Genotype–phenotype databases: challenges and solutions for the post-genomic era

Gudmundur A. Thorisson, Juha Muilu‡ and Anthony J. Brookes**

Abstract | The flow of research data concerning the genetic basis of health and disease is rapidly increasing in speed and complexity. In response, many projects are seeking to ensure that there are appropriate informatics tools, systems and databases available to manage and exploit this flood of information. Previous solutions, such as central databases, journal-based publication and manually intensive data curation, are now being enhanced with new systems for federated databases, database publication, and more automated management of data flows and quality control. Along with emerging technologies that enhance connectivity and data retrieval, these advances should help to create a powerful knowledge environment for genotype–phenotype information.

Screen-scraping
The automated process of extracting data from web pages intended for human viewing.

Genotype-to-phenotype
(G2P). The relationship between genetic variation in an organism and how this affects its observable characteristics.

**Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK.*
‡Institute for Molecular Medicine Finland, University of Helsinki, Haartmaninkatu 8, FIN-00290, Helsinki, Finland.
Correspondence to A.J.B
e-mail: ajb97@leicester.ac.uk
doi:10.1038/nrg2483
Published online
9 December 2008

The World Wide Web has become an indispensable tool for biomedical researchers who are striving to understand how genes cause disease. Web sites such as the PubMed literature-search service[1], the Ensembl[2], University of California Santa Cruz (UCSC)[3] and National Center for Biotechnology Information (NCBI)[1] genome browsers, and the BLAST[1] sequence-search service, are examples of the Internet resources that many biologists use on an almost daily basis. Behind the scenes these resources are based upon similar technologies and design principles (that is, they have standard 'architectures'), but their user interfaces differ widely in terms of style, functionality and content. This diversity complements the diverse needs of the field, but to investigate a given biological question a user might need to browse many web sites and still never feel sure they have tracked down all the information they might need.

Although the proliferation of data resources can be frustrating for traditional biologists, it presents an even bigger challenge for 'omics' researchers who need to automate large-scale data aggregation across many different sites. Historically, such users were forced to write software to automatically surf web sites to extract information that was originally designed for human consumption. As noted by Stein[4], this 'screen-scraping' approach has numerous disadvantages. Instead, there are better ways to interconnect large sets of related information so that they can be searched and downloaded from a single portal.

In this Review, we consider how this has been tackled in the past for genotype-to-phenotype (G2P) data, and look at how the relevant technologies are currently being improved. We discuss some of the technical issues surrounding database development, and the recent trend towards an increased emphasis on federated database solutions, which can link independent databases through a central portal and be married with the proven benefits of traditional central databases in which related data is stored all in one place. Looking further into the future, we consider even more revolutionary approaches to data integration and utilization, and discuss potential challenges that need to be addressed.

Lessons from the past

To understand the obstacles and the opportunities surrounding modern G2P databases, it is helpful to consider how the field has grown and evolved into its current state. Until recently, online stores of genetic data tended to be built as 'centralized databases' (FIG. 1), and this model has served the field well given its previous requirements.

Sequence databases. The earliest databases of prominence in genetics were designed to hold DNA-sequence data. In the early 1980s, as soon as the use of commercial technologies for DNA sequencing became widespread, such depositories were needed to facilitate exchange and comparison of DNA sequences. Three major central databases were constructed for this purpose: the DNA

© 2009 Macmillan Publishers Limited. All rights reserved

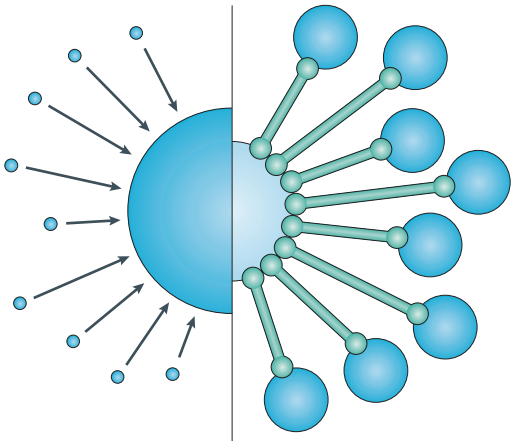


Figure 1 | **Extreme models for database integration.** Radical forms of centralized and federated database networks are shown. In the centralized model, outstations or 'nodes' (small blue circles) gather and prepare data for transfer to a massive central 'hub', where it is stored, integrated and made available for searching and presentation. In the federated model, the outstations are replaced by fully functional databases (large blue circles) that gather and expertly curate data, that provide various means for human- and machine-based searching and accessing of this information, and that offer a range of data presentation options. In the federated model, the hub receives no data from the nodes, but undertakes the important job of coordinating the nodes and organizing searches across the databases. The genotype-to-phenotype database network of the future will probably be a hybrid of these two extreme models.

databank of Japan (DDBJ)[5], GenBank (based in the United States)[6] and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database[7]. In the mid 1980s the International Nucleotide Sequence Database Collaboration (INSDC) was established to promote full content exchange between these databases.

The INSDC is a large-scale central database project, and it provides an excellent example of how effective the central databasing strategy can be. Arguably, however, this project was such a success because DNA-sequence information is simple to represent as a directly annotated string of letters and sequence regions (that is, it is one-dimensional) and, despite a massive growth in data volume, the scale of the problem did not exceed the capabilities provided by concomitant advances in computer technology.

Model organism databases. Model organism databases (MODs) provide a second example of how the central database model can be used effectively. These databases specialize in capturing genomic, phenotypic and other information for a model organism. Examples include Wormbase[8], the Rat Genome Database[9] and the Mouse Genome Informatics Database[10]. A single group or a few groups working closely together, armed with expert knowledge on their model organism of interest, were the typical creators of these early central G2P databases. The resulting web sites provide a focal point for information gathering and access, as well as centralized services such as a genome browser interface and tools for comparative genomics.

A simplistic assessment of MODs would put their success down to the limited volume of data they have to manage, compared with the amounts flowing into a global nucleotide-sequence database. But this view fails to allow for the fact that the data contained in a MOD are far more diverse and complex than mere one-dimensional sequence strings (that is, genetic and phenotypic two-dimensional information). This complexity makes it far more difficult to organize the data within a single depository. The MODs overcame this hurdle through good leadership and the relatively small community sizes that made it possible for agreements to be reached on matters such as data model standards, laboratory protocols, terminologies and curation practices. Having these standards in place ensured that effective data management, interpretation and exchange could occur between the central database and many different laboratories. The MOD experience thus emphasizes the absolute need for robust and universal standards in order to aggregate and integrate G2P data. Extending this principle, MODs for many different species are now working together as part of the Generic Model Organism Database (GMOD) project to further harmonize and standardize their activities.

Central databases for human 'Mendelian mutations'. The databasing of human G2P relationships has lagged behind what has been achieved in other species. There are many reasons for this, one of which is the far larger size and the dispersed and diverse nature of the underlying research community that necessarily includes biologists, clinicians, epidemiologists and statisticians. This has made it difficult to agree and organize a full series of universal standards. Furthermore, the standards themselves are difficult to devise for human G2P relationships, owing to the complexity of the data. These complexities include: the full spectrum of medical diagnoses and clinical test results that are often open to subjective interpretation; a wide range of normal traits that might vary with age and between populations; and a myriad of sequencing, genotyping and other laboratory procedures that are used for the generation of primary data, which can be analysed and utilized in a plethora of different ways.

However, attempts have been made to capture a broad picture of human Mendelian G2P knowledge via the centralized database model. Good progress has been made by the Online Mendelian Inheritance in Man[11] (OMIM) database, which provides a genotype–phenotype catalogue of human genetic disorders, and which first appeared in book form over 40 years ago[12]. The project went online in 1990, and is now maintained by a substantial team of curators who manually extract experimental findings from the literature. This has resulted in a compilation of high-quality records on ~11,000 genes and diseases (on 1 April 2008). But this is a long way from being a fully comprehensive summary of all knowledge on human G2P relationships, even

© 2009 Macmillan Publishers Limited. All rights reserved

for Mendelian disorders. The task is simply too large for one team to manage in a centralized fashion, and given the complexity of the source information, OMIM is forced to package its content in narrative form. This makes it unsuitable for automated mining or deep integration with other database resources. Similar data collection issues are being faced by the Human Gene Mutation Database[13] (HGMD), which uses manual curation to extract from the literature a list of mutations that underlie Mendelian disorders and then places these in a structured and readily searchable format. Another project with data collection challenges is the Pharmacogenetics and Pharmacogenomics Knowledge Base[14] (PharmGKB), which collates extensive knowledge about the relationships between drugs, diseases and genes, to assist pharmacogenomics research. These examples illustrate some fundamental limitations to the central database concept, primarily relating to data complexity and quantity, especially in the context of human G2P information.

Locus-specific databases for human 'Mendelian mutations'. Taking an alternative approach, many groups have collated primarily Mendelian G2P information for just one or a few genes of relevance to one or a few diseases of interest. The first of these locus-specific databases (LSDBs) was published in 1976, in the form of a catalogue of human globin mutations[15]. Following a slow but steady increase in the number of online LSDBs, ~700 databases are now listed at the Human Genome Variation Society web site[16]. LSDB entries tend to be rich in information content and enhanced by domain-specific expert curation. As well as published information they typically include unpublished DNA variation data along with evidence linking the variation to pathogenicity. Unfortunately, these databases are created independently, with little coordination or harmonization, and with little or no dedicated funding.

LSDBs range from simple non-networked spreadsheets to fully fledged online databases. Consequently, they represent a fragmented network of silos that are full of rich information (FIG. 2), across which it is not possible to efficiently exchange or integrate G2P information. In their current disjointed form, LSDBs are essentially at the opposite end of the spectrum from central G2P databases that provide a shallower but genome-wide perspective. Neither is ideal, and so another approach — or a combination of approaches —is needed.

Challenges for modern G2P databases

MODs, LSDBs and related databases have taught us that it will not be straightforward to computationally organize all human G2P information. Furthermore, the challenge is growing: high-throughput genomic data generation is now within the reach of many laboratory budgets. In addition, it is now or soon will be possible to explore phenomena such as structural variation, rare or unique alleles, DNA methylation and somatic genome changes in a comprehensive manner.

Currently, genetic association studies, especially genome-wide association (GWA) studies, are a particularly prolific source of G2P data. The principal databases that were set up to store and organize GWA data include the dbGaP archive in the United States[17], the European Genotype Archive (EGA) and the GWAS Database of Japan. Other related projects include HGVbaseG2P, the Genetic Association Database (GAD)[18], the Type 1 Diabetes Genetics Consortium, and disease-specific efforts AlzGene[19] for Alzheimer disease, PDGene for Parkinson disease and SZGene[20] for schizophrenia. At present, all of these use the centralized database model, although it is unclear whether this approach will suffice in the long term. As the field advances, these databases will have to grapple with complex issues, such as: increasingly convoluted data governance issues pertaining to different countries and legislatures; rapid changes in the scale, depth and format of the primary and processed data; and the probable addition of other forms of variation and levels of etiologic complexity. The overriding need is to achieve a sufficient degree of global comprehensive coverage to make two-dimensional G2P databases as successful as one-dimensional DNA-sequence databases.

Database creators, their patrons and their funding bodies are acutely aware of the need for more and better human G2P databases. Based on past experiences as elaborated above, and in light of recent technology developments, we suggest six key areas that need attention if improved G2P databases are to be built.

Data quantity. The scale of current and future G2P research means that data sets will keep getting larger and more numerous. This acceleration in the rate of data production might even start to outstrip the ability of database technologies to handle the information. For example, results might be repeatedly split and

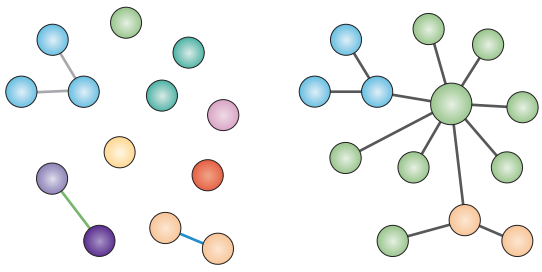


Figure 2 | **Databases and database networks.** Early genotype-to-phenotype (G2P) databases were based on many different designs (represented by circles of different colours) with few connections between them. As the field develops, databases will instead be built on more standardized design and operation principles, enabling extensive interconnectivity between projects. Each resource in the resulting network will have an emphasis on being a data-storage 'node' (smaller circles) or a data-searching 'hub' (larger circle), or a combination of the two. It is hoped that emerging semantic web technologies will develop the network into a powerful and seamless G2P knowledge environment.

Genome-wide association study (GWA study). Examination of DNA variation (typically SNPs) across the whole genome in a large number of individuals who have been matched for population ancestry and assessed for a disease or trait of interest. Correlations between variants and the trait are used to locate genetic risk factors.

© 2009 Macmillan Publishers Limited. All rights reserved

merged, and re-examined — for example, GWA studies that share control materials and cross-compare their primary data sets, and that are being extended by further empirical work and by statistical reanalysis. The SNP data underlying GWA results are also now being used to extract information on structural variation; and all of the above can be repeated *ad infinitum* for every phenotype and subphenotype characterized. The data volume will be further increased owing to the addition of other forms of variation, other areas of bioscience and the emergence of routine whole-genome sequencing[21]. Therefore, data quantity must be a key consideration in database design.

Data quality. Even though database records will never be completely error-free, efforts must be made to avoid inaccuracies wherever possible. Such activities can be split into two activities: manual data curation efforts involving reading and redrafting of data that involves knowledge and concepts, for example, from the literature into databases such as OMIM; and automated validation of generally larger data sets with more straightforward content, such as markers, genotypes and sequences. Quality control should be optimized from the stage of data generation onwards, but databases can only become involved from the stage of guiding researchers in the preparation of accurate and appropriate data submissions.

After data is received, databases should then use their own quality assurance measures to check for internal consistency and completeness of the submission. In scenarios in which this requires manual curation, domain experts are invaluable and they will often interact with the data submitters in performing their task. Future federated and community curation efforts (for example, wiki systems) will need to be carefully managed if they are to match the high standards achieved by current manual curation activities. The responsibility for that will lie, at least partly, with each database in the federated system, although oversight might be applied by stakeholders such as funders and international consortia, and by feedback systems from the community.

Currently, databases try to ensure a high level of data quality, and they know that this is a challenging task[22], but perhaps in the future their obligations should go even further. For example, consistency with other data sources could be assessed, such as comparing SNP allele frequencies with previous data sets to identify fundamental laboratory or data management errors, or to identify cases in which the wrong DNA strand has been referenced. Across the full breadth of G2P data there are many features that could be checked to ensure accuracy, and standards and guidance need to be developed to underpin data curation from generating data to placing it in public databases. Ideally, software support for this will increasingly be provided.

Data complexity. Although data quantity is a matter for concern, it will hopefully be overcome by improvements in data processing algorithms and innovations in computer science. More indomitable, however, will be the matter of data complexity. Biological data, especially the G2P data of the future, differs from that of data from most other 'big sciences' (such as astronomy) by its high level of complexity[23]. Consider phenotypes, for example, studies such as the UK Biobank and the Framingham Heart Study collect thousands of phenotypic variables in a prospective manner, with each item supported by extensive metadata (information that describes the primary data), for tens or hundreds of thousands of subjects. The phenotype definitions used might change as knowledge advances, and patient phenotype categorizations might change with age and treatment. Given this data complexity, standardization of the 'phenotype' parameter is needed, and this is one of the goals of several projects, such as the Public Population Project in Genomics (P3G)[24], the Human Genome Epidemiology Network (HuGENet)[25] and the PhenX project. The information that describes how genotypes connect to phenotypes, that is, the '2' in G2P, is even more complex. A plethora of constantly evolving methods, strategies and analyses that offer varying levels of precision might be used to work out how DNA sequences control phenotypes. The results only provide clues to the underlying etiologic processes, sometimes of a contradictory nature. Environmental effects (such as those considered by the Genes, Environment and Health Initiative), a person's genetic background and chance also feed into this complexity. Thus, the complexity of the analytical methods and the experimental results make it difficult to store G2P information in a structured way, or to fully optimize the integration and presentation systems.

Knowledge representation. As more and more analyses are performed on ever more extensive and cross-domain data sets, it will become increasingly difficult to comprehensively gather and present all the resulting hypotheses and conclusions, a process known as knowledge representation. The issue of how to present conclusions is distinct from the question of what tools and systems are developed to generate those ideas, and how the systems interface with databases. Traditionally, scientific journals have been the principal vehicle for distributing the interpretation of data, but it is not clear whether their current *modus operandi* will enable them to keep pace as the rate of new discoveries continues to grow exponentially. The human-readable narrative format of journals does not easily lend itself to the storage of 'interpretations' and 'concepts' in databases. However, some databases, such as OMIM, the scope of which extends into knowledge representation, do use narrative text. This resource illustrates the value of handling this kind of information beyond journals, but it also shows that it is limiting to store this data without any rigid structure. G2P knowledge representation therefore needs to become better structured and anchored on appropriate ontologies if databases are to be more than just high-tech lists of primary experimental data.

Data access. As G2P data sets become larger and more diverse it will become increasingly difficult to locate any particular data item. To tackle this issue, more powerful tools for database searching will obviously be needed,

Knowledge representation
Structured presentation of information that facilitates the drawing of inferences or conclusions, often giving predictive abilities.

© 2009 Macmillan Publishers Limited. All rights reserved

but it is equally important that those improved search engines are connected to all the relevant data that needs to be searched. Although this could be seen as an argument for widespread adoption of the central database model (detailed above), data size and complexity make it impossible to gather all the information in one central depository. Instead, complementary ways to consolidate the tasks of data access and data presentation across many different databases (for example, LSDBs) are needed, so that the interrogated information never needs to leave its remote source. Such single point of access, or federated, database solutions that tap into large volumes of diverse data are technically feasible. An example from outside the G2P domain is the ENCODEdb portal[26], which offers a simple query interface that searches across all the ENCODE experimental data deposited in several public databases.

Data sensitivity. After the probable phenotypic consequences of carrying a particular sequence variation are established, knowing one's genotype at a given locus becomes meaningful. It also becomes meaningful for one's relatives, who share various fractions of your genome, and it probably is something that health providers, employers, insurers and even governments and the criminal justice system might want to (rightly or wrongly) know about. This raises complex ethical dilemmas[27–30]. Even if genome data are anonymized (as they usually are in epidemiological studies), there is a risk of re-identification of persons based on their genome variation profile, and/or their phenotype and environmental profiles[31]. Exemplifying this, a recent paper has shown that an individual's involvement in a genetic study can be reliably established from just summary level allele frequency data (that is, no individual genotypes) if this is available for two matched sample groups (such as cases and controls in a GWA study), and if the genome profile of the subject of interest is known[32].

A full discussion of the myriad questions surrounding 'data sensitivity' are reviewed in REF. 33, but a few points are worthy of mention. Currently, most databases try to avoid showing sufficient genotype or phenotype information to enable re-identification. This might not, however, be as easy as it seems. For example, LSDBs in which rare mutations or diseases might be reported along with geographical data could be used for re-identification. When G2P data do raise the possibility of re-identification, the current default position is to not make it publicly available, and pass access requests to the original custodians of the information. This stance was immediately adopted for summary level allele frequency data sets, once it became clear that they could be used for individual identification[34]. But when even summary level data cannot be shared for unfettered research access, maybe it is time to start questioning when and where the protection of an individual's privacy becomes overly paranoid or too onerous to implement, given that it detracts from the wider research benefits of making data freely available. If, in reality, it will not be possible to completely ensure the anonymity of all research participants, then perhaps the optimal way forward would be to accept this, to make data more freely available, and concentrate instead on preventing and punishing abuse of the data. Given the existence of such perplexing privacy issues, an ethics advisory voice should arguably be an integral part of every G2P database.

The future: the untapped power of federation

Given the above considerations, federated databases can be expected to play an increasingly large part in the future of G2P information management. Before we discuss that role in more depth, it is useful to reflect on extreme versions of the federated and centralized models (FIG. 1). The fully centralized model would involve all generated information being automatically piped into one large data centre, from which all search and presentation activities are managed. A completely federated solution would involve the information from all the domains being organized into geographically discrete packets (databases), with no regular data flows between them. Global searches would be mediated by portals that scan all available contents, and data presentation would be powered by each database for each item of its own content. Neither of these extremes is a realistic option for the G2P domain, owing to the limitations of each model (see below), so a hybrid model would seem to be the best solution. So far, however, most successful databases have been based on the centralized model. This probably reflects both the newness of a field that only began with the emergence of the Internet, and the fact that the current pressing need for more advanced solutions is relatively recent. Now, as Internet and database technologies rapidly advance, federated systems are emerging alongside and are intermingled with the existing established central databases.

Both central and federated systems have advantages and disadvantages. The main advantages of central databases include cost efficiency, which is due to economies of scale, ease of management and reliable archiving of the community's data. By contrast, federated databases are a more complicated solution in terms of the required technologies, but they bring certain advantages that cannot be endowed by a centralized database. Largely, the advantages concern 'ownership' and accreditation for the database teams, with the potential result that more and higher quality data can be gathered in a federated system, owing to the reward gained by the workers involved. Federated and central database systems both provide centralized search capabilities, although federated alternatives can also offer more sophisticated search options via direct interrogation of the source databases.

Taking into account the challenges facing the G2P field as outlined above, and the pros and cons of each database model, it seems evident that a purely centralized model cannot fulfil all the requirements of an optimal G2P databasing solution, and that some degree of federation is vital to the success of this enterprise. So what would a group of databases need to do to become usefully federated to an optimum degree? The first decision would involve the level of federation to be achieved. Essentially, this equates to deciding what portion of their

ENCODE
(Encyclopedia of DNA Elements). An international research project to identify all functional elements in the human genome.

© 2009 Macmillan Publishers Limited. All rights reserved

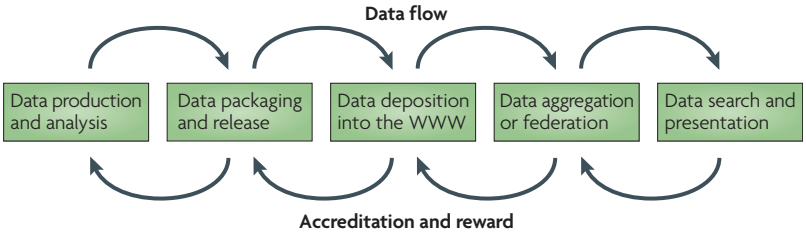


Figure 3 | **Success depends upon recognition and reward.** The value of any future genotype-to-phenotype database network and its supporting infrastructure will be dependent on how effectively the information gets into that system. Individuals responsible for establishing and operating this data flow — from the bench scientist that produces the raw data right through to the people that make the integrated data sets available for searching and access — will all need to be recognized, rewarded and thereby motivated to play their part. Mechanisms for achieving this in the context of databases (as opposed to data publication via journals) are yet to be put in place. WWW, World Wide Web.

content each database would wish to make available for other computers to read over the Internet. They might choose to provide no content, and instead transfer to one or more common search centres some pre-agreed 'core' data elements for each record they hold, along with links back to those entries in their database. The search system would then use that assemblage of minimal data items to enable multi-database searches, and report search results as a series of annotated links pointing back to the source databases. This partly centralized and partly federated solution is being piloted by several closely collaborating initiatives as a way to begin federating LSDBs[35]. Alternatively, and with more effort, the search platform could obtain the full details for each record of interest from the different sources (perhaps even by screen-scraping if necessary) and compile this into a uniform data set for presentation.

A more elegant way of federating would involve making some or all of the record details from each remote database directly searchable by other computers. This approach removes the need for sending in and regularly updating core data sets, thereby ensuring that searches through the central portal always query the latest data sets. It also addresses the scalability problems outlined above, as any new LSDB needs only to register its existence with the central portal to become part of the multi-database search catalogue. Another advantage is that it minimizes the workload of the central search system, as it no longer has to chase up and manage ever-changing core data sets every 24 hours or so. Finally, it alleviates many of the data complexity issues faced by central databases, as each nodal database can provide and customize (at the final display stage) whatever additional record details it deems appropriate above and beyond the common data items made available as part of the federated search. Achieving this 'complete' federation, however, requires all participating databases to accept certain rules. For example, the level of autonomy of each team, in terms of database design, system execution and the degree of association with the rest of the federation, must not be so high as to make the whole federation ineffective. Furthermore, all nodal databases must either adhere to certain standards so that their records can be easily integrated with those of others, or place advanced 'translation' software on top of their database so that search requests and resulting data sets can be freely communicated between remote and local computers.

Finally, certain other advantages of federation are also worthy of specific comment. The first relates to empowering and rewarding database creators (FIG. 3). It takes effort to design, build, fund and continuously manage and curate a database — and it is all too often a thankless task. The federated model, however, places a lot more control and recognition in the hands of those running the individual databases. Federated databases have complete control over what records, and what details per record, are made available to different users at any point in time. This can be important in the case of commercial databases, and it is very important in the context of data sensitivity (as mentioned above). Second, the federated structure distributes data management and curation work among many individuals, making the most of the expert knowledge of these individuals. A third advantage is that the federated structure enables new search portals to be set up quickly and easily, potentially offering unique new perspectives — for example, a gene-centric view for researchers specializing in a single gene, a disease-centric view for clinicians or a genome browser-based view for genomics researchers. Fourth, federated networks operate as democracies, so unilateral changes cannot be imposed on common aspects of the federated system (for example, data models). This does not mean that innovation becomes stifled, but rather that new ideas will be widely debated, piloted and validated before they are implemented.

The G2P database network

Today, the components that are needed to create a powerful and highly integrated system, based on a partially federated and partially centralized model, are either already available or in advanced stages of development. The key missing components that are needed to bring the G2P network to life are expanding technology awareness, establishing recognition and reward systems, and targeting the appropriate allocation of sufficient funding. In Europe, many of these issues are being tackled by initiatives such as the GEN2PHEN project, whereas consortia such as the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) and the European Life Sciences Infrastructure for Biological Information (ELIXIR) are actively planning for the investment of up to several billion Euros into bioscience database and biobanking infrastructures (TABLE 1).

Various technologies, such as web services and ontologies, that will underpin future G2P databasing have been discussed elsewhere[36,23] (see BOX 1 for a summary of the main components). Most importantly, the field will have to become increasingly standardized in order for a global network of G2P databases to interoperate effectively. This standardization concerns syntax and semantics.

Biobanking
Assembling large collections of biosamples and associated information, for the purpose of biomedical investigation.

Syntax
The syntax of information is concerned with how the data is organized, ordered and structured.

Semantics
The semantics of information is concerned with the meaning of the data elements, such as words.

© 2009 Macmillan Publishers Limited. All rights reserved

A core syntax challenge involves designing and validating robust data models for different biomedical domains, so that the models are intercompatible. Typically, these models are also accompanied by standard specifications for data-exchange formats, providing a basis for data exchange between systems. Various existing data models are currently being cross-compared and harmonized to enable more widespread data integration within a research domain, and even across different domains (BOX 2).

Semantic challenges involve ensuring that data items are represented in a way that conveys the same meaning to each and every person (or computer) that reads them. For example, a field named 'sample' might mean 'blood sample' in one database, but 'an individual sampled from a population' in another database. The goal is to structure and specify all of this in ontologies, and to build software and support systems that ensure the terms are used correctly[37]. Semantic standardization is difficult to achieve[38], and tackling this issue across all bioscience subfields involves precisely defining a complete domain lexicon. To break this mammoth task down, researchers are working on ontologies for clearly demarcated subjects (such as 'Gene', 'DNA_sequence' or 'Anatomy'). A large and highly collaborative network of ontology groups has now grown into the Open Biomedical Ontologies (OBO) consortium[39], which is currently tackling what could be the biggest ontology challenge of all: 'Phenotypes'. Encouragingly, much progress has already been made in this area, especially by the MOD communities.

Table 1 | **Genotype-to-phenotype database infrastructure and coordination projects**

Project	Description
BBMRI	European Biobanking and Biomolecular Resources Research Infrastructure. This is a programme by ESFRI, and it aims to create a pan-European research infrastructure for biobanking
caBIG	The Cancer Biomedical Informatics Grid. A data integration network and application infrastructure developed for the cancer research community
CASIMIR	Coordination and Sustainability of International Mouse Informatics Resources. An EU-funded project on coordination and integration of mouse model organism databases
EATRIS	The European Advanced Translational Research Infrastructure in Medicine. An ESFRI project aimed at translating research findings into improved diagnosis, disease prevention and treatment
ECRIN	European Clinical Research Infrastructures Network. An ESFRI programme aimed at integrating national clinical research facilities into a pan-European infrastructure
EGEE	Enabling Grids for E-sciencE. A large EU-funded multidisciplinary infrastructure project
ELIXIR	European Life-Science Infrastructure for Biological Information. An EU-funded ESFRI bioinformatics infrastructure programme for life-science research
EMBRACE	European Model for Bioinformatics Research and Community Education. Collaboration network in the area of Grid computing and databasing in biomolecular research
ESFRI	The European Strategy Forum on Research Infrastructures is an EU-funded framework for developing scientific infrastructures in Europe
EuroGenetest	An EU-funded Network of Excellence fostering standardization and harmonization of genetic testing across Europe, including informatics, ethics, new technologies, education and quality management
GEN2PHEN	An EU-funded project aiming to unify human and model organism genotype to phenotype databases in a drive towards increasingly holistic views into this information
GMOD	Generic Model Organism Database project. A collection of open-source software tools for creating and managing genome-scale biological databases
HL7	Health Level 7. A standardization organization operating in the health-care arena
HuGENet	Human Genome Epidemiology Network. An international collaboration of individuals and organizations in the field of genetic epidemiology
HVP	Human Variation Project. An open organization that is helping to catalogue all human Mendelian genetic variation, making that information freely available to researchers, clinicians and patients worldwide
MIBBI	Minimum Information for Biological and Biomedical Investigations[52]. A project and web resource promoting the development and use of minimum information specifications and checklists
Obiba	An open-source project that aims to build open-source software infrastructure applications and software components for biobanking. One of the P3G core projects
OBO	The Open Biomedical Ontologies. An international community for supporting the development and use of ontologies in the biomedical domain
OpenEHR	A data standards and modelling framework for managing electronic health-care data
P3G	The Public Population Project in Genomics. An international consortium promoting collaboration between population genomics researchers. The P3G observatory provides a central repository of relevant tools and information

EU, European Union.

© 2009 Macmillan Publishers Limited. All rights reserved

Box 1 | **Technologies in genotype-to-phenotype databasing**

Databases and database networks will be key to organizing, storing and providing access to the wealth of biomedical data already produced and yet to be generated. The task of building the necessary databases is primarily a technological construction effort, as the required technology solutions are already well developed or are at least identified in principle. Some of the core concepts behind these technologies are outlined below.

Object (data) model
A formalized conceptualization of how data elements, or objects, are structured and organized, and how they are connected to other data elements. This might include semantic information on those objects and connections, by way of references to ontologies (see below). An example is the Microarray Gene Expression Object Model[46] (MAGE-OM), which standardizes the representation of microarray information, spanning experiment design and data.

Exchange format
The specifications of the syntax, or physical representation, of data complying with the model. This is essential for unambiguous transmission of data between computers. Examples range from the simple FASTA format used to exchange DNA- and protein-sequence data, to the elaborate XML-based Microarray Gene Expression Markup Language (MAGE-ML) for MAGE-OM-compliant microarray data.

Ontology
A controlled vocabulary of terms for concepts, including their meaning and well-defined relationships between them. Ontologies enable the representation of domain-specific knowledge and, when used properly, make database searches far more powerful. Examples include Gene Ontology[47] (GO) for annotating gene products from many species, and Functional Genomics Investigation Ontology (FuGO)[48] for functional genomics investigations.

Globally-unique identifier (GUID)
A digital object identifier, which is guaranteed to be unique and persistent across the intended usage domain. GUIDs solve data integration problems that result from ambiguity in names, or identity, of biological concepts and objects, such as genes and proteins. GUIDs are key components of semantic web technologies, such as the Resource Description Framework (RDF). Examples currently being evaluated include the Persistent Uniform Resource Locator (PURL) and Life Science Identifiers[49] (LSIDs).

Web services
A series of standard protocols for facilitating machine-to-machine interaction over the Internet. Web services simplify the task of 'plumbing together' distributed data retrieval or analysis services over the network, forming the basis of the service-oriented architecture (SOA). An example of a service-oriented 'Grid' is provided by the National Cancer Institute's caGrid[50].

As syntactic and semantic standards start to fall into place, more groups are likely to start considering building federated databases. Technologies will then be required that can broadcast, deliver, receive or interrogate (locally or remotely) the available data sets. An early example of one such technology is the Distributed Annotation System (DAS)[40] — a simple protocol for exchanging annotations on genomic sequences. Many databases already make their records available via their own DAS server to DAS clients such as the Ensembl browser. Those third-party data sets are then overlaid on other DAS-supplied information or locally available annotations, such as reference sets of genes, demonstrating the power of a federated system.

Hopefully, as databases, search platforms, visualization interfaces and analysis tools start to become fully federated and seamlessly joined together, the system will be transformed from a collection of cleverly connected data into a universal G2P 'knowledge environment' — a place in which new questions can be asked, new types of experiments can be performed *in silico* and new knowledge can be created. That, at least, is the vision of the semantic web[41], the proponents of which claim it will comprise a highly sophisticated and powerfully connected series of servers and client computers (the 'Grid') that together provide highly automated data retrieval and analysis 'web services' across the Internet. When this Grid-enabled G2P knowledge environment becomes a reality, we might find ourselves in a situation in which the distinction between database entries and research manuscripts has become blurred, and new paradigms like web publishing and real-time community mark-up of databased information are commonplace. Initial forays into this world of merged 'database–journal' publication vehicles are already taking place, one example being a partnership between the *Human Genomics and Proteomics* journal and the FINDbase database. Indeed, traditional journals might become a thing of the past, as they evolve to become an integral part of this unified Grid of biomedical knowledge.

Future challenges and opportunities

Despite the optimistic future of G2P databases, a number of basic issues, mostly non-technical ones, remain to be solved.

One issue involves getting enough people sufficiently well trained in the relevant technologies to build and connect all of the G2P databases. To achieve this, educators must decide to fund and organize such training. In parallel, software engineers can drastically reduce the level of competence required of the users by devising off-the-shelf solutions. This philosophy lies at the heart of the GEN2PHEN project, which is producing empty 'database-in-a-box' installation packages along with training, and open-source complete genetic association database systems for download — both of which are based upon the Phenotype and Genotype Experiment Object Model (PaGE-OM) data model, giving them the option of federation.

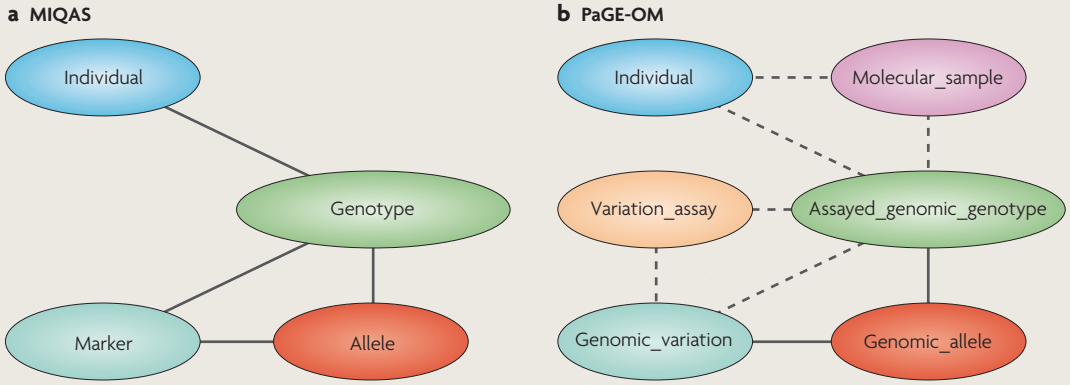
Another issue is the question of tracking who is building G2P databases and populating them with useful data. Several initiatives have recently been launched to look into this. For example, the idea of 'microattribution' has been proposed, whereby database systems would track the interest in each database entry[42], record this interest in relation to the original submitter of the entry, and thereby steadily assemble a metric for the value of each person's database contributions. Similarly, database creators would be able to extrapolate from this kind of information something akin to a journal impact factor. This will require cooperation of journals, database creators and funders, who would all need to use an agreed tracking system. Additionally, success would be dependent on there being a way to assign globally unique identifiers (GUIDs) to individual data packets (for example, database records) on the Internet. Because the semantic web will also need such identifiers (not only for data,

Semantic web
An extension of the World Wide Web that embeds semantics, or meaning, in documents, in links between documents and in descriptions of web services, thereby enabling navigation and reasoning by automated agents.

Genetic association database
A catalogue of reported genetic associations between genotype and phenotype.

© 2009 Macmillan Publishers Limited. All rights reserved

Box 2 | **Harmonizing data models**



Data models can often be aligned or 'mapped' to each other to identify similarities and differences. Once this is done, and equivalent concepts and relationships thereby identified, it is then possible to specify a consensus model and/or derive a data-exchange format with which both models will be compatible.

This can be demonstrated with subportions of two related data models, the Minimum Information for QTLs and Association Studies specification (MIQAS), and the Phenotype and Genotype Experiment Object Model (PaGE-OM) (see the figure). Equivalent entities in these models are in some cases named differently (for example, 'Marker' and 'Genomic_variation'), and so to highlight the corresponding item pairs they have been placed in the same relative positions in the diagrams and are shown in the same colour. Naming discrepancies are problematic, especially when the same name is used to mean different things between models (for example, 'Sample' for a person or for a reagent). Such confusion is eliminated when models include semantic information on their components, by references to ontologies. Solid lines indicate relationships (for example, a 'Marker' has an 'Allele'), and the dotted lines in the PaGE-OM indicate that relationships can be optional to allow for data elements that might not be specified (for example, the 'Variation_assay' used in a genotyping experiment). On the basis of such mapping diagrams, data-exchange formats can be specified that support only the common components from the models, or they can be extended to include some non-common items, whereupon those data elements would be declared optional.

Similar intermodel mapping can be done between subdomains, in which the underlying models might be quite different, so long as the models have at least some common concepts or attributes. Data sets produced according to those models can then be connected together, rather than fully merged, by linking through those common fields on data rows in which the values are identical (for example, a SNP-marker identifier), as explained by Wang *et al.*[51].

but also for services, concepts, metadata and so on), several solutions to this problem are now being evaluated. An alternative approach to accreditation has been proposed — the Bio-Resource Impact Factor (BRIF)[43], the scope of which would include G2P databases. BRIF is more directly akin to the journal impact factor. As journals might be overtaken steadily by databases as the preferred means for getting data into the public domain, BRIF, or something like it, will be needed to demonstrate researchers' productivity and the importance of their work. It will also help in making it evident to funding bodies that database creation efforts are worthy of support — a message that currently needs some re-enforcement[44].

Considerable sums of money are being spent on G2P research to improve our understanding of health and disease so that medical care can advance. It follows, therefore, that the G2P databases should remain tightly focused on the needs of the medical community. The problem, however, is that although imprecise knowledge and uncertain data are an essential part of research, the clinical world requires more straightforward, reproducible information upon which to base its decision making. Perhaps one of the real unmet challenges for G2P databases is that of distilling from basic research data the firm conclusions and predictions that would help physicians diagnose a patient. Close attention to how this information is presented is essential, as researchers and clinicians typically have very different expectations when searching for information. These important issues are well known to those who work at the interface of bioinformatics and medical informatics. Examples of key efforts designed to close the gap between the two include the development of electronic health-care records (EHRs), and the work of the Health Level Seven (HL7) organization towards genomics standards in medical information.

In summary, the field of G2P databasing is at a significant stage in its development, taking into account lessons from the past, and being challenged by the exponentially growing and rapidly changing data sets of the present and the future. This Review discusses the technical solutions and the logical ways forward for the field, all of which are supported by extensive open-source collaboration and contribute to the emerging "cyberinfrastructure for the biological sciences"[45]. We can therefore expect G2P databasing to advance significantly in the near future, enabling research and clinical practitioners to make the best use of the wealth of G2P data now being generated.

© 2009 Macmillan Publishers Limited. All rights reserved

1. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).
2. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
3. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
4. Stein, L. Creating a bioinformatics nation. *Nature* **417**, 119–120 (2002).
5. Miyazaki, S. *et al.* DDBJ in the stream of various biological data. *Nucleic Acids Res.* **32**, D31–D34 (2004).
6. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **36**, D25–D30 (2008).
7. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33**, D29–D33 (2005).
8. Chen, N. *et al.* WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.* **33**, D383–D389 (2005).
9. Twigger, S. N. *et al.* The Rat Genome Database, update 2007 — easing the path from disease to data and back again. *Nucleic Acids Res.* **35**, D658–D662 (2007).
10. Bult, C. J. *et al.* The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* **36**, D724–D728 (2008).
11. Hamosh, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
12. McKusick, V. A. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, 1966).
13. Ball, E. V. *et al.* Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.* **26**, 205–213 (2005).
14. Altman, R. B. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nature Genet.* **39**, 426–426 (2007).
15. Lehmann, H. & Kynoch, P. A. M. *Human Haemoglobin Variants and Their Characteristics* (North-Holland Publishing, Amsterdam, 1976).
16. Horaitis, O. *et al.* A database of locus-specific databases. *Nature Genet.* **39**, 425 (2007).
17. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
18. Becker, K. G. *et al.* The Genetic Association Database. *Nature Genet.* **36**, 431–432 (2004).
19. Bertram, L. *et al.* Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genet.* **39**, 17–23 (2007).
20. Allen, N. C. *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genet.* **40**, 827–834 (2008).
21. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
22. Howe, D. *et al.* Big data: the future of biocuration. *Nature* **455**, 47–50 (2008).
23. Goble, C. & Stevens, R. State of the nation in data integration for bioinformatics. *J. Biomed. Inform.* **41**, 687–693 (2008).
This paper describes many of the technologies and challenges in data integration; in particular, different methods ranging from 'heavyweight' data warehousing approaches to loose-touch data 'mashups'.
24. Knoppers, B. *et al.* Population Genomics: The Public Population Project in Genomics (P3G): a proof of concept? *Eur. J. Hum. Genet.* **16**, 664–665 (2008).
25. Ioannidis, J. P. A. *et al.* A road map for efficient and reliable human genome epidemiology. *Nature Genet.* **38**, 3–5 (2006).
26. Elnitski, L. L. *et al.* The ENCODEdb portal: simplified access to ENCODE Consortium data. *Genome Res.* **17**, 954–959 (2007).
27. Hoyweghen, I. V. & Horstman, K. European practices of genetic information and insurance: lessons for the Genetic Information Nondiscrimination Act. *JAMA* **300**, 326–327 (2008).
28. Diergaarde, B. *et al.* Genetic information: special or not? Responses from focus groups with members of a health maintenance organization. *Am. J. Med. Genet. A* **143**, 564–569 (2007).
29. Gilbar, R. Patient autonomy and relatives' right to know genetic information. *Med. Law* **26**, 677–697 (2007).
30. Knoppers, B. M. *et al.* The emergence of an ethical duty to disclose genetic research results: international perspectives. *Eur. J. Hum. Genet.* **14**, 1170–1178 (2006).
31. Godard, B. *et al.* Data storage and DNA banking for biomedical research: informed consent, confidentiality, quality issues, ownership, return of benefits. A professional perspective. *Eur. J. Hum. Genet.* **11** (Suppl. 2), S88–S122 (2003).
32. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
33. Cambon-Thomsen, A., Rial-Sebbag, E. & Knoppers, B. M. Trends in ethical and legal frameworks for the use of human biobanks. *Eur. Respir. J.* **30**, 373–382 (2007).
34. Zerhouni, E. A. & Nabel, E. G. Protecting aggregate genomic data. *Science* **322**, 44 (2008).
35. Giardine, B. *et al.* PhenCode: connecting ENCODE data with mutations and phenotype. *Hum. Mutat.* **28**, 554–562 (2007).
36. Stein, L. D. Integrating biological databases. *Nature Rev. Genet.* **4**, 337–345 (2003).
37. Stevens, R., Goble, C. A. & Bechhofer, S. Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.* **1**, 398–414 (2000).
38. Quackenbush, J. Standardizing the standards. *Mol. Syst. Biol.* **2**, 2006.0010 (2006).
39. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnol.* **25**, 1251–1255 (2007).
40. Dowell, R. D. *et al.* The Distributed Annotation System. *BMC Bioinformatics* **2**, 7 (2001).
41. Berners-Lee, T., Hendler, J. & Lassila, O. The Semantic Web — a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci. Am.* **284**, 34–43 (2001).
42. Compete, collaborate, compel [Editorial]. *Nature Genet.* **39**, 931 (2007).
43. Kauffmann, F. & Cambon-Thomsen, A. Tracing biological collections: between books and clinical trials. *JAMA* **299**, 2316–2318 (2008).
44. Merali, Z. & Giles, J. Databases in peril. *Nature* **435**, 1010–1011 (2005).
45. Stein, L. D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Rev. Genet.* **9**, 678–688 (2008).
This is a recent comprehensive review of current and emerging components of informatics infrastructure for modern biological research.
46. Spellman, P. T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, research0046.1–00469 (2002).
47. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
48. Jones, A. R. *et al.* The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotechnol.* **25**, 1127–1133 (2007).
49. Clark, T., Martin, S. & Liefeld, T. Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* **5**, 59–70 (2004).
50. Saltz, J. *et al.* caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics* **22**, 1910–1916 (2006).
51. Wang, X., Gorlitsky, R. & Almeida, J. S. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nature Biotechnol.* **23**, 1099–1103 (2005).
This paper describes the potential of semantic web standards and technologies for describing and integrating biological data.
52. Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnol.* **26**, 889–896 (2008).

Acknowledgements

The authors acknowledge the valuable ideas, advice and funding provided by the GEN2PHEN project as part of the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754, which enabled the preparation of this Review.

FURTHER INFORMATION

AlzGene: http://www.alzforum.org/res/com/gen/alzgene
Biobanking and Biomolecular Resources Research Infrastructure (BBMRI): http://www.biobanks.eu
Cancer Biomedical Informatics Grid (caBIG): https://cabig.nci.nih.gov
Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR): http://www.casimir.org.uk
dbGaP: http://view.ncbi.nlm.nih.gov/dbgap
Enabling Grids for E-sciencE (EGEE): http://www.eu-egee.org
ENCODEdb: http://research.nhgri.nih.gov/ENCODEdb
EuroGenetest: http://www.eurogentest.org
European Advanced Translational Research Infrastructure in Medicine (EATRIS): http://www.eatris.eu
European Biobanking and Biomolecular Resources Research Infrastructure (BBMRI): http://www.bbmri.eu
European Clinical Research Infrastructures Network (ECRIN): http://www.ecrin.org
European Genotype Archive (EGA): http://www.ebi.ac.uk/ega
European Life Sciences Infrastructure for Biological Information (ELIXIR): http://www.elixir-europe.org
European Model for Bioinformatics Research and Community Education (EMBRACE): http://www.embracegrid.info
European Network of Genomic and Genetic Epidemiology (ENGAGE): http://www.euengage.org
European Strategy Forum on Research Infrastructures (ESFRI): http://cordis.europa.eu/esfri
FINDbase: http://www.findbase.org
Framingham Heart Study: http://www.framinghamheartstudy.org
GEN2PHEN project: http://www.gen2phen.org
Generic Model Organism Database (GMOD): http://www.gmod.org
Genes, Environment and Health Initiative: http://www.gei.nih.gov
Genetic Association Database (GAD): http://geneticassociationdb.nih.gov
GenomEUtwin: http://www.genomeutwin.org
GWAS Database, Japan: https://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi
Health Level Seven (HL7): http://www.hl7.org
HGVbaseG2P: http://www.hgvbaseg2p.org
Human Gene Mutation Database (HGMD): http://www.biobase-international.com/pages/index.php?id=hgmddatabase
Human Genome Epidemiology Network (HuGENet): http://www.cdc.gov/genomics/hugenet
Human Genome Variation Society: http://www.hgvs.org/dblist/glsdb.html
***Human Genomics and Proteomics* journal:** http://www.sage-hindawi.com/journals/hgp
Human Variation Project (HVP): http://www.humanvariomeproject.org
International Nucleotide Sequence Database Collaboration (INSDC): http://www.insdc.org
Minimum Information for Biological and Biomedical Investigations (MIBBI): http://www.mibbi.org
Minimum Information for QTLs and Association Studies specification (MIQAS): http://miqas.sourceforge.net
Obiba: http://www.obiba.org
Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
Open Biomedical Ontologies (OBO): http://obofoundry.org
OpenEHR: http://www.openehr.org
PDGene: http://www.pdgene.org
Persistent Uniform Resource Locator (PURL): http://www.purl.org
Pharmacogenetics and Pharmacogenoics Knowledge Base (PharmGKB): http://www.pharmgkb.org
Phenotype and Genotype Experiment Object Model (PaGE-OM): http://www.omg.org/spec/PAGE-OM
PhenX: http://www.phenx.org
Public Population Project in Genomics (P3G): http://www.p3gconsortium.org
Public Population Project in Genomics observatory: http://www.p3gobservatory.org
Resource Description Framework (RDF): http://www.w3.org/RDF
Service-oriented architecture (SOA): http://www.w3.org/TR/2003/WD-ws-arch-20030514
SZGene: http://www.schizophreniaforum.org/res/sczgene
Type 1 Diabetes Genetics Consortium: http://www.t1dgc.org
UK Biobank: http://www.ukbiobank.ac.uk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2009 Macmillan Publishers Limited. All rights reserved