G2P databasing: (1) devising data standards and pipelines for submitting and collecting data, (2) designing and deploying federated and interoperable modular components for storing and curating diverse datasets, and (3) solutions for exchanging, integrating, and extracting information from the resulting network of federated and centralized databases. Consortium partners all have solid track records in some or all of these areas and are well connected with the broader G2P community. This latter point is essential for aligning and codeveloping GEN2PHEN solutions with those of other allied projects, often via close collaboration. Indeed, consultation, outreach, and dissemination involving the wider G2P community and beyond was prioritized from the very outset of the project.

Details of specific GEN2PHEN objectives, planned and completed deliverables, ongoing activities, and other related information are all published online (http://www.gen2phen.org/about) and so they will not be elaborated here in detail. Instead, this section briefly summarizes the main areas where GEN2PHEN is currently focusing its effort, listing several projects as examples.

Standards Development

To facilitate resource interoperability and enable seamless G2P data exchange and integration, it is vital to increase the overall level of standardization in the field. To this end, GEN2PHEN has worked extensively with others toward developing, refining, and promoting key G2P domain data standards. This includes conceptual models, ontologies and nomenclature conventions, with an overall focus that entails coordinated "bottom-up" standards creation by the community [Brazma et al., 2006; Quackenbush, 2006], rather than "top-down" impositional approaches and formal standardization procedures. Therefore, GEN2PHEN has much in common with, and has connections to, related initiatives such as the Reporting Structure for Biological Investigations Working Groups (RSBI WGs; http://www.mged.org/Workgroups/rsbi/index.html) and Minimum Information for Biological and Biomedical Investigations (MIBBI; http://www.mibbi.org), which promote collaborative development of "omics" reporting standards [Sansone et al., 2008; Taylor et al., 2008].

Examples that embody GEN2PHEN's collaborative approach and success in the area of standards development include close partnering with the groups behind the PaGE-OM for G2P data [Brookes et al., 2009] (http://www.pageom.org), imminent publication of new core data models for phenotype data and locus-specific database (LSDB) content, and a joint effort with the NCBI that has produced the Locus Reference Genomic (LRG: http://www.lrg-sequence.org) framework for standardized reporting of gene variants [Dalgleish et al., 2010] (see Box 1).

Box 1. GEN2PHEN-Sponsored Data Standards Projects

Variation Ontology (VariO). Systematic description of consequences and effects of variation at the DNA, RNA and protein level. http://www.variationontology.org

Variation data model (VarioOM) and Variation Markup Language (VarioML). Minimal data model and XML-based format for describing LSDB content. http://www.varioml.org

Phenotype Object Model (Pheno-OM). Minimal data model to describe phenotypes and other observations. http://www.gen2phen.org/document/pheno-om-2010-03-10

Locus Reference Genomic (LRG). Framework for standardized reporting of gene variants. http://www.lrg-sequence.org

Toward a Unified G2P Data Infrastructure

The main thrust of GEN2PHEN's infrastructural work is the creation of a range of reusable databases and software tools, with an emphasis upon federation and Web services. Naturally, these components are all standards compliant, and they provide the G2P community with a suite of technological building blocks for creating new (or augment existing) data systems that can be incorporated into the globally emerging online network of G2P resources. Thereby, in combination with other databases, a fully interconnected, interoperable, and transparently searchable universe of G2P resources can be assembled, for manual and automated data discovery and analysis, as represented in Figure 1.

Similar to the approach taken for standards development, GEN2PHEN favor collaborative, open-source software development and reuse/adaption of existing software where possible. The power of this open, community-oriented approach, is shown by bioinformatics software initiatives such as BioPerl [Stajich et al., 2002] (http://www.bioperl.org) and BioJava [Holland et al., 2008] (http://biojava.org). Examples of GEN2PHEN projects in this arena include software packages for easy creation of LSDBs and close partnership with the team developing Molgenis (http://www.molgenis.org), an open-source platform for rapid prototyping of genomics database software [Swertz et al., 2010] (see also Box 2).

Data Flow, Data Access, and Data Integration

GEN2PHEN is also creating a variety of solutions for search, retrieval, and integration across the G2P information space. This work builds on, and will demonstrate the utility of, databases and software tools created in the project. Initial work has focused on integration and advanced data provision via existing centralized resources, notably the Ensembl genome browser (http://www.ensembl.org). This employs established technologies such as the BioMart data integration system for large-scale data querying [Smedley et al., 2009] (http://www.biomart.org), and the DAS protocol for exchanging record annotations [Dowell et al., 2001; Jenkinson et al., 2008] (http://www.biodas.org), as well as by building new Web services on top of various project databases and by the construction of data discovery platforms.

Box 2. GEN2PHEN-Sponsored Software Development Projects

Café for Routine Genetic data Exchange (Café RouGE). A central data "clearinghouse" for streamlining the flow of single-locus variation data from clinical diagnostic laboratories to LSDBs. Can be redeployed to support the safe advertising of many types of data. http://www.caferouge.org

DiseaseCard. Integration of genetic and medical information for health applications. http://www.diseasecard.org

GWAS Central. Global study catalog providing rich visualization and query tools for for comparing and contrasting multiple study datasets. http://www.gwascentral.org

Human Genome Mutation Database (HGMD). Genome-wide mutation database specialized in cataloguing variants of clinical utility [Stenson et al., 2009]. http://www.hgmd.org

Human Splicing Finder (HSF). Online tool for predicting the effect of mutation on splice signals. http://www.umd.be/HSF/

Leiden Open (source) Variation Database (LOVD). "In-a-box" software for creating LSDBs [Fokkema et al., 2005]. http://www.lovd.nl

Mutalyzer. Online tool for checking sequence variants reported according to the HGVS nomenclature guidelines [Wildeman et al., 2008]. http://www.mutalyzer.nl

MUTbase. "In-a-box" software for creating LSDBs [Riikonen and Vihinen, 1999]. http://bioinf.uta.fi/MUTbase/

SNP Effect Predictor. Extension to the Ensembl system to provide a Web-based tool and API for deriving variation consequences [McLaren et al., 2010]. http://www.ensembl.org

Universal Mutation Database (UMD). "In-a-box" software for creating LSDBs [Béroud et al., 2005]. http://www.umd.be

Web Analysis of the Variome (WAVe). Integration application for LSDBs and genome-wide databases. http://bioinformatics.ua.pt/WAVe/

Highlights of work undertaken to date include data exchange and integration between LSDBs and Ensembl, facilitated by the aforementioned LRG standard. Also, exemplifying the power of the hybrid federation/centralization approach, GEN2PHEN has built HGVbaseG2P [Thorisson et al., 2008a] (http://www.hgvbaseg2p.org)—recently rebadged as "GWAS Central" (http://www.gwascentral.org)—to provide powerful graphical and textual modes for comparing and contrasting multiple datasets from published, unpublished, and private user-uploaded GWAS studies. Finally, as an illustration of how data can be openly exposed yet still shared in a controlled manner, the GEN2PHEN project offers Mendelian gene mutation data via the Café for Routine Genetic data Exchange ("Café RouGE": http://www.caferouge.org)—an innovative "clearing house" concept that could be easily redeployed to support the safe advertising of many types of data. See Box 2 for a more detailed listing.

Beyond the above practical projects concerned with creating or extending mostly traditional data-centric online resources, GEN2PHEN is working on cultural and policy issues, such as: ethicolegal considerations around G2P data collection and sharing; the idea of providing a BioResource Impact Factor (BRIF) metric for biobanks and databases [Cambon-Thomsen, 2003; Kauffmann and Cambon-Thomsen, 2008]; and designing, creating, and piloting the use of digital IDs for researchers via involvement in the newly formed ORCID initiative (http://www.orcid.org), so that a researcher's online G2P activities and contributions can be discovered, recognized, rewarded, and encouraged. All these different aspects of the GEN2PHEN work program progress in parallel, with links and crossfertilization opportunities being exploited wherever possible. But there is one overriding activity that seeks to bring virtually all the other subprojects together: the G2P Knowledge Centre (KC), a virtual "Center of Excellence" designed to provide a range of new, innovative services to support G2P research.

The G2P Knowledge Centre

The overriding goal of the KC is to provide a central platform amalgamating direct access to distributed G2P data with specialist

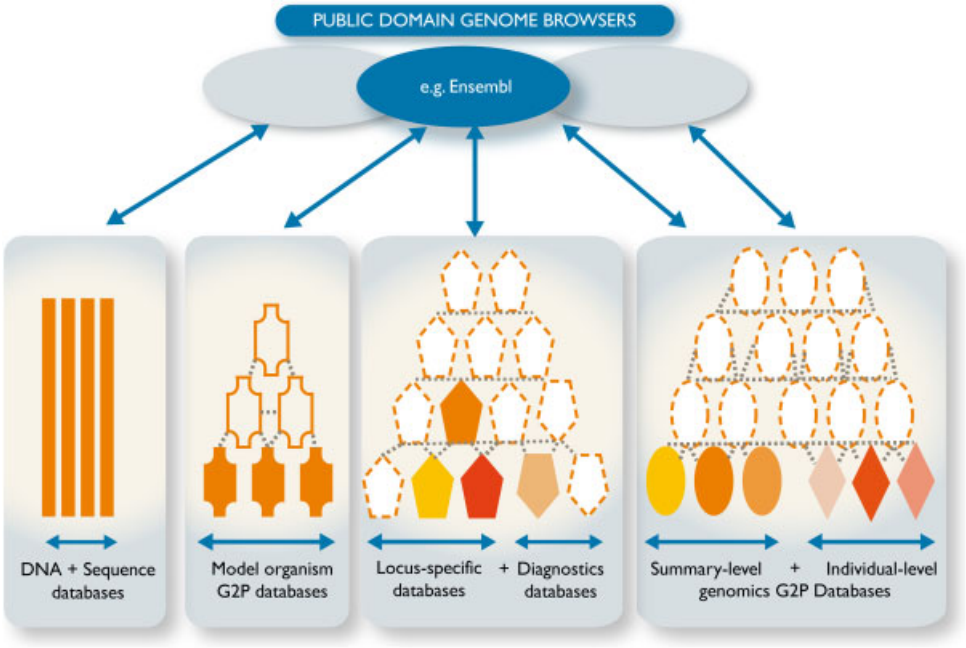

Figure 1. The ultimate fully integrated network of G2P resources. This figure illustrates the "pre-GEN2PHEN" status of G2P databases set against the "post-GEN2PHEN" arrangement the project seeks to help create. The former comprises very few extant databases and great diversity of design (the shapes filled with variously colored patterns) with essentially no interoperability connections (adjoining lines) between them. This provides no convenient way to populate the databases, no easy way to exchange or compare or integrate the different resources, and absolutely no way to search the totality of gathered information. In contrast, the future vision entails one of a broad array of G2P databases (shown in dashed outlines), all constructed from common principles and standards via open-source software (hence, all uniformly colored white), so enabling widespread interconnectivity in the resulting G2P Knowledge network. The ultimate unified system will naturally assume a hierarchical arrangement wherein "basal" databases will tend to hold more detailed and diverse information (e.g., individual-level data, study results from single institutions, expert manual annotations), whereas the databases above them will bring many such datasets together with a degree of simplification per record. At all levels, but particularly from the larger data "warehouses" on top of each hierarchy, information will be channeled and/or served to universal search platforms and graphical browsers, such as Ensembl (the central browser in GEN2PHEN) at http://www.ensembl.org. [Color figure can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

knowledge, all encompassed within a collaborative scientific online workspace (Fig. 2).

Other scientific disciplines have already embraced this kind of online collaborative data enrichment resource. For example, in the field of nanotechnology, the nanoHUB facility (http://www.nanohub.org) provides direct access to powerful simulation tools coupled with extensive community-driven features such as downloadable lectures and presentations, online seminars, events listings, and mechanisms for rapid publication of data and other results. The nanoHUB project has been hugely successful, with its scope extending to some 1,600 resources involving 600 contributors, and over 100,000 users per year. Such initiatives are far less common in the biomedical sciences, although some notable smaller scale examples do exist, such as Alzforum [Kinoshita and Clark, 2007] (http://www.alzforum.org), which combines access to data from Alzheimer's disease research, discussion forums, event listings, and virtual conferences. In going beyond the remit of simple data portals, such sites can help the scientific endeavor by bringing experts together around common problems and concrete data. So far, however, no such tool has existed for the genotype–phenotype field in general—an oversight the G2P Knowledge Centre seeks to address.

Although not its central mission, one section of the KC serves as the GEN2PHEN project Website. This gives GEN2PHEN a way to leverage the KC environment to disseminate information on the project, to provide full and immediate access to all the project deliverables/outputs and training activities, and to furnish a comprehensive listing of every GEN2PHEN-related tool, Website, and database (see Box 3).

Box 3. Project Resources Available at the GEN2PHEN Section of the G2P Knowledge Centre

Project deliverables. The majority of completed GEN2PHEN deliverable documents covering all 10 work packages are available to the general public to view or download. Many of these highly detailed reports will be of interest to both G2P researchers and bioinformaticians. Details on each work package, complete with deliverable documents and any other relevant documents, can be found at http://www.gen2phen.org/about-gen2phen/work-packages

GEN2PHEN Resource List. GEN2PHEN's outputs are both many and diverse. To emphasize the project's involvement in shaping the future of G2P data flows, we provide a comprehensive listing of all GEN2PHEN-supported activities. This includes not only fully fledged database systems, but also links to software source-code and data specifications for various projects. http://www.gen2phen.org/resources

Training tools and materials. As part of GEN2PHEN's overarching strategy, the project produces myriad training tools such as user manuals, tutorials, and videos for GEN2PHEN-related tools and resources. The KC provides a central access point for these tools, which can be either hosted locally or on external Websites, and can be accessed via the training section at http://www.gen2phen.org/training

This information thereby contributes to the KC's far broader set of data listings and search capabilities, in turn coupled into an innovative system whereby users may provide per-record annotations for remotely hosted G2P data. Finally, superimposed upon all of this, the KC provides an array of tools for establishing and nurturing active online research communities.

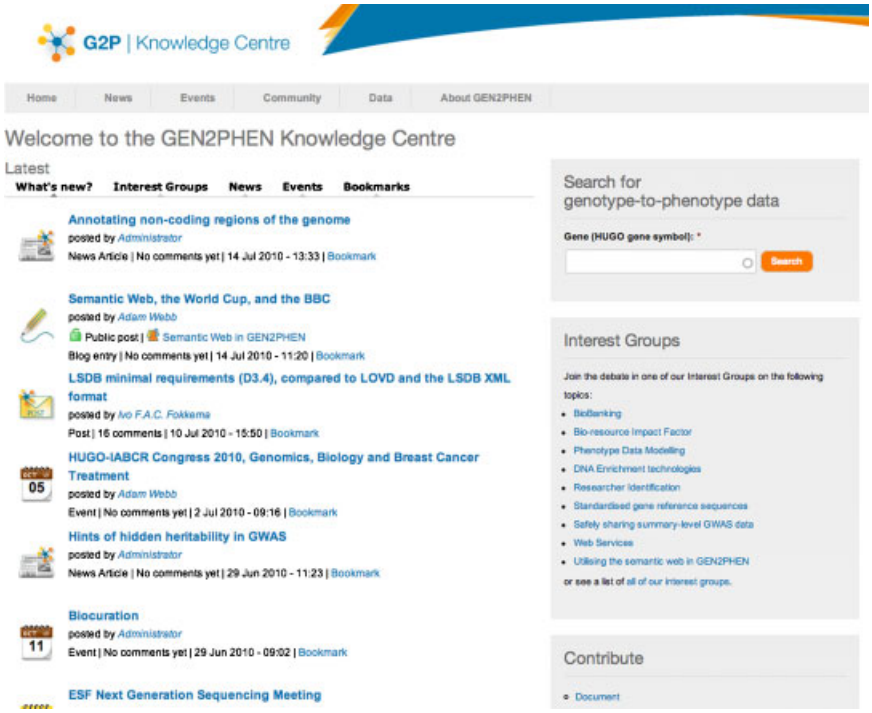

Figure 2. Homepage of the G2P Knowledge Centre. Via this Web page users are presented with a list of most recent content, latest news, full events listings, lists of interest groups, and access to bookmarked content. A gene search box is also provided. Tabbed menus lead to main site sections, namely, news, events, community, data and about GEN2PHEN. [Color figure can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

The main features of the KC will now be discussed in more detail below.

A Central Search for G2P Data

Databases holding G2P data are both many and diverse, ranging from per-gene locus-specific databases, to GWAS catalogs and G2P archives. A researcher hoping to track down G2P data for a given locus would need to visit several databases and negotiate differing user interfaces and data formats merely to see what data are available, let alone retrieve, integrate, and analyze those data. The KC seeks to reduce this workload by providing a single access point to the wealth of data stored throughout an extensive federated G2P database network. The principle is simple—a KC search will return a summary of available G2P data organized by source database, with result entries linking back to original records. Relevant records within the aforementioned Café RouGE will also be made available via this central search tool.

This holistic searching is carried out "live"—that is, searches do not serve up old results from an internal database updated periodically by scanning client databases; instead, the system interrogates the many source databases directly, in real time. Therefore, data are always up to date (within a few hours), subject to caching mechanisms put in place both to prevent overloading client databases and to still provide results in the event of a source database being inaccessible.

Crucially, in addition to the broad search capability, the KC aims to go yet further by providing a novel annotation system, whereby users can directly comment on and flag search results from remote databases. The idea here is that these user-supplied annotations will be made available both to database maintainers and to the wider G2P community, both on the KC Website and in machine-readable form via an application programming interface (API), thus providing a community-edited annotation layer for distributed G2P data. This innovative new system, inspired by recent wiki-like community annotation projects like WikiProteins [Mons et al., 2008] (http://www.wikiproteins.org) and the RNA WikiProject [Daub et al., 2008] (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_RNA), allows annotations to be anchored to specific database records or resources, perhaps sparking further community-based debate and discussion. In this way, G2P database content is taken beyond the realm of static database records so that they become "living" entities, enhanced and evolved by user and producer comments.

A Catalog of Locus-Specific Databases

As a particular service to researchers and clinicians interested in inherited disease risk, the KC provides a comprehensive listing of LSDBs, which is further enriched by incorporating extensive metadata from a recent survey into this field [Mitropoulou et al., 2010]. Collaborations are being sought to continually update this list from various sources, by both manual and automated means. In addition to a fast, searchable Web interface, these data are supplied in numerous formats including comma separated value (CSV) text, Microsoft Excel spreadsheet, Atom feeds, and JSON/JSONP. Such formats allow use of these data by individual end-users, or incorporation into third-party Websites and tools. To enable this latter option, a simple JavaScript copy-and-paste plugin is provided that allows Website maintainers to insert the LSDB listing, complete with sorting, paging, and filtering functionality, into their own pages. The plugin can be fully styled via CSS to suit the hosting Website. Further information is available online (http://www.gen2phen.org/data/lsdbs).

The LSDB catalog provides a perfect example of the KC's philosophy of providing content not only through human-readable

Web pages, but also via alternative machine-readable formats, allowing greater integration with other tools and resources. This provision will be expanded upon and enhanced with future incremental KC updates.

News, Blogs, and Event Listings

The KC provides a broad-scope information portal for the G2P community, encompassing not just data and analytical resources, but also useful day-to-day features such as news items, events listings, and blog posts. For the news section, abstracts and short summaries are gathered from relevant journals and other online resources and manually selected for particular relevance. Each abstract or article is linked to the original full article on the source Website. Visitors to the KC can post comments on articles, and read comments posted by others. Additionally, visitors can utilize the site-wide bookmarking system to track updates, alterations, and comments on these items, providing a personalized listing of interesting content via simple Web interfaces and an RSS feed. In particular, the specialist editorial selection of articles is likely to be of high interest to scientists working on genotype-to-phenotype relationships and allied fields. As with most KC content, the regular news digest is available not only through the Web interface, but also RSS feeds. Finally, users are strongly encouraged to contribute news articles that they find interesting, either by submitting full stories or simply by suggesting useful links. These articles or useful links will then be published at an editor's discretion.

Complementing the KC's aggregated news provision, a number of contributor blogs are provided for the dispensation of timely opinion pieces and short rapidly disseminated articles, often of a less formal nature than the aforementioned news articles. These are intended to quickly highlight both scientific and technological developments and provoke healthy debate by the G2P community. Again, blog posts can be obtained via blog-specific and site-wide RSS feeds, and these can be monitored via the site's intuitive bookmarking system. Enquiries from users interested in running their own blog on the KC are welcomed.

The KC also features listings of upcoming G2P-related events that may be of interest to its users, including conferences, symposia, and training events. The community can also use the facility to advertise their own events to others in the G2P field via a simple submission form. Events are displayed in a simple listing, an interactive calendar and on a map.

In general, users can comment upon, and sometimes update, almost any item within the KC. This reflects one of the site's principal goals, that content is not simply posted to be viewed or remain stagnant, but instead it should be allowed to, and encouraged to, evolve so that it drives community debate and hence advances science in the G2P domain.

Interest Groups

The "Interest Groups" section of the KC provides self-contained areas of the site ("mini-KCs") dedicated to particular fields or projects in a manner similar to commonplace Internet forums. Unlike regular forums, however, where users are typically restricted to simple thread-based text messages, users may contribute documents (which will be then be viewable within the Web browser, or downloadable), regular posts, create wiki pages (which may be edited by other group members thus easing the production of collaborative documents), news articles, and events, as part of the content in these groups (Fig. 3). These features provide a flexible and powerful workspace for collaborative groups, and hence are used both within the GEN2PHEN project and by the wider G2P community. Each group is maintained by a dedicated group administrator, whose job it is to manage group posts and

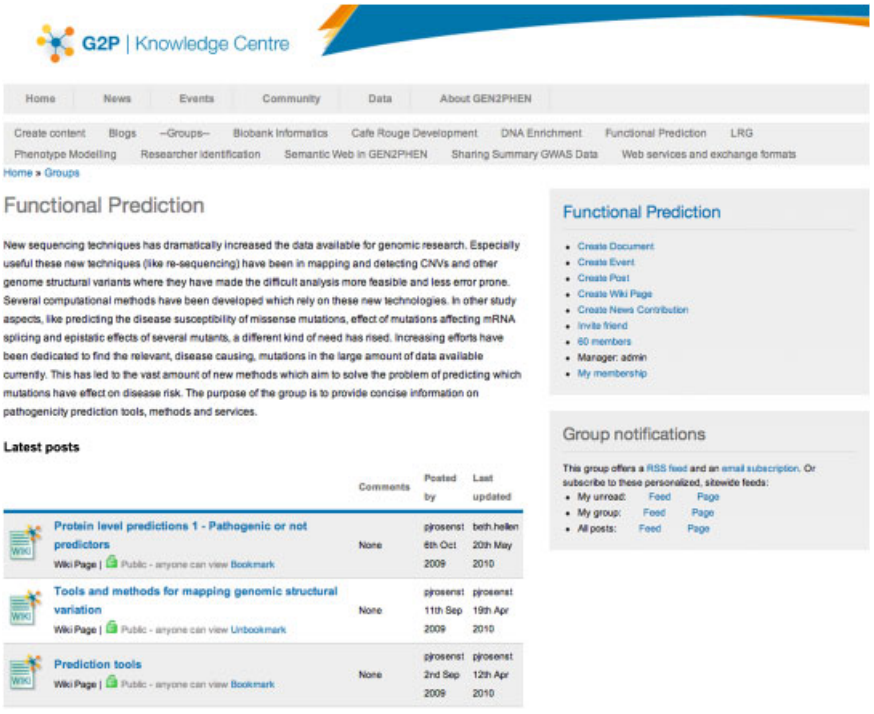

Figure 3. An example of an Interest Group. Each group operates like a small self-contained Knowledge Centre and can contain posts, wiki pages, documents, events, news, and blogs. Membership for each group can be open to the public or restricted to private workgroups. [Color figure can be viewed in the online issue, which is available at www.wiley.com/humanmutation.]

memberships. Posts within groups may be restricted to the group, or made visible to all users of the site.

Interest groups further allow collaborative workspaces to be tightly coupled to the resources available and accessible from the KC. Active interest groups as of January 2011 (see http://www.gen2phen.org/community) cover the following topics:

- Bio-Resource Impact Factor (BRIF).
- Locus Reference Genomic standard (LRG).
- Phenotype data modeling.
- Researcher identification.
- Utilizing the semantic Web.
- Web services and exchange formats.

Most of these groups are available to the general public, whereas a few require users to receive authorization from a group administrator to participate. The KC welcomes and encourages applications from members of the G2P community who wish to utilize the interest group facilities for their own topics and projects. Such proposals can be made via a simple request form available at the site.

System Design and Implementation

The KC has been constructed using the Drupal content management system (CMS) (http://www.drupal.org). Despite a steeper development learning curve compared to other popular CMSs, Drupal was selected for its robust and extremely flexible code base with which to build sophisticated Web applications. The standard platform has since been considerably extended using a combination of both public domain contributed modules and several custom coded modules specific to the KC implementation. As with other GEN2PHEN software packages, these extensions will be made available for download as open-source software.

Future Directions

As summarized in this article, even though only two-thirds of the way through its funding period, the GEN2PHEN project has already generated many key resources and paved the way to a fully integrated, community-enhanced G2P network. In its current form, the KC provides the G2P community with a central hub for data, other useful information, and community interaction. To further leverage these powerful tools and resources, the project generally (and the KC in particular) will continue to explore new methods to distribute its contents besides the human-readable HTML-based Website. Besides commonly used formats for Web content syndication such as RSS and Atom, technologies and formats such as the resource description framework (RDF) may be employed to provide data and leverage the potentially immense power of the Semantic Web [Berners-Lee and Hendler, 2001; Berners-Lee et al., 2001]—a self-describing Web-based global network of linked data (http://linkeddata.org). An associated expansion of Web services both on the KC and as part of other project resources will simultaneously help to facilitate Web-based data integration or "mash-ups" [Cheung et al., 2008], and machine-oriented knowledge generation.

Most ambitiously of all, the GEN2PHEN project has recently begun exploring how G2P data and related tools might be adapted or newly created to move G2P knowledge beyond the research domain and into the healthcare environment. Clearly, this raises many challenges that are far too large and way beyond the scope of GEN2PHEN itself. But the very successful philosophy the project has followed, with its emphasis upon integrated community development work toward fully interoperable information networks and the bidirectional flow of information to/from the user community, probably represents a good template for future projects seeking to integrate G2P and other bioscience realms, especially if aiming at delivering improved healthcare.

In summary, even though we argue that GEN2PHEN has made a good start in a range of directions, we are fully aware that a massive amount of further work needs to be done if scientists are to fully meet the challenge set out at the start of this article: that is, to effectively organize, integrate, share, and interpret the wealth of new life science information in sophisticated and effective ways. We believe this challenge can and will be met, and foresee many exciting and revolutionary years ahead as this is achieved.

References

Berners-Lee T, Hendler J. 2001. Publishing on the semantic web. Nature 410: 1023–1024.

Berners-Lee T, Hendler J, Lassila O. 2001. The Semantic Web. Sci Am 284:34–43.

Béroud C, Hamroun D, Collod-Béroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184–191.

Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, Lopez R, Gbloe CA. 2010. BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res 38(Suppl):W689–W694.

Bourne P. 2005. Will a biological database be different from a biological journal? PLoS Comput Biol 1:179–181.

Bourne PE. 2010. What do I want from the publisher of the future? PLoS Comput Biol 6:e1000787.

Brazma A, Krestyaninova M, Sarkans U. 2006. Standards for systems biology. Nat Rev Genet 7:593–605.

Brookes AJ, Lehvaslaiho H, Muilu J, Shigemoto Y, Oroguchi T, Tomiki T, Mukaiyama A, Konagaya A, Kojima T, Inoue I, Kuroda M, Mizushima H, Thorisson GA, Dash D, Rajeevan H, Darlison MW, Woon M, Fredman D, Smith AV, Senger M, Naito K, Sugawara H. 2009. The phenotype and genotype experiment object model (PaGE-OM): a robust data structure for information related to DNA variation. Hum Mutat 30:968–977.

Buetow KH. 2005. Cyberinfrastructure: empowering a "third way" in biomedical research. Science 308:821–824.

Buetow KH. 2009. An infrastructure for interconnecting research institutions. Drug Discov Today 14:605–610.

Cambon-Thomsen A. 2003. Assessing the impact of biobanks. Nat Genet 34:25–26.

Cannata N, Merelli E, Altman RB. 2005. Time to organize the bioinformatics resourceome. PLoS Comput Biol 1:e76.

Cheung K, Kashyap V, Luciano J, Chen H, Wang Y, Stephens S. 2008. Semantic mashup of biomedical data. J Biomed Inform 41:683–686.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully R, Proctor G, Chen Y, McLaren W, Larsson P, Vaughan B, Béroud C, Dobson G, Lehväslaiho H, Taschner PE, den Dunnen JT, Devereau A, Birney E, Brookes AJ, Maglott DR. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2:24.

Daub J, Gardner PP, Tate J, Ramsköld D, Manske M, Scott WG, Weinberg Z, Griffiths-Jones S, Bateman A. 2008. The RNA WikiProject: community annotation of RNA families. RNA 14:2462–2464.

Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. 2001. The Distributed Annotation System. BMC Bioinformatics 2:7.

Editors. 2008. Prepare for the deluge. Nat Biotechnol 26:1099.

Fokkema I, den Dunnen J, Taschner P. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63–68.

Goble C, Stevens R. 2008. State of the nation in data integration for bioinformatics. J Biomed Inform 42:687–693.

Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, DeRoure D. 2010. myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Res 38(Suppl):W677–W682.

Hey T, Trefethen AE. 2005. Cyberinfrastructure for e-Science. Science 308:817–821.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Holland RCG, Down TA, Pocock M, Prlić A, Huen D, James K, Foisy S, Dräger A, Yates A, Heuer M, Schreiber MJ. 2008. BioJava: an open-source framework for bioinformatics. Bioinformatics 24:2096–2097.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, Pierre SS, Twigger S, White O, Rhee SY. 2008. The future of biocuration. Nature 455:47–50.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. 2006. Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34: W729–W732.

Jenkinson A, Albrecht M, Birney E, Blankenburg H, Down T, Finn R, Hermjakob H, Hubbard T, Jimenez R, Jones P, Kähäri A, Kulesha E, Macías JR, Reeves GA, Prlié A. 2008. Integrating biological data—the Distributed Annotation System. BMC Bioinformatics 9:S3.

Kauffmann F, Cambon-Thomsen A. 2008. Tracing biological collections: between books and clinical trials. JAMA 299:2316–2318.

Kaye J, Boddington P, de Vries J, Hawkins N, Melham K. 2010. Ethical implications of the use of whole genome methods in medical research. Eur J Hum Genet 18: 398–403.

Kaye J, Heeney C, Hawkins N, Vries JD, Boddington P. 2009. Data sharing in genomics—re-shaping scientific practice. Nat Rev Genet 10:331–335.

Kinoshita J, Clark T. 2007. Alzforum. Methods in molecular biology 401:365–381.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070.

Mitropoulou C, Webb AJ, Mitropoulos K, Brookes AJ, Patrinos GP. 2010. Locus-specific database domain and data content analysis: evolution and content maturation toward clinical usea. Hum Mutat 31:1109–1116.

Mons B, Ashburner M, Chichester C, Mulligen EV, Weeber M, Dunnen JD, Ommen GV, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E, Roes PJ, Borner K, Bairoch A. 2008. Calling on a million minds for community annotation in WikiProteins. Genome Biol 9:R89.

Neylon C. 2009. Head in the clouds: Re-imagining the experimental laboratory record for the web-based networked world. Automat Exp 1:3.

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20:3045–3054.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: disastrously deficient. Trends Genet 21:333–338.

Povey S, Al Aqeel AI, Cambon-Thomsen A, Dalgleish R, den Dunnen JT, Firth HV, Greenblatt MS, Barash CI, Parker M, Patrinos GP, Savige J, Sobrido MJ, Winship I, Cotton RG, Ethics Committee of the Human Genome Organization (HUGO). 2010. Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). Hum Mutat 31: 1179–1184.

Quackenbush J. 2006. Standardizing the standards. Mol Syst Biol 2:2006.0010.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852859.

Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, Kher M, Manisundaram A, Shanbhag K, Covitz P. 2006. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics 22:1910–1916.

Sansone S, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S, Members of the RSBI Working Group. 2008. The First RSBI (ISA-TAB) Workshop: "Can a Simple Format Work for Complex Studies?" OMICS 12:143–149.

Shotton D, Portwin K, Klyne G, Miles A. 2009. Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS Comput Biol 5: e1000361.

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart—biological queries made easy. BMC Genomics 10:22.

Stafford N. 2010. Science in the digital age. Nature 467:S19–S21.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl Toolkit: Perl modules for the life sciences. Genome Res 12:1611–8.

Stenson P, Mort M, Ball E, Howells K, Phillips A, Thomas N, Cooper D. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13.

Swertz M, Dijkstra M, Adamusiak T, van der Velde J, Kanterakis A, Roos E, Lops J, Thorisson G, Arends D, Byelas G, Muilu J, Brookes AJ, de Brock EO, Jansen RC, Parkinson H. 2010. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. BMC Bioinformatics 11:S12.

Taylor CF, Field D, Sansone S, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz P, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, JulianJr RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, StoeckertJr CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26:889–896.

Thorisson GA, Lancaster O, Free R, Hastings R, Sarmah P, Dash D, Brahmachari S, Brookes A. 2008a. HGVbaseG2P: a central genetic association database. Nucleic Acids Res 37:D797–D802.

Thorisson GA, Muilu J, Brookes AJ. 2008b. Genotype–phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 10:9–18.

Walport M, Brest P. 2011. Sharing research data to improve public health. The Lancet Published online 10 January ahead of print. Available at: http://dx.doi.org/10.1016/S0140-6736(10)62234-9.

Wellcome Trust. 2011. Sharing research data to improve public health: full joint statement by funders of health research. Available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm [Accessed 24-01-2011].

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13.