

**Database federation, resource interoperability
and digital identity, for management and
exploitation of contemporary biological data**

Thesis submitted for the degree of Doctor of
Philosophy at the University of Leicester

Gudmundur A. Thorisson

Department of Genetics

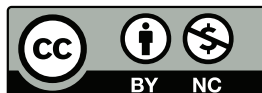
gt50@leicester.ac.uk

November 23, 2010



**University of
Leicester**

© Copyright 2010, Gudmundur A. Thorisson. Some rights reserved.



**This work is licensed to the public under the Creative Commons
Attribution Non-Commercial 3.0 Unported License.
<http://creativecommons.org/licenses/by-nc/3.0/>**

Formatted for L ^A T _E X by MultiMarkdown

Abstract

Database federation, resource interoperability and digital identity, for management and exploitation of contemporary biological data.

Gudmundur A. Thorisson

Modern research into the genetic basis of human health and disease is increasingly dominated by high-throughput experimentation and routine generation of large volumes of complex genotype to phenotype (G2P) information. Efforts to effectively manage, integrate, analyse and interpret this wealth of data face substantial challenges. This thesis discusses informatics approaches to addressing some of these challenges, primarily in the context of disease genetics.

The genome-wide association study (GWAS) is widely used in the field, but translation of findings into scientific knowledge is hampered by heterogeneous and incomplete reporting, restrictions on sharing of primary data, publication bias and other factors. The central focus of the work was design and implementation of a core informatics infrastructure for centralised gathering and presentation of GWAS results. The resulting open-access HGVbaseG2P genetic association database and web-based tools for search, retrieval and graphical genome viewing increase overall usefulness of published GWAS findings. HGVbaseG2P conceptual modelling activities were also merged into a collaborative standardisation effort with international partners. A key outcome of this joint work is a minimal model for phenotype data which, together with ontologies and other standards, lays the foundation for a federated network of semantically and syntactically interoperable, distributed G2P databases.

Attempts to gather complete aggregate representations of primary GWAS data into HGVbaseG2P were largely unsuccessful, chiefly due to concerns over re-identification of study participants. This led to a separate line of inquiry which explored - via in-depth field analysis, workshop organisation and other community outreach activities - potential applications of federated identity technologies for unambiguously identifying researchers online. Results suggest two broad use cases for user-centric researcher identities - i) practical, streamlined data access management and ii) tracking digital contributions for the purpose of attribution - which are critical to facilitating and incentivising sharing of GWAS (and other) research data.

Acknowledgements

First of all I would like to thank my supervisor Tony for taking me on as a student those four, long (and yet short) years ago, and for sticking with me through the good times and the bad, right until the end of the journey and past the finish line. As I have learned a great deal from you, I hope you have learned perhaps a little bit from me too.

I also want to thank Owen, Rob Hastings, Rob Free and Adam in the bioinformatics group in Room 210 Adrian building, who have been with me on much of this journey. Towards the end of the project, later additions to the group, Sirisha and Tim, also provided support in the form of stimulating discussions on various topics. Others in the Brookes group residing downstairs in Adrian also deserve thanks, including my fellow Phd student Katherine, and also Colin, Reshma and others, of whom I have seen far too little during those four years.

Special thanks go to the author of Scrivener¹, the crucial piece of software which has proved invaluable to me in the writing phase, and to the entire L^AT_EX² open-source community for creating and maintaining a wonderful system for preparing and typesetting very large documents.

Last, but not least, I want to thank my lovely wife Lýdía and my children Dagbjört and Ísak for agreeing to move with me from Iceland to start a new life in a new country (again!). Your support throughout this project was tremendously important to me, especially during those last few months of writing in winter and spring 2010 when you did not see a great deal of me.

Leicester, England, 22nd of November 2010

¹<http://literatureandlatte.com/scrivener.html>

²<http://www.latex-project.org/>

Contents

Abstract	iii
Acknowledgements	iv
1 Introduction	1
1.1 Thesis aims and organisation	2
1.2 Genetic variation and studies of human disease	3
1.2.1 Hunting for disease genes	4
1.2.2 Positional cloning and linkage analysis	5
1.2.3 Common diseases, common variants	6
1.2.4 Genome-wide associations studies	7
1.2.5 GWAS success, criticism and the missing heritability	9
1.2.6 Sequencing	12
2 Genotype to phenotype databases: challenges and opportunities	15
2.1 The “omics” data problem	17
2.2 A survey of G2P databases	19
2.2.1 Genetic sequence archives	20
2.2.2 Core genomic variation databases	22
2.2.3 Databases for model organisms	28
2.2.4 Association study catalogs and knowledgebases	37
2.2.5 Databases for phenotype-associated mutations in humans	40
2.2.6 Disease-specific portals	46
2.2.7 Archives for primary and aggregate G2P data	47
2.3 Challenges for modern G2P databases	51
2.3.1 Data quantity	52
2.3.2 Data complexity	54
2.3.3 Data quality	55
2.3.4 Data discovery	58
2.3.5 Data access	59
2.3.6 Knowledge representation	62
2.4 The untapped power of federation	64
2.5 Building blocks of a federated G2P database network	68
2.6 The critical role of standards	73

2.6.1	Domain standards and dynamic software infrastructure . . .	74
2.6.2	Linked Data and the Semantic Web	78
2.7	Distributed grids in the life sciences	81
2.7.1	Biomedical grids for cancer research and neuroscience . . .	82
2.7.2	Distributed web services and workflows	84
2.7.3	Towards a holistic G2P knowledge environment	85
3	Designing a data model for genetic association studies	89
3.1	The HGVbaseG2P model	91
3.1.1	The SAMPLE domain	94
3.1.2	The GENOTYPE domain	95
3.1.3	The SEQUENCE domain	97
3.1.4	The PHENOTYPE domain	99
3.1.5	The EXPERIMENT domain	101
3.1.6	The COMMON domain	104
3.2	Validating the model	106
3.2.1	Marker data from dbSNP	107
3.2.2	Study organisation and metadata from CGEMS	109
3.2.3	Marker and population data from DGV	112
3.2.4	Phenotype data and metadata from dbGaP	114
3.3	Discussion	118
3.3.1	The HGVbaseG2P model	119
3.3.2	Future development: formalising the model	122
4	Comparing related G2P domain data models	126
4.1	Aligning with FuGE-OM	127
4.2	Aligning with XGAP	131
4.3	Pheno-OM: creating an improved object model for phenotypes . .	134
4.3.1	Describing protocols and provenance	135
4.3.2	Describing features	137
4.3.3	Describing the outcome of observations	138
4.4	Discussion	139
4.4.1	HGVbaseG2P versus FuGE-OM and XGAP	140
4.4.2	Pheno-OM and “pluggable” G2P information models	142
5	Creating a federated database of genetic association studies	147
5.1	Constructing the HGVbaseG2P system	148
5.1.1	System overview	150
5.1.2	Database design and content	150

5.1.3	Database APIs	158
5.1.4	Software tools for data transformation and loading	164
5.1.5	Web-based applications for online access to database contents	167
5.2	Using the HGVbaseG2P website toolkit	173
5.3	Gathering reference variation data for HGVbaseG2P	178
5.4	Gathering association study data for HGVbaseG2P	183
5.5	Discussion	188
5.5.1	Core informatics infrastructure	188
5.5.2	Website features and usability	190
5.5.3	Data content and usefulness	195
5.5.4	Towards a model-driven, semantics-enabled architecture	197
6	Exploring the role of digital identity in data publication	200
6.1	Identity on the Internet	201
6.2	Federated identity technologies	206
6.3	A digital identity for researchers	217
6.3.1	Attribution and accreditation for scientific contributions	219
6.3.2	Managing access to G2P data	226
6.4	Use cases for identity-enabled access management and attribution	233
6.4.1	Basic access management with local user IDs	233
6.4.2	Access management with digital IDs	234
6.4.3	Access management with digital IDs and semantic authorisation	235
6.4.4	Attribution with ORCID IDs and DOIs	237
6.5	Discussion	238
6.5.1	Federated identity and research	238
6.5.2	Identity-based use cases for Cafe RouGE	241
6.5.3	Cafe RouGE as an experimental platform for digital IDs	244
7	Final summary, conclusions and future work	246
7.1	The HGVbaseG2P project	247
7.1.1	Informatics infrastructure	247
7.1.2	A global, semantically enhanced GWAS catalog	248
7.1.3	Centralised GWAS data gathering and unpublished studies	250
7.2	G2P data models	252
7.2.1	An implementation model for genetic associations studies	252
7.2.2	A minimal data model for phenotypes	253
7.2.3	A modular G2P object model architecture	255
7.3	The role of digital identity in research	257
7.3.1	A scholarly identity and contributor recognition	257

7.3.2	Identity-based security in distributed G2P data sharing . . .	260
7.4	Future perspective	262
263section*.192		
Appendices		265
A	Research methods	266
A.1	Technical specification	266
A.2	Processing and loading marker data with dbSNP-lite	268
A.3	XML-based data loading tools	281
B	Supplementary materials for modelling chapters	287
B.1	PaGE-OM logical model diagrams	287
B.2	HGVbaseG2P object class definitions	299
C	Supplementary materials for identity chapter	302
C.1	Use cases for an identity-enabled Cafe RouGE system	302
D	DVD-ROM contents	323
References		324

1. Introduction

For much of the 20th century, studies of single genes and their role in disease aetiology, or medical genetics, received little acknowledgement by medical practitioners, despite having for some time played “a large role in the health care of a few patients and a small role in the health care of many” (Guttmacher and Collins, 2002). With the completion of the draft human genome sequence in 2001 (The International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), molecular biology entered what has been called the *genomic era* (Guttmacher and Collins, 2003). Traditional medical genetics is rapidly making way for the broader, emerging field of genomic medicine, which is concerned with translating knowledge of the complete human genome and the molecular basis of disease into clinical practice, through disease prevention, intervention, better drugs, genetic therapy, prediction of susceptibility and personalised health care (Guttmacher and Collins, 2005; Sander, 2000).

Genomic medicine is expected by many to improve human health, and consequently large amounts of public and private funding are being invested in attempts to advance the field. This has fuelled major research programmes focusing on the first step of the process: investigating the correlation between genotype and phenotype in human and model organisms for human disease. These genotype to phenotype (G2P) investigations increasingly involve high-throughput experimental technologies and large numbers of biological samples, and generate large volumes of complex data which need to be analysed, interpreted and integrated with other data in order to create knowledge (Louie *et al.*, 2007). For this to happen in an optimised way and ensure that valuable research data are fully exploited, effective database strategies and technologies are required. It is this *databasing* aspect of G2P research that is the thrust of this thesis.

1.1 Thesis aims and organisation

The work presented in this thesis is focused on the G2P databasing challenges noted above, specifically in the context of disease genetics and genome-wide association investigations.

The overall aims of the work are as follows:

- To develop conceptual data models, with a focus on data generated by, and metadata describing, genetic association studies. This will address the general need for data standardisation in the field and help facilitate integration and exchange of this information.
- To develop online database systems for disseminating and managing association data, and to explore federation as a strategy for connecting many distributed G2P databases on the Internet. This will address acute problems with accessibility of data from published and unpublished genetic association studies.
- To explore the potential role of digital identity on the Internet as a means to streamline existing G2P research activities on the Web and enable new ones. Particular areas studied include the use of federated identity technologies to manage access to sensitive biomedical data, and as a basis for incentive-based accreditation schemes for database submissions and other digital contributions.

The overall structure of the thesis takes the form of seven chapters. In order to contextualise the work and provide background, the rest of this introductory chapter briefly introduces the history of medical genetics leading up to the modern era of high-throughput genotyping and next-generation sequencing technologies. **Chapter 2** provides an in-depth survey of mainstream databases of G2P information, identifies several important challenges facing such databases and discusses database federation and technologies underpinning such distributed database networks as a key strategy for tackling these challenges. **Chapters 3 and 4** describe work carried out to develop data models for G2P information, whilst **Chapter 5** reports on the construction of a centralised genetic association database based

on some of those models. **Chapter 6** introduces concepts and technologies relating to identity on the Internet and presents results from an exploratory study into the potential role of digital identity in G2P research. Finally, **Chapter 7** gives an overall summary of findings and conclusions of the preceding chapters, and also discusses the relevance of the work to the field and potential future avenues of research. Appendices contain details on research methods, data model definitions and other supplementary information referenced as appropriate in the results chapters.

1.2 Genetic variation and studies of human disease

An organism's *phenotype* arises from a complex interplay between an organism's genetic makeup and environmental factors. This interplay has long been of scientific interest as we strive to understand the natural world and evolution by studying the genomes of organisms, including our own. Though humans vary considerably in weight, height, colour and other traits, our genomes are remarkably similar compared to the genetic diversity of many other species. This, in part, is the result of our evolutionary history (Reich and Goldstein, 1998). Early estimates based on genomic and cDNA gene sequence data indicated that two randomly chosen human genomes will differ on average by about 1 basepair out of every 1000, or merely 0.1% (Li and Sadler, 1991). A decade later, Kruglyak and Nickerson (2001) added that ~ 11 million sites out of ~ 3 billion, or less than 0.4%, could be expected to vary in at least 1% of the population, based on the extent of single nucleotide polymorphism (SNPs) in the human genome. Subsequently, larger scale, complex "structural variation" has been recognised as a major contributor to human genetic diversity, and more recent estimates are at least an order of magnitude higher, or up to 12% by some estimates (Redon *et al.*, 2006).

Many human diseases have a substantial genetic component and such diseases often run in families. Motivated by this notion, early 20th century pioneers like Garrod and Bateson investigated hereditary pathological traits such as chemical abnormalities in urine (Cox, 1999). The actual molecular basis for inherited disease would not be known for several

more decades. In the 1980s, deletions in the alpha-globin gene cluster on chromosome 16 were found to be the main cause of the serious disease alpha-thalassemia (Higgs *et al.*, 1989). According to the Human Genome Epidemiology (HuGE) Watch¹ (Yu *et al.*, 2008c), in only the past decade over 45,000 publications described population-based studies of genes associated with human diseases. This immense literature reflects a great scientific interest in *genetic epidemiology*, which Burton *et al.* (2005) describe as a “discipline closely allied to traditional epidemiology that focuses on the familial, and in particular genetic, determinants of disease and the joint effects of genes and non-genetic determinants”.

1.2.1 Hunting for disease genes

In the past twenty years, thousands of genes have been implicated in human disease. Most of the ~2,000 diseases listed in the Online Mendelian Inheritance In Man (OMIM) catalog² (Amberger *et al.*, 2009) are single-gene, or monogenic, “Mendelian” disorders - so called because of their simple familial inheritance patterns. Mendelian disorders are caused by one or a few highly-penetrant mutations present in affected individuals. For example, Huntington’s disease, a neurodegenerative disorder caused by variable-length, expanded trinucleotide repeat in the HT gene, affects only 5-7 in 100,000 white people, with a lower incidence in non-white populations (Walker, 2007). Many mutations implicated in Mendelian disorders, however, are common - the average person is believed to be a heterozygous, unaffected carrier of ~25 such alleles - but as these mutations tend to be recessive, the disorder manifests itself only in homozygotes who have inherited two mutant alleles. The incidence of these diseases is therefore low and they tend to be persistent in the population. For example, cystic fibrosis (CF), one of the first Mendelian disorders to be characterised (Kerem *et al.*, 1989; Riordan *et al.*, 1989; Rommens *et al.*, 1989), is caused

¹<http://hugenavigator.net/HuGENavigator/startPageWatch.do>

²<http://www.ncbi.nlm.nih.gov/omim/>

by mutations in the CFTR gene on chromosome 7. CF affects approximately one person in 3,000, and most other monogenic disorders are much less common.

1.2.2 Positional cloning and linkage analysis

The majority of currently known disease genes were discovered following advances in experimental techniques in the 1980s, which enabled isolation, sequencing and functional characterisation of many human genes and their mutations via an approach known as positional cloning. The discovery of restriction-fragment length polymorphism (RFLP) markers (Botstein *et al.*, 1980) and subsequent identification of far more abundant, highly-polymorphic variable-length mini- and micro-satellite repeats (Weber and May, 1989) facilitated construction of dense genetic maps of polymorphic markers. This enabled linkage studies, a strategy which involves typing up to several thousand markers across the genome in family pedigrees, and subsequently analysing the segregation patterns to find loci that co-segregate with the disease, thus revealing the genetic interval likely to harbour the disease-causing mutation.

Linkage analysis was a breakthrough compared to previous candidate-gene based methods, whereby only one or several known genes are investigated based on some *a priori* hypothesis of their involvement in the disease (Hirschhorn and Daly, 2005). As only a fraction of the possible risk factors or causal variants are tested, success of the candidate-gene method depends entirely on correctly predicting the candidate genes (Tabor *et al.*, 2002). Linkage analysis, on the other hand, provided a way to scan the whole genome, without any preconceived notion of where the disease-associated gene or genes are located or the biochemical nature of the mutation. As linkage analysis relies solely on a genetic map - i.e. genetic distance between markers - as a frame of reference, the method could be applied to interrogate the whole genome at a time when sequencing even a single gene was very expensive and labour-intensive, let alone sequencing the whole genome.

High-throughput genotyping methods developed in the 1990s were utilised in large-scale linkage studies using large, extended families to study diseases amenable to this approach.

As an example of a particularly productive enterprise in this arena, the biotechnology company deCODE Genetics Inc.³ was founded in Iceland in 1996 to leverage the unique properties of the country's small, relatively homogenous population for genetics research. Access to well-characterised disease phenotypes in extended family pedigrees and reliable genealogical records dating back a millennium, combined with industrial-scale genotyping facilities and other resources, enabled deCODE to use linkage analysis to identify genes linked with diseases such as familial essential tremor (Gulcher *et al.*, 1997), stroke (Gretarsdottir *et al.*, 2002) and many more.

1.2.3 Common diseases, common variants

Identification of the causal gene and knowledge of the protein product has led to greatly increased understanding of many Mendelian diseases where the molecular and physiological basis was previously poorly understood, if at all (Botstein and Risch, 2003). Beyond the benefits from improved diagnosis and therapy to patients suffering from many of these conditions, in some cases this has also resulted in new insights into the aetiology of related, non-Mendelian disorders which are far more common in the population. As Guttmacher and Collins (2002) note, the overall impact of Mendelian disease gene discoveries on public health has thus been far greater than the low prevalence of Mendelian diseases would seem to indicate.

Linkage mapping in families has also been used to study the genetic basis of several common genetic disorders with more complex inheritance patterns, in the hope that successes from the rarer monogenic disorders could be replicated and leveraged to bring the same benefits to the far greater number of people afflicted by these common diseases. However, the high-risk, rare mutations that have been implicated in Mendelian versions of some of these diseases contribute little to overall population-attributable risk for the broader disease group, and, as Hirschhorn and Daly (2005) discuss in their review, linkage analysis has generally been relatively ineffective as an investigative strategy for these disorders.

³<http://www.decode.is>

This is largely due to inherent limitations of the method, in that it can only detect with confidence a small number of loci with simple inheritance patterns and relatively strong effects, making linkage studies ill-suited to detecting common variants with modest effect sizes (Risch, 2000). Research attention therefore shifted towards the “common disease, common variant” (CD/CV) model (Chakravarti, 1999; Collins *et al.*, 1997), which posits that many allelic variants, each with a relatively high (5% or greater) frequency in the population and conferring a small individual increase in disease susceptibility, collectively contribute to the overall heritability of common diseases.

1.2.4 Genome-wide associations studies

The draft human genome assembly facilitated new sequence-based methods for disease gene discovery which address many of the limitations of linkage analysis for studying common disorders. Large-scale, systematic variation discovery projects (The International SNP Map Working Group, 2001; Wang *et al.*, 1998) have created genome-wide maps of millions of polymorphic SNP markers. SNPs are present in far higher densities in the genome than mini- and micro-satellites, and SNP genotyping assays are more readily automated, miniaturised and multiplexed. This in turn has enabled rapid advances in microarray-based SNP genotyping technology in the span of only a few years. Further improvements in our knowledge of genetic variation have come from the HapMap project⁴ (The International HapMap Consortium, 2005), a large-scale effort to comprehensively map the extent of linkage disequilibrium (LD) between neighbouring SNPs.

Taken together, these key factors - availability of dense SNP maps, information about genome-wide LD patterns and advances in SNP genotyping technology - have enabled targeted design of genotyping microarrays which directly or indirectly survey most common variation in the genome. High-throughput platforms from several commercial suppliers now allow simultaneous typing of millions of SNPs across the genome at very low cost. It is this experimental capability and affordability that lies at the heart of the current

⁴<http://www.hapmap.org>

mainstream methodology for studying the genetics of common disease - the *genome-wide association study*, or GWAS - which facilitates comprehensive, unbiased, genome-wide scans for common variants associated with diseases, or with other phenotypic traits. Unlike linkage analysis which examines segregation of markers in families and correlates with disease status, association analysis tests specific versions of each marker for co-occurrence with the disease in a large number of unrelated individuals (Carlson *et al.*, 2004). If a SNP allele is observed at a significantly higher frequency in the group of individuals affected by the disease compared to the control group, the allele is said to be associated with the disease. The inference is that the variant either contributes risk of disease directly, or is in LD with one or more nearby pathogenic variants.

An important advantage of the GWAS approach compared to linkage analysis is that family pedigrees are not required, and so application is not limited to diseases occurring in families which are accessible to researchers. Furthermore, this allows large population samples of unrelated individuals to be collected and tested, thereby increasing statistical power to detect the smaller effect sizes predicted by the CD/CV model. The first published GWAS investigation by Klein *et al.* (2005) used a set of ~116,000 SNPs genotyped in 146 individuals to find a common variant of the CFH gene associated with age-related macular degeneration (AMD) disease of the eye. Since this initial GWAS publication, the cost of SNP genotyping has fallen considerably and the number of probes on the microarrays has increased many-fold, with the result that genome-wide scans with over one million SNPs typed in tens of thousands of individuals have become affordable.

Technological advances in genotyping have gone hand in hand with a sharp increase in the size of study cohorts. Gathering and storing of blood, tissue and other specimens linked to extensive annotations of large numbers of individuals - an activity commonly referred to as *biobanking* (Manolio, 2008) - is driven by the need for larger sample sizes (for increased statistical power) and better phenotype characterisation. A key factor is the decreasing cost of SNP genotyping, sequencing and other experimental techniques, which is now enabling large-scale studies not financially feasible before.

Increasingly, large-scale biobanks are being created not as raw materials for particular

disease studies but rather as prospective studies in their own right; i.e. as long-term biomaterial and data resources intended for future use in epidemiological research into genetic, environmental and lifestyle factors in human health and disease. Examples of biobanks now under construction include the Estonian Genome Project⁵ which will include a large proportion of the population, and the UK Biobank⁶ (Elliott *et al.*, 2008) with a target of ~500,000 participant cross-section of the UK population. According to a catalog⁷ maintained by the international Public Population Project in Genomics (P3G)⁸, which promotes harmonisation of biobanks and population-based studies, as of Feb 2010 ~150 large-scale biobanks are underway or in the planning stages.

1.2.5 GWAS success, criticism and the missing heritability

In the five years that have passed since the Klein *et al.* (2005) study, the GWAS approach has been used to identify common variants associated with a wide range of diseases. As of June 2010, the NHGRI GWAS Catalog⁹ (Hindorff *et al.*, 2009a) lists ~500 primary publications describing investigations into hundreds of diseases and other traits, ranging from prostate and breast cancer to Type 2 diabetes and Alzheimer's disease. Some of these associations confirm previous findings from linkage analysis of Mendelian disorders, whilst others have identified genes not previously suspected of a role in the disease.

However, the results produced so far (and the GWAS method itself) has not gone uncriticised. For example, many findings have supplied tantalising clues about underlying biological pathways, but very few associated genes have yet been concretely connected to a disease mechanism with follow-up functional studies or other means, and in most cases the causal variant is yet to found. Notable exceptions do exist: for example, the studies by Gretarsdottir *et al.* (2008) and Gudbjartsson *et al.* (2007), who discovered two

⁵<http://www.geenivaramu.ee>

⁶<http://www.ukbiobank.ac.uk>

⁷<http://www.p3gobservatory.org/studylist.htm>

⁸<http://www.p3g.org>

⁹<http://www.genome.gov/gwastudies/>

variants on 4q25 associated with atrial fibrillation (AF) and ischemic stroke, have led to the development of the deCODE AF diagnostic test¹⁰. The AF laboratory test detects which SNP alleles are present in two locations in a person's genome. Presence of either of the two associated alleles indicates that the person is at increased risk of stroke. This is valuable information which helps decision-making regarding medication, monitoring and other preventative measures. Nevertheless, the overall scarcity of causal variants and potential drug targets has prompted many to doubt that GWAS results can be translated into clinical practice, although as Dermitzakis and Clark (2009) point out the studies arguably “only nominate candidate villains”, and that further research will be required to decipher their contribution to disease aetiology.

Small effect sizes and a low attributable fraction, or potential impact on public health, of most GWAS associations has also prompted criticism; for example, in a recent commentary Goldstein (2009) questions the usefulness of the common variants detected in GWASs for predicting individual disease risk, or for providing meaningful biological insights. However, in his riposte Hirschhorn (2009) cites several examples, including lipid levels and Type 2 diabetes, supporting the claim that small effect sizes do not preclude GWAS discoveries from giving clues about biological underpinnings of phenotypic traits, and that the value of GWASs need not be limited to individual risk prediction.

A common observation of nearly all GWAS results produced to date is that small effect sizes of the numerous disease-associated variants identified collectively explain only a small proportion of overall heritability of the trait. Manolio *et al.* (2009) discuss a number of explanations which may account for this “dark matter” of genome-wide association, such as epigenetic factors, unknown interaction between multiple genes, allelic heterogeneity, rarer variants with larger effect size, and structural variants.

Much of the missing heritability is thought to reside in rarer variants which are either known but not directly assayed or indirectly “tagged” by genotyping microarrays, or have yet to be discovered. Current SNP microarrays are designed to maximise the proportion

¹⁰<http://www.decodediagnostics.com/AF.php>

of common variation surveyed by the limited set of SNP probes that can be placed on the array. Given that the number of probes that can fit on a single array keeps increasing as the technology advances, future SNP microarrays will be able to incorporate rarer variants, thus addressing the first part of this problem. To address the latter limitation, large-scale projects to expand catalogs of sequence variation are now underway. For example, by employing the latest technologies to sequence hundreds of individual genomes from around the world at varying levels of coverage, the 1,000 Genomes Project¹¹ (The 1000 Genomes Project Consortium, 2010) aims to extend our knowledge of human genetic diversity to variants with a population frequency below 1%. Once known, these rare variants can be incorporated into GWAS genotyping microarrays.

Structural variation in genome structure is also thought to explain some of the missing heritability. Genomic rearrangements have long been known to be the direct cause of a number of diseases such as Down's syndrome (trisomy of chromosome 21) and a subtype of Charcot-Marie-Tooth disease (duplication of the *PMP22* gene (Lupski *et al.*, 1991)), via processes such as disrupting gene transcription and altering gene copy number. Until recently, such variation was thought to be relatively rare and it was considered that SNPs were the dominant form of genetic variation. Fredman *et al.* (2004b) discovered that many regions harbour more complex forms of variation masquerading as SNPs, and that such variants were much more common than previously believed. After initial genome-wide surveys by Sebat *et al.* (2004) and Iafrate *et al.* (2004), further discovery studies by these authors and others have reported several thousand CNVs and other structural variants ranging in size from fine-scale (1-500bp) to large-scale(>100Kbp) across the genome (see Wain *et al.* (2009) for a recent review).

It is now well established that CNVs and other structural variants constitute a much greater proportion of human genetic variation than previously suspected, and that at least some of these variants influence normal phenotypic traits and disease susceptibility. Many copy-variant regions are now assayed by the latest genotyping microarrays and are gradually

¹¹<http://1000genomes.org>

being incorporated into GWAS investigations (McCarroll, 2008), most recently in a major study by The Wellcome Trust Case Control Consortium (2010) who analysed over 3,000 common CNVs in 16,000 cases and 3,000 controls. However, according to these findings and a recent CNV discovery study by Conrad *et al.* (2010), the missing heritability “void” is not likely to be filled by common CNVs. This further emphasises the need for more research into rarer variants, more complex forms of structural variation not easily measured with existing array-based platforms, gene interactions, epigenetics, micro-RNAs and other factors.

1.2.6 Sequencing

Advances in sequencing technology are on the verge of transforming biology. The so-called “next-generation” instruments include commercial platforms such as Illumina’s Genome Analyzer¹², Roche’s 454¹³ and ABI’s SOLiD¹⁴. These platforms use massively parallel techniques to generate up to hundreds of millions of sequence reads in a single run, achieving orders of magnitude higher throughput compared to “first-generation” capillary sequencing methods. Though instrument and per-run costs are still considerable compared to array-based methods, these technologies are being applied in domains as diverse as transcriptomics (transcription profiling via RNA sequencing), metagenomics (sequencing of environmental samples) and mutation discovery (Mardis, 2008). Sequencing has also very recently been used in epigenomics to create genome-wide, single-base resolution maps of methylation status of the human genome, or the methylome (Lister *et al.*, 2009). Whole-genome sequencing of hundreds of samples is already underway in the 1,000 Genomes Project, and Gilad *et al.* (2009) note that such strategies for studying natural variation (including structural variants) are free of many of the biases of previous discovery efforts. Sequencing thus offers a powerful means for exploring the population genetics and

¹²http://www.illumina.com/systems/genome_analyzer_iix.ilmn

¹³<http://www.454.com>

¹⁴<http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html>

evolutionary history of humans and other species.

Sequencing is also being aggressively applied in disease investigations and clinical genomics. A number of pilot studies such as the ClinSeq project¹⁵ (Biesecker *et al.*, 2009) and the Personal Genome Project¹⁶ (Church, 2005) are exploring whole-genome sequencing for clinical applications. Complete sequences for lung and skin cancer genomes published by Pleasance *et al.* (2010b) and Pleasance *et al.* (2010a), respectively, have revealed thousands of accumulated somatic mutations. The International Cancer Genome Consortium (ICGS)¹⁷ is undertaking whole-genome sequencing of normal and cancerous tissue from thousands of patients suffering from 50 different types of cancer (The International Cancer Genome Consortium, 2010). Indeed, with the recent development of a new technique called Personalised Analysis of Rearranged Ends (PARE) (Leary *et al.*, 2010), which uses next-generation sequencing to detect fingerprints of tumour cells in blood, personalised medicine may soon make its first significant mark in cancer diagnosis and therapy.

In conclusion, immense progress has been made in the study of common, complex disease and other traits in the span of only a few years, and thousands of gene variants contributing to rare and common diseases have been identified. The GWAS method has emerged as an important research tool to dissect the genetic basis of these disease, as high-throughput SNP genotyping technology and large-scale biobanks now under construction are used to study ever-larger cohorts. Sequencing of whole genomes with new massively parallel sequencing technologies will soon be commonplace, and decreasing costs of next-generation sequencing and further technology advances will likely in the next few years replace array-based genotyped technologies in many applications.

All this progress would not have been possible without the availability of the human

¹⁵<http://genome.gov/20519355>

¹⁶<http://www.personalgenomes.org>

¹⁷<http://www.icgc.org>

genome sequence at the start of the decade, and the accumulated information about natural variation in the human genome. Online databases cataloging this information have played a critical role here, and will continue to do so as efforts to discover rare and structural variants expand our knowledge of genetic variation. However, there are mounting challenges from the volume and complexity of data already produced and will be produced in the future. These challenges are the topic of the next chapter.

2. Genotype to phenotype databases: challenges and opportunities

The process of disseminating and debating research findings, or scientific discourse, is an inexorable part of the modern scientific method and traces its beginnings to the establishment of the Philosophical Transactions of the Royal Society in 1665 (Oldenburg, 1665). But this centuries-old model of publishing formal articles in printed journals as the primary means of communicating science is being transformed by advances in information and communication technology. Electronic mail and the World Wide Web have perhaps been the most pervasive aspects of this digital revolution, and in the span of only a decade these technologies have transformed the way we access scientific publications and communicate with colleagues. Virtually all scientific journals are now accessible online, empowering scientists to search for and retrieve almost any publication in any journal anywhere in the world, aided by the latest Web applications which offer new, powerful ways to exploit the information contained in these digital libraries (Hull *et al.*, 2008).

Online access to journals has greatly accelerated and streamlined - and may ultimately transform - scientific communication. But technological progress has had profound impact on science on a more fundamental level. The advances in genetics and genomics noted in Chapter 1 reflect a broader trend: in the last three to four decades, the life science enterprise has increasingly involved production, gathering and analysis of increasingly voluminous and complex experimental and observational data. Space telescopes orbiting the Earth look back in time to observe massive quasars and other interstellar objects. Detectors connected to high-energy particle accelerators observe the very smallest units of matter. Sequencing devices decipher entire microbial genomes in a matter of days. These are all examples of large-scale, data-driven, “Big Science” undertakings which are evolving in parallel

with, and being facilitated by, the digital revolution. The Internet in particular has been instrumental in these developments as a means for broad dissemination of scientific data, and thus this recent period in the history of science has been aptly described recently by Smith (2009) as follows: “If digital technologies are the engine of this revolution, digital data are its fuel”.

For biologists, the Web has become an indispensable research tool by empowering them to easily access this expanding array of genomic and other research data and publications via a simple browser program on their computers. Websites such as the PubMed literature search service (Wheeler *et al.*, 2008), the Ensembl (Flicek *et al.*, 2010), UCSC (Rhead *et al.*, 2010) and NCBI (Wheeler *et al.*, 2008) genome browsers, and the BLAST sequence search service (Wheeler *et al.*, 2008) are examples of important Internet resources that biomedical researchers rely on in their quest to understand the relationship between genes and disease. However, the large degree of heterogeneity in the way information is accessed and retrieved across the many hundreds of biological databases now accessible on the Web (Galperin and Cochrane, 2009) presents problems. A biologist investigating a given biological question is required to browse many websites and learn to use many different user interfaces, and still never feel sure that he has tracked down all the necessary information to address the question.

While this is a long-standing source of frustration for traditional biologists, the focus on humans as data consumers in Web site design is also a major stumbling block 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 websites to extract information originally designed for human consumption. As noted by Stein (2002), this “screen scraping” approach has numerous disadvantages, such as “brittleness” (user interface tends to change frequently, so scraping-code must also change), lack of reliability and duplication of effort. In recent years there has been a growing trend in the field towards more sophisticated methods for connecting many different databases on the Internet in such a way that data can be navigated, queried and retrieved by automated “agents” - i.e. programmatically and without human intervention.

My treatment in this chapter begins with a brief introduction to the broader “omics” data problem, followed by a survey of key mainstream online database resources supporting G2P research. The next section introduces a core set of challenges facing many of these and other databases. This will be followed by a section discussing recent developments towards increased emphasis on federated database solutions which can link independent databases through a central portal, and how this can be married with the proven benefits of traditional central databases. The technological basis of database federation will then be introduced, followed by a discussion of key Internet and domain standards, and advanced semantic technologies which promise even more powerful data integration and exploitation in the future. Several examples of projects in the biosciences which utilise this technology will be briefly discussed.

2.1 The “omics” data problem

As noted in the Introduction, the large-scale datasets generated by microarray platforms and massively-parallel sequencing instruments present formidable challenges in data management and exploitation. Data generated by other high-throughput experimental techniques used to study biological systems, such as protein/DNA and protein/protein interaction studies, gene expression profiling, RNA interference, determination of protein structure and digital imaging are also increasing in scale and dimensionality. Huge advances in experimentation, coupled with advances in computational methodology and other areas of the life sciences, are enabling a shift from investigating individual mechanisms or partial biological systems in isolation towards systems-level approaches (Hood *et al.*, 2004; Kitano, 2002). In *systems* or *organismal biology*, such high-level, holistic views of cells and whole organisms - based on sophisticated mathematical models of the structure and dynamics of biological systems - require as input large numbers of accurate measurements in order to predict the behaviour of these systems and their environments in normal and perturbed states. With the required technology now becoming available, the time may indeed soon be ripe for comprehensively exploring how genomes

produce organisms, how organisms interact with the environment, and other “grand challenges” in organismal biology (Schwenk *et al.*, 2009).

The data integration problem. Integration of all relevant information across all the various “omics” research activities to address biological questions is one of the big bioinformatics challenges in contemporary biology (Ge *et al.*, 2003). For example, tasks such as combining disease-associated genomic regions from GWAS scans with functional annotations of gene products, information on up/down-regulated genes and possible roles in biochemical pathways are critical to understanding how these genetic elements may contribute to disease. But such integration tasks are far from trivial to undertake. Fragmentation of data across tens or hundreds of heterogeneous databases; lack of standardisation in data formats and semantics, inconsistent identification of biological objects and concepts; the sheer scale and complexity of the data produced; all these factors combine to make organisation, integration, analysis and interpretation of “omics” data extraordinarily difficult. The data integration problem in biology has been discussed by many authors (see e.g. Goble and Stevens (2008)) and much current research in bioinformatics is focused on this problem.

The data sharing problem. Other related issues compound the overall problem of integrating “omics” data. The task of locating relevant published datasets can be difficult, especially for the “long tail” of small but potentially useful datasets generated by much smaller-scale projects. A large number of these datasets resides in the long tail, but these are not given nearly as much attention to ensure data preservation and enable reuse as large-scale datasets (Heidorn, 2008). Many important datasets, especially genetic and medical data on human subjects, are by necessity published exclusively via controlled-access databases which currently makes data access and retrieval cumbersome. More generally, primary data supporting published conclusions in peer-reviewed publications is frequently not made available. All these are long-standing problems concerning data access and availability in the biosciences, where many intertwined issues such as intellectual

property, licensing, data ownership, lack of incentives for sharing, and lack of effective tools for data publication contribute to an overall unsatisfactory state of affairs (Costello, 2009; Smith, 2009). Awareness is growing in the community about the severity of these problems, however, as evident from several papers which appeared in a recent special issue of the journal *Nature*¹ (see e.g. Nelson (2009)).

Though interesting as a research problem in itself, the data sharing problem will not be discussed in further detail here. But due to its high relevance to the central aims of this thesis, several aspects of the problem will receive further treatment as appropriate in the more specific context of the chapters to follow.

2.2 A survey of G2P databases

The field of G2P databasing covers a vast number of online data resources, including genetic sequence archives, global catalogs of genetic variation and a diverse range of datasources relating genotype to the phenotype at varying levels of resolution. Until recently, nearly all such online stores of G2P data tended to be built as “central” databases. In the centralised model, outstations or “nodes” (e.g. sequencing centres or individual laboratories) gather and prepare data for transfer to a large central “hub”, where the data are stored, integrated and made available for searching and presentation, primarily with the human data consumer in mind. This section reviews a cross-section of this diverse “ecosystem” of G2P databases. Numerous important differences and common trends will be highlighted and discussed, in particular as these relate to the effectiveness (or otherwise) of a centralised database strategy in this domain.

2.2.1 Genetic sequence archives

The need for preserving and disseminating large, complex scientific datasets in a structured format predates the Web, and specialised online databases - *digital repositories* - have been

¹<http://www.nature.com/news/specials/datasharing/>

constructed in data-intensive scientific disciplines such as physics and astronomy. The earliest such databases of prominence in molecular biology were designed to hold nucleic acid sequence data. As soon as the use of commercial technologies for DNA sequencing became widespread in the 1970s, it quickly became apparent that the printed-journal format was ill-suited for publishing increasing volumes of sequence data. This drove the creation of the first computer-based repositories to facilitate exchange and comparison of DNA sequences: instead of including the sequence data in the manuscript itself as page after page of A, C, T, and G, authors could now use permanent accession identifiers which acted as pointers to the actual sequences submitted separately to a database along with biological annotation.

The three major central sequence databases initially constructed were located in Japan, the US, and Europe: the DDBJ² (Sugawara *et al.*, 2008), GenBank³ (Benson *et al.*, 2008) and EMBL⁴ (Cochrane *et al.*, 2008), respectively. In the mid 1980s the International Nucleotide Sequence Database Collaboration (INSDC)⁵ was established to promote full content exchange between these databases, making them effectively function as a single, master archive of sequence data. This tripartite database collaboration is often collectively referred to as simply GenBank, a convention that will be used here.

The impact of the establishment of a primary archive for sequence data was profound. A host of different sequence analyses were made possible because all published sequence data were now available in a single place in accessible form. This facilitated and drove development of data standards and analysis tools which would later prove indispensable to the success of the Human Genome Project, and along the way spawned an entire branch of computational biology now known as *bioinformatics*. Previously non-existent research areas such as transcriptomics and comparative genomics that emerged in the mid-1990s (see e.g. Andrade and Sander (1997)) would not have been possible without unrestricted

²<http://www.ddbj.nig.ac.jp>

³<http://www.ncbi.nlm.nih.gov/Genbank/>

⁴<http://www.ebi.ac.uk/embl/>

⁵<http://www.insdc.org>

access to all available sequence data. The whole became bigger than the sum of its parts, and such synergy is indeed a major rationale for creating scientific databases.

Given the success of GenBank, the INSDC partners have elected to follow the same strategy for archiving data generated by next-generation sequencing platforms. In an effort to extend and re-engineer existing centralised repositories for primary sequencing data, the Sequence Read Archive (SRA)⁶, the DDBJ Read Archive (DRA)⁷ and the European Read Archive (ERA)⁸ have been established by the NCBI, DDBJ and EMBL, respectively.

The story of GenBank over the past quarter century is proof that a large-scale centralisation strategy can be very effective. It illustrates how a large number of little pieces of data which have relatively low value on their own are immensely useful once gathered into one place, albeit at the expense of increased effort on behalf of data creators to package the data into the appropriate format and submit them to the database. However, the success of sequence archives can arguably be ascribed to two main factors: i) DNA sequence information is relatively simple to represent as a directly annotated string of letters and sequence regions, and ii) despite a massive growth in sequence data volume (see Fig. §2.1), the scale of the problem did not exceed the capabilities provided by parallel advances in computer technology. But as the following subsections will show, not only is much of G2P research data far more diverse and complex than sequence data, but an explosion in sequence data volumes is on the horizon which even the new generation of centralised sequence archives will find difficult to cope with.

2.2.2 Core genomic variation databases

The principal role of *core variation databases* is to serve as a central “backbone” of information on known genomic variation in humans and other species, irrespective of any association with phenotype. This treatment will focus on database resources for the

⁶<http://www.ncbi.nlm.nih.gov/Traces/sra/>

⁷http://trace.ddbj.nig.ac.jp/dra/index_e.shtml

⁸<http://www.ebi.ac.uk/embl/Documentation/ENA-Reads.html>

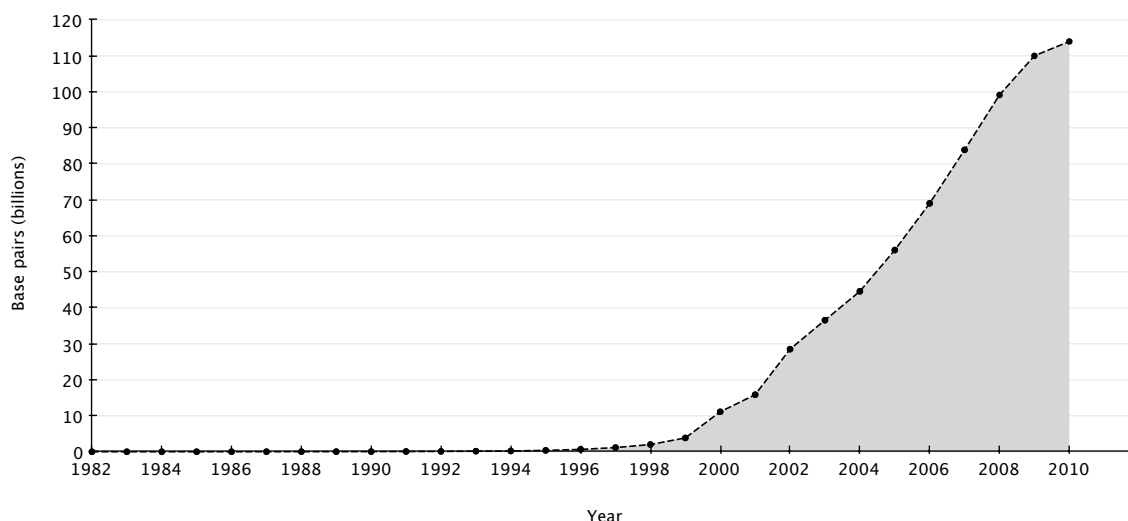


Fig. 2.1: Growth of GenBank from 1982 to the present. Summary statistics from <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>.

categories of variation shown in the upper half of Fig. §2.2, which will serve as a useful reference for the rest of the subsection.

Micro-satellite markers and other sequence-tagged sites. The discovery of short tandem-repeat polymorphic sites in the human genome facilitated the construction of dense genetic maps. These polymorphic markers are characterised by the PCR primer pairs used for amplifying sections of the genome for further analysis. Sequence-tagged sites (STSs), both polymorphic and invariant, served as important genome landmarks in the Human Genome Project by providing a “common language” for unifying the diverse collection of linkage maps, radiation-hybrid maps and other maps used to construct the physical map or scaffold which guided sequencing efforts (Olson *et al.*, 1989).

At present, NCBI’s UniSTS⁹ is the main global repository of markers and mapping data from various sources, including GenBank’s STS¹⁰ division, the now-decommissioned Genome Database (GDB) (Letovsky *et al.*, 1998) and many other human and mouse genetic

⁹<http://www.ncbi.nlm.nih.gov/sites/entrez?db=unists>

¹⁰<http://www.ncbi.nlm.nih.gov/dbSTS/>

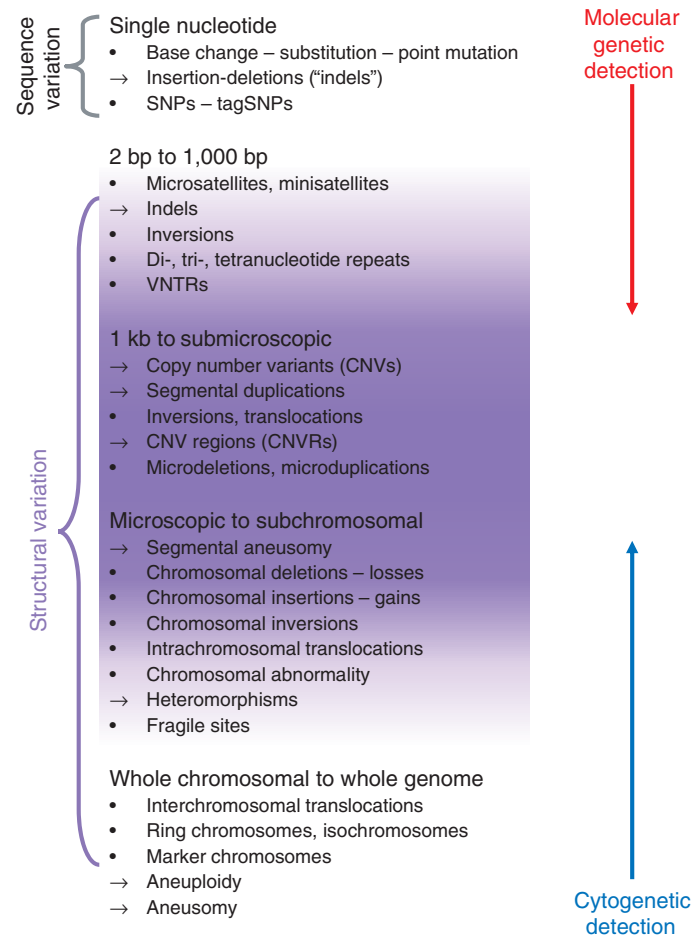


Fig. 2.2: The spectrum of genomic variation. From Scherer *et al.* (2007).

map resources. As of Feb 2010, UniSTS contained ~324,400 total markers in the human genome. STSs are still used as a research tool, but as explained in Chapter 1 they have in recent years been largely supplanted by SNPs as genome landmarks in high-throughput genetic studies.

SNPs and other small sequence variants. A major factor in the advances in disease genetics discussed in Chapter 1 was the emergence of central databases of SNPs, insertions/deletions (indels) and other small-scale sequence variation, populated with data from large-scale discovery projects. These small variants are characterised not solely by a PCR primer pair (as for STSs), but rather by the sequences flanking a variant site and

the allelic alternatives at that site. Knowledge of the variation on the level of the sequence itself enables the design of genotyping assays which underpin modern high-throughput SNP analysis.

One of the first databases to catalog this information in the late 1990s, the Human Genome Variation Database (HGVbase) (Fredman *et al.*, 2004a, 2002), focused on gathering variant reports submitted by researchers and undertaking high-quality curation of these variants and their functional consequences. However, manually curating each entry became unfeasible as discovery techniques were automated and the number of variants grew to the millions. This scalability issue, coupled with a lack of long-term financial backing, led to much-reduced relevance of HGVbase (A. J. Brookes, personal communication), while at the same time the database of Single Nucleotide Polymorphisms¹¹ (dbSNP) (Sherry *et al.*, 2001), established by the NCBI, instead became the *de facto* central SNP repository. The current dbSNP release (build 130) contains information on nearly 18 million distinct polymorphic sites in the human genome, and millions more in over 50 other organisms (with a significant overlap of STS entries already cataloged in UniSTS), as well as individual- and population-level genotype information from several thousand population samples from around the world.

dbSNP provides an interesting perspective on some of the strengths and weaknesses of the centralised database model. Just as GenBank optimised the utility of primary sequence data, dbSNP provides tremendous utility of a public archive of variation data as a research tool. This success is in no small part due to the realization by the parent funding body (the US National Institute of Health (NIH)) that dbSNP is an essential piece of research infrastructure that necessitates long-term, financial stability. But dbSNP also exemplifies some of the limitations of the centralised model. For example, given dbSNP's primary remit as a primary variation archive, most emphasis is placed on keeping the production data "pipeline" running reliably, with little effort devoted to sustained development of dedicated data access and visualisation tools. Most such functionality is provided via

¹¹<http://www.ncbi.nlm.nih.gov/SNP/>

connections with other NCBI data resources (Wheeler *et al.*, 2008), such as RefSeq (gene annotations) and MapView (genome browsing), and this is indeed appropriate and useful for a genomics-oriented perspective of this information.

However, other areas of research are not as well served, such as anthropological and evolutionary studies of human populations. The Allele Frequency Database (ALFRED)¹² (Rajeevan *et al.*, 2005) was created in the late 1990s to cater for the needs of anthropologic geneticists who require genotype data from individuals sampled from around the world. ALFRED provides a number of tools and data retrieval options geared to the needs of those researchers, with a special focus on comparing allele frequencies from selected polymorphic loci across many geographically-dispersed populations. The bulk of the variant loci in ALFRED are catalogued in dbSNP also (and cross-links are provided), whereas population frequency data are acquired via submissions and via extraction from the literature. dbSNP also contains a wealth of population data on the same and many other variants, but the critical difference here is that to this specific community of researchers, population data are made far more useful as presented in ALFRED (e.g. allele frequencies for specific population samples overlaid on a geographical map). The HapMap website is another example of this; although all genotyping data generated in the project were submitted to dbSNP, the HapMap Data Coordination Centre elected to provide facilities for data display, retrieval and analysis on its own project website¹³ (Thorisson *et al.*, 2005).

Structural variation. Another, and perhaps more serious, disadvantage of the centralised model exemplified by dbSNP is a tendency to “stagnate” when it comes to adapting to changing needs of the field, including those resulting from advances in biological knowledge. When dbSNP and similar genetic variation databases were created in the mid-1990s, conventional wisdom held that SNPs were the dominant form of variation in the genome, and thus database and software tools were optimised for information describing simple sequence-level variants. But as copy number variants and other, more

¹²<http://alfred.med.yale.edu>

¹³<http://www.hapmap.org>

complex forms of variation were found to be widespread in the human genome, existing SNP-oriented variation databases were not up to the task of cataloging this new kind of information. In response to this, numerous alternative databases were created. dbRIP¹⁴ (Wang *et al.*, 2006) contains reports of over 2,000 polymorphic retrotransposons in the human genome. The Human Structural Variation Database¹⁵ (no longer maintained) catalogs larger segmental duplications discovered by Sharp *et al.* (2005) and in the two landmark CNV surveys by Sebat *et al.* (2004) and Iafrate *et al.* (2004). The Database of Genomic Variants (DGV)¹⁶ (Zhang *et al.*, 2006) contains (as of Feb 2010) ~50,000 CNVs, indels and inversions from several dozen published studies undertaken in the past 5 years, and is currently the most comprehensive catalog of structural variation in normal individuals.

Despite this proliferation of structural variation databases, there are still many important unresolved issues relating to data standardization and data sharing which limit the usefulness of the cataloged information. Some of these were highlighted by Scherer *et al.* (2007) as they discussed various challenges associated with a lack of standards for variant characterisation, data collection, quality assessment and reporting in this new field, including several key databasing issues. The most vexing problem is ill-defined boundaries of reported structural variants. Most experimental platforms employed in the field are oligonucleotide- or BAC-array based which cannot resolve variant breakpoints at the sequence level. This has lead to inconsistent reporting of variants and variant-harboring regions; a study may find several overlapping CNVs in one genome region and report them as a single, over-arching copy-variant region, even if they may in fact be distinct mutations in different individuals. This and various other technical and biological factors greatly complicate development of robust genotyping assays for structural variants and their incorporation into mainstream studies of common disease (McCarroll and Altshuler, 2007). Furthermore, the lack of data standardization creates a barrier to effective data

¹⁴<http://dbrip.brocku.ca>

¹⁵<http://humanparalogy.gs.washington.edu/structuralvariation/>

¹⁶<http://projects.tcag.ca/variation/>

integration. For example, it will be necessary to integrate data from normal individuals catalogued in DGV with data from resources focusing on structural variants linked with a clinical phenotype, such as DECIPHER¹⁷ (Firth *et al.*, 2009) and ECARUCA¹⁸ (Feenstra *et al.*, 2006), as well as disease-specific resources such as the Autism Chromosome Rearrangement Database¹⁹ (Marshall *et al.*, 2008).

Unfortunately, the current crop of online resources are not very helpful in addressing these acute problems. For instance, DGV only represents variants at a high level as reported in the original paper (i.e. as features on the reference genome assembly), with no attempt to capture, archive, display, re-analyse or standardise raw study data which would help to address some of these issues. Nonetheless, as not to devalue the important work of the DGV creators, it should be noted that any efforts to globally re-analyse primary data in this field are presently hampered by the unpleasant fact that raw data from many of the studies are not publicly available. It thus seems clear that there is an urgent need for standardised reporting and gathering of structural variants, including raw primary data.

In an attempt to remedy this, NCBI has recently created the Database of Genomic Structural Variation (dbVar)²⁰ for archival of all non-SNP variation. According to the dbVar submission information page²¹, dbVar will accept submissions containing reported variant regions and other supporting high-level information, assign these variants unique, stable identifiers and, crucially, require submitters to also submit raw experimental data to the appropriate primary archives (e.g. ArrayExpress or GenBank). According to a recent publication by the NCBI, EBI and creators of DGV (Church *et al.*, 2010), dbVar will exchange data bilaterally with a new database named DGVa located at the EBI and provide a primary repository for archival and accessioning of structural variation studies (thus replicating the proven GenBank/DDBJ/EMBL model). In this new collaborative

¹⁷<https://decipher.sanger.ac.uk>

¹⁸<http://www.ecaruca.net>

¹⁹<http://projects.tcag.ca/autism/>

²⁰<http://www.ncbi.nlm.nih.gov/dbvar>

²¹<http://www.ncbi.nlm.nih.gov/projects/dbvar/submission.html>

scheme, DGV will serve as a higher-level resource for curation and interpretation, fed by data from the primary archives.

The challenges outlined above clearly show that structural variation is more difficult to describe and store than sequence data and simpler sequence variants like SNPs, and also underlines the difficulty of managing such data in a centralised manner in the face of rapid knowledge advances. Nevertheless, it would seem fundamental to progress in the field that essential baseline information on genome sequences and natural genome variation, simple and complex alike, be centrally managed and accessible via comprehensive, long-term archives such as GenBank, dbSNP and dbVar. Another noticeable trend is for such primary archives to be augmented by specialized tools and databases such as ALFRED which provide alternative views of the primary data and often contain a wealth of additional related, high-quality curated data.

2.2.3 Databases for model organisms

Another example of effective database centralisation is provided by *model organism databases*, or MODs, which specialise in capturing genomic, phenotypic and other information relating to a particular model organism, or several closely related organisms. MODs are typically created by a single or a few research groups who leverage their expert knowledge of the organism(s) of interest to gather into one place and curate relevant research data and scholarly literature.

An MOD is usually not the primary repository for sequences, gene expression or other experimental data for the organism. Database curators collect this information from the primary sources and organise it into useful collections, check for consistency, curate and link the data to (and extract information from) the scientific literature (Hirschman *et al.*, 2010). This enriches the data and makes the data more accessible and useful to biologists. Such relatively small-scale operations have tended to evolve into major web portals dedicated to providing important online resources to the community of researchers

studying the organism. An MOD web site is therefore, often, the main access point for various online services which help researchers make use of the data, such as sophisticated genome browsers and tools for comparative genomics, be it hosted on the MOD site itself or elsewhere on the Web.

Early MODs for traditional invertebrate experimental systems. One of the earliest MODs to be constructed was a database for *Caenorhabditis elegans*, a nematode worm studied by geneticists since the 1960s. The software, aptly named ACeDB²² (a *C. elegans* database), was an object-oriented database, designed in the late 1980s by Richard Durbin and Jean Thierry-Mieg to handle the complexity of biological data. It was particularly useful in coordinating the *C. elegans* sequencing effort and facilitating its integration with genetic and physical map information. Initially developed and maintained by the Sanger Centre, ACeDB was distributed as a non-networked, standalone package distributed on a CD containing the database software pre-loaded with data. Over time, ACeDB evolved into a fully-fledged online database, becoming WormBase²³ (Harris *et al.*, 2010) which is today a major online resource for worm biology, operated by a consortium of research groups in the US, Canada and UK. The site is home to large amounts of highly curated data ranging from genomic sequences to high-throughput expression data and RNAi-knockout phenotype data, as well as various analysis and visualisation tools.

Another favourite of geneticists, and one of the very first model organisms, is the fruit fly, *Drosophila melanogaster*. Due to its small size, its ease of breeding and potential for genetic manipulation, *D. melanogaster* has been used to model genetic systems of higher animals for over a century. Not surprisingly MODs serving the fly research community also have a long history dating back to before the modern World Wide Web; in the early 1990s the FlyBase Consortium released a database containing *D. melanogaster* genetic and molecular data which could be accessed over the Internet via Gopher and FTP (The

²²<http://www.acedb.org>

²³<http://www.wormbase.org>

FlyBase Consortium, 1994). In its current form the FlyBase web site²⁴ (Tweedie *et al.*, 2009) offers access to full genomes of all 12 sequenced fly strains, expression data, mutant phenotype data and more, as well as genome browsers and query tools.

MODs for simpler organisms and pathogens. MODs exist for simpler organisms as well. The unicellular budding yeast (*Saccharomyces cerevisiae*) has long been used as a system for studying biological processes relevant to higher eukaryotes, and its small genome was one of the first to be fully sequenced in 1996. The *Saccharomyces* Genome Database (SGD)²⁵ (Hong *et al.*, 2008) has been in operation for over a decade to serve the yeast research community. More recently, the Comprehensive Yeast Genome Database (CYGD)²⁶ (Guldener *et al.*, 2005) for *S. cerevisiae* and sequenced genomes of related yeasts was launched in 1999. It is worth noting that CYGD is actually part of a battery of four MODs for fungal genomes which are all hosted at <http://mips.gsf.de/projects/fungi> and all run on the same software platform. This streamlines maintenance for these databases and leverages economies of scale, because further MODs can be added to the site with minimal effort. As another example of this, *Schizosaccharomyces pombe* (fission yeast) and three other fungal MODs use the GeneDB platform (Hertz-Fowler *et al.*, 2004) hosted centrally²⁷, alongside dozens of other organisms sequenced by the Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute (WTSI)²⁸.

There is great scientific interest in studying genomes of organisms which cause disease in humans and livestock. Due to its large impact on human health in many parts of the world, the malaria-causing protozoan *Plasmodium falciparum* and its mosquito vector *Anopheles gambiae* have been intensively studied for a long time and the parasite/vector pair was an early sequencing target. The genome sequences of both were published in

²⁴<http://flybase.org>

²⁵<http://www.yeastgenome.org>

²⁶<http://mips.gsf.de/genre/proj/yeast/>

²⁷<http://www.genedb.org>

²⁸<http://www.sanger.ac.uk/Projects/Pathogens/>

2002 (Gardner *et al.*, 2002; Holt *et al.*, 2002) and a MOD for malaria parasites called PlasmoDB²⁹ (Aurrecochea *et al.*, 2009) appeared online two years previously.

Numerous other pathogens have now been sequenced and continuing the trend mentioned above for fungal genomes, database resources for these organisms are typically provided in shared-hosting fashion. MODs for these organisms are hosted on the same web server along with similar resources for related organisms, often grouped by taxon (e.g. PlasmoDB) or family (e.g. TriTrypDB³⁰, for *Trypanosomatidae* pathogens), and tend to be powered by the same specialised software. A good example of this is the EuPathDB portal³¹ points to seven distributed sites specialising in important eukaryotic pathogens in the *Apicomplexa* phylum. Each participating site hosts a database for up to several species, and all databases are built with the ApiDB software (Aurrecochea *et al.*, 2007). The logistical advantages of this arrangement are the same as described above for fungal MODs. An important additional point here (to be revisited later in this chapter) is that this level of standardisation enables uniform queries across all these databases from the central portal.

Animal models of human disease. The organisms mentioned so far are useful for a variety of biological research, whether it be to gain insight into aspects of human biology or non-human phenomena in their own right. But in order to study mammalian-specific human traits, and human disease in particular, closer relatives are required. The most commonly used mammalian experimental models are the mouse (*Mus musculus*) and the rat (*Rattus norvegicus*). Both have been extensively studied for decades, and as expected both have a rich MOD tradition with major emphasis on facilitating the use of these organisms as models for human disease.

The Mouse Genome Database (MGD) (Blake *et al.*, 2009) and several other key mouse resources are integrated via the Mouse Genome Informatics (MGI) portal³², the main

²⁹<http://plasmodb.org>

³⁰<http://tritrypdb.org>

³¹<http://eupathdb.org>

³²<http://www.informatics.jax.org>

mouse databasing centre of gravity located at Jackson Laboratories in the USA³³. As an authoritative source of curated information on mouse genes, strains and more, MGI provides many query and retrieval tools, some of which are semantically enhanced (e.g. via ontologies, discussed further below) and optimised for tasks such as finding information on mouse strains with mutations in orthologous genes which makes them suitable models for a particular disease. The Rat Genome Database (RGD)³⁴ (Dwinell *et al.*, 2008) similarly has tools tailored to specific areas of human disease research in the form of disease “portals” (currently cardiovascular, neurological, cancer and obesity / metabolic syndrome are offered). These portals offer a genome-wide view of genes and quantitative trait loci (QTLs) associated with disease phenotypes in the respective category, links to each gene displayed, and an annotation summary for the genes in the set.

RGD, MGD and related resources such as the Mouse Phenome Database (MPD)³⁵ (Grubb *et al.*, 2009) contain extensive information on the hundreds of strains available for each species, with emphasis on helping researchers find a mouse strain with suitable characteristics (e.g. high blood pressure) to use in experiments as a model for the human condition. Much of the data originates from efforts such as EUMODIC³⁶ which collect extensive phenotype information for a large number of strains, from systematic gene knockouts and phenotypic screens.

Common software platforms and software reuse. In many respects, more recent MOD projects have taken their cue from the success of established projects like WormBase and FlyBase. This is clearly visible in areas such as web site design and general organisation of data, and the trend increasingly extends into software reuse; e.g. the ACeDB software was used in the early versions of SGD and several other databases, and examples from fungal and pathogen MODs were given above. This follows a general “common sense” principle

³³<http://www.jax.org>

³⁴<http://rgd.mcw.edu>

³⁵<http://www.jax.org/phenome>

³⁶<http://www.eumodic.eu>

of focusing developer effort into one or a few high-quality, useful software packages that can be reused by many projects (i.e. “don’t reinvent the wheel”), rather than many groups solving essentially the same problems over and over again. Some obvious advantages of software reuse for MOD developers include greatly reduced up-front effort and time to construct the MOD (compared to building own software from scratch), and also subsequent lowered maintenance costs (many groups working to fix bugs and add useful features). These two factors are often key for smaller groups which may not have the funds and/or expertise to construct and maintain a MOD without deploying existing software.

It should be noted, however, that many larger groups or consortia do indeed have resources to undertake major software projects, and such communities may have specific needs not fulfilled by off-the-shelf software, e.g. some of the pathogen resources described above. In such cases it is sometimes a better strategy to construct (or keep developing existing) custom software. Furthermore, a single software package will not fulfil the needs of every project, and so having some options provides developers with a choice of the most appropriate tool for the job.

The Generic Model Organism Database. In the late 1990s, as more and more genomes were sequenced and MODs proliferated, funding agencies began to recognise the problem that funds were being repeatedly allocated to solving the same databasing problems over and over again. It was, of course, not feasible to require grant recipients to use off-the-shelf software if suitable software did not exist, or existing software (e.g. from established MODs) was not usable in its current state. To address this problem, several funding agencies have funded the Generic Model Organism Database project (GMOD)³⁷, an initiative to develop a complete set of reusable, interoperable software components for “creating and managing genome-scale biological databases”. Participating organisations and projects include all of the MODs previously described, and dozens more for organisms ranging from human influenza through bees to fleas and soybeans (see <http://gmod.org/wiki/MOD> for an up-to-date list).

³⁷<http://gmod.org>

Central to the GMOD mission is a collection of reusable software components, each optimised for a certain common task (e.g. storing and accessing genome annotations). From this collection database developers can choose the components they need, and then build from these “LEGO bricks” a system to suit their needs. A major role played by GMOD is to coordinate the various development activities for existing or new components to minimise duplication of work, and to help ensure that software components work together. Perhaps the best known of these is the Generic Genome Browser (GBrowse)³⁸ (Stein *et al.*, 2002), a web-based graphical tool for visualising genomic annotations. GBrowse has reached a level of near-ubiquitousness; almost all the MODs have converged on GBrowse for this task, as have numerous major, non-MOD resources. Similarly, the BioMart data mining tool³⁹ (Smedley *et al.*, 2009) is widely used among MODs and many non-MOD websites as well.

In addition to the general advantages from software reuse described above, this standardisation results in consistent user interfaces across many MOD websites, and so users who need to use multiple sites do not have to re-learn many tools for the same task (a major plus for users routinely comparing many species). But to illustrate the previous point that one size does not necessarily fit all, some MODs have instead chosen to use the Ensembl genome browser and automated genome annotation pipeline⁴⁰ (Hubbard *et al.*, 2009), a software platform jointly developed by EBI and WTSI.

A prime example of the broader value of GMOD is provided by a MOD for a single-cell eukaryote, the ciliate *Paramecium tetraurelia* which has been studied for over 50 years as a model for a number of multi-cellular organism functions. ParameciumDB⁴¹, a dedicated MOD for this organism containing the recently-sequenced 40,000 gene macro-nuclear sequence and annotations linked with genetic data, was only relatively recently constructed (Arnaiz *et al.*, 2007). However, this late arrival on the MOD scene had the

³⁸<http://gmod.org/ggb>

³⁹<http://www.biomart.org>

⁴⁰<http://www.ensembl.org>

⁴¹<http://paramecium.cgm.cnrs-gif.fr>

side benefit that a number of ready-made components could therefore be used for the job. ParameciumDB developers were able to assemble their MOD largely from ready-made components, saving time and cost compared to creating equivalent software tools from scratch.

Community-developed ontologies. A major facet of MOD activities is the exchange of various sorts of data, be it import of data from submissions, or import and export of data to and from analysis tools and other databases. Historically, emphasis has been placed on the *syntactic* aspect of this process, i.e. the creation of tools and file formats to ensure that the structure and organisation of data transmitted from one database can be correctly parsed and loaded into the receiving database. However, over time it has become clear that in order to integrate concepts and make sense of large amounts of diverse and complex biological data, attention must be given to consistent ways of transmitting the meaning of the information, i.e. its *semantics*.

An established approach to addressing the problem of semantic integration is to create ontologies - structured, controlled vocabularies of terms for concepts, their definitions and well-defined relationships between them. An ontology formalises domain knowledge in such a way that it can be used as a “semantic layer” to convey the meaning of data. This then enables the use of sophisticated informatics technologies to drive query tools and knowledge-discovery applications (Bard, 2003; Stevens *et al.*, 2000). Ontologies and the issue of syntax and semantics in data exchange will be further discussed in §2.5.

Bio-ontologies were pioneered by MODs, and the best known ontology in the biological domain is the Gene Ontology (GO)⁴² (The Gene Ontology Consortium, 2008), the fruit of a decade-long collaboration between the major MODs and other stakeholders in the GO Consortium. GO, along with companion resources such as the GO Annotation database (GOA)⁴³ (Barrell *et al.*, 2009), has become a reference source of annotations for genes and gene products and is routinely used in genome analysis. Another

⁴²<http://www.geneontology.org>

⁴³<http://www.ebi.ac.uk/GOA/>

example is the Mammalian Phenotype Ontology (MPO) (Smith *et al.*, 2005), developed primarily by the rat and mouse communities, and frameworks for describing phenotypes with both compound MPO terms and via combinatorial annotation (using terms from multiple ontologies) (Beck *et al.*, 2009). These and other technologies underpin major bioinformatics projects, such as EuroPhenome⁴⁴ (Mallon *et al.*, 2008) which is building systems for storing, managing and analysing high-throughput mouse phenotype data generated by the aforementioned EUMODIC project.

From this brief survey of the MOD landscape, one can recognise a number of important trends and lessons which are highly relevant to the human G2P genetics/genomics database world. First and foremost, further to the issue of centralised databasing, it would be easy to ascribe the success of many MODs merely to the limited volume of data they contain. But it is important to note that data contained in a given MOD are far more diverse and complex than the simpler annotated sequence data in the far larger, global GenBank sequence archive, which makes the task of organising them in a central databases more difficult. A significant factor in the success of many MODs, however, is likely good leadership and relatively small community sizes which simplifies the task of presenting, discussing and reaching agreement on issues such as data standards, curation practices and gene naming conventions.

Several other key trends and concepts are also worth noting. First, concerted efforts to develop and reuse specialised software for common bio-databasing tasks can clearly pay big dividends. Second, some of the sophisticated informatics solutions and semantic tools being developed by MODs, for phenotype data in particular, have direct implications for human G2P databases. Third, the MOD experience illustrates the benefits that can be attained when smaller research communities work together towards a common goal, and at the same time raises the question of whether this can be replicated in the far larger, more diverse human genetics community.

⁴⁴<http://www.europenome.org>

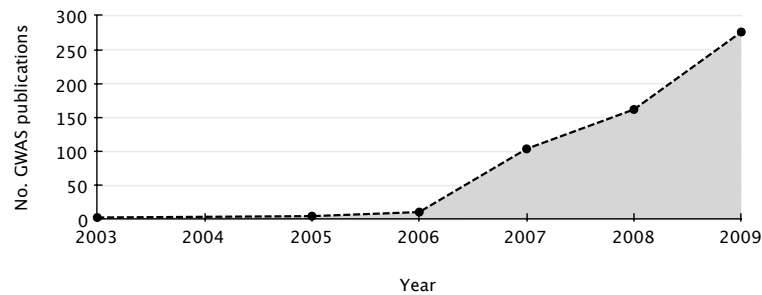


Fig. 2.3: GWAS publications per year, growing from a negligible number to nearly 300 in 2009 alone. Summary statistics from the HuGE Watch.

2.2.4 Association study catalogs and knowledgebases

As noted above, there is a vast amount of peer-reviewed literature on disease genetics from the past several decades of research. For an investigator in the field, identifying the relevant publications and synthesising accumulated and emerging knowledge for a given disorder is a big challenge, due to the sheer number of studies in the given sub-field and also due to diversity of study designs, quality, statistical power and other factors relevant to the interpretation of results (Little *et al.*, 2009; Pearson and Manolio, 2008; von Elm *et al.*, 2009). This problem has become particularly acute with the explosion of GWAS publications in the past 3-4 years, as previously discussed (see Fig. §2.3). The expert-curated, comprehensive disease-specific resources described later in this section are an enormous help with this for the specific disorders covered, but alas not nearly all areas of disease genetics research have such good support systems.

High-level G2P knowledgebases The general problem of keeping up with accumulating knowledge as reported in scholarly publications is not a new one. Specialised initiatives for navigating the G2P literature have been developed, such as the HuGE Navigator⁴⁵ (Yu *et al.*, 2008b). Based on abstracts indexed in PubMed, expert HuGE curators categorise and extract several key elements from publications in genetic epidemiology and populate

⁴⁵<http://hugenavigator.net>

a high-level *knowledgebase* of disease terms linked to genes, along with various other parameters such as study type (e.g. GWAS, clinical trial) and more. Various web-based tools, such as the HuGE Literature Finder, the Gene Prospector, HuGEPedia (Phenopedia and Genopedia) (Yu *et al.*, 2010) and the HuGE Watch, query this knowledgebase to create lists and views of the published literature and associated genes, filtered or prioritised by key parameters (e.g. all studies involving the BRCA2 gene, plots of trends over time, and keyword searches for gene symbol or phenotype). Such tools are designed to steer researchers towards a suitable starting point for a systematic literature review, and subsequent critical assessment of reported G2P associations. However, many of the generated displays are extensively cross-linked with other resources, such as Entrez Gene⁴⁶ and many of the databases described below, and thus they are highly useful on their own as an “encyclopaedic” knowledgebase. The underlying database has also been used as substrate for sophisticated software tools for literature screening, such as GAPscreeener (Yu *et al.*, 2008a).

GWAS catalogs. As useful as they are for their intended tasks, the HuGE Navigator suite and similar tools ultimately only help to narrow down the list of publications that the researcher must ultimately read, digest and synthesise, and are furthermore limited to information available in abstracts only. The desire for finer granularity of information, in particular in response to rapidly rising numbers of published GWAS reports, has spurred the creation of a number of efforts to gather further details from the literature. Curators of these catalogs collect, beyond the elements listed above, individual markers (usually dbSNP identifiers) found to be significantly associated with the disease or trait, along with the reported level of significance (p-value) and odds ratios (indicating the relative risk conferred by the risk allele). This information is then presented in relatively simple tabular views, with hyperlinks to the original papers and (similar to the HuGE Navigator tools) links to OMIM or other resources containing further disease information. Several of these association study catalogs of varying quality and usability have been created in

⁴⁶<http://www.ncbi.nlm.nih.gov/gene>

recent years, such as the Genetic Association Database (GAD)⁴⁷ (Becker *et al.*, 2004) and the NHGRI GWAS catalog already mentioned in §1.2.5.

The NHGRI catalog has garnered most attention, and as such it demonstrates how a useful balance can be struck with respect to study coverage, quality of curation and, importantly, design of user interface. The studies indexed in the catalog are presented as a table on a single web page, with filters for phenotype, p-value and several other criteria, which despite its simplicity is a boon to researchers navigating the GWAS literature. Because the catalog has data resolution down to the variant level, the collected data can be used as a basis for genome-wide analysis of disease-associated variants and their correlation with genome features, as demonstrated by the NHGRI catalog creators in a recent publication (Hindorff *et al.*, 2009a). However, the NHGRI catalog also has numerous drawbacks when used in this way, some of which Hindorff *et al.* acknowledge in their discussion. For example, various kinds of reporting bias from the original publications - e.g. which associated variants are included/excluded and which populations are studied - will be carried over directly into, and thus distort, any downstream analyses. This is in addition to the inherent study-level bias: given that only GWAS results published as journal articles are included in the catalog, this implies a bias against negative, “uninteresting” results which are much less frequently published in journals than positive findings (Blomqvist *et al.*, 2006; Brookes and Prince, 2005; Shields, 2000). The amount of information captured from each paper is also minimal and not presented in a consistent way, limiting the main utility of the catalog to simple perusal by a human.

2.2.5 Databases for phenotype-associated mutations in humans

The databasing of information relating Mendelian mutations and other genomic variants to various aspects of human phenotype has historically lagged behind what has been achieved by the MOD communities. This is due to a variety of reasons. Community size is certainly a factor, as noted in the previous section on MODs; the human G2P research community

⁴⁷<http://geneticassociationdb.nih.gov>

is a large, heterogeneous mix of biologists, clinicians, epidemiologists, statisticians and many other specialists, all of whom have their own view of how data should be structured, semantically encoded and interpreted, which makes agreeing on data standards difficult. The complexity of human G2P information is also a significant source of difficulty: apart from the general characteristics of biological information as noted previously, complexities relating to G2P information include (but are not limited to); primary data are generated by a wide range of sequencing, genotyping and other laboratory techniques; primary data can be analysed, re-analysed and interpreted in many different ways; and methods used to observe and define clinical and normal phenotypic traits are extremely diverse and difficult to standardise.

Central mutation databases. A number of catalogs aim to capture a broad picture of existing and emerging knowledge on Mendelian mutations and human disease. The best known of these is OMIM, already mentioned in Chapter 1. OMIM is a catalogue of human genetic disorders which first appeared in book form over 40 years ago (McKusick, 1966) under the name MIM and has been available online since 1990. Its success has spawned the parallel OMIA catalog⁴⁸ for non-human animals (Lenffer *et al.*, 2006). Central to the OMIM/A project model is a team of curators who manually extract G2P relationships from the literature to create a compendium of high-quality, narrative records for genes and disease phenotypes with a known or suspected genetic basis.

This impressive and highly useful collection of records on ~13,000 genes (of which ~2,700 harbour disease-causing mutations (as of February 2010)) may at first glance seem to speak to the strength of such centralised knowledge gathering. However, OMIM does not represent a comprehensive view of all human G2P knowledge; the task is simply too large for any one team to manage. Furthermore, as OMIM records are presented in narrative form in the style of traditional publications and does not make use of controlled vocabularies, the scope for advanced queries and deep integration with other data resources is limited. It should be noted, however, that even in this narrative form, OMIM is highly valuable as a

⁴⁸<http://www.ncbi.nlm.nih.gov/sites/entrez?db=omia>

key reference information source for text mining tasks, e.g. to classify phenotypes based on similarity (van Driel *et al.*, 2006) and as a foundation for constructing ontologies of disease phenotypes (further discussed below).

Similar challenges to centralised gathering of G2P knowledge are facing the Human Gene Mutation Database (HGMD)⁴⁹ (Stenson *et al.*, 2009). Like OMIM, HGMD is largely populated by manual extraction of information from published reports on disease-causing Mendelian mutations, but it also incorporates information on mutations with non-clinical functional consequences. Mutation data are gathered not only from published papers but also include unpublished mutations collected from other mutation databases (see below), albeit with the caveat that only the first report of a given mutation is included⁵⁰. This would seem to significantly limit HGMD's usefulness, as subsequent reports which may independently validate/refute or withdraw erroneous associations with a phenotype will not be represented. In contrast to OMIM's narrative style, database records in HGMD are structured, which facilitates more specific database queries and other functionality.

Potential reuse value of HGMD is, however, hampered by the sustainability model employed: in the absence of public financial support, funding for HGMD operations comes from paid subscription to the full "Professional" version of the database from academic/non-profit and industry users. The Professional version includes additional query functions, bulk data export options and, importantly, an up-to-date version of the full database which (as of Sep 2009) holds over 90,000 mutation entries. Free access to non-commercial users is restricted to records that have passed out of a 2 1/2 year embargo (<70,000 as of Sep 2009) since initial incorporation into the database, and bulk data downloads are not possible unless explicitly allowed on a collaborative basis, following the signing of a confidentiality agreement.

On the whole, these restrictions on HGMD data access and reuse run contrary to the spirit of open scientific dissemination and prevent researchers with limited funds from accessing

⁴⁹<http://www.hgmd.org>

⁵⁰http://www.hgmd.cf.ac.uk/docs/new_back.html

the latest mutation data. Nevertheless, it is acknowledged that databases such as HGMD, which have only partial or no institutional backing, need to raise funds in some way in order to continue to operate. HGMD's business model is merely one way of achieving this.⁵¹ It should also be noted that, although HGMD's current implementation for controlling access to its intellectual property severely limits integration of HGMD content with other G2P data, there is potential for addressing these shortcomings by leveraging emerging Web 2.0 technologies. These will be further discussed in later chapters.

Knowledge of G2P relationships, particularly association of genetic variation with differential drug response, is important to the drug development process (Kalow, 2002). An example of a G2P database resource supporting such activities is the Pharmacogenomics Knowledge Base (PharmGKB)⁵² (Sanguhl *et al.*, 2008) which provides access to primary data from published clinical trial studies deposited by the community. PharmGKB has a dual remit as a specialized web portal to support pharmacogenomics research and as an "educational portal" for non-specialists. Like OMIM and HGMD above, and many MODs, incorporated data on the effect of variants (e.g. gene expression or protein three-dimensional structure) are curated by PharmGKB staff and linked with other data and published literature. Relevant publications are gathered using a combined strategy of manual curation on one hand, and automated retrieval and annotation on the other. It is worth noting that the Pharmspresso literature mining tool (Garten and Altman, 2009) used for the automated part of this is based on Textpresso (Muller *et al.*, 2004) which was originally created for the *C. elegans* literature. Pharmspresso thus represents an excellent example of human-focused G2P research tools being built on previous work in the MOD community.

⁵¹A detailed discussion of challenges and potential solutions for ensuring long-term maintenance and development of biological databases and other resources is out of scope for this work. Chandras *et al.* (2009) provide a useful discussion on this topic.

⁵²<http://www.pharmgkb.org>

Locus-specific databases (LSDBs). In contrast to genome-wide G2P resources described above which are “shallow” with respect to the level of detail in the information collected for each entry, many groups have collated in detailed fashion reported mutations and phenotypic consequences for just one or a few genes of relevance to one or a few related diseases. The first of these LSDBs was published in 1976, in the form of a printed list of ~200 human globin mutations (Lehmann and Kynoch, 1976), followed by other printed compendia which were later included in the Globin Gene Server⁵³ (Hardison *et al.*, 1998) and the HbVar database⁵⁴ (Giardine *et al.*, 2007b). According to a listing maintained on Human Genome Variation Society (HGVS) website⁵⁵ (Horaitis *et al.*, 2007), as of March 2010 over 1,400 LSDBs are now in operation. Most of these are accessible over the Internet.

Similar to their species-specific MOD counterparts, LSDBs tend to be rich in information content and are enhanced by domain-specific expert curation. As well as containing published information mined from the literature, LSDBs typically include unpublished DNA variation along with evidence concerning pathogenicity, or consequence of the reported mutation (Horaitis and Cotton, 2004). Consequently, a given LSDB is potentially a rich source of extensive, deep and curated information of high value to researchers on its own, and even more so when linked with other, related LSDBs and datasets from large-scale genome projects. Pilot projects aiming to place LSDB information on genome browsers have recently been undertaken, such as the PhenCode project⁵⁶ (Giardine *et al.*, 2007a).

Unfortunately, LSDB integration projects such as PhenCode have been quite difficult to undertake. One major reason for this is a severe lack of interoperability. LSDBs have historically been created independently of one another, with little coordination

⁵³<http://globin.cse.psu.edu>

⁵⁴<http://globin.cse.psu.edu/hbvar/>

⁵⁵<http://www.hgvs.org/dblist/glsdb.html>, Date accessed: 2010-04-12. Archived by WebCite® at <http://www.webcitation.org/5ow6K1lqe>

⁵⁶<http://www.bx.psu.edu/phencode>

or harmonisation, and with little or no dedicated funding. Database implementations range from simple, non-networked spreadsheets through to fully-fledged online databases. Consequently, the LSDB “world” represents a heterogeneous, fragmented network of data-rich “silos”, across which it is difficult or impossible to exchange or integrate G2P information. Another obstacle to data integration is reluctance in the LSDB community to share data. den Dunnen *et al.* (2009) list a number of concerns expressed by LSDB curators about data sharing, regarding issues such as data quality, provenance and ownership. Notably, curators fear that LSDB data aggregated by central databases will be displayed without proper attribution, used for commercial purposes without permission and/or sharing of profit, or that clinical information such as mutation consequence will be misrepresented or in some other way inappropriately interpreted. However, LSDB curators are gradually realising that in order to make full use of their data, some degree of cross-database sharing is necessary and would bring advantages (such as increased recognition) to themselves and other stakeholders involved. Partially under the auspice of the GEN2PHEN project (further described below), the LSDB community is currently developing recommendations for data sharing between LSDBs and central data repositories (den Dunnen *et al.*, 2009).

To summarise, a common feature of central mutation databases is that they rely directly or indirectly on harvesting published G2P information in a centralised fashion. However, as noted above, the sheer size and complexity of the task of gathering all G2P knowledge in this way is beyond the reach of any single group, and furthermore such an approach fails to capture important datasets that are not published in journals. This suggests fundamental limitations to relying on this databasing strategy alone. Nevertheless, OMIM, HGMD and other databases, though not comprehensive, are important research tools, especially when cross-linked with other databases such as LSDBs.

At the opposite end of this spectrum, LSDBs provide a detailed look at only a small number of variants in which their curators are interested. Due to their small size and direct links to the community investigating particular genetic disorders, LSDBs are better suited than the

central archives to responding to the needs of those researchers, thereby complementing the global catalogs. Because of the diverse nature of the various disorders and disease genes studied, and the methods and strategies employed by researchers, it would be difficult or impossible to usefully centralise those activities. Critically, the decentralised approach can also address numerous hard-to-grapple issues such as data governance and data ownership. On the other hand, as a consequence of their largely-uncoordinated activities, the diverse organisation and architecture of LSDBs makes data integration amongst LSDBs themselves and with central databases difficult or impossible. The LSDB world indeed illustrates, on a smaller scale, the more general problem resulting from proliferation and diversification of biological databases as noted above. However, this situation appears to be taking a turn for the better in recent years. Most LSDBs now use a common method - the HGVS nomenclature⁵⁷ (den Dunnen and Antonarakis, 2000) - for describing mutations in journal papers and database records. A recently-initiated project in the LSDB community aims to anchor these descriptions in so-called Locus Reference Genomic sequences, or LRGs⁵⁸ (Dagleish *et al.*, 2010), in order deal with long-standing problems relating to the use of non-standard and/or obsolete sequences as a reference for reporting mutations. LSDBs are also gradually converging on a small number of off-the-shelf software packages - the Leiden Open Variation Database (LOVD)⁵⁹ (Fokkema *et al.*, 2005), MutBase⁶⁰ (Riikonen and Vihinen, 1999) and the Universal Mutation Database (UMD)⁶¹ (Beroud *et al.*, 2005) - which are striving towards interoperability. These moves towards increased standardization and software reuse closely resemble that described for MODs in the previous section, and this is likely to be crucial to facilitating future connectivity between these databases.

⁵⁷<http://www.hgvs.org/mutnomen/>

⁵⁸<http://www.lrg-sequence.org>

⁵⁹<http://www.lovd.nl>

⁶⁰<http://bioinf.uta.fi/MUTbase/>

⁶¹<http://www.umd.be>

2.2.6 Disease-specific portals

A number of human G2P data resources are something of a hybrid between an LSDBs and an MOD: they focus on a particular well-studied disorder, but do so in a comprehensive way by including not just mutations in associated genes but all aspects of genetic research into that disorder. An example of this is T1DBase⁶² (Hulbert *et al.*, 2007) which supports the Type 1 diabetes (T1D) research community. The array of resources offered by T1DBase includes genome browsers for visualising T1D-associated genes, analysis tools and data downloads, as well as links to information on rat strains which are models for the disorder. Some disease-specific web resources go far beyond the remit of a genetic research tool and aim to be a fully comprehensive web portal for researchers, clinicians and patients. The best known of these is the Alzheimer Research Forum (AlzForum)⁶³, a community web portal created 13 years ago in the early days of the Web. The AlzForum portal offers a plethora of resources, including discussion forums, curated papers from the vast literature on this genetically-complex neurodegenerative disorder and news on the development of drugs and other treatments. An important part of the Alzforum arsenal is the AlzGene database⁶⁴ (Bertram *et al.*, 2007), a comprehensive collection of disease-gene association data from several hundred genetic association studies of Alzheimer's disease undertaken in the past two decades. As such, the AlzGene catalog is similar to high-level G2P knowledgebases, such as the HuGE Navigator discussed in §2.2.4, but it adds further value by also presenting results from systematic meta-analyses of pooled data from many studies.

Drawing parallels with some of the MOD websites described above, a major reason for the success of Alzforum, T1DBase and similar resources is undoubtedly a sizeable research community focused on one over-arching goal: improving the lives of patients suffering from one disorder, and to, ultimately, find a cure. But a key additional factor is that Alzforum is continually maintained and developed, an enviable situation which is

⁶²<http://www.t1dbase.org>

⁶³<http://www.alzforum.org>

⁶⁴<http://www.alzgene.org>

made possible by sustained financial support from funders who appreciate the importance of such a resource to research into the disease.

Similarities with the MOD world do not end there. Demonstrating yet again the value of software reuse, the software and analysis tools created by Alzforum also underpin other web portals built for a similar purpose for other disorders; PDGene⁶⁵ and SZGene⁶⁶ (Allen *et al.*, 2008), for research into Parkinson’s disease and schizophrenia, respectively. Similarly, the open-source GDxBase software⁶⁷ supports the T1DBase website and several other portals, including HDBase, a community website for Huntington’s disease⁶⁸.

2.2.7 Archives for primary and aggregate G2P data

Funding bodies and other stakeholders are gradually realising that in order to maximise the utility of G2P research data and accelerate scientific progress, primary data need to be disseminated more widely than has traditionally been the case. Large-scale collaborative research initiatives such as the Genetic Association Information Network (GAIN) in the US (Manolio *et al.*, 2007) and the Wellcome Trust Case-Control Consortium (WTCCC)⁶⁹ in the UK (The Wellcome Trust Case Control Consortium, 2007) now mandate broad sharing of research data to any “qualified” researcher, i.e. beyond the small group of collaborators participating in the project.

In their discussion of this “changing landscape” of primary data sharing in genetic research, Kaye *et al.* (2009) highlight several challenges to progress in the field, notably the use of more restrictive data release policies than have been traditionally applied in high-throughput genomics projects. Unrestricted sharing of genotypes, phenotypes and other person-specific data on study participants over the Internet can lead to breach of anonymity, so the personal privacy implications are considerable. Therefore, the open-access data

⁶⁵<http://www.pdgene.org>

⁶⁶<http://www.schizophreniaforum.org/res/sczgene/>

⁶⁷<http://www.gdxbase.org>

⁶⁸<http://www.hdbase.org>

⁶⁹<http://www.wtccc.org.uk>

release policies (as applied in the Human Genome Project, HapMap and other large-scale “community resource” projects in genomics) are not appropriate in this setting.

In response to these concerns, a new breed of digital repositories - *G2P archives* - have recently emerged, designed for secure, long-term storage of primary G2P research data and for controlled dissemination of these data to researchers and clinicians with appropriate data access permissions. Two main centralised G2P archives are now in operation, one in the USA and the other in Europe.

dbGaP. The Database of Genotype and Phenotype (dbGaP)⁷⁰ is operated by the NCBI and was launched in 2007, in close collaboration with dbSNP and other NCBI data resources. As described by Mailman *et al.* (2007), dbGaP employs a tiered access model, whereby the database is divided into open-access and controlled-access sections. The open-access tier contains a wealth of documents describing how a GWAS was conducted, details on each phenotypic variable measured and study metadata, all of which are available for open access for users to peruse and evaluate whether or not the protected, individual-level data is relevant to their research. In the controlled-access tier, genotypes, phenotypes and other data on individuals, as well as summary-level frequency data and association analysis results, are only accessible to users who have been authorised for access.

In order to access the protected data for a given study, a researcher must submit a request via the dbGaP authorised access portal⁷¹. dbGaP forwards the request to the appropriate Data Access Committee (DAC) at the NIH institute which sponsored the study. If the request is approved by the DAC, the researcher is assigned user account credentials on the dbGaP website, which he can then use to sign in and retrieve the individual-level data for the study.

dbGaP is also engaged in disseminating non-GWAS data released under a controlled-access policy, such as sequencing data from the Cancer Genome Atlas project⁷². More generally,

⁷⁰<http://www.ncbi.nlm.nih.gov/gap>

⁷¹<http://view.ncbi.nlm.nih.gov/dbgap-controlled>

⁷²<http://cancergenome.nih.gov>

the primary aim of dbGaP is to serve as a central archive for controlled-access data from all large-scale G2P studies funded by the NIH, and for this and a variety of legal and other reasons not discussed here it is in effect a US-only solution. Other repositories are therefore needed to archive data from G2P studies conducted elsewhere in the world.

EGA. Currently the other main G2P archive outside the US is the European Genome-phenome Archive (EGA)⁷³ which went online in July 2008. EGA comprises a parallel infrastructure being developed by the EBI for secure, encrypted storage of data which cannot be deposited into the EBI/EMBL public databases (I. Lappalainen, personal communication). This includes, at present, mainly GWAS data, but will in the future also include DNA sequences and other high-throughput data. EGA's controlled-access protocol is similar to that of dbGaP, in that authorisation decisions are not made by EGA, but instead delegated to the appropriate external body which governs access to study data. Importantly, in contrast to dbGaP this governing body (the data access-granting organisation, or DAO) is normally formed from the organisation which was responsible for generation of the original data (e.g. WTCCC), not from the funding body.

Apart from the aforementioned difference regarding who authorises access, dbGaP and EGA both fulfil the same over-arching purpose of facilitating deposition and sharing of sensitive G2P study data for analysis in a secure way. Beyond these commonalities, the two controlled-access archives are quite different with respect to the features offered, which no doubt reflects their different remits and, to some extent, different stage in the development cycle. EGA at present functions as a no-frills data “clearinghouse”, with only minimal study information and limited information on individual samples beyond genotypes. This sharply contrasts with the multitude of web-based tools, integration with a genome browser and rich, detailed study information provided by the much more comprehensive dbGaP. As of February 2010, dbGaP and EGA together contain nearly 100 studies in various stages of completion, most of which are GWASs. Outside of these two main archives, a

⁷³<http://www.ebi.ac.uk/ega/>

number of smaller, project- or country-specific archives exist. Some have the remit of wide dissemination of study data, including the Case Control GWAS Database (GWAS DB)⁷⁴ (Koike *et al.*, 2009) which contains results on several disorders studied in Japanese cohorts. Others are constructed to facilitate data sharing only among partners in a consortium, such as the European Network for Genetic and Genomic Epidemiology (ENGAGE)⁷⁵ for several studies into cardiovascular disease which involve ~600,000 individuals from several European populations (see e.g. (Aulchenko *et al.*, 2009)), and other biobank-based projects.

Whilst these various archives do indeed facilitate secure access to primary G2P data, their overall usefulness is degraded by the fact that datasets relevant to a given research problem (e.g. all published GWAS results for a particular disease) will often be fragmented across two or more such data silos, with varying requirements for data access (some are completely closed to non-partners). This clearly impedes, or prevents, useful cross-database data integration. These impediments may be alleviated in part by the planned exchange of study metadata between EGA and dbGaP, which (with participation from other G2P archives) could facilitate the creation of a global, cross-archive study catalog (S. Sherry and I. Lappalainen, personal communication). However, this plan will only help with study data deposited in the participating, centralised archives. The myriad genetic association study generated to date, are by and large, not published in comprehensive fashion, neither in journals nor in databases, and as noted in §2.2.4 negative studies are often not published at all. The Human Genome Variation database of Genotype to Phenotype information (HGVbaseG2P)⁷⁶ (Thorisson *et al.*, 2008) aims to address these issues by providing a comprehensive view of all association study findings and access to aggregate study data via a central web portal. HGVbaseG2P is comprehensively documented and discussed in Chapter 5 and will thus not be further described here.

⁷⁴<http://gwas.lifesciencedb.jp>

⁷⁵<http://www.euengage.org>

⁷⁶<http://www.hgvbaseg2p.org>

2.3 Challenges for modern G2P databases

The diverse assortment of online G2P data resources surveyed in the previous section illustrates that comprehensively organising the whole of the G2P information space is not a straightforward task. This survey is by no means complete, in that it samples only a minority of the vast, heterogeneous collection of G2P databases that currently exist. Nevertheless, for the purpose of this discussion, it provides a useful cross-section of mainstream data resources in the field and, notably, of experiences of centralised databasing. Based on these experiences, and also considering recent technology developments, it is evident that several key areas need attention if G2P research data are to be effectively exploited.

2.3.1 Data quantity

An exponential rate of increase in data volumes has moved the bottleneck in the knowledge-generating process from the data production stage to the data management and analysis stage. For example, primary GWAS data generated with the latest SNP microarrays for tens of thousands of samples contain several billion data points. Management of these datasets is further complicated by repeated splitting and merging of results, based on the multitude of phenotypes or sub-phenotypes characterised, or by pooling of data across multiple studies, creating yet more derived datasets. The data volume will further increase as greater numbers of SNPs and other forms of variation are routinely assayed, and with the lowering costs of large-scale experiments in other areas of bioscience, such as transcriptomics, proteomics and metabolomics.

Also, a veritable deluge of data is now coming from next-generation sequencing projects, as whole-genome resequencing, RNA sequencing and other applications of the new technologies become commonplace. On a local level, the data volumes produced by the current generation of these devices presents major challenges for smaller laboratories who purchase and operate them, but then lack the required computational infrastructure and in-house expertise for storing and processing the data. Building infrastructure and expertise

to deal with the sequencing “data glut” takes time and is expensive, and this is presently hindering many smaller laboratories in adopting and taking full advantage of the new technology (Editors, 2008b; McPherson, 2009).

On a global level, efforts to create central archives for petabyte-scale data from next-generation platforms are already stretching available data centre resources to the limit (Doctorow, 2008), even with the relatively small number of genomes fully sequenced to date. Mass transfer of sequence data is starting to strain the backbone of the Internet itself and necessitates the use of specialized, proprietary network protocols such as Aspera⁷⁷ for efficient transfers between data centres. Meanwhile, many data submitters and data consumers without sufficient network bandwidth are relegated to shipping sequence data on portable hard drives. This does not bode well for a future where an individual’s genome can be sequenced in less than a day and tens of thousands, if not millions, of whole-genome sequences will need to be archived and analysed.

It is nevertheless encouraging that current increases in data volume have not yet outstripped advances in computer technology. At present, building and operating data centres for petabytes of research data requires significant financial investment and is a challenging, but not insurmountable, task requiring considerable expertise. For the future, innovations in computer technology will continue to be driven by multiple forces outside the biosciences. For example, large scale experimental facilities such as the Large Hadron Collider (LHC)⁷⁸ gather terabytes of data every day of operation, requiring specialized equipment to store and process the data generated (Deatrich *et al.*, 2008).

In the private sector, massive computing farms operated by IT companies such as Google are driving innovation in computer hardware, by optimising design and operation of data centres. New software techniques are being developed for highly-redundant, distributed and scalable databases, parallel filesystems and programming models. Examples include Google’s MapReduce (Dean and Ghemawat, 2008) and the open-source Hadoop engine⁷⁹.

⁷⁷<http://www.asperasoft.com>

⁷⁸<http://lhc.web.cern.ch/lhc/>

⁷⁹<http://hadoop.apache.org>

These technologies are already being applied to problems in bioinformatics, such as sequence read mapping (Schatz, 2009) and variant discovery (Langmead *et al.*, 2009). Also, intensive scientific computation is increasingly performed in the “cloud” - that is, using on-demand, rent-as-you-go computing capacity over the network, provided by commercial vendors (Bateman and Wood, 2009; Sansom, 2010). So, while Moore’s law⁸⁰ continues to hold, it would seem that the field can continue to “piggyback” on these technological advances and, where necessary, adopt new distributed computing approaches (Dudley and Butte, 2010; Stein, 2010). Data quantity must nevertheless be a key consideration in database design and operation.

2.3.2 Data complexity

Whilst the challenge of data quantity is one that will in principle be addressed by advances in information technology and analysis algorithms, the matter of data complexity is far less tractable. Biological data differs from that of most other “Big Science” by its high level of complexity. For instance, in astronomy research, datasets comprising high-resolution digitised images are becoming larger and more homogeneous (Szalay *et al.*, 2000). A similar trend can be observed in certain sub-disciplines in the biosciences, e.g. in large-scale genomics, as the field settles on a limited number of high-throughput experimental platforms and standards for exchanging such data are increasingly available (Editors, 2008c). But in many other respects the complexity of biological data is increasing, and in the G2P domain especially so. For example, major biobanking projects, such as UK Biobank mentioned in §1.2.4 initiated in 2008 and the Framingham Heart Study (FHS)⁸¹ (Cupples *et al.*, 2007) initiated in 1948, collect thousands of phenotypic variables and other observations in a prospective manner for tens or hundreds of thousands of subjects. The specific variables collected in these longitudinal studies may change over time as knowledge advances and new types of observations become possible (e.g. techniques for

⁸⁰Moore (1965) described a trend in the computer industry whereby processing power and storage increases exponentially, doubling approx. every two years per unit cost.

⁸¹<http://www.framinghamheartstudy.org>

genome interrogation, as already discussed in §1.2). The generated datasets are rich and dynamic (e.g. individual phenotype categorisation may change with age and treatment). A further layer of complexity is added by information about *how* genotype influences phenotype. This connection - the “2” in G2P - is often far from clear-cut; numerous and constantly evolving analytical methods are used to characterise the complex interplay between the DNA “blueprint” and the environment, lifestyle and other factors, and how this interplay results in the final emergent phenotype. Research results therefore often provide only clues (sometimes contradictory ones), rather than facts, as to the underlying etiological processes, which complicates the incorporation and utilisation of this information into a clinical decision-making process. Thus, the complexity of both experimental results and other observations on one hand, and analytical methods on the other, combine to make storage, exchange, integration and presentation of G2P information a difficult task.

2.3.3 Data quality

The data available to researchers today and in the future will be only useful if they are of sufficient quality. This principle was recognised early on in the genomic era with the adoption of the Bermuda principles⁸², which stipulated that sequences generated in the Human Genome Project would be categorised as either “finished” (less than one error in 10,000 bp, the “gold standard”) or “draft”. To meet the demands of data generated by the latest sequencing technologies, more fine-grained schemes with several levels of completeness and quality have recently been proposed (Chain *et al.*, 2009). It should be noted, however, that “quality” is a rather general concept, encompassing a number of more specific measures which depend on the type of the data, context and other factors. For example, Macmullen (2007) lists in his analysis the following quality measures, or “facets”, which are important for GO annotations in MODs; consistency, reliability, specificity, completeness, accuracy and validity. For this discussion, however, these nuances are not

⁸²http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml

critical, and so merely the broader term of overall data quality will be used.

Even though database records will never be completely error free, efforts to avoid inaccuracies and errors, and ensure a certain level of quality, are arguably an important aspect of any biological databasing enterprise. This work falls under the broad remit of *biocuration*, defined by Howe *et al.* (2008) as “the activity of organising, representing and making biological information accessible to both humans and computers”. A major aspect of biocuration involves the extraction of knowledge from the literature and other sources, as previously discussed. This important work, and many other curation tasks are well illustrated by Shimoyama *et al.* (2009) in their account of the work undertaken by Rat Genome Database expert curators, in particular the overseeing of data acquisition and resolving data conflicts.

Quality control should naturally be optimised from the stage of data generation, but databases can only become involved from the stage of guiding researchers in the preparation of accurate and appropriate data submissions. Once data are received, databases should then deploy their own quality assurance measures to check for internal consistency and completeness of the submission. Links to external database entries can be checked for validity and consistency with other datasets assessed; for example, by comparing SNP allele frequencies with previous datasets to identify fundamental laboratory or data management errors, or cases where the wrong DNA strand has been referenced. Such automated validation will become increasingly important as dataset sizes grow and manual inspection of individual data elements is not practical. Across the full breadth of G2P data there are many features that could be checked to ensure accuracy. Standards and guidance need to be developed to underpin data curation throughout the path of generating data through to placing it in public databases.

Automated validation checks do not solve all problems, however. For example, Ioannidis *et al.* (2009) demonstrated that primary microarray datasets can pass automated quality checks after submission to a central database, and yet fail to reproducibly support published results when re-analysed. The general problem of quality in data generation and missing, incomplete or insufficiently annotated datasets is beyond the scope of this discussion.

Nevertheless, it is worth noting that the potential for data reuse and repurposing is heavily reliant on *data provenance*. Provenance refers to information on how and when a given dataset was generated, by whom, in what context, how it was captured or computed, and more. Another name for this information is *metadata* (i.e. data which comes after, or describes, the data) and it is often on the basis of metadata that a scientist will not only locate a dataset of interest, but also judge whether a given dataset is suitable for reuse for his research and of sufficient, and known, quality. Data provenance is thus an important aspect of research data management, and a key role of databases is to ensure that a minimal level of provenance information accompany the actual dataset upon submission. Increased use of minimal information checklists (Taylor *et al.*, 2008) will play a key role in this regard, as further discussed in §2.6 below.

Looking to the future, the rising number, size and complexity of G2P datasets argues that expert curation is becoming more important than ever before. Some authors have suggested that the gap between expert manpower and the flood of data that requires curation can be closed, at least in part, with so-called community curation, as described by Salzberg (2007) in the context of genome re-annotation. As Osborne *et al.* (2007) point out, the success of Gene Reference Into Function (GeneRIF)⁸³, a long-running collaborative resource linked to the NCBI Entrez Gene database⁸⁴ (Maglott *et al.*, 2007), indicates that principle has some traction in the community.

A number of recently launched community curation projects are based on the well-known wiki model (Waldrop, 2008), best exemplified by the hugely popular Wikipedia⁸⁵. Examples of promising projects in this arena include WikiProteins⁸⁶ (Mons *et al.*, 2008) and WikiGenes⁸⁷ (Hoffmann, 2008), notable for the emphasis placed by those projects on

⁸³<http://www.ncbi.nlm.nih.gov/projects/GeneRIF>

⁸⁴<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>

⁸⁵<http://www.wikipedia.org>

⁸⁶<http://www.wikiprofessional.org>

⁸⁷<http://www.wikigenes.org>

recognised authorship and discrete, fine-grained contribution tracking. A community of experts has also been mobilised with good effect to annotate RNA families in the RNA WikiProject within Wikipedia itself⁸⁸ (Daub *et al.*, 2008). However, as Kyrpides (2009) argues fiercely, community efforts by volunteers, while useful on their own merit, perhaps should not be relied on as the sole source of data curation and should at best complement and augment (rather than replace) high-quality data curation by professional experts.

In a recent paper, representatives of the main MODs and several other prominent figures in the field (Howe *et al.*, 2008) discuss the need for far greater emphasis on professional data curation to grapple with the “data avalanche”, particularly with respect to making it attractive for domain experts to undertake a career in curating. The recently-established International Society for Biocuration⁸⁹ and international conferences on biocuration in recent years are steps in the right direction. The UK Digital Curation Centre (DCC)⁹⁰, established in 2004, is an example of a more general initiative in this area. Another key enabler is likely be an incentive-based system for systematically crediting curators for their contributions to new or quality-enhanced database content (Seringhaus and Gerstein, 2007), which will be further discussed in Chapter 5.

2.3.4 Data discovery

Published G2P information has limited value if it cannot be relocated and reused by the researchers who need to access it. There are several intertwined issues relating to the challenge of making G2P information accessible. The first of these is *data discovery*. The task of finding relevant data is becoming increasingly difficult for researchers as G2P datasets grow larger, more diverse and spread across a large number of databases. Indeed, as Cannata *et al.* (2005) point out, merely finding the appropriate databases to search in an expanding bioinformatics “resourceome” is a growing problem for researchers and has

⁸⁸http://en.wikipedia.org/wiki/Wikipedia:WikiProject_RNA

⁸⁹<http://www.biocurator.org>

⁹⁰<http://www.dcc.ac.uk>

prompted initiatives such as the Database Description Framework (DDF)⁹¹ (Smedley *et al.*, 2010) to enhance discoverability and quality-assessment of bioinformatics data resources. Finding the right data resources is thus a challenge in and of itself, but to properly tackle this problem one obvious strategy is to develop more powerful database search tools and, equally importantly, to ensure that these improved search tools are connected to the relevant data.

At first glance this may be seen as an argument for the central database model. But as the previous sections amply illustrate, data size, complexity and sensitivity (and other issues surrounding data warehousing) make gathering and managing all the relevant G2P information in a central repository an imperfect solution at best. Therefore, searching and retrieval of data across many different databases will be required, such that the information itself never needs to leave its remote source. An example of this from outside the G2P domain is the ENCODEdb portal⁹² (Elnitski *et al.*, 2007), which offers a simple query interface that searches across all primary high-throughput experimental data from the Encyclopaedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium, 2004) deposited in several public databases. This sort of distributed search and retrieval is the essence of federated database solutions (as discussed below), which are now technically feasible and being developed and deployed in the life sciences.

2.3.5 Data access

The next issue to consider once relevant data have been located is *data access*. This relates to whether the discovered data are released for open-access (i.e. can be downloaded by anyone), or whether there are conditions for access. Ideally, all G2P research data should be released for open-access to maximise the potential for reuse, but as previously discussed this may not always be possible for a number of reasons, e.g. intellectual property issues, privacy concerns or reluctance on behalf of researchers to share.

⁹¹http://www.casimir.org.uk/casimir_ddf

⁹²<http://research.nhgri.nih.gov/ENCODEdb/>

In the context of sharing data about human subjects participating in disease research and genomics in general, privacy is a thorny issue. Online access to so-called “identifiable” data (i.e. data that can be used to re-identify anonymous sample donors) is typically strictly controlled. Current practice for most primary G2P archives is to implement such access controls independently of one another, with little or no coordination, requiring researchers to obtain a separate username and password for each site, and to subsequently request access to each dataset separately. Needless to say, as the number of G2P archives grows, this quickly becomes burdensome for investigators who need to combine and analyse increasing numbers of distributed datasets. A further challenge is posed by the dataset sizes; data consumers typically need to download each of possibly many datasets in its entirety and process them locally, even if only a subset or “slice” of the data or data of lower resolution are required.

Until recently, a compromise which addressed some of the privacy and data size issues was the practice of disseminating without restriction aggregate data, or summary-level representations of large-scale genotype datasets (i.e. allele and genotype frequencies, but no data elements on individuals). The National Cancer Institute (NCI) Cancer Genetic Markers of Susceptibility (CGEMS) project⁹³ was one of the first projects to do this when they published aggregate data from some of the very first GWAS scans conducted (further discussed in Chapter 3). The same practice was adopted by dbGaP on its launch in early 2007. In addition to the benefit to individual researchers, this enabled secondary data providers such as HGVbaseG2P to gather aggregate GWAS data and redistribute them via web portals, perform high-level meta-analysis and develop visualisation tools, and thus provide added-value to researchers (see further discussion in Chapter 5). However, such activities were derailed by the findings of Homer *et al.* (2008) who showed that it was, in some circumstances, possible to predict whether an individual participated in a GWAS based on only aggregate frequency information on study subjects (e.g. GWAS case and control groups). The technique, originally motivated by forensic applications, has since

⁹³<http://cgems.cancer.gov>

been confirmed and continues to be debated in the literature (see e.g. Braun *et al.* (2009); Jacobs *et al.* (2009); Sankararaman *et al.* (2009)).

The findings of Homer *et al.* prompted an immediate flurry of response from data providers, funders and investigators (see e.g. Zerhouni and Nabel (2008)) and rapidly resulted in complete withdrawal of aggregate data available in the public domain at the time from CGEMS, dbGaP, WTCCC and a number of other download locations. These responses have been criticised by many as “knee-jerk” and premature, and it has been said that the lengthy approval procedures now required for access to any genome-wide data - aggregate or not - impedes research (Gilbert, 2008). Several authors have welcomed a side-effect of these events, which has been to draw to the forefront a long-standing debate over how best to balance the benefits of data sharing in genomic research versus protecting privacy and confidentiality (Lowrance and Collins, 2007).

The issue of *data identifiability* - the potential for linking genetic, phenotypic and other data to individual study participants - is becoming increasingly important as such data are collected not only in large-scale medical studies, but also by personal genomics companies who provide commercial genotyping and sequencing services to individuals (Editors, 2008a; Prainsack *et al.*, 2008). Given that the increased volume of personal data makes it more and more likely that an anonymous participant can be re-identified, consent and the promise of anonymity are no longer clear-cut issues. For example, O’Brien (2009) argues that “the concept and promise of absolute anonymisation should be dropped since it cannot be guaranteed and precludes recontact of research participants for significant future personal medical discovery that would benefit them”. Proponents of the Personal Genome Project have advocated for open-consent models which explicitly stipulate full disclosure of study data (Lunshof *et al.*, 2008). For most studies, however, protection of privacy of participants is important.

In conclusion, there are many challenging issues regarding privacy, confidentiality and consent which greatly complicate G2P data sharing and reuse. These issues clearly indicate the need for an ethics advisory voice as an integral part of every G2P database. Regarding

the issue of data access, the present default position by data providers is to apply the same, onerous access restrictions to all potentially identifiable data, irrespective of actual risk of identification. This extreme stance encumbers effective data reuse. But data release policies could be modified to alleviate these problems, and some of the technological solutions which may ultimately play an important role in streamlining data access will be discussed in Chapter 5. Finally, if it will not be possible to completely ensure the anonymity of all research participants, then perhaps the optimal way forward may be to accept this, to make data more freely available if consent allows, and concentrate instead on preventing and punishing abuse of the data.

2.3.6 Knowledge representation

As more analyses are performed on ever more extensive, cross-domain datasets, it will become increasingly difficult to comprehensively gather and present all the resulting hypotheses, tests and conclusions. The issue of how to present this knowledge is distinct from the question of which tools and systems are developed to generate this knowledge, and how the systems interface with databases. The bulk of published scientific knowledge is published as narratives or “stories” in journals, a form of presentation that humans are comfortable with, but which has important limitations: the sheer volume of information, as well as the fact that the narrative is not amenable to processing, analysis and understanding by computers.

Expert literature curation and knowledge extraction is part of the answer to these problems. But as noted above the volume of knowledge makes this impractical unless the level of curation is “shallow”, i.e. limited information is extracted (e.g. HuGE Navigator, HGMD) or if “deep” curation is limited to specific sub-areas of research, such as a particular disorder (e.g. Alzheimer’s) or a model organism. This has led many to conclude that a different approach is needed to explicitly capture knowledge “upstream” in the process, when a manuscript is prepared for publication. Indeed, there is a desire to capture not just the essence of an article - i.e. the “knowledge” end product - but also to record the

experimental and other methodology (or workflow) that led up to the results (Neylon, 2009). This, then, is an extension of the knowledgebase principle of HuGE Navigator, OMIM and others, albeit with the much more ambitious aim of capturing all scientific knowledge and to do so in a structured, machine-processable form.

Some authors have questioned the ability of journals and the “human-readable aliquots” of traditional articles to adequately capture findings from modern high-throughput research and have called for an overhaul of the publishing process (Seringhaus and Gerstein, 2007). Some have even gone so far as to call for the “death” of the scientific article (Seringhaus and Gerstein, 2006). Predictions for the demise of scientific journals altogether go back the pre-Web era; Laporte *et al.* (1995) surmised rather amusingly that the “clicks, beeps, and whirrs of computers linked to the Internet” may be the “death knell of biomedical journals as we know them”. But the narrative story presents an important device for us as humans to connect with the research being presented, and so there will probably always be a role for the journal article in some form.

A rather less drastic strategy than replacing journals with databases is to endow the narrative with more logical structure and explicit meaning. As argued by Shotton (2009), such enhancements should optimally be an integral, routine part of online *semantic publishing*, which he defines as “anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers”. Semantic publishing could bring enormous benefits in several areas. The first is greatly enhanced ability for researchers to survey, discover, analyse and synthesise relevant publications in their field. Embedded meaning from ontology terms, links to databases and other enhancements, combined with tools to utilise them, would “augment the unique effectiveness of natural language narrative” and thus support and enhance, rather than replace, the strategic reading that researchers have always done (Renear and Palmer, 2009). Second, as Howe *et al.* (2008) point out, improved representation of knowledge in literature will also improve the efficiency and usefulness of curation. Third, semantic markup of publications would enable processing

and reasoning by automated agents to infer previously unknown knowledge (or “meta-research”) (Shotton, 2009).

Finally, on the database side, the fact that knowledgebases such as OMIM employ free text to capture the full nuances of G2P knowledge illustrates the challenge of presenting complicated concepts and connections as traditional, structured database records. Indeed, much of current cutting-edge bioinformatics and computer science research is focused on just this problem of knowledge representation. Major biological databases are beginning to experiment with new data publication approaches; for example, the Universal Protein Resource⁹⁴ (The UniProt Consortium, 2010) provides alternative views of its data in semantic form (see more below). With online journals and online databases being increasingly cross-linked and indexed, the line between database entries and research articles may indeed begin to blur in the near future (Bourne, 2005).

2.4 The untapped power of federation

Given the pressing challenges considered above, it would seem unlikely that a purely centralised database model will fulfil all the requirements of an optimal G2P databasing solution for the future. For this reason, *federated databases* are emerging alongside and intermingled with existing, established central databases. But before discussing the potential and implications of federation in detail, it is useful to reflect on extreme versions of the federated and centralised models depicted in Fig. §2.4. In the fully centralised model, all generated data are piped automatically into one large data centre, from where all search and presentation activities are managed. This contrasts with a completely federated system, where all information in the domain is organised into many geographically separated, distributed databases where data gathering and expert curation takes place, but which do not exchange data amongst themselves. Global search and data retrieval across the federation is mediated by specialised *data portals* which may not themselves hold any data.

⁹⁴<http://www.uniprot.org>

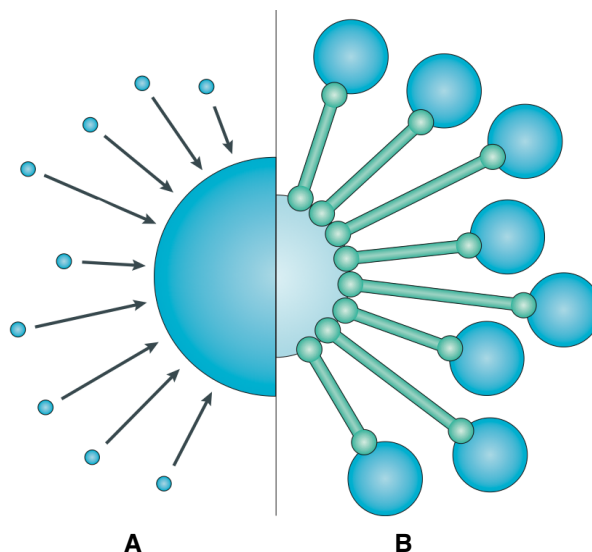


Fig. 2.4: Two extreme models of database integration contrasted: fully centralised (A) vs fully federated (B). From Thorisson *et al.* (2009).

Due to the limitations of the fully centralised and the fully federated model, neither of these extremes is a realistic option for the G2P domain, and therefore a hybrid model would seem to offer the best way forward. However, previously, most successful databases in this domain have been based on the centralised model. One can speculate that this reflects the relative newness of a field which really came into its own with the emergence of the Web (compared to some other disciplines with a longer history of data management), and the fact that the current pressing need for more advanced solutions is relatively recent.

Other disciplines with a longer history of “Big Data” have made good progress in dealing with similar data problems. Astronomy is a good example of this: hundreds of terabytes of image data from observations of billions of celestial objects from ground- and space-based telescopes are available to astronomers worldwide via a global network of “virtual observatories”⁹⁵ (Szalay and Gray, 2001). This network enables thousands of “extra pairs of eyes” to analyse, and greatly increase the scientific value of, these expensively-gathered

⁹⁵<http://www.ivoa.net>

data. An extreme example of this is Galaxy Zoo⁹⁶, a “citizen science” project in which members of the public help with analysis of photographs from the Hubble space telescope by classifying distant galaxies based on their shape.

Centralisation vs federation - pros and cons. Both central and federated systems have advantages and disadvantages. The main advantages of centralisation include cost efficiency due to economies of scale, ease of management and reliable archiving of the community’s data. In contrast, federated databases represent a more complicated solution in terms of the required technologies (see next section), but they do however bring certain advantages that cannot be endowed by a central database. In large part this relates data “ownership” and accreditation for the database teams, with the potential result that more and higher quality data can be gathered in a federated system, due to the reward gained by the workers involved. Federated and central database systems can both provide centralised search capabilities, although federated alternatives can also offer more sophisticated search options via direct interrogation of source databases.

Levels of federation. Given that a group of databases in a particular domain wishes to become federated, an important first decision to make involves the level of federation to be achieved; i.e. what portion of data content each database wishes to make available for remote querying. The databases might choose not to make any data available directly, and instead transfer a pre-agreed “core” set of data elements for each record they hold (e.g. identifier, human-readable label, short description, keywords / ontology terms, cross-links to other databases such as gene symbols), along with links back to those entries in their database. The central search system would then create an index from these minimal data items to enable cross-database distributed searches, and report search results as a series of annotated links pointing back to the source databases.

A downside of a partly federated, partly centralised - or centralised indexing - solution is the administrative overhead of transmitting data to the central hub and keeping the data

⁹⁶<http://www.galaxyzoo.org>

up to date, and so it shares some of the disadvantages of data warehousing as noted by Stein (2003). Nevertheless, this strategy brings many benefits of federation with limited technological overhead required from participating databases. An excellent example is provided by the EB-eye search engine⁹⁷ (Valentin *et al.*, 2010), which indexes some 400 million records from 62 disparate datasets hosted at various units within the EBI/EMBL organisation. Centralised indexing is being explored by several initiatives as a way to begin federating LSDBs.

A more elegant and sophisticated means 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 transmitting and regularly updating core data in a central index which ensures that searches through the central portal always query the very latest datasets. It also addresses the scalability problems outlined in the previous section, since any new LSDB merely needs to register its existence with the central portal to become part of the multi-database search catalogue.

Another advantage which helps to deal with increased overall data volume (e.g. next-generation sequencing data) is that federation greatly minimises the workload and processing requirements of the central search system, since data are stored on each individual node and the actual search takes place there. Finally, federation alleviates many of the data complexity issues faced by central databases, in that each nodal database can provide and customise (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. The ENCODEdb portal mentioned in §2.3.4 is an example of this. Alternatively, the nodal databases can be devoid of data presentation functions, with the central portal responsible for retrieval and display of all the data found in the search.

Achieving these higher levels of federation, however, requires all participating databases to accept certain rules of engagement. For example, the level of autonomy that each team can enjoy, in terms of database design, system execution, and the degree of association

⁹⁷<http://www.ebi.ac.uk/ebisearch/>

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 result datasets can be freely communicated between remote and local computers.

Special advantages of federation. Certain other advantages of federation are also worthy of specific comment. The first relates to empowering and rewarding database creators. 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 may be very important in the case of commercial databases, as illustrated by HGMD, as well as in the context of sensitive, identifiable data.

Second, the federated structure distributes data management and curation work among many individuals (rather than a small, central team), 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 on existing datasets: for example, a gene-centric view for researchers specialising in a single gene, a disease-centric view for clinicians and a genome browser-based view for genomics researchers.

Fourth, federated networks by default operate as democracies, so unilateral changes cannot be imposed on common aspects of the federated system (such as data models and exchange formats, see next section). This does not mean that innovation becomes stifled, but rather that new ideas will be widely debated, piloted and validated before they are implemented, in true community fashion.

2.5 Building blocks of a federated G2P database network

The components that are needed to create a G2P database network, based on a hybrid, partially federated and partially central model, are either already available or in advanced stages of development, and have in the past few years began to be widely adopted to build so-called *cyberinfrastructure* or *e-infrastructure* for science (Buetow, 2005; Hey and Trefethen, 2005). Various technologies which facilitate cyberinfrastructure have been described in more detail elsewhere (see e.g. Stein (2008) and Goble and Stevens (2008)). This section will focus on the specific component of cyberinfrastructure that Stein (2008) refers to as the “communication infrastructure”, or the standardization required for members of the network to connect and exchange data effectively, and introduce key technologies that are critical to this.

Web services and grids. A set of standard, high-level protocols for facilitating machine-to-machine interaction over the Internet - collectively known as web services - simplify the task of “plumbing together” distributed data retrieval or analysis services over the network. Web services form the basis of service-oriented architecture (SOA)⁹⁸, which is a cornerstone of interoperable, distributed networks of computers, or grids. The term “grid” is frequently associated with high-performance computing (HPC) clusters which provide processing power for highly CPU-intensive tasks, such as modelling complex structures in protein research, or simulating supernova explosions in astrophysics. But a grid in the general sense can be considered as any network of computers connected via a standard means of communication (including the Internet itself). Most commonly, the term is used to denote a grid organized around a specific purpose and/or within a specific domain, for the purpose of sharing distributed resources on the network amongst individuals and institutions within a virtual organisation (VO) (Foster *et al.*, 2001). Many such service-oriented grids, conveniently referred to as federations (as has been done here), are being constructed to support collaborative research on an institutional, national or international

⁹⁸<http://www.w3.org/TR/ws-arch/>

level, and are a cornerstone of e-infrastructure (Foster, 2005).

An important feature of SOA is that the nodes in the federation can be heterogeneous with respect to computer hardware, programming languages, operating systems or database organisation, and yet are able to communicate at a high level because they share a common technology which enables them to interoperate. Such *loosely-coupled*, distributed systems thus contrast with *tightly-coupled* federations such as TwinNET, a G2P database network developed in the GenomEUTwin project⁹⁹ (Litton *et al.*, 2003; Muilu *et al.*, 2007) to support biobanking activities spanning several institutions in Europe and Australia. TwinNET relies on low-level database protocols for network communication and identical software and databases on each participating node, so participant databases cannot easily evolve individually to meet changing local needs without breaking interoperability.

An early example of the use of web services in biology is the Distributed Annotation System (DAS)¹⁰⁰ (Dowell *et al.*, 2001), a simple protocol and format for exchanging annotations on genomic sequences. Many genomics databases make their records available via their own DAS server to DAS clients such as the Ensembl browser. These third-party datasets are then overlaid on other DAS-supplied information or locally available annotations, such as reference sets of genes, thus demonstrating the power of a federated system.

Syntax and object models. Whilst web services standardise the way nodes in a grid exchange messages, another key aspect of standardisation relates to the data contained within those messages - data representation - and concerns both syntax and semantics. A core syntax challenge involves designing and validating *object models* which are formalised conceptualisations of how data elements, or objects, are structured and organized, and how they are connected to other data elements. Such models are typically the basis of standard specifications for data exchange formats which are essential for unambiguous transmission of data between computers. Examples in the biosciences range

⁹⁹<http://www.genomeutwin.org>

¹⁰⁰<http://www.biodas.org>

from the simple FASTA format used to exchange DNA and protein sequence data, through to the elaborate XML-based MAGE-ML format based on the Microarray Gene Expression object model (MAGE-OM) (Spellman *et al.*, 2002) which standardises representation of microarray information.

By definition, even informal exchange formats for biological data are always based on some notion of data structure, whether explicitly described and documented or not. In recent years, conceptual data models are increasingly used within a formal software engineering framework known as *model-driven architecture* (MDA) (Mellor *et al.*, 2002). In MDA, conceptual models form the basis from which standard specifications for exchange formats, and often also databases and software, are generated. Numerous formal object models have been, or are being, developed in the biosciences in this fashion, which reflects a drive towards increased standardisation in the field to support sharing and reuse of research data (Field *et al.*, 2009). Examples of this are the Functional Genomics Experiment object model (FuGE-OM)¹⁰¹ (Jones *et al.*, 2007), a high-level model for biological investigations, and the ISA-TAB exchange format specification¹⁰² (Sansone *et al.*, 2008) which are intended to capture commonalities between various kinds of “omics” research study designs. In the G2P domain, the Phenotype and Genotype Experiment Model object model (PaGE-OM)¹⁰³ (Brookes *et al.*, 2009) was recently published, with contributions from our group. FuGE-OM and PaGE-OM will be discussed in Chapters 3 and 4.

An important function of object models is to facilitate data integration through inter-model alignment or “mapping” to identify commonalities and differences. This involves identifying equivalent concepts or relationships which can be used as a basis of a consensus model and/or derive a data exchange format with which both models will be compatible, and thus provide a *lingua franca* for interoperability. This is extremely useful for not only integration of the same kind of data from different sources within a domain, but in particular for cross-domain integration of different kinds of data where the underlying models might

¹⁰¹<http://fuge.sourceforge.net>

¹⁰²<http://isatab.sourceforge.net>

¹⁰³<http://www.pageom.org>

be quite different, as long as the models have at least some common concepts or attributes.

Semantics and ontologies. Syntactic standardisation is a necessary requirement for effective data exchange and interoperability, but it does not fully address the problem of semantic ambiguity which is frequently a significant barrier to meaningful integration of biological data. Formal object models, and even informal, *ad hoc* data formats always carry a certain level of implied or tacit semantics. Data elements will thus necessarily have a certain meaning, and indeed this is why simple data formats such as FASTA for sequence data are useful. However, given that object models and formats are designed to be generic in order to support a wide range of data within a domain, such semantics are frequently conveyed at an abstract level (e.g. “gene”). Furthermore, the meaning of data elements is frequently conveyed as simple textual labels, rather than formal agreed-on definitions. This leads to substantial room for ambiguity in data transfer, as the same term can mean different things to different people, and different terms can refer to the same thing. For example, a field named “sample” may mean “blood sample” in one database, but be taken to mean “set of individuals sampled from a population” in another database.

Shared ontologies are an established strategy to tackle the problem of semantic ambiguity. Ontologies facilitate formalisation of domain knowledge independent of any data structures (which may be sub-domain or application-specific). This principle of decoupling semantics from data is being leveraged to great effect in the biosciences by focusing efforts of domain experts towards creating a compendium of ontologies, each of which captures knowledge in a clearly demarcated subject area. Each ontology can then be used in a variety of contexts to embed meaning to data or data-driven applications, by itself or in combination with other ontologies. Examples of ontologies of relevance to this discussion include the Gene Ontology previously mentioned for annotating gene products, the Sequence Ontology (SO)¹⁰⁴ (Eilbeck *et al.*, 2005) for annotating DNA and protein sequences, and the Ontology for Biomedical Investigations (OBI)¹⁰⁵ for describing biomedical investigations. These and

¹⁰⁴<http://www.sequenceontology.org>

¹⁰⁵<http://obi-ontology.org>

many other ontologies are being developed under a common “umbrella” organisation - the Open Biomedical Ontologies (OBO) Foundry¹⁰⁶ (Smith *et al.*, 2007) - a key role of which is to facilitate a common architecture and syntax for bio-ontologies, ensure quality and foster research into ontology design and curation in the biosciences.

Unambiguous identification of data objects. Separate from the issue of syntax and semantics is the problem of naming data objects across many databases. One of the peculiarities of biological databases is the proliferation and ambiguity of names used for things of interest such as genes and proteins. For example, a protein will often go by different names in different databases, and the same gene symbol may be used to refer to a gene, the transcript of a gene and the protein product¹⁰⁷. These problems of synonymy and homonymy, respectively, greatly complicate cross-database data integration and database searches. Furthermore, a lack of commonly employed, persistent and non-reusable identifiers in many domains for database entries and other entities of interest is a major hindrance to adoption of advanced semantic technologies.

Initiatives such as the Shared Names project¹⁰⁸ have been created to encourage community use of shared, community-maintained Universal Resource Identifiers (URIs)¹⁰⁹ to identify abstract or physical resources on the Web. A URI, like other globally-unique identifiers (GUIDs) is a type of identifier which is guaranteed to be unique and persistent across the intended usage domain. As such, GUIDs solve the name ambiguity problem. A further crucial advantage is that automated agents can resolve the URI (i.e. follow the link) to retrieve information relating to the data object. However, as Goble and Stevens (2008) note, a long-standing debate over which particular GUID scheme to use across the biosciences,

¹⁰⁶<http://obofoundry.org>

¹⁰⁷A more philosophical question, out of scope for this discussion but nevertheless worth mentioning, is whether a given identifier refers to the object itself (e.g. a protein) or a record describing the object (e.g. a UniProt entry)

¹⁰⁸<http://sharedname.org>

¹⁰⁹<http://tools.ietf.org/html/rfc3986>

as well as related issues such as long-term funding of central naming authority services where such services are required, currently hampers progress on this front.

2.6 The critical role of standards

As evident from the previous section, standardisation is crucial to the federated database model as it provides a means to fit together the various nodes of the network to make a useful, integrated whole. More generally, standards underpin the Internet itself, with established Internet standards such as the Hypertext Transport Protocol (HTTP)¹¹⁰ facilitating communication between computers on the global grid. HTTP, web service protocols and other low-level, content-agnostic communication standards are distinct from the specialised ontologies, conceptual models and exchange formats - sometimes collectively referred to as *data standards* - created for a particular field of study.

2.6.1 Domain standards and dynamic software infrastructure

Minimal information standards. In the past decade there has been a trend towards increased standardisation in the life sciences. Several pillars underpinning domain-specific reporting standards have already been introduced - ontologies, object models and exchange formats. The final pillar is *minimal information checklists* - lists of well-defined data elements required to be present for interpretation of a given type of biological study. Such standard, community-developed checklists were pioneered by the microarray community who devised the the Minimal Information About a Microarray Experiment (MIAME) standard in the late 1990s¹¹¹ Brazma *et al.* (2001). Checklists are based on the notion that, for a given type of “omics” study, a lot of different kinds of information could conceivably be reported, but in practice usually only a subset of this is useful when shared with others, and often an even smaller subset is absolutely critical to interpret (and

¹¹⁰<http://www.w3.org/Protocols/>

¹¹¹<http://www.mged.org/Workgroups/MIAME/miame.html>

potentially reproduce) the experiment. Therefore, formalising this list of i) essential and ii) useful but non-essential data elements is useful for specifying which information should be transmitted in a data exchange, in order for the receiving party to make use of the data. Minimal information checklists are thus an essential aid for enforcing data sharing policies and ensuring that data are shared in a useful form (see also §2.3.3). For example, most journals now require authors of microarray papers to provide the minimal MIAME-mandated provenance information for their experiments, as well as mandating submission of primary data to public microarray databases.

Checklists are also crucial to the overall standards development process. In their review of standards development in systems biology, Brazma *et al.* (2006) stress that without a clear scope and goals, there is a risk of “over-modelling” and ending up with a very complicated model which incorporates every possible data elements and handles even many uncommon “edge” cases, rather than a simpler, more practical model with broad utility. For example, the aforementioned MAGE-OM model can be used to describe extremely complicated microarray experiment designs, but as a result the model and the corresponding MAGE-ML XML-dialect are very complex. This complexity makes the standard difficult to support in software implementation and data processing, in part because the same information can be represented in multiple ways, leading to a “Tower of MAGE-ML Babel” (Maier *et al.*, 2008). This has driven the development of a simpler standard for reporting microarray experiments - the MicroArray Gene Expression Tabular (MAGE-TAB) model and tab-delimited format¹¹² (Rayner *et al.*, 2006) - which covers the majority of use cases and complies with MAGE-ML on a conceptual level.

Bottom-up vs top-down. A recent example of community-based standards development is the Minimum Information for Biological and Biomedical Investigations (MIBBI) project¹¹³ (Taylor *et al.*, 2008), an “umbrella” community project modelled on the OBO Foundry. Several initiatives focused on minimal information checklists for various “omics”

¹¹²<http://www.mged.org/mage-tab/>

¹¹³<http://www.mibbi.org>

technologies are self-organising under the MIBBI banner. The various MIBBI-affiliated initiatives and many of the other standards mentioned above are part of a broader Reporting Structure for Biological Investigations (RSBI) community initiative¹¹⁴, which promotes collaborative development of “omics” reporting standards (Sansone *et al.*, 2008, 2006). RSBI and related initiatives reflect a general preference in the biosciences for community-developed, or “bottom-up” grassroots standards which are accepted by the community on merit. For example, PaGE-OM and some of the other object models mentioned above have been formally submitted to, and approved by, standards organisations such as the Object Management Group (OMG)¹¹⁵. However, official standardisation has not historically proved critical in the biosciences; many key standards (such as GO) have not undergone such a formal standardisation process and yet have been adopted by the community as *de facto* standards (Brazma *et al.*, 2006).

This bottom-up approach contrasts with the “top-down” strategy proposed by Cassman (2005), who argues for a move away from a “cottage industry” style of software development in systems biology, and instead focus efforts and funding on a formal validation, documentation and standardisation process managed by a central organisation. Cassman’s proposal was summarily rebuffed by Quackenbush *et al.* (2006) and colleagues on several grounds, citing factors such as difficulty in pre-engineering software and database solutions ahead of time for rapidly evolving fields of research, a preference by researchers for pragmatic solutions which get the job done over computationally-elegant approaches¹¹⁶ and, crucially, the inherent democratic nature of community-based development as already mentioned. As subsequent chapters will show, this thesis is firmly placed in the bottom-up camp of this debate.

Dynamic, standards-based software. Cassman’s proposal for software as a means of standardisation does have some merit, at least where suitable standards have emerged

¹¹⁴<http://www.mged.org/Workgroups/rsbi/index.html>

¹¹⁵<http://www.omg.org>

¹¹⁶Interested readers are referred to “How Perl Saved the Human Genome Project” by Stein (1997)

(or are emerging). Indeed, as Quackenbush (2004) notes, “having a standard is no real use unless there is a way to implement it”. Though this comment was made in relation to the aforementioned flexible, but complex, MAGE-ML, it has some validity even for much simpler standards. In effect, this boils down to the following question: if there is no immediate gain to be had, why would a busy bioinformatics developer working on an in-house project spend extra time on making a piece of software standards-compliant (instead of simply creating a “one shot”, disposable tool with no eye for data or software reuse)?

A major factor in the widespread use of *de facto* domain standards, such as FASTA, is that the software tools which support them are widely available and easy to use. One may therefore conclude that if more off-the-shelf software tools supported newer, more sophisticated standards such as FuGE-OM, then adoption would follow. Following this line of reasoning, Swertz and Jansen (2007) stress that the field must move from an “expensive, almost one-at-a-time, ‘cottage-industry’ towards twenty-first-century engineering practice”, and leverage sophisticated, standardised software components which can be customised and evolve over time to meet the diverse needs of researchers.

Although Swertz and Jansen discuss the general principles of the approach in a broad scope, at the core of their thesis is a specific implementation called Molgenis¹¹⁷ (Swertz *et al.*, 2004), a framework which enables a collection of standard software and database components to be combined with variable, custom components in a “mix-and-match” way to create an integrated toolset. A key feature of their system is an established software engineering technique called “generative software development”, whereby components which require customisation (such as database tables or user interface elements) are automatically generated and combined with standard, reusable components, or “assets”, according to a specification written in a minimal domain-specific language (DSL) or “infrastructure blueprint”. This modular approach results in minimal or no programming effort required to create a custom system to meet local needs, and a far more rapid development cycle compared to conventional software development strategies.

¹¹⁷<http://www.molgenis.org>

An example of the above put into practice is a case study by Smedley *et al.* (2008), who used a Molgenis-based system as one of several components in a web service-based analysis workflow integrating mouse variation data, genomic annotations and metabolic pathway data. Another example is the eXtensible Genotype And Phenotype (XGAP) system¹¹⁸ (Swertz *et al.*, 2010) recently created for management and analysis of various kinds of G2P data. The the XGAP system and the underlying conceptual model will be further discussed in Chapter 3 and 4.

2.6.2 Linked Data and the Semantic Web

A particularly important set of domain-agnostic standards are those that will underpin the *Semantic Web*, a futuristic vision originally ventured almost a decade ago by Tim Berners-Lee, the creator of the Web. According to this vision, the Semantic Web will be a medium for publishing information and its associated meaning on the Web, enabling data to be linked other data in a uniform, machine-interpretable way that facilitates navigation and reasoning (i.e. deduction of inferences) by computers (Berners-Lee and Hendler, 2001; Berners-Lee *et al.*, 2001). This notion of a *Web of data*, which computers can comprehend, navigate and manipulate, contrasts with the existing Web of documents, which is primarily intended for humans to read and largely semantically opaque to machines. For example, even with all its computing power and sophisticated result-ranking algorithm, the Google search engine cannot distinguish between the distinct concepts of blood sample, a population sample, or digital audio sample when it encounters the word “sample” mentioned in a web page.

The building blocks of the Semantic Web are many of the same standard formats and protocols that make the regular Web work, augmented with additional software standards and tools¹¹⁹ (the so-called Semantic Web stack¹²⁰, see Fig. §2.5), most of which are

¹¹⁸<http://www.xgap.org>

¹¹⁹<http://www.w3.org/2001/sw/>

¹²⁰<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

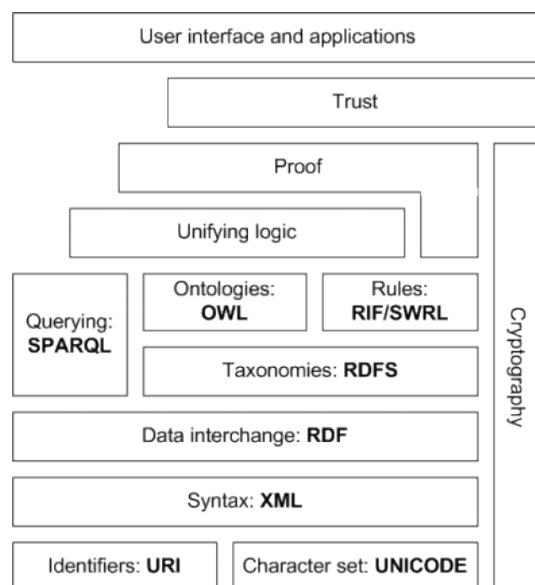


Fig. 2.5: The hierarchy of standard languages, protocols and concepts which together will comprise the architecture of the future Semantic Web. The bottom, foundational layers are the same, or extensions of, established technologies that underpin the current, hypertext-based Web of documents. From http://en.wikipedia.org/wiki/Semantic_Web_Stack.

developed under the auspices of the World Wide Web Consortium (W3C)¹²¹. A full treatment is outside the scope of this thesis (see e.g. Sagotsky *et al.* (2008) for an excellent review), but for the purpose of discussions in later chapters, a subset of the relevant technologies which have a bearing on the aforementioned problems of syntactic and semantic data integration will now be introduced.

Linked Data. The various semantic standards technologies are now mature or reaching maturity. However, uptake in the life sciences, where the technology may have enormous implications for data integration, has so far been slow. There are various reasons for this, including the proverbial chicken-and-egg problem; there is little immediate gain for a biological database resource from adopting the required technologies unless a critical mass of cross-linked databases also do so. Indeed, Semantic web proponents acknowledge that the full vision will not be realised until the required standards are widely adopted and used

¹²¹<http://www.w3.org>

on a global scale (Shadbolt *et al.*, 2006). Good and Wilkinson (2006) have coined the term “semantic creep” to describe the piecemeal adoption of the standards and technologies by major online bioinformatics resources, and argue that at present the obstacles are primarily social in nature. Nevertheless, the full range of technologies that comprise the Semantic Web technology stack is far more sophisticated than that required for traditional HTML publishing on the Web, and inherent complexity of these technologies is a barrier to mainstream adoption at present.

Realising that the full vision of the Semantic Web is perhaps too ambitious and not achievable in the near term, proponents have adopted a strategy with more limited scope, commonly referred to as *Linked Data*¹²². Linked Data, described by some as a “pragmatic Semantic Web”, involves leveraging a subset of the full range of Semantic Web standards for publishing and linking together structured data on the Web in a way that facilitates global integration. The general strategy is summarised by the following four rules from Tim-Berners Lee’s Linked Data Principles¹²³:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs. so that they can discover more things.

Everything is a triple. The common Linked Data integration layer is the Resource Description Framework (RDF)¹²⁴, a simple, graph-based abstract data model in which all information and knowledge is represented as simple statements known as *triples*. Triples (as illustrated in protein example below) are simple statements of the form [subject, predicate, object] where the subject and the predicate are Web resources identified by URIs, and the object is either also a resource or a literal (e.g. a number or a text string). Triples

¹²²<http://linkeddata.org>

¹²³<http://www.w3.org/DesignIssues/LinkedData.html>

¹²⁴<http://www.w3.org/RDF/>

are an elegant, simple way of describing things and the properties of those things, and to represent typed relationships between things.

Shared URIs are fundamental to data interoperability conferred by Linked Data publishing, in that they facilitate merging of disparate datasets describing the same things based on their identifiers. Shared identifiers for things, combined with the “universal solvent”¹²⁵ properties of RDF as a single, canonical data model, makes Linked Data an ideal medium in which almost any conceivable type of data can be “dissolved” and integrated with other data.

Ontologies for imposing order on chaos. In and of themselves, the RDF data model and URI identifiers are a powerful mechanism for “omics” data integration across diverse domain data models with only partial overlap (Wang *et al.*, 2005), even in the absence of semantics. However, as some authors have pointed out, transforming all data into RDF triples does not necessary solve the data integration problem and introduces challenges of its own (Quackenbush, 2006). Due to the low-level nature of the model, merging vast numbers of triples from many sources into a single graph results in a vast, homogenised mass of interlinked information which is difficult to navigate. Ontologies provide a “scaffold” for endowing this sea of triples with structure and meaning, thus adding a layer of semantics on top of the universal data integration layer.

To illustrate this, the RDF representation of the protein A1BQS5 published at <http://www.uniprot.org/uniprot/A1BQS5.rdf> contains a triple describing the “type” of the protein by way of a reference to a resource identified as <http://purl.uniprot.org/core/Protein>. A machine can follow this URI link to look up a term in an ontology described in the Web Ontology Language (OWL)¹²⁶ and retrieve additional domain knowledge regarding the kind of thing being described (e.g. the various properties which characterise a protein), and how this concept relates to other concepts. OWL provides a vocabulary for formally defining classes, class properties and relationships in a rigorous way which

¹²⁵<http://www.mkbergman.com/483/advantages-and-myths-of-rdf/>

¹²⁶<http://www.w3.org/TR/owl-features/>

facilitates automated inferencing. The OWL language is represented in RDF, and thus both the data and the ontology can be described in the same, uniform format over which machines can navigate and compute.

2.7 Distributed grids in the life sciences

Service-oriented, loosely-coupled grid architectures are already pervasive in a number of data-intensive scientific disciplines, and they are now beginning to make an impact in the life sciences. This section introduces several exemplar projects which illustrate some of the principles and technologies introduced in previous sections.

2.7.1 Biomedical grids for cancer research and neuroscience

Several large-scale cyberinfrastructure projects have been constructed in recent years, notably in areas of neuroscience and cancer research. For example, the Biomedical Informatics Research Network (BIRN)¹²⁷, initiated in 2001 with a focus on sharing large volumes of neuroimaging data, has grown into an extensive network of several dozen US neuroscience research institutes. Another prominent example is the NCI cancer Biomedical Informatics Grid (caBIG)¹²⁸ (Saltz *et al.*, 2006) which connects over 80 cancer research centres across the US. caBIG enables researchers to collaborate and securely use distributed data and analysis services, such as clinical trial data and patient registration and tracking. Both of these large-scale biomedical grid projects leverage enterprise-level technologies, including SOAP-based web service protocols and Globus Toolkit¹²⁹ open-source grid software for communication between nodes, and both rely on common ontologies for semantic interoperability.

Although they share many architectural features, an interesting contrast between BIRN and

¹²⁷<http://www.birncommunity.org>

¹²⁸<https://cabig.nci.nih.gov>

¹²⁹<http://www.globus.org>

caBIG is the different levels of openness. BIRN is primarily a closed, virtual community between participating centres, with close integration achieved by distribution of standard, pre-configured hardware “racks” which come pre-loaded with all the software required for a research centre to hook up to the grid. caBIG, on the other hand, is more open, with participating centres only required to meet a certain number of requirements in order to become compatible with the grid. Centres can achieve compatibility with grid software custom-built locally, or deploy the open-source caCORE software development kit (Phillips *et al.*, 2006) developed in the project to more easily create caBIG-compliant data and analysis services.

The substantial amounts of funding expended over several years and the focused, centrally-managed development effort is no doubt a large factor in the success of these projects. However, although the end result is a system that works, this style of building biomedical grids has some disadvantages. In his review, Stein (2008) notes that the top-down management style employed in caBIG incurs substantial overhead and enforces a level of organisation and discipline that many bioinformatics researchers are not comfortable with. Also, many of the advanced, grid-enabled software tools created in the project have yet to find a broad following amongst end users in the US cancer research community, to many of whom the benefits of grid is not immediately apparent in their daily work.

A further disadvantage concerns the broader utility of the grid infrastructure software and other components created in such enterprise-level projects. The caCORE toolkit, controlled data vocabularies, metadata registration tools and other components require substantial developer expertise to deploy, configure and maintain. For example, it would be a substantial challenge for a small research group without expert informatics support to connect their systems to caBIG. On the other hand, larger collaborations with more IT resources, such as the UK CancerGrid¹³⁰ have been able to integrate with the caBIG network, in part by utilising the caCORE software and also via sharing and harmonisation of data models.

¹³⁰<http://www.cancergrid.org>

Another example is the recently created Biomedical Research Informatics Centre for Cardiovascular Science (BRICCS)¹³¹ at University Hospitals of Leicester (UHL). BRICCS is constructing a new clinical research database, with several key building blocks sourced from the caBIG project and two others not previously mentioned: the OBiBa biobanking software project¹³² and the Informatics for Integrating Biology and the Bedside (I2B2) project¹³³ (Murphy *et al.*, 2010).

2.7.2 Distributed web services and workflows

The BioMoby¹³⁴ (The BioMoby Consortium, 2008; Wilkinson and Links, 2002) and myGrid¹³⁵ (Stevens *et al.*, 2003) projects are representative of a different breed of distributed cyberinfrastructure, built around public registries of large numbers of distributed web services for data retrieval or analysis. These web services, typically developed in loosely-organised, community-driven manner without a central coordination effort, can be marshalled by the appropriate tools to create complex *in silico* experimental workflows, or execution chains, of data retrieval and analysis sub-processes. An important tool in this regard is Taverna¹³⁶ (Hull *et al.*, 2006; Oinn *et al.*, 2004), an interactive application for constructing workflows from local and remote web services. Examples of this approach in the G2P domain are provided by a study into African trypanosomiasis (sleeping sickness) resistance in the mouse, where Fisher *et al.* (2007) were able to reuse a Taverna workflow previously created to identify a G2P correlation in a study of trypanosomiasis in cattle.

There is a strong community aspect to these bottom-up, grassroots initiatives. From the community outreach strategies of the research groups who create Taverna and other myGrid

¹³¹<http://www2.le.ac.uk/projects/bru/researchers/briccs>

¹³²<http://www.obiba.org>

¹³³<https://www.i2b2.org>

¹³⁴<http://www.biomoby.org>

¹³⁵<http://www.mygrid.org.uk>

¹³⁶<http://www.taverna.org.uk>

tools (De Roure and Goble, 2009), to the myExperiment community website¹³⁷ created for sharing scientific workflows and models Goble *et al.* (2010) - all this is indicative of a rather different mindset from the centrally-coordinated BIRN and caBIG projects. The barrier for entry is low; anyone can download Taverna and start creating sophisticated workflows to answer biological questions, and share them with the community. Similarly, anyone can create a web service interface or an analytical tool or database and expose it to the Web for others to discover and use, simply by adding the service to a public registry. Such registries address the common problem of discovering suitable web services for a given task amidst potentially thousands available.

Semantic service discovery. A particularly interesting development in this regard is the recent launch of BioCatalogue¹³⁸ (Bhagat *et al.*, 2010; Goble *et al.*, 2008), a new, unified registry of web services in the life sciences which incorporates web services previously held in the BioMoby, myGrid and EMBRACE¹³⁹ registries. A key feature of BioCatalogue is semantic annotation of web service types and their inputs, outputs and other parameters, which facilitates automated discovery of suitable services given a particular type of data (for example, from within Taverna). Several thousand expert-curated services are listed in the BioCatalogue, for diverse tasks ranging from retrieving and aligning nucleotide sequences to text mining.

2.7.3 Towards a holistic G2P knowledge environment

There is no doubt that enterprise-level federated grid technology will in the future be critical for effective data sharing and integration of research data in many areas of scientific research. Progress made in the last several years to create reusable cyberinfrastructure components is promising. For example, caBIG technology is being

¹³⁷<http://www.myexperiment.org>

¹³⁸<http://www.biocatalogue.org>

¹³⁹<http://www.embraceregistry.net>

adopted for the CardioVascular Research Grid (CVRG)¹⁴⁰ and the Nationwide Health Information Network (NHIN)¹⁴¹ in the US, and to support an international cancer research collaboration involving institutions in the US, UK, China and India (Buetow, 2009). Privacy and intellectual property protection, as well as numerous other requirements, demand the security, overall robustness and other high-end features provided by the sophisticated grid architectures employed in these clinically-focused large-scale projects.

The rise of open, distributed grids. The importance of the “light touch”, bottom-up brand of distributed bio-computing should not be underestimated, however. NCBI has provided programmatic access to most of their resources for several years via their eUtils service¹⁴², and EBI provides over 300 web services¹⁴³ as an alternative access mode to their range of data and analytical services (McWilliam *et al.*, 2009). In addition to web service provisioning by EBI, NCBI and other major bioinformatics centres, a myriad smaller service providers specialise in certain types of data or analysis methods (e.g. the TreeBuilder service¹⁴⁴ for calculating evolutionary distances). A great deal of interesting *in silico* analysis can be performed with these freely-available resources on the open, distributed grid. Furthermore, as reported by Tan *et al.* (2008), recent upgrades to the Taverna tool open up the possibility of incorporating secure caBIG web services into Taverna-built workflows alongside non-caBIG services, thus providing a means of bridging the two “worlds” of distributed grid computing in the biosciences.

Grid-enabling the long tail. Many smaller database resources containing valuable G2P information remain inaccessible to these new grid-based methodologies, however, which limits their potential utility. Assuming that a given database project decides to join the grid at some level (see §2.4), an important tactical decision is how to do it. For many

¹⁴⁰<http://www.cvrgrid.org>

¹⁴¹<http://www.nhin.com>

¹⁴²<http://eutils.ncbi.nlm.nih.gov>

¹⁴³<http://www.ebi.ac.uk/Tools/webservices/>

¹⁴⁴<http://www.biocatalogue.org/services/1917>

smaller projects, deploying enterprise-level caBIG technology, would arguably be the equivalent of using a sledgehammer to crack the proverbial nut, and certainly beyond the reach of teams of one or perhaps two developers. A better strategy in many cases is to adapt existing infrastructure and aiming for “just enough” grid connectivity, using the simplest standards possible. The aforementioned DAS standard illustrates this nicely. The protocol and exchange format that comprise the standard are simple and well-supported by several open-source software libraries¹⁴⁵. As a result, for a bioinformatics developer, the task of turning an existing database containing genome annotations into a DAS server is a relatively straightforward one.

An example of a grid-based integration project where DAS plays a central role is the BioSapiens project¹⁴⁶ (The BioSapiens Network of Excellence, 2005; Thornton and the BioSapiens Network, 2009), which leverages the protocol for lightweight integration of genome and proteome annotations across 25 institutions based in 14 countries in Europe. Another example is adoption of the OpenSearch standard¹⁴⁷ by UniProt protein knowledgebase¹⁴⁸, the EBI and several other data providers in the domain to describe their search engines and search results in a standard, machine-readable way, thus enabling cross-site, distributed searches from a single application.

Performance and reliability. For all its advantages, an inherent weakness of distributed grid computing is the reliance on analytical and data resources residing on remote servers. A key benefit of the centralised data warehousing model is that all datasets and computational resources required for a given analysis are available at the central site. Conversely, a distributed, multi-step analysis on the open grid is entirely dependent on the computational resources available on each of the remote sites. If just one of the multiple services takes a long time to respond or is unavailable, the workflow as a whole runs slowly

¹⁴⁵http://www.biodas.org/wiki/Everything_DAS#Setting_up_a_DAS_Server

¹⁴⁶<http://www.biosapiens.info>

¹⁴⁷<http://www.opensearch.org>

¹⁴⁸<http://www.uniprot.org>

or not at all - in other words, the chain of analytical steps is only as strong as the weakest link. It follows that the more steps there are in a complex analytical workflow and the more numerous remote servers are being relied upon, the more potential points of failure are introduced.

In the case of services hosted at EBI, NCBI and other major bioinformatics centres, this tends not to be a big problem, as these providers are committed to long term provision of a certain level of service performance and server uptime. However, many important web services are created by research groups or individuals who have other priorities and/or do not have sufficient funding or local IT resources to operate high-performance, highly-reliable, professionally managed services for the benefit of others than themselves or others in their own organization.

A number of solutions are being considered to deal with this problem. For example, by using cloud computing (see §2.3.1), research organizations could operate web services without having to invest in physical IT infrastructure, and possibly even outsource such operation to private companies. Cloud-based storage and computation has also been suggested as strategy for long-term preservation and online access to important datasets and services after time-limited research project funding runs out, at nominal cost (see Dudley and Butte (2010) for a recent treatment of this topic). Tiered service provision is another option; lightweight, non-commercial use of a given service could be free, whereas more serious use by, for example, pharmaceutical companies could be charged for¹⁴⁹.

The GEN2PHEN project. Recent international initiatives have embraced the bottom-up philosophy to grid building. One of these is GEN2PHEN¹⁵¹, a 5-year project funded by the European Commission which launched in 2008. With over 20 participating institutions and private companies, GEN2PHEN aims to significantly improve G2P databasing infrastructure in Europe. Rather than trying to build a caBIG-like pan-European federated grid, a key project strategy is to produce a set of small, reusable, standards-

¹⁴⁹150

¹⁵¹<http://www.gen2phen.org>

based and interoperable databases, software tools and technologies. The purpose of these technological building blocks is to enable the evolution of today's diverse and unconnected G2P databases into a seamless, grid-linked "G2P biomedical knowledge environment", with a particular emphasis on integration via genome browsers. Our group plays a coordination role in this project and much of the work presented in the chapters to follow has been undertaken under the auspices of - and with funding from - GEN2PHEN.

3. Designing a data model for genetic association studies

Essentially, all models are wrong, but some are useful.

George E. P. Box and Norman R. Draper. Empirical Model-Building and Response Surfaces (1987)

The work described in this chapter concerns conceptual modelling motivated by the needs of the genetic association database called HGVbaseG2P mentioned in Chapter 2. An early strategic decision in the HGVbaseG2P project was to use a model-driven approach (see §2.5) in software and database development, and to make use of existing data standards where possible and practical. As discussed in detail in Chapter 5, a wealth of models and software tools could be reused and modified for the genomics-focused aspects of the project. However, suitable models for phenotype information and investigations into the link between genotype and phenotype were lacking.

A number of data standards with potential utility had been published at the time when work commenced. The most promising of these was the caBIG Common Data Element (CDEs) available via the Cancer Data Standards Registry and Repository (caDSR)¹ which lie at the heart of the model-driven caBIG software architecture. Based on my initial exploration and assessment of caDSR, I concluded that the overall complexity of the caBIG models (and the tools required to use them) was such that adapting and extending them for my work would not be practical.

The PaGE-OM model for G2P investigations. The most relevant standardisation activity at the time was work by an international consortium of ~20 groups, including our group, towards the creation of the aforementioned PaGE-OM standard for G2P investigations.

¹<https://cabig.nci.nih.gov/concepts/caDSR/>

The PaGE-OM consortium had previously developed a standard specification - the Polymorphism Markup Language (PML)², focused on the genotype aspects of the domain - which was submitted for a formal standardisation with the OMG and approved in 2006. This standard was subsequently revised and extended to cover phenotype information and G2P correlation experiments, eventually resulting in the PaGE-OM specification. Version 1.0 Beta 3 of this specification was approved as an OMG standard in 2009³ and a paper describing the model was recently published (Brookes *et al.*, 2009).

When my PhD work commenced, PaGE-OM was still being actively developed and the specification far from finalised, and thus PaGE-OM was deemed unsuitable as a foundation for HGVbaseG2P development. Instead, a decision was taken to semi-independently pursue development of a standalone implementation model focused on the needs of this project, but with full intent to later align and harmonise this model with PaGE-OM, and in time align this with other emerging domain models (see next chapter). Through participation of our group in the PaGE-OM project, the two models have in many ways evolved in parallel and the design of each has been influenced by the other. My work on the HGVbaseG2P model has therefore indirectly and non-trivially contributed to the development of the PaGE-OM standard.

In this chapter, the current version of the HGVbaseG2P model for genetic association studies, as developed and refined by real-world usage during the course of the project, will be presented first, set in the context of the PaGE-OM reference model. This is followed by results from several model validation exercises, which typify the continuous testing undertaken during development and real-world use of the model.

²<http://www.openpml.org>

³<http://www.omg.org/spec/PAGE-OM/1.0/Beta3>

3.1 The HGVbaseG2P model

The HGVbaseG2P object model derives from and builds on previous modelling work done by our group in support of HGVbase, the predecessor of the HGVbaseG2P project. As noted in §2.2.2, HGVbase was focused on gathering and curating simple sequence variants. By the start of my PhD, an early untested prototype version of a conceptual model for genetic association studies had been devised by way of extending the core of the HGVbase model. In addition to reference variation data, the scope of this new model spanned two broad categories of information pertaining to association studies: i) various descriptive metadata about the investigation (e.g. study design, protocols used, summary of results) and ii) an aggregate representation of genotype and phenotype data and the results from G2P association analysis. My initial work was therefore focused on building on these previous efforts to create a fully-working, practical model, which would provide a conceptual foundation for the various HGVbaseG2P database and software components. My role in this development work was that of the main designer and coordinator of the model creation.

PaGE-OM as a reference. As noted above, the HGVbaseG2P model has many similarities with the PaGE-OM model. The overall rationale for PaGE-OM and an overview of its main features are described in Brookes *et al.* (2009) and will thus not be repeated here. Instead, this discussion will focus on the HGVbaseG2P as an implementation model created for a specific task and how it relates to PaGE-OM, which has a far broader scope and covers numerous other potential application areas in the G2P domain. Even though the HGVbaseG2P model is not specified as a formal extension of PaGE-OM (see Discussion), it is nevertheless useful to consider the model as a specialisation of the more general, and more detailed, PaGE-OM reference model. Importantly, the present version of PaGE-OM includes many refinements made, based on practical experiences from developing the HGVbaseG2P system. One purpose of this section is therefore to highlight similarities and differences between the two models and try

and identify areas where improvements are needed, whether in the HGVbaseG2P model, PaGE-OM, or both.

Model presentation. The conceptual data elements or classes which make up the HGVbaseG2P model are organised into largely the same set of modules or domains as used in the PaGE-OM model: SAMPLE, GENOTYPE, PHENOTYPE, EXPERIMENT, SEQUENCE and COMMON. Fig. §3.1 shows the overall high-level organisation of the whole model (excepting the COMMON domain). Each of the core model domains will be illustrated and described in further detail in the subsections to follow.

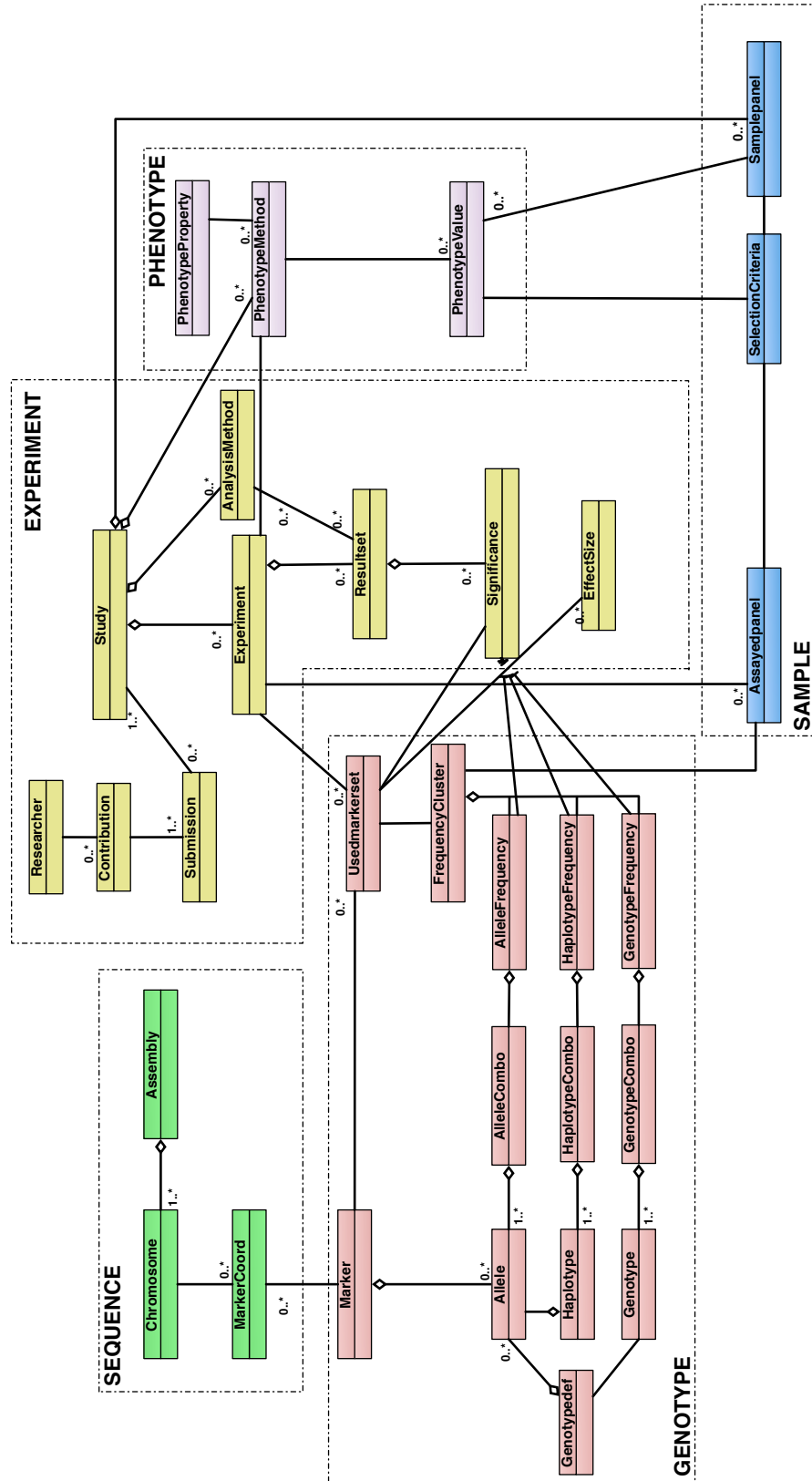
Unless otherwise noted, logical model diagrams are presented according to the Unified Modelling Language (UML)⁴ conventions, using the same notation and colour coding scheme as the simplified PaGE-OM diagrams presented in Brookes *et al.* (2009). Data examples follow the same colour-coding scheme, but use a slightly different notation. HGVbaseG2P class names in diagrams are written with no prefix (e.g. *Study*). Names of classes from other models are written with a corresponding prefix label followed by two semicolons (e.g. *PaGE::Panel*). Where classes in the HGVbaseG2P model can be mapped to PaGE-OM classes, this will be clearly indicated in logical diagrams. In the main text, all class names are italicised.

To simplify the logical diagrams, only a subset of attributes are shown for most classes. The full set of attributes for all classes can be seen in the relational database diagrams in §5.1.2. A full list of HGVbaseG2P class definitions is provided in Table §B.1.

For convenience, the simplified PaGE-OM diagrams and corresponding fully-detailed diagrams from the PaGE-OM website are reproduced in full in Appendix §B.1 and referenced below as appropriate.

⁴<http://www.uml.org>

Fig. 3.1: Overview of the HGVbaseG2P object model in UML notation. Conceptual elements, or object classes, are represented by boxes, colour-coded by domain. Logical connections between classes are represented by lines, decorated with standard UML cardinality indicators ("1..*" = one or more, "0..*" = zero or more).



3.1.1 The SAMPLE domain

The SAMPLE domain, depicted in Fig. §3.2, is concerned with groups of study subjects drawn from a study population. Since the focus of HGVbaseG2P is on aggregate data and not individual-level data, no attempt is made to model individuals. The intent is to accurately describe, at the level of the group, how individuals are organized into groups or panels based on the source population and/or some shared characteristics (such as disease affection status, age or ethnicity).

The *Samplepanel* class models a group of individuals sampled for a study. Association study findings (see below) are reported in terms of one or more instances of the *Assayedpanel* class. *Assayedpanel* represents a group of test subjects derived by splitting and/or merging one or more *Samplepanels* or *Assayedpanels* to create new subject collections, on the basis of some explicit phenotype criteria such as disease affection status or severity/subclass of disease (as modelled by the *SelectionCriteria* class). The ability to explicitly describe how panels of study subjects are created by splitting and merging is critical for accurately capturing complex genetic study designs.

The corresponding classes in PaGE-OM are shown in Fig. §B.1 and Fig. §B.2. *PaGE::Panel* represents a group of study subjects and thus has a direct mapping to the two HGVbaseG2P panel classes. PaGE-OM also provides the *PaGE::Individual* and *PaGE::Molecular_sample* classes which represent individual study subjects (who may belong to panels) and biological samples taken from individuals, respectively. As explained in Brookes *et al.* (2009), PaGE-OM makes use of abstract superclasses to represent generalisation of more specialised subclasses a way to describe scenarios where several subclasses can be used interchangeably. For example, the *Abstract_observation_target* construct generalises the subject of an observation, which can in practice be either a single individual, a sample drawn from an individual, or a group of individuals. The latter two are represented by the secondary abstract class *Abstract_population*.

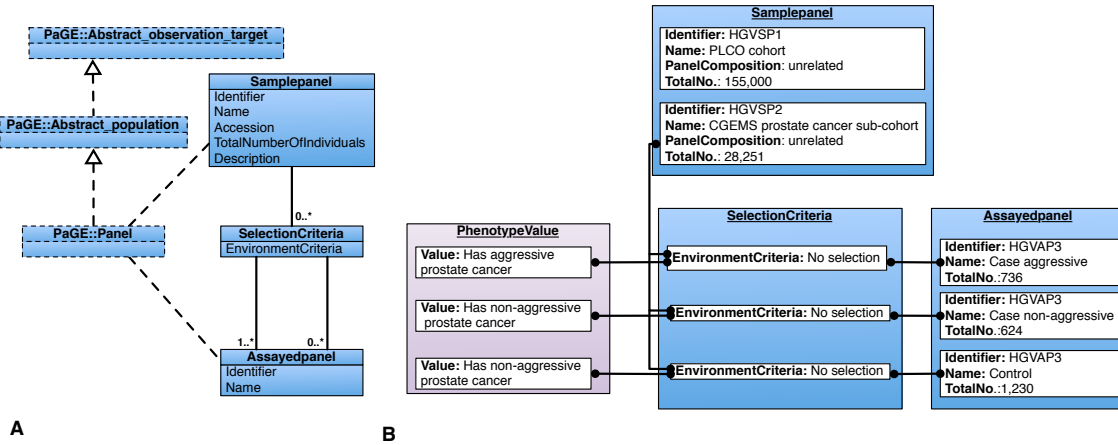
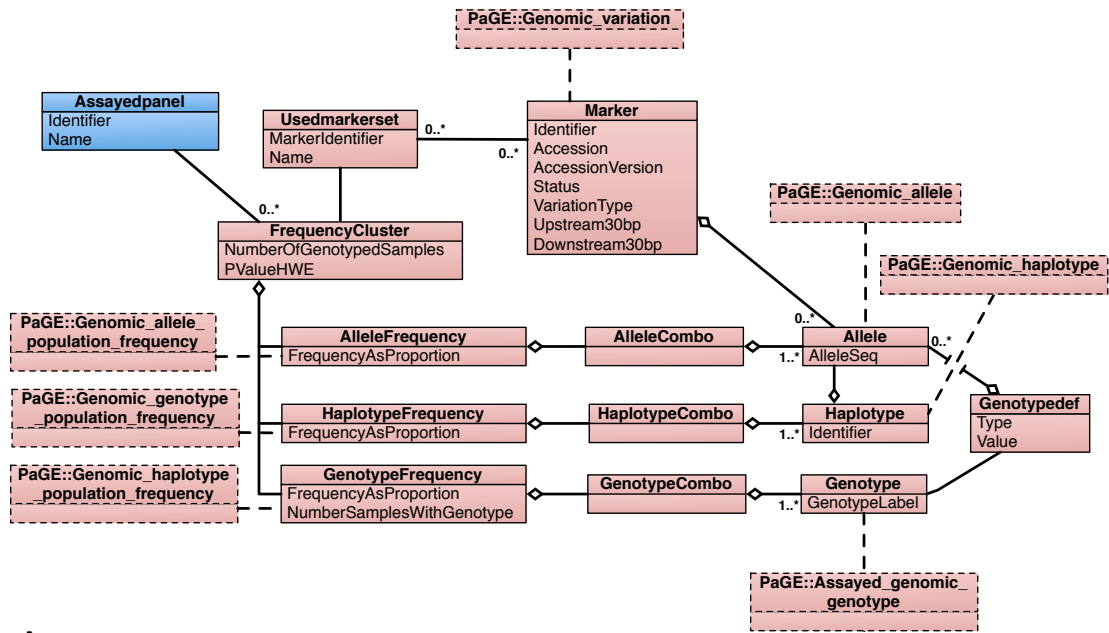


Fig. 3.2: The SAMPLE domain of the HGVbaseG2P model (A) and data example (B).

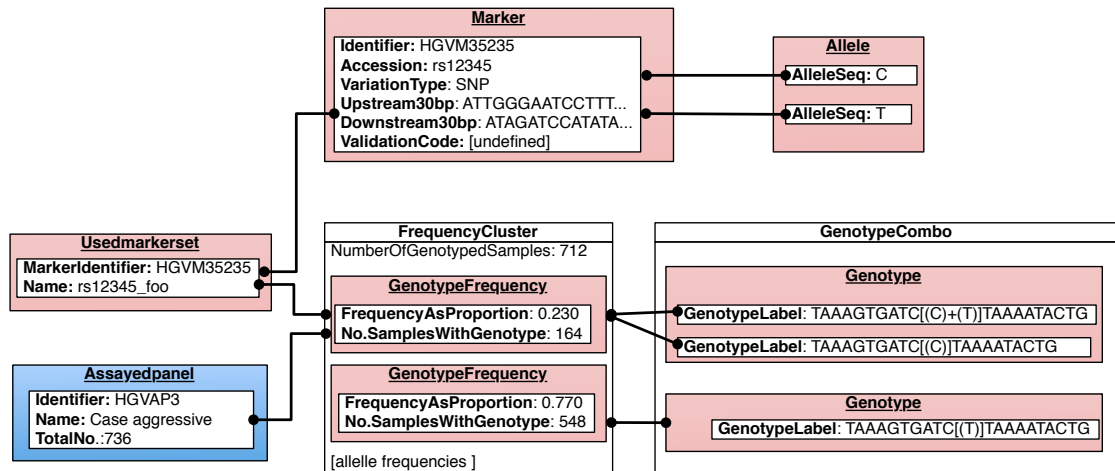
3.1.2 The GENOTYPE domain

One of the two main purposes of the the GENOTYPE domain of the model, shown in Fig. §3.3, is to describe variant sites in the genome as represented in reference variation archives (see §2.2.2). The central concept of this part of the model is the *Marker*, which is operationally defined as “a DNA sequence for which identical or highly similar instances exist at one or more locations in a genome” (see §B.1). A *Marker* is typically characterised by up to several *Alleles* which represent the set of alternative sequences, or versions, that have been reported for the given *Marker*. A combination of *Alleles* across multiple *Markers*, located on the same contiguous strand of DNA, constitute a *Haplotype*. As indicated in Fig. §3.3A, these classes all have direct mappings to equivalent PaGE-OM classes (see Fig. §B.3, §B.4 and §B.5).

The other main purpose of the GENOTYPE domain is to facilitate aggregate representations of genotype data generated in modern genetic and genomic experiments. The *FrequencyCluster* class represents the collection of allele, genotype and haplotype frequency data generated for a given *Marker* on a given *Assayedpanel* in a particular study. *Usedmarkerset* serves as a “proxy” or abstraction layer that sits between between frequency data generated in association studies and reference variation data, and as such it



A



B

Fig. 3.3: The GENOTYPE domain of the HGVbaseG2P model (A) and data example (B).

can represent either a single marker as genotyped in a study (the common case), or several markers in a set (e.g. for describing multi-marker tests for association).

Each type of frequency data element in the HGVbaseG2P model has a direct mapping to corresponding PaGE-OM constructs. The HGVbaseG2P model further adds an organisational layer called “combos” which are used to group together multiple instances of *Allele*, *Genotype* or *Haplotype* for the purpose of analysis. *GenotypeCombo* facilitates grouping of several distinct *Genotypes* into one genotype class. The data example in §3.3B shows frequencies reported for two SNP genotype classes: for “C” allele carriers on one hand, and for non-carrier “T” homozygotes on the other. Similarly, an *AlleleCombo* allows grouping of alleles into allele classes for the same locus (not shown in data example).

The *Genotypedef* construct provides a way for describing observed genotypes in greater detail, for example by specifying the absolute or relative number of alleles detected in a genotyping experiments. However, datasets from mainstream association studies do not as a rule include such detail, nor is there much added value in this for SNP genotype data at the aggregate level, as is the focus here. This part of the model has therefore been little tested or used. But as noted in the previous chapter, capturing such experimental details for reported genotypes will increase in importance as CNVs and other class of structural variation become more routinely assayed in association studies.

Nearly all of these classes for describing genotype data have straightforward mappings to equivalent classes in PaGE-OM, with the exceptions mainly relating to implementation conveniences (see Chapter 4). PaGE-OM further adds *Variation_assay* and several other classes for describing experimental assay details, but since capturing this information is not a priority of HGVbaseG2P, such constructs do not feature in the HGVbaseG2P model.

3.1.3 The SEQUENCE domain

The SEQUENCE domain of the HGVbaseG2P model describes how a variant sequence is located on, or mapped to, a sequence assembly. As discussed in Chapter 2, data models, databases, software tools and exchange formats for genetic sequences and sequence

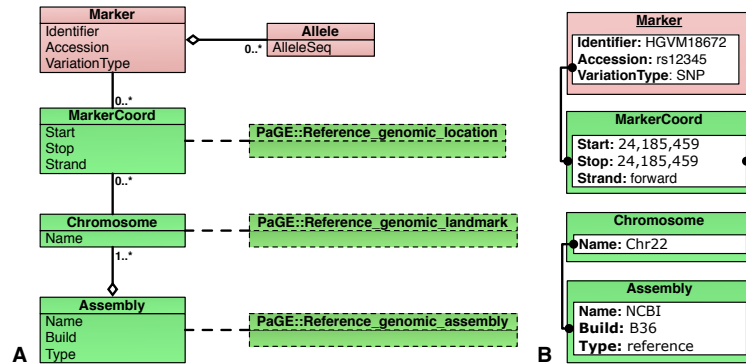


Fig. 3.4: The SEQUENCE domain of the HGVbaseG2P model (A) and data example (B). dbSNP entry rs12345 has also been mapped to the Celera and HuRef alternative assemblies, but only the mapping to the NCBI reference assembly is shown.

annotations have a long tradition stretching back to the origins of GenBank and other sequence databases in the 1980s. Key concepts concerning genomic mapping information are thus quite mature by now and can be applied with little difficulty here.

Genomic coordinates for markers are described in a straightforward way with the set of classes shown in Fig. §3.4: a *Marker* can have up to several *MarkerCoords*, or locations, on a *Chromosome*. Multiple genome sequence assemblies are supported via the *Assembly* class, so mappings to non-reference assemblies can be handled if required. As further discussed in Chapter 4, all these constructs have straightforward mappings to various standard relational databases and software tools for manipulating sequence feature information.

The corresponding part of the PaGE-OM model provides rich structures for describing locations of variants on genetic, cytogenetic and sequence maps, as well as for describing effects of variants on gene transcription and translation (see Fig. §B.6). Given the restricted scope of the HGVbaseG2P model, only PaGE-OM classes concerned with sequence mappings are of relevance here. However, the HGVbaseG2P model could be easily be extended by adding these PaGE-OM classes as a means to handle non-sequence mapping information if required in the future.

3.1.4 The PHENOTYPE domain

The concepts and relationships in the domains described thus far are relatively well-recognized and have representations which are more or less agreed. With minor exceptions, concepts such as “genotype”, “sequence” and “allele” have the same or similar meaning across the G2P community, and domain models (such as PML mentioned above) are well-developed. This has been helped by the relatively simple nature of information describing sequences and sequence variation, which makes it easier to standardise how this information is represented. However, as discussed in §2.3.2, the hugely diverse phenotypes and other observational data collected in biomedical investigations present a far greater standardisation challenge.

A further complication arises from different interpretations of the term “phenotype” itself. For example, to a geneticist studying a particular gene, the phenotypic consequence of a mutation in a particular gene is conceptually different from weight, disease affection status and a myriad other variables of interest to an epidemiologist studying disease in human populations. Furthermore, even within a given research sub-community the term “phenotype” is used interchangeably to refer to what was measured, how it was measured, the outcome of the measurement, or all of the above.

The PHENOTYPE domain of the HGVbaseG2P model shown in Fig. §3.5 takes an inclusive view of the elusive phenotype concept, and approaches the problem by dividing phenotype into three distinct sub-components or elements. The *PhenotypeProperty* class represents the abstract concept of the character or trait investigated, which may be defined at various levels of granularity, such as in the context of particular anatomical structures (e.g. nose size) or categorisation (e.g. disease affection status). The second element is the *PhenotypeValue* which represents a particular observation result produced by measuring a given trait, such as size of nose=1.7cm or disease case/control status. Finally, the *PhenotypeMethod* class describes the measurement method, for example by measuring the nasal septum in centimetres to the first decimal place with a ruler or by determining via a standard clinical protocol whether a person is affected by a given disease. The

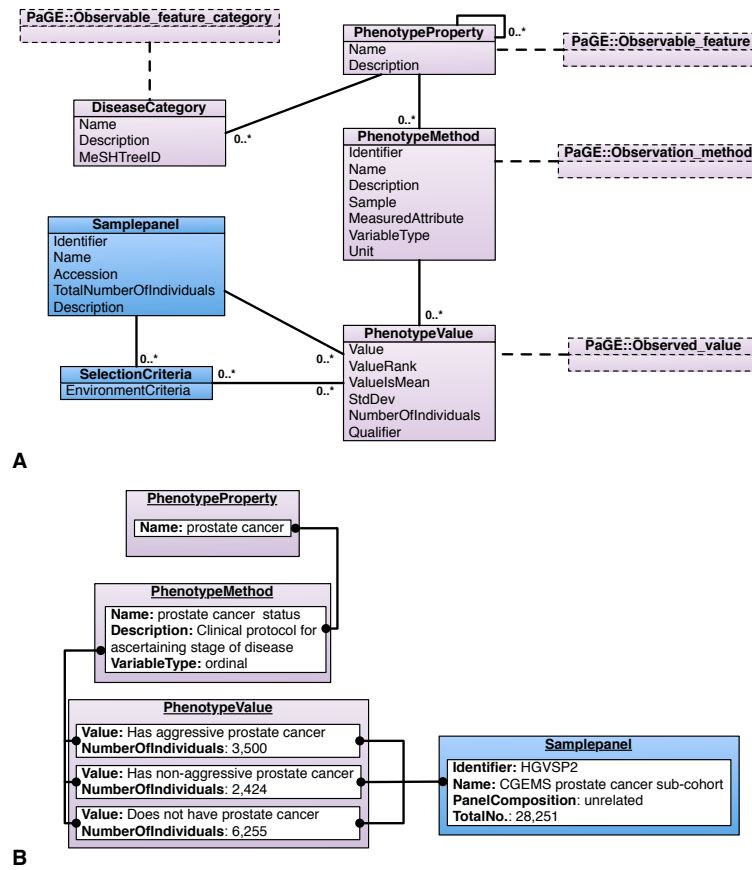


Fig. 3.5: The PHENOTYPE domain of the HGVbaseG2P model (A) and data example (B).

information captured by *PhenotypeMethod* is critical to meaningful interpretation of the resulting observational data, in particular when assessing whether a set of results from one study are comparable from results from another study (for example when pooling data from multiple studies).

As a general framework, this tripartite system is a flexible way of describing a wide range of phenotype observations, including values for discrete ordinal or nominal variables, or categories (e.g. disease affection status) and quantitative variables such as weight. As with the other domains of the HGVbaseG2P model, the focus here is on characterising groups of individuals, and so an instance of *PhenotypeValue* for a non-continuous variable describes the number of individuals observed with that value on a specific *Samplepanel* (see data

example in Fig. §3.5B). For a continuous variable, on the other hand, the *PhenotypeValue* contains the mean of values (and other descriptive statistical metrics) across the group of individuals. Besides describing groups of study subjects, the *PhenotypeValue* class can also be used to specify criteria for selection of individuals on panels (see Fig. §3.2 in the previous section). Most real-world use of this has involved simple case/control disease status categories, but the model also allows specifying thresholds for continuous variables (e.g. statements like “99 individuals in panel ‘MyGroup’ weigh 100kg or more”).

The corresponding PaGE-OM domain shown in Fig. §B.7 and §B.8 shares the same tripartite division of the phenotype concept, but does so in a more general way to also capture non-phenotype observations. Additional PaGe-OM phenotype-related features include a flexible, detailed model for describing different kinds of observation values (see Fig. §B.12), and a mechanism for describing inferred values, or conclusions, from any number of primary observations (see also §4.3.3). Given the more limited scope of the HGVbaseG2P model, such advanced constructs were not deemed necessary, and so the single *PhenotypeValue* class is used to represent all values.

3.1.5 The EXPERIMENT domain

The EXPERIMENT domain ties all the previous model domains together into a complete description of a genetic association study. The overall high-level organisation of the model follows established conventions across the empirical sciences, in that various packets of information are organized much like the Results section of journal manuscript. As the upper half of Fig. §3.6 illustrates, the *Study* class comprises various components or assets which are used in up to several G2P correlation *Experiments* conducted as part of the overall study. Most of these high-level metadata constructs have close mappings to similar or equivalent concepts in PaGE-OM and other models in the domain, notably FuGE-OM as discussed later in this chapter. Within an experiment, the three main sets of classes important here - those modelling genotype data, phenotype data and analysis results - are

demarcated by broken lines in areas labelled A-C in the lower half of Fig. §3.6, and further described below.

Genotype information. An *Experiment* in the HGVbaseG2P model contains aggregate representations of data generated in a genotyping experiment. As discussed in the GENOTYPE section above, this information is represented by one *FrequencyCluster* per marker, for each group of individuals or *Assayedpanels* (typically cases and controls) which were genotyped in the study, and which form the basis of the specific analytical question being asked. The classes shown in Fig. §3.6A align to the more general *Genomic_observation* class in PaGE-OM (see Fig. §B.5).

Phenotype information. Although subjects in an association study may have a wide range of phenotype and other observations associated with them, the results presented in a given *Experiment* relate to a question about one specific trait only. If multiple phenotypes are investigated during the course of a study, the results are organized into several *Experiments* within the study. The single phenotypic trait of interest in the analysis and, importantly, the method used to measure it in the study, is indicated with a direct link from *Experiment* to *PhenotypeMethod*, as shown in Fig. §3.6C.

Association analysis results. This part of the EXPERIMENT domain, shown in Fig. §3.6B at the intersection of genotype and phenotype, is the heart of the whole model and describes the final product of the G2P experiment. This product is the outcome of an analysis which tests the hypothesis that the observed genetic variation correlates with the observed phenotypic variation. The *Significance* class represents the level of significance of the outcome from a specific statistical test applied at a given site in the genome, typically a p-value from a single-marker test of independence in a traditional case-control study design. The model allows for an instance of *Significance* to link to the genotype data which underlie the analysis, via an association to *FrequencyCluster*. The *EffectSize* class models in a generic way various measures of risk associated with the tested marker given

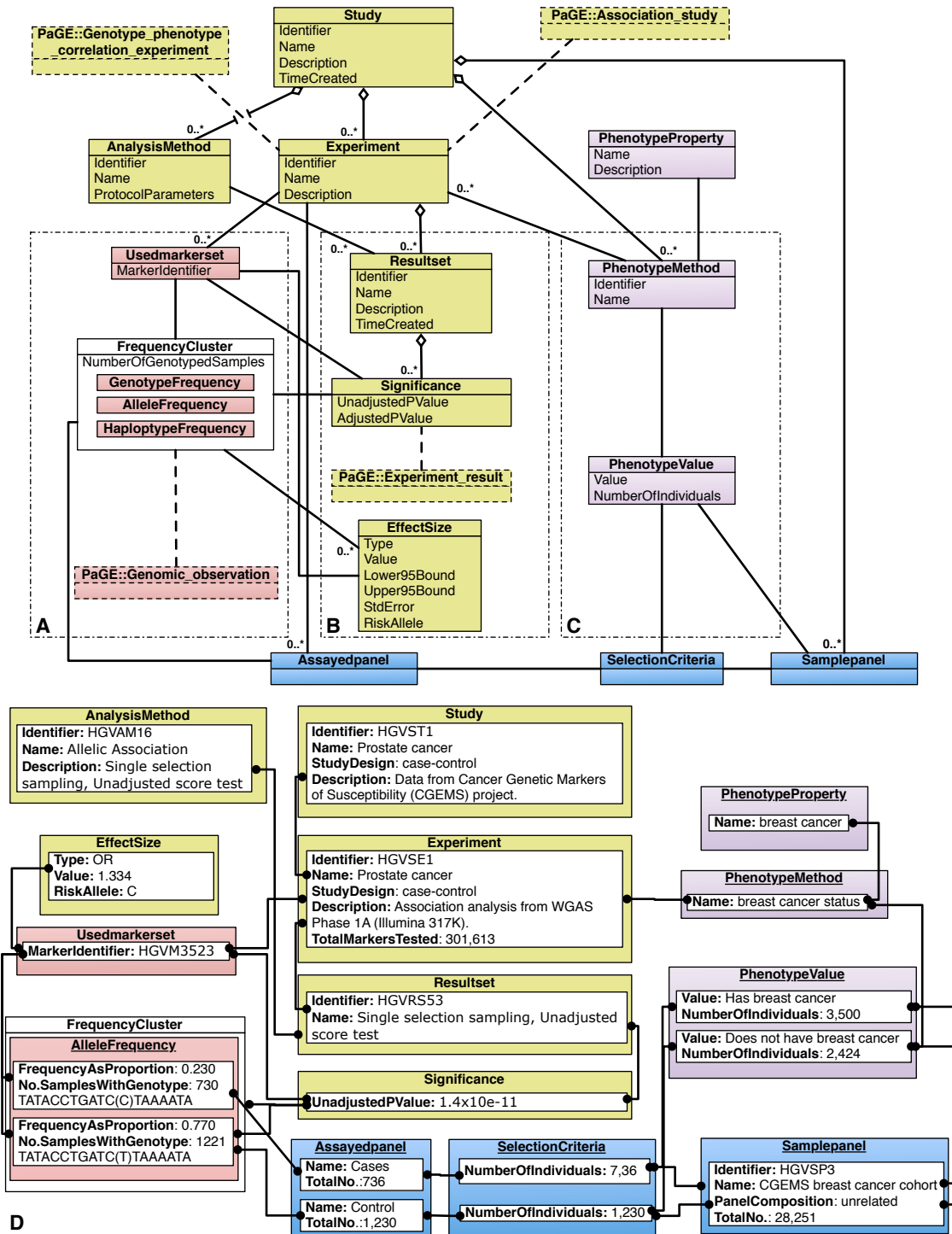


Fig. 3.6: The EXPERIMENT domain of the HGVbaseG2P model (A-C) and data example (D).

the data (Sistrom and Garvan, 2004), though in mainstream GWAS case-control study designs normally only odds ratios are an appropriate measure (typically reported in terms of increase in risk with each copy of the risk allele carried).

The above closely matches PaGE-OM constructs (see Fig. §B.10). Apart from adding the *EffectSize* class, the HGVbaseG2P model goes beyond PaGE-OM by incorporating an additional layer - the *ResultSet* class - which represents the process of conducting a statistical analysis using a specific method on an input genotype dataset. Information on the statistical method used, if available, is described by the *AnalysisMethod* class.

Submissions and author contributions. A notable feature of the HGVbaseG2P not matched by PaGE-OM is a set of classes for tracking contributions to studies in a fine-grained way such that various contributor roles are represented. Fig. §3.7 illustrates how a combination of *Submission*, *Contribution* and *Researcher* can be used to explicitly state that a particular person was one of several authors of a GWAS publication, whilst another person gathered or submitted the study data to the HGVbaseG2P database. The model also provides a basic means for capturing provenance for discrete data import batches, important in the common scenario when a HGVbaseG2P study comprises data from several GWAS publications submitted at different times. The need for both of these features was identified through practical use of the HGVbaseG2P system in GWAS data gathering (see §5.4).

3.1.6 The COMMON domain

The COMMON domain contains several general-purpose utility classes which are used in connection with many of the classes previously described. Fig. §3.8 illustrates how the *Crossref* class is used to describe cross-references from instances of the principal *Study* and *Marker* classes to entries in external databases, which are typically accessible over the Internet. The *Hotlink* class provides a non-redundant way to store the URL, which is combined with a local identifier in the external database and presented as an outgoing hyperlink in a web page. Similarly, the *Citation* class can be used to associate instances of

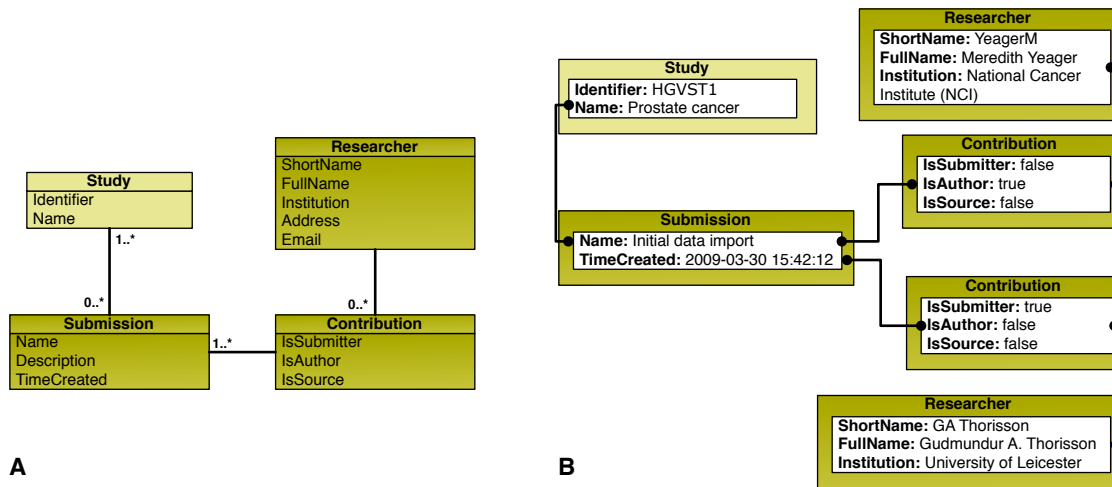


Fig. 3.7: Classes for capturing provenance in the HGVbaseG2P model (A) and data example (B).

several other classes with a bibliographic reference.

The general principle of such common, shared facilities is used in many related models, including PaGE-OM. But the HGVbaseG2P COMMON classes are not as well-aligned with the PaGE-OM reference model as most of the classes discussed thus far. For example, the *PaGE::Db_xref* class contains attributes spread across three classes in the HGVbaseG2P model: *Crossref*, *Hotlink* and *DataSource*. Furthermore, some of the more powerful facilities provided by the PaGE-OM COMMON domain have no equivalents in the HGVbaseG2P model. For example, *PaGE::Annotation* (see Fig. §B.11) can be used to describe any data object via generic attribute/value pairs. The omission of such a generic annotation capability does not detract from the usefulness of the HGVbaseG2P model in current practical use, but does limit the ability of the model to capture information that does not fit into the predefined class attribute slots. A further discussion on model flexibility and extensibility is presented in the next chapter.

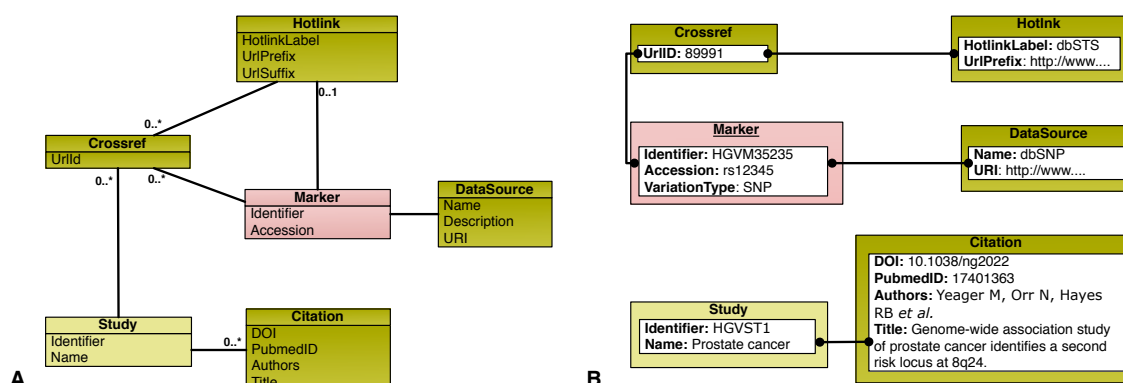


Fig. 3.8: The COMMON domain of the HGVbaseG2P model (A) and data example (B).

3.2 Validating the model

As the model above evolved, it was implemented and hence I could assess how well it worked when applied to real-world data. The results presented in the subsections to follow, though not exhaustive, are representative of the model validation which took place on an ongoing basis during HGVbaseG2P system development and data loading. Model validation is something of an inexact science, inasmuch as it is frequently performed by “populating” the abstract object model with test data, followed by a subjective evaluation of whether or not the object model adequately captures the data. The first two subsections below report results from such informal validation exercises using reference variation data and association study data from dbSNP and CGEMS, respectively. The data examples shown are typical of the datasets which currently populate the HGVbaseG2P catalog (see also §5.3 and §5.4). The validation exercises presented in the latter two subsections, on the other hand, use variation data from DGV and study data from dbGaP which are not routinely gathered into the HGVbaseG2P catalog at present, but which serve a useful test for other parts of the model not stressed by data in the previous category.

3.2.1 Marker data from dbSNP

Due to its importance as a reference source of sequence variation data in the context of SNP-based association studies, representing dbSNP reference SNP entries, or rs#'s (see also §A.2.2), was a key requirement of the HGVbaseG2P model. By way of its origin as a database focused on SNP variation, HGVbaseG2P inherited the foundation for the simple yet useful model of genetic variation discussed in the previous section. It was not the intent of the model to represent the full extent of variation details archived by the dbSNP database. Rather, the intent was to capture a concise, summary view of a dbSNP entry, with the minimal information necessary to the interpretation of results from a study where the SNP was tested for association.

The data example already provided in Figure §3.3B for rs12345⁵ illustrated the simplest case of a SNP polymorphism represented in the HGVbaseG2P model. The following data examples illustrate the capability of the model for handling data for various other types of variation cataloged by dbSNP, and by extension data from other datasources where marker information is organised in a similar way.

Insertions and deletions. The rs# identified by rs4186⁶ is characterised by a 4bp insertion-deletion currently mapped to the reverse strand of the reference genome assembly between 12,508,855bp and 12,508,856bp on Chr11. Figure §3.9A shows a representation of this marker in the HGVbaseG2P model. The variation type for this entry in dbSNP is described as “DIP: deletion/insertion polymorphism” or “in-del” for short, whereas in HGVbaseG2P the type is specified as “indel” which is a standard controlled vocabulary term (stable ID: SO:1000032⁷) taken from the Sequence Ontology. Other dbSNP classes are similarly mapped to SO terms (see also §A.2.4).

⁵http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs12345.

Accessed: 2010-01-22. Archived by WebCite® at <http://www.webcitation.org/5n1MhkhxD>

⁶http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs4186.

Accessed: 2010-01-22. Archived by WebCite® at <http://www.webcitation.org/5myHcFuxe>

⁷http://www.sequenceontology.org/miso/current_release/term/SO:1000032

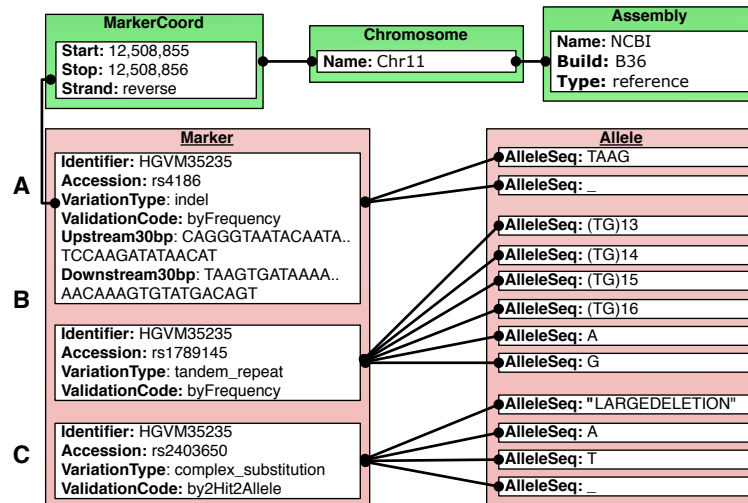


Fig. 3.9: Marker information from dbSNP reports for three variants represented in the HGVbaseG2P model. Genomic mapping information is easily captured, as is the list of reported allele sequences encoded according to the HGVbaseG2P nomenclature rules. For brevity, mapping information to the reference genome assembly is shown for only one of the three markers, and not all attributes are shown for all markers.

Microsatellites. The microsatellite marker identified as `rs1789145`⁸ and depicted in Fig. §3.9B is characterised by four allele sequences containing the TG sequence repeated in different numbers and two single-base alleles A and G. The HGVbaseG2P nomenclature rules (see Chapter 5) are used here to indicate repeat numbers in the sequence string, which in this case are identical to dbSNP conventions for representing tandem repeats.

Complex variation. The final dbSNP data example is `rs2403650`⁹, a complex polymorphism characterised by allelic heterogeneity. Two single-bp alleles have been reported for this sequence by several independent submitters, as well as an insertion-deletion which is not further described at the rs# level. However, further inspection of

⁸http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs1789145.

Accessed: 2010-01-22. Archived by WebCite® at <http://www.webcitation.org/5myHlqDE3>

⁹http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs2403650.

Accessed: 2010-01-22. Archived by WebCite® at <http://www.webcitation.org/5myHqpFoi>

the submitter SNP report for the insertion-deletion¹⁰ reveals that the submitter has indeed provided the full inserted sequence, via the “Comment” field in the dbSNP submission file. This is no doubt due to restrictions on the dbSNP submission format which specifies a total length of less than 255 characters for the field used to report observed alleles (see the dbSNP “How to Submit” page¹¹).

Reflecting on the criticism of dbSNP in §2.2.2, this further illustrates the apparent inability of this central archive - initially focused on simple, short sequence variation - to change and adapt to handle data describing more complex forms of variation. By contrast, the length of allele sequences is not restricted by the HGVbaseG2P model, and in this particular case the full-length insertion sequence would have been accommodated.

3.2.2 Study organisation and metadata from CGEMS

One of the earliest genome-wide scans were undertaken under the auspices of the US Cancer Genetic Markers of Susceptibility (CGEMS) project¹² to identify common variants associated with susceptibility to several cancers. Via the CGEMS data portal¹³, the project released aggregate GWAS datasets for the prostate and breast cancer studies pre-publication in late 2006 and early 2007, followed by peer-reviewed publications by Yeager *et al.* (2007) and Hunter *et al.* (2007), respectively. These two GWASs served as the primary test datasets in early HGVbaseG2P development, of which the prostate cancer study was selected as the focus of the validation exercise presented here.

Much of my model development and validation work was undertaken before the publication of Yeager *et al.* (2007). Preliminary study methodology details had, however, been published informally via the CGEMS data portal alongside the aggregate datasets. This

¹⁰http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ss.cgi?subsnp_id=8476988. Accessed: 2010-01-22. Archived by WebCite® at <http://www.webcitation.org/5myXxyt7g>

¹¹http://www.ncbi.nlm.nih.gov/SNP/how_to_submit.html#REPORTING_ALLELES

¹²<http://cgems.cancer.gov>

¹³<http://cgems.cancer.gov/data/>

document (PDF now available as supplementary materials¹⁴), along with documentation distributed with the aggregate data release, provided sufficient information to reconstruct the study organisation. Briefly, in Phase 1A of the project, an initial genome-wide association scan with 317,000 SNPs undertaken with the Illumina HumanHap300 array. In the Phase 1B of the project, another 240K SNPs were genotyped using the Illumina HumanHap240 array, and the data combined with data generated in Phase 1A to make up an overall 540K SNP GWAS. Each dataset was analysed using two statistical methods and two sampling strategies, resulting in four sets of results per dataset. Several key aspects of this GWAS investigation are captured by the model as described below.

Study organisation. As depicted in Fig. §3.10, multiple discrete experiments undertaken within the same overall investigation are easily accommodated. The first two *Experiments* contain, respectively, genotyping data and association analysis results from Phase 1A. The second two *Experiments* contain, respectively, genotyping data generated in Phase 1B combined with the Phase 1A data, and results from association analysis. The multi-stage CGEMS experimentation strategy can thus be described adequately, and this design is also useful for describing multiple experiments involving multiple phenotypes within the same study (e.g. prostate cancer subtypes).

The report by Yeager *et al.* (2007) also includes results from several follow-up replication studies, which tested highly-significant SNPs identified in the initial genome-wide scans in a number of other cohorts. These follow-up studies were not included in this analysis, but could be easily represented in the model as additional instances of *Experiment* nested within the same parent study (one for each replication cohort, or combination thereof, depending on the analysis undertaken).

Analysis results. Each set of results from a series of analysis over the same input dataset are modelled as an instance of *Resultset* linked to an instance of *AnalysisMethod*, which enables aggregation and comparison of analysis results by protocol. This contrasts with

¹⁴<http://www.nature.com/ng/journal/v39/n5/extref/ng2022-S5.pdf>

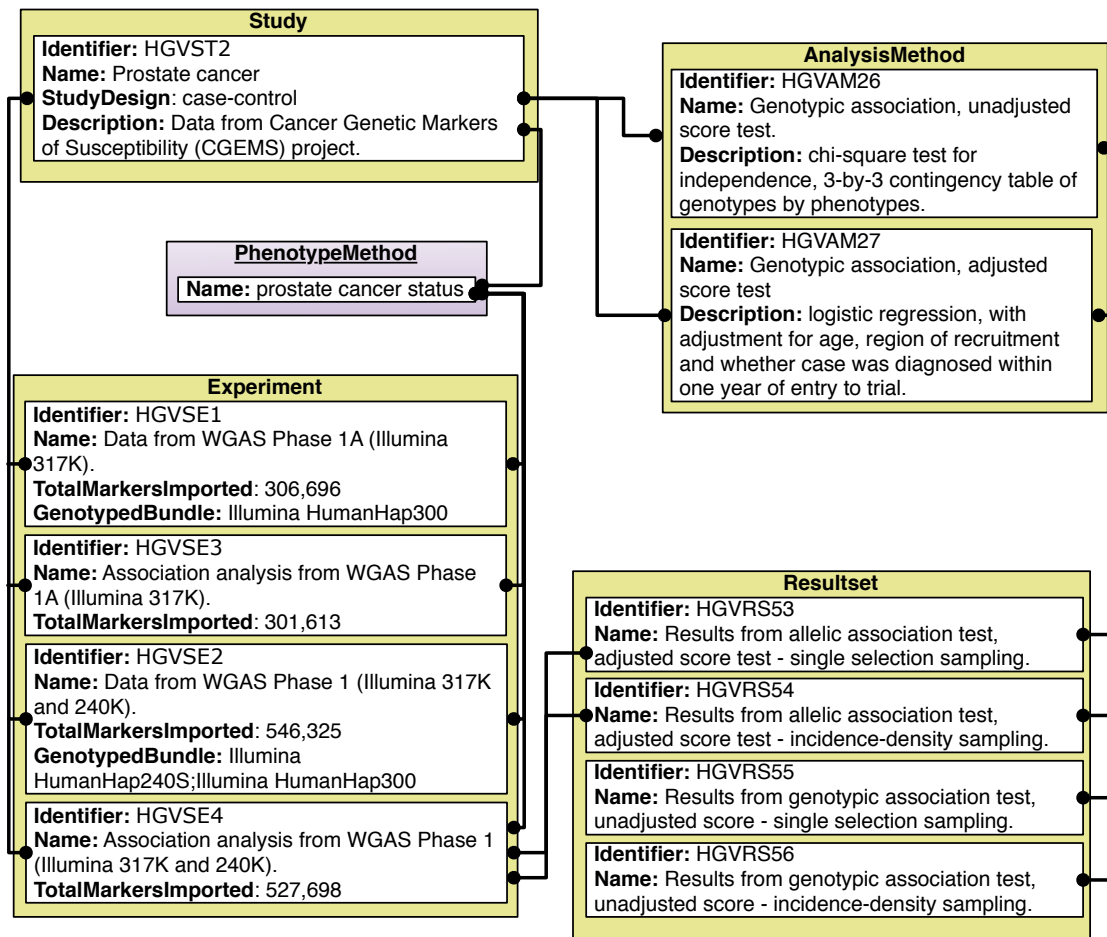


Fig. 3.10: Overall organisation of the CGEMS prostate cancer study, as represented in the HGVB G2P model. For brevity, several elements from Fig. §3.6 have been omitted, and not all possible *Experiment/Resultset combinations* are shown.

the section of the model previously shown in Fig. §3.6A, via which genotyping results can only be organised on a by-experiment basis, making handling of multiple sets of genotypes within an experiment difficult (see also discussion in next chapter).

Phenotypes. The phenotype aspect of study information is handled adequately in the model, as already illustrated in Fig. §3.5B. The primary trait of interest is prostate cancer status, more specifically which of the three categories a study subject belongs to: i) control (i.d. non-diseased), ii) non-aggressive cancer or iii) aggressive cancer. Importantly, the

model is not intended to capture in a structured way the full details of *how* subjects are classified - for example, ‘non-aggressive cases are men with a tumour score <7 and tumour stage $<III$ ¹⁵ - but rather the characteristics of each panel as this pertains to the association analyses undertaken and as presented by the study authors.

3.2.3 Marker and population data from DGV

CNVs and other structural variants are increasingly included in genome-wide investigations and establishing reference structural variation archives will be crucial to proper integration and interpretation of study data in the near future. The most comprehensive source of this information is currently the Database of Genomic Variants (DGV), and despite the shortcomings of DGV as a reference archive (already mentioned in §2.2.2), it is nevertheless useful as a source of data with which to test the HGVbaseG2P model.

Copy-number variation. The data example in Figure §3.11 shows data extracted from the DGV report for the CNV identified as `Variation_48525`¹⁶ represented in the HGVbaseG2P model. As in the examples described in the previous section, the standard SO term “CNV” is used here, whereas DGV uses the non-standard term “CopyNumber”. In contrast to the sequence-level variation archived by dbSNP, CNVs are not characterised by multiple observed alleles. Instead, different number of copies of the variant region are observed in different samples, frequently reported relative to a reference sample which may or may not be diploid for the locus (as opposed to absolute copy numbers). Therefore, although the 13,6Kbp genomic sequence corresponding to the interval spanned by this CNV can indeed be represented in the HGVbaseG2P model as the allele, this is not

¹⁵the definitions of these criteria and other details are provided in the primary paper by *Yeager et al.* or supplementary materials.

¹⁶http://projects.tcag.ca/cgi-bin/variation/xview?source=hg18&view=variation&id=Variation_48525. Accessed: 2010-01-22. (Archived by WebCite® at <http://www.webcitation.org/5myGTxZKO>)

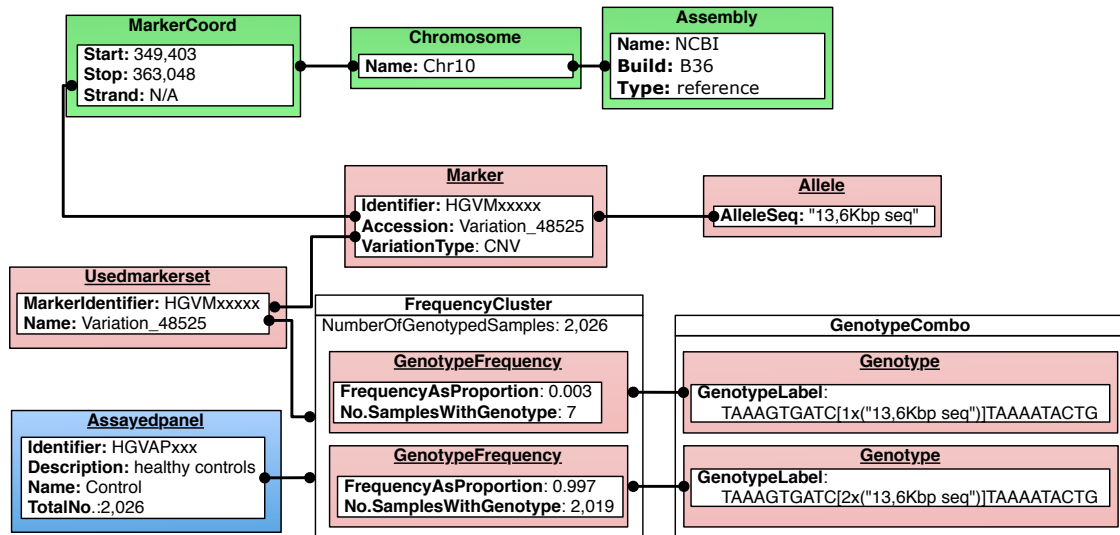


Fig. 3.11: Marker information from a Database of Genomic Variants report for a copy-number variant represented in the HGVbaseG2P model.

sufficient to characterise this marker beyond merely stating that the specified region is copy-number variant.

The observation that 7 out of the total 2,206 individuals tested, or 0.3%, show a loss in copy-number at this locus is the key characteristic of this CNV, and the HGVbaseG2P model can indeed capture this as aggregate study genotype data as shown in Figure §3.11. However, this arrangement does not fit at all well with the notion of a “reference” variant in the dbSNP rs# mould, which is a key premise of the HGVbaseG2P model and many other models for genetic variation. That is, a variant site in the genome sequence is discovered with a given technique (e.g. reduced-representation shotgun sequencing and subsequent read alignment), and is subsequently incorporated into routine, high-throughput genome-wide investigations which generate genotype data (e.g SNP genotyping arrays). The model does not provide a mechanism for stating, at the summary level, that `Variation_48525` is characterised by multiple observed genotypes (2x copies vs 1x copies), rather than by multiple observed alleles.

Whilst adding a new class to the model to list known genotypes for a marker (analogous

to the *Allele* class for known alleles) may partly address this issue, further complications remain regarding boundaries of the genomic region harbouring the CNV. The sequence mapping information extracted from DGV indicates that this CNV is located on Chr10 from 349,403bp to 363,048bp on the reference genome assembly. However, unlike SNPs and indels in dbSNP, which are known at the sequence level and thus map (if they can be mapped) to an absolute genomic location or locations, due to the detection techniques currently employed to discover and type CNVs reported genome coordinates are usually only approximate. For example, the methodology used to detect this particular CNV, originally gathered from the CNV project at the Children’s Hospital of Philadelphia (CHOP)¹⁷ (Shaikh *et al.*, 2009), uses signal intensities from raw SNP microarray data to infer copy-number gain or loss. The resulting CNV calls are thus not ascertained at basepair-level resolution. From the perspective of HGVbaseG2P, it is not clear how to reconcile such “fuzzy” delineation of CNVs and regions harbouring them with an object model centred on a sequence-based *Marker* concept.

3.2.4 Phenotype data and metadata from dbGaP

A key requirement of the HGVbaseG2P model was the ability to incorporate phenotype descriptions from other GWAS resources. The data example previously shown in Figure §3.5B demonstrated the capability of the model to represent simple categorical variables commonly used to represent simple disease classification in the common case/control study design. But there was also a desire to represent more detailed, complex phenotypes when available, including quantitative data. At the time when this work described in this section was done, dbGaP (see §2.2.7) was the most easily-accessible source of structured phenotype information. Due to the initial emphasis placed on gathering GWAS datasets from dbGaP (see Chapter 4), assessing compatibility with dbGaP phenotype data was an important test for the HGVbaseG2P model.

¹⁷<http://cnv.chop.edu>

Acquiring dbGaP phenotype information. In addition to individual-level study data and findings which are subject to access control, dbGaP contains a comprehensive collection of study documents which can be freely browsed via the website. This includes detailed metadata for all study variables, irrespective of whether they have relevance to the main GWAS findings, as well as aggregate phenotype data describing study subjects at the group level. The bulk of the variable metadata is provided in a free-text, unstructured format, but core metadata have been compiled into a structured format and published as a set of XML-files, alongside additional XML-files containing aggregate phenotype data. It is these core variable metadata and aggregate phenotype data which is of main interest here.

In order to test the HGVbaseG2P model, XML-files corresponding to the first study published via dbGaP - the National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS), dbGaP accession phs000001.v1.p1¹⁸) - were downloaded from the dbGaP FTP-site¹⁹. Out of the 174 total variables available for this study, two were chosen as the main focus for this analysis: i) the continuous variable `weight00` (dbGaP accession phv00000048.v1.p1) representing the weight at follow-up year 1, and ii) the unordered categorical, or nominal, variable `amdstat` (dbGaP accession phv00000173.v1.p1) which relates to the actual AMD disease phenotype investigated in this study. The relevant sections of the dbGaP XML-files are shown in Listings §?? and §??.

a39673918e608ab4aef05cf1586a8e80

Continuous variables. The `weight00` variable metadata provided in the dbGaP XML-file can be transformed in straightforward manner into the HGVbaseG2P model (Figure §3.12A). dbGaP characterises the type of the variable as “Num” for numerical or continuous, which provides the type attribute for the *PhenotypeMethod* class. Free-text

¹⁸http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v1.p1

¹⁹<ftp.ncbi.nlm.nih.gov/dbgap/NEI/AREDS/phs000001.v1.p1/>

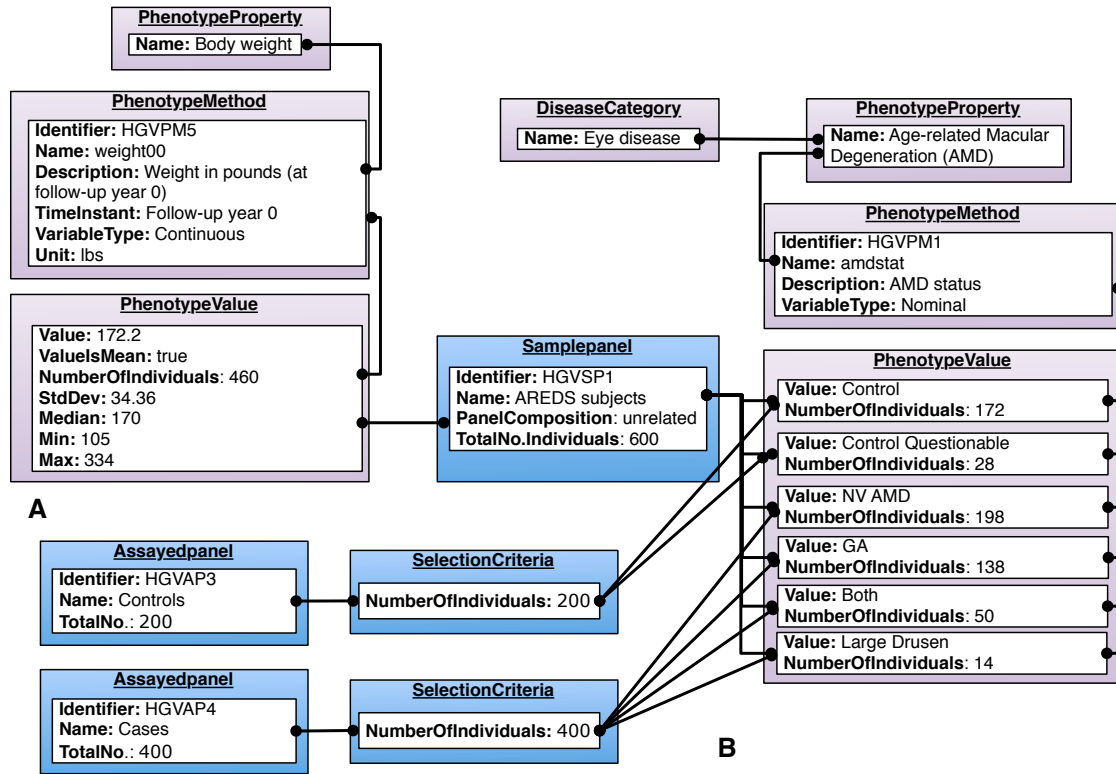


Fig. 3.12: Phenotype information from the dbGaP AREDS study represented in the HGVbaseG2P model. A) The continuous `weight00` variable. B) The categorical `amdstat` variable. Abbreviations: NV AMD = neovascular age-dependent macular degeneration, GA = geographic atrophy.

description, short name and variable type are handled by the respective attributes in *PhenotypeMethod*. The “follow-up year” parameter can be mapped to the *TimeInstant* attribute, albeit after manual intervention which is required to infer this from the free-text variable description. Further enhancement is provided by the *PhenotypeProperty* instance named “Body weight” which represents the measured trait at an abstract level. The concept “Body weight” is inferred (again, via manual intervention) from the dbGaP variable description rather than this being stated explicitly, and indeed dbGaP makes no attempt to harmonise or semantically annotate study variables in any way, whether within a study or across studies. Thus, in this particular study, there are 13 variables named `weight00` through `weight13` which are all concerned with weight at various years of

follow-up, but dbGaP provides no link between these variables to indicate that they all relate to the same trait. This focus on capturing core variable metadata in a systematic way but at the syntactic level only is a stated priority of dbGaP, with the intention that within- and cross-study harmonisation would be undertaken later (S. Sherry, personal communication).

With respect to phenotype data for the `weight00` variable, the commonly-used descriptive statistics provided in the dbGaP XML-file, such as mean and standard deviation across the AREDS subjects panel, are easily modelled as an instance of the *PhenotypeValue* class. Not all 600 study subjects have weight measurements, and this can also be represented by the model by specifying how many individuals on the panel contribute to this particular value (530, as 70 values are missing). The dbGaP XML additionally provides a set of values representing the discrete distribution of values, which is useful for rendering descriptive graphs (such as the graph shown on the “Variable” tab on the dbGaP report page for the AREDS study). This extra information cannot currently be represented in the HGVbaseG2P model, but could (if deemed necessary in the future) conceivably be modelled with by allowing self-recursion of *PhenotypeValue* instances - that is, by “nesting” *PhenotypeValues* one within the other. Alternatively, the model could be extended by adding a new class derived from *PhenotypeValue* (see also Pheno-OM discussion in §4.3.3).

Categorical variables. The dbGaP data example shown in Figure §3.12B relates to the actual AMD disease phenotype investigated in the AREDS study. As with the `weight00` variable, the `amdstat` variable is described quite succinctly with a name and description, and the type “Char” to indicate that it is a discrete, non-continuous variable. No attempt is made to link this variable to the AMD disease phenotype. This basic descriptive variable information is again handled by the *PhenotypeMethod* instance, whereas the *PhenotypeProperty* instance which represents the AMD disease concept is added as an enhancement, following manual inspection of supplementary, non-structured study documents.

As in the CGEMS data example above, a crucial part of the summary-level phenotype data provided for this variable is the classification of study subjects into case and control groups, according to the results of various eye exams conducted during the course of the AREDS study. Full details on these procedures are accessible as free-text descriptions via various study documents held by dbGaP, but as with the continuous variable above, the focus is here on the structured representation of the results as provided in the dbGaP XML. The five categories can easily be represented in the HGVbaseG2P model as five distinct *PhenotypeValues*, each specifying the name of the category and the number of individuals in that category. Furthermore, the *SelectionCriteria* class provides a way to explicitly state that individuals in the first two categories are classified as controls, whereas individuals in the remaining four categories (which show symptoms of the AMD disease) are classified as cases. This method for describing panel selection criteria or rules in a structured way thus contrast with the free-text, unstructured descriptions provided in the AREDS study documents.

In contrast to the `amdstat` variable, several other categorical variables in the AREDS study specify a list of pre-defined numeric or alphabetical codes. Such codes are often used as “shorthand” substitutes for disease classification (e.g. “1”=affected by disease X, “0”=unaffected), questionnaire answers (e.g. “Y”=yes, “N”=no) and similar purposes. At present the HGVbaseG2P model does not support such codes, which in the case of data gathering from dbGaP is not a problem since the aggregate phenotype data include both codes and category descriptions. However, this lack of expressiveness means that the model would not be able to fully capture phenotype data where only such codes are used.

3.3 Discussion

This chapter reported on my work to construct a conceptual data model for describing genetic association studies and aggregate study data. The resulting model provided an important conceptual grounding for the HGVbaseG2P database and software system, and as such I consider it complete given its original purpose and current use. Although

not explicitly derived from the more general and more detailed PaGE-OM model, this implementation model can be considered a specialised extension of PaGE-OM and was presented above in those terms.

In this final section, I summarise the main features of the model and highlight several areas where improvements could be made. I then conclude the chapter by proposing a formal model development process more suited to the collaborative modelling activities which are the topic of the next chapter.

3.3.1 The HGVbaseG2P model

Describing groups of study participants. The SAMPLE domain describes groups of individuals participating in an association study as two types of panels. A distinguishing feature of the model (albeit one that has seen limited practical use so far, see §5.4) is the ability to describe how study participants are selected from *Samplepanels* based on phenotype criteria to create *Assayedpanels* for experimental study.

After significant real-world usage of the SAMPLE domain, it is becoming clear that the distinction between *Samplepanel* and *Assayedpanel* is in many ways an artificial one. These two constructs are conceptually all but identical, and in essence they only differ in the way they are used. Furthermore, there is sometimes a need to create *Samplepanels* based on other *Samplepanels* based on certain criteria (see for example the two cohorts in the data example in Figure §3.2B), but the model does not support this at present. Therefore, a better strategy here is arguably to take cue from PaGE-OM and instead merge these two classes into single, generic *Panel* class for increased simplicity, while losing none of the flexibility.

Describing markers, alleles and aggregate genotypes. The GENOTYPE domain describes genetic variation and aggregate results from genotyping experiments. This part of the model has worked well in practical use in the intended target domain, where the centre of focus is a polymorphic SNP marker characterised by two or more known alleles

for which GWAS subjects are tested. After significant real-world use of this part of the model for mainstream GWAS datasets and small-scale model validation using structural variation data, two main areas have been identified where improvements are needed. The first concerns handling of frequency data for alleles, genotypes and haplotypes and overall complexity of the model. Through the use of “combos”, the HGVbaseG2P model accommodates groupings of genotypes, alleles and haplotypes for the purpose of analysis, and as such the model is flexible and relatively “future proof” in this regard. However, this support for “edge cases” (for which real-world test datasets have not been gathered), as currently structured in the model, introduces unnecessary complexity for the simpler, common case (mainstream GWAS genotype data) and results in increased implementation complexity (see also §5.1.2). A simpler design - one that facilitated representation of such complex cases by way of optional extensions - would be advantageous, for example a minimal model like that used in XGAP (see next chapter).

The second aspect of the GENOTYPE domain that requires attention pertains to the core concept of a “marker”, and its utility in describing structural variation. The validation exercise presented in §3.2.3 illustrated the limitations of the current HGVbaseG2P model (and, by extension, PaGE-OM) for describing copy-number variant regions, and similar difficulties arise when considering translocations, inversions and other complex rearrangements. With minor extensions to the model, however, it may be possible to usefully represent structural variation at a more abstract level as a genomic interval for which the sequence basis of variation is not necessarily known.

Imprecisely-defined sequence coordinates for regions identified as harboring structural variants also present a modelling challenge. The very nature of complex structural variation may render the concept of a reference marker useless, and instead focus needs to be placed on standardised reporting of structurally-variant regions in terms of the experimental and analysis techniques used (i.e. an “audit trail” for CNVRs, as suggested by Scherer *et al.* (2007)), at least until structural variation has been fully characterised at the sequence level in a reference set of individuals. Specific suggestions regarding how to address this problem are beyond the scope of this thesis and will thus not be further discussed here. But given

that this area of research is progressing rapidly and is in a state of flux, it would seem prudent to wait until reporting standards begin to emerge and can inform work to refine reference domain models such as PaGE-OM.

Describing study organisation and findings. The EXPERIMENT domain provides several high-level concepts for organising study contents into a coherent whole, modelled after a conventional journal manuscript. Since related models in the biosciences (such as the XGAP and FuGE-OM models analysed in the next chapter) tend to follow the same general paradigm, many of the same concepts are found in these models as well, and this should in principle ensure a certain level of compatability between the models. A key feature of the HGVbaseG2P model which needs to be explored further is the ability to track individual data contributions to a study (see also Chapter 5 for further discussion on this and related issues).

At lower levels of organisation, classes for grouping association analysis findings by analysis instance, or “run”, are very similar to that used in related models for describing the general process of applying a protocol. Such facilities for capturing provenance information and data organising in a more fine-grained way are lacking for genotype data and phenotype data. Given the scope of HGVbaseG2P as a summary-level database and the common case of GWAS publications where only a single genotyping platform was deployed, this omission was not deemed crucial. Nonetheless, more flexibility here will be important in the future for accurately describing findings from more complicated analysis workflows, such as GWAS meta-analyses using multiple genotype datasets produced by different genotyping platforms.

Association findings are modelled via a single class representing the outcome, or significance, of a single- or multi-marker statistical test for association, optionally linking to an aggregate representation of the underlying genotype data. This has served well in practical use, given the needs of the HGVbaseG2P project - i.e. no individual-level genotype data, limited phenotype data (see next chapter) - but one limitation of this simple model is that it is not possible to directly link to the phenotype data used in the

analysis. Due to this and the lack of support for provenance, a significant amount of inference is required to resolve exactly which phenotype observations (possibly amongst many) underlie a given set of findings. This limitation does not, however, affect simplified representations of study findings, where panels are not characterised beyond simple case/control status (currently the common case in the HGVbaseG2P database). But this would seem to significantly reduce the ability of the model to handle more detailed phenotype data and appropriately relate these to analysis outcomes.

Describing phenotypes. The tripartite organisation of the “phenotype” concept into its core sub-elements is the cornerstone of the PHENOTYPE domain of the model. The general principle of distinguishing between the property that was measured, the method used to measure it, and the results from the measurement is emerging as a powerful modelling pattern for phenotype and non-phenotype observations alike. The three classes representing these core concepts in the HGVbaseG2P model are, however, currently associated in such a way that limits the functionality of the model in fundamental ways, as discussed in the next chapter.

3.3.2 Future development: formalising the model

The HGVbaseG2P model in its current form is intimately tied to the HGVbaseG2P system for which it was created, and as such it works well in this capacity as a model for current mainstream genetic association studies. However, the G2P field is advancing rapidly. Knowledge of genetic variation is expanding, and the SNP-based GWAS approach are likely to be supplanted by large-scale sequencing-based studies within a few years. It is therefore vital that the HGVbaseG2P system (and, consequently, its conceptual underpinnings) evolve to accommodate new study designs and expanding biological knowledge.

The current HGVbaseG2P model is, however, not well suited as a starting point for further work in this direction. The comparative model analysis and discussion presented in the

next chapter will elucidate some of the rationale for this conclusion. But first, I will close the current chapter by highlighting a specific technical issue which needs to be addressed in future development of the model.

An informal modelling process. At present, no formal modelling language specification exists for the HGVbaseG2P model. The logical diagrams presented in this chapter were created by hand, specifically for the purpose of illustration. These and several other diagrams, combined with a loose collection of textual documents, comprise the informal documentation of the model. The HGVbaseG2P relational database schema (see §5.1.2), meanwhile, serves as the master specification for classes (tables), class attributes (columns) and class associations (table relations). This mixture of informal graphical visualisations or “sketches” and documents worked adequately in early phases of the project, when modelling was focused on the HGVbaseG2P system and its relational database underpinnings. However, as reported in the next chapter, model development in the project has increasingly involved collaborative work with GEN2PHEN partners and others, in a broader context beyond the HGVbaseG2P system proper. It is now clear that in order to reap maximum benefits from such joint development work, and from combining and extending related object models, sophisticated modelling tools need to be deployed. Specifically, there is a need for progressing from using quasi-UML diagrams as an informal communication tool (analogous to hand drawings), to using UML notation as a formal “blueprint” for describing the object model structure. Such formalised UML models are now commonly produced and published as part of standardisation initiatives in the biosciences, as noted in §2.5.

Benefits of a formal modelling framework. Many of the suggested model improvements outlined above and in the next chapter are best implemented via inheritance, a standard technique commonly used to create extensible information models (Jones and Paton, 2005). This is not straightforward to do in an informal modelling framework, due to difficulties in keeping track of which subclass is derived from which superclass, even

within the same model. For example, *Samplepanel* and *Assayedpanel* in the HGVbaseG2P model are implicitly derived from a more general “panel” concept, but this fact is not stated formally. By contrast, the formal PaGE-OM model explicitly declares that the *PaGE::Lifestyle_feature*, *PaGE::Phenotype_feature* and *PaGE::Environment_feature* classes are derived from the more general *PaGE::Observable_feature* class. A formal model notation also facilitates declaring an explicit derived-from relationship between a class in an external model (e.g. *PaGE::Panel*) and a more specialised subclass in the target model (e.g. *Samplepanel*). Indeed, the majority of classes in the HGVbaseG2P model align in this way to either PaGE-OM or FuGE-OM, and declaring these mappings formally would help considerably with understanding various cross-model relationships and semantics.

Another advantage of formal notation relates to documentation of the model and communicating it to HGVbaseG2P developers and others outside the project. Low-level documentation (notably class and attribute definitions) would be embedded in the model itself, thus largely eliminating the need for manually updating documentation located elsewhere (see also Chapter 4 Discussion). Moreover, standard modelling tools typically have facilities for auto-generating logical diagrams which are linked directly to the formal model, a big advantage over the labour-intensive, error-prone process of manually creating logical diagrams, as has been done here.

A formal model can also be used as input for code generation in a model-driven framework. Though not universally liked by software developers, generating sourcecode from object models has some advantages, in particular for saving repetitious coding of low-level functionality (e.g. object methods for getting/setting attribute values). The DSL-based Molgenis platform described in §2.6.1 is a good example of such a framework. Code generation from a relational database schema is also strategically used in certain parts of the HGVbaseG2P system (see Chapter 5).

Finally, a formal model notation is a prerequisite for creating a RDFS/OWL representation of the object model as a formal ontology. This will be important in the context of future Semantic Web developments in the project, as such an ontology representation of the

model can endow HGVbaseG2P data content published as Linked Data with structure and semantics.

4. Comparing related G2P domain data models

As my work on the HGVbaseG2P project progressed, it became clear that many aspects of the original conceptual model could have broader implications in the context of other modelling efforts in the domain. The special relationship with the PaGE-OM standardisation activities, explained in the previous chapter, is an example of this. Through the involvement of our group in the PaGE-OM consortium, several refinements to the HGVbaseG2P model were fed back to - and thus influenced the design of - the more general-purpose and detailed PaGE-OM standard. Beyond this connection with PaGE-OM, I recognised the importance of aligning my efforts with other modelling initiatives. Two initiatives were identified as particularly important in this regard, both of which published formal models after my PhD work commenced:

FuGe-OM: A high-level model for “omics” investigations. In contrast to many other domain standards developed in the life sciences which have tended to be technology-specific, the Functional Genomics Experiment object model already mentioned in §2.5 was designed to capture and generalise the core concepts that are common to the various “omics” technologies (Jones *et al.*, 2009). Specifically, FuGE-OM aims to describe i) the investigation design and experimental variables, ii) procedures or protocols for experiments and data analysis, iii) materials used in experiments and iv) the data generated by instruments or from analysis.

XGAP: A G2P domain model. The eXtensible Genotype and Phenotype object model (XGAP-OM) is the underlying conceptual model for the XGAP system introduced in §2.6.1. XGAP-OM was initially created to address practical databasing problems in

model organism-based G2P research, such as the management and analysis of mouse gene expression datasets and metabolic network reconstruction. The system has recently also been used for human GWAS datasets, e.g. a study of gene expression phenotypes by Stranger *et al.* (2007). See the XGAP dataset listing page¹ for details. The XGAP-OM model thus shares many features with the HGVbaseG2P model, with respect to both the common G2P target domain and common roots as implementation models.

The first two sections in this chapter will present results from a cross-model mapping and comparative analysis of the HGVbaseG2P model versus FuGE-OM and XGAP-OM, respectively. This sets the stage for the final section which introduces Pheno-OM, a new object model for describing phenotypes and other observations. Pheno-OM is a key outcome from joint G2P modelling work in which I was centrally involved, facilitated by the partnership of our group in the GEN2PHEN Consortium. Key features of this model and a comparison with the HGVbaseG2P model will be presented and discussed.

4.1 Aligning with FuGE-OM

The primary motivation for establishing mappings from the HGVbaseG2P object model to FuGE-OM relates to enhancing data accessibility and data exchange. First, in the broader context of “omics” investigations, it would be advantageous if association study metadata in the HGVbaseG2P system were compatible with metadata from other kinds of investigations in other databases. Emerging standard frameworks such as the FuGE-OM based Investigation/Study/Assay (ISA) Infrastructure² could then be leveraged to retrieve, process and otherwise manipulate HGVbaseG2P study metadata. For example, in order to increase visibility of an HGVbaseG2P study, metadata could be submitted to central “yellow pages” investigation registries such as the BioInvIndex (BII)³ created by the EBI

¹<http://www.xgap.org/wiki/DataSets>

²<http://isatab.sourceforge.net>

³<http://www.ebi.ac.uk/bioinvindex>

Nutrigenomics, Environmental genomics and Toxicogenomics (NET) programme⁴.

Second, in the specific context of association studies, standardising on high-level study representations would facilitate exchange of study metadata amongst primary G2P archives, secondary GWAS resources such as HGVbaseG2P, and other parties. This would also enhance study discoverability and make it easier for researchers to retrieve all relevant data which may be fragmented across several archives (see §2.2.7 and Chapter 5). Exchange of standardised sample metadata is also becoming important in biobanking activities where, in order to boost study sample sizes, researchers need to source biosamples with suitable characteristics from multiple biobank repositories from around the world. An example of a system which enables this is the Sample Availability System (SAIL)⁵, created for sharing of biosample annotations and availability information in ENGAGE and several other major biobanks in Europe.

Study organisation. As noted in the previous section, the high-level concepts in the EXPERIMENT domain of the HGVbaseG2P follow established conventions with respect to how research manuscripts are organised. This is reflected in the straightforward mappings of these concepts to equivalent or more general classes in the `FuGE::Bio::Investigation` package. As illustrated in Fig. §4.1, the highest level organisation unit in the HGVbaseG2P model, *Study* maps to the equivalent concept `FuGE::Investigation` and the second-tier *Experiment* class corresponds to `FuGE::Study`. *Samplepanel* and several other study components can be considered as specialised `FuGE::InvestigationComponent`.

Protocols and provenance. The EXPERIMENT domain of the HGVbaseG2P model has some similarities with FuGE-OM concepts relating to capturing protocol information. Although it is not the intent of the model to describe in great detail how an association study was conducted, nevertheless it is useful to represent certain study workflows at a

⁴<http://www.ebi.ac.uk/net-project/>

⁵<http://www.ebi.ac.uk/Tools/sail/>

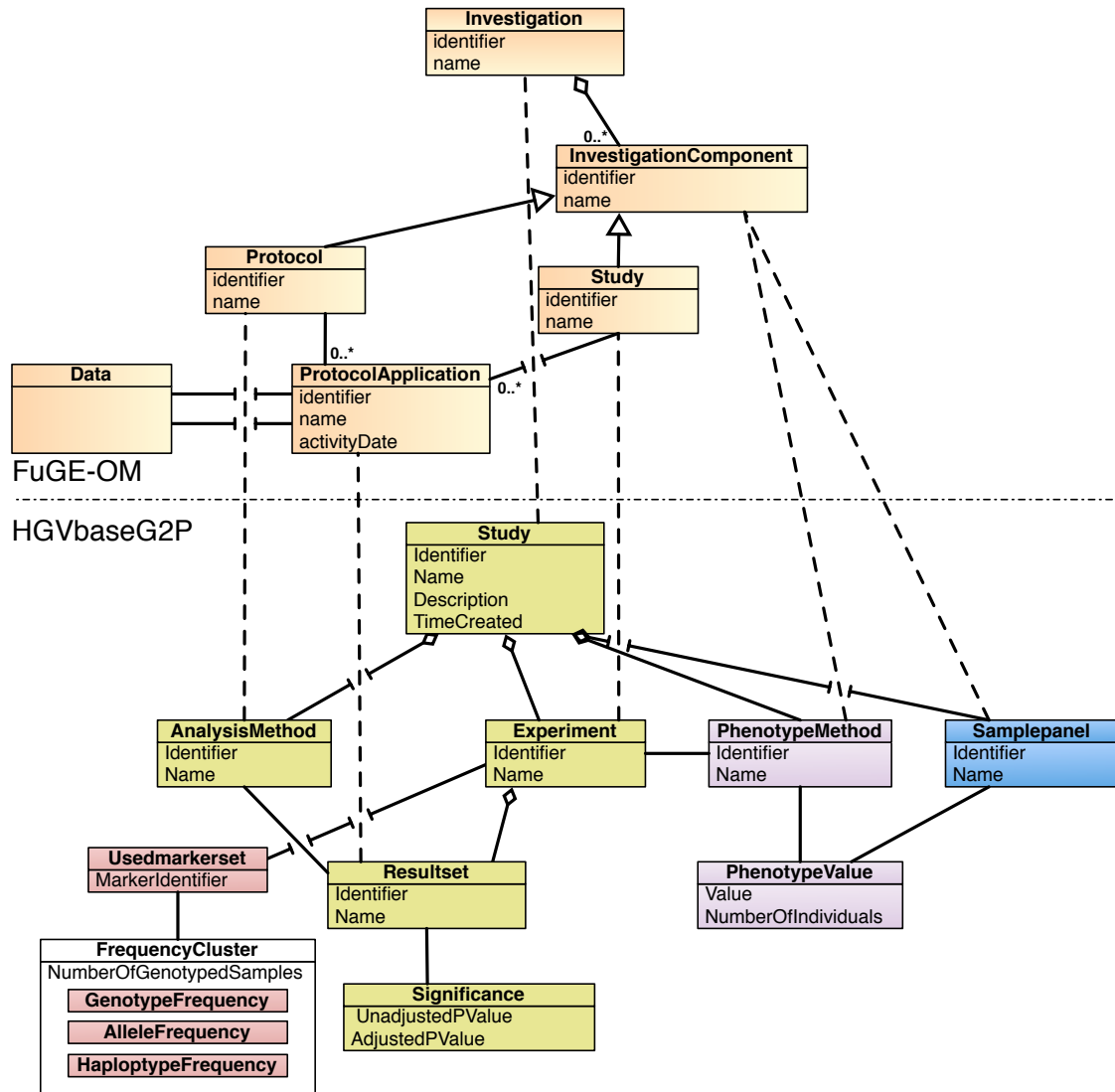


Fig. 4.1: Aligning the HGvbaseG2P model with FuGE-OM. The simplified representation of core FuGE-OM classes (top) is derived from original figures in the FuGE specification v1 (<http://fuge.sourceforge.net/dev/V1Final/FuGE-v1-SpecDoc.doc>). The HGvbaseG2P model (bottom) shows key classes previously discussed in §3.1.

high level. Notably, representing up to several instances of data analysis conducted within an *Experiment* relating to a particular phenotype question was deemed important, such as when multiple statistical analysis methods were used on the same input data, or when the same method was applied several times with different parameters each time. As noted in §3.1.5, the *Resultset* class and its association to *AnalysisMethod* represents such an instance of an analysis. As such, *Resultset* maps to the *FuGE::ProtocolApplication* class which, in connection with *FuGE::Protocol*, models in a general way the process of applying a protocol and capturing provenance information. FuGE-OM also provides additional classes for describing workflows in terms of inputs and outputs, such as experimental manipulation of physical materials (e.g. sample preparation) or data generation (e.g. genotyping, analysis).

In the HGVbaseG2P model, the useful protocol and provenance capabilities endowed by *Resultset* and *AnalysisMethod* only apply to the specific process of turning data into analysis results, and other parts of the model relating to data generation are currently lacking in such features. As noted in the previous chapter, there is no concept of a genotyping protocol or technology platform to describe in detail how a given genotype dataset was generated, nor is there a way to state that a set of aggregate genotype data within an experiment were generated together in a particular genotyping experiment. This complicates representation of scenarios where, for example, genotype data from multiple experiments are combined for meta-analysis. Although not an issue in current use of the model for describing mainstream GWASs, this nevertheless limits the utility of the model for handling more complex study designs in the future.

Similar limitations apply to phenotype data organisation. The HGVbaseG2P model does not provide a way to capture provenance, since *PhenotypeValues* can only be aggregated solely based on the *Samplepanel* they are associated with. Thus, it is not possible to specify that, for example, a set of blood pressure measurements were taken at 10 min intervals during a 1 hour visit to a clinic, and subsequently another set of measurements were taken during a separate visit the following week. Another issue relates to how phenotyping protocol information is captured. Presently, all information regarding how a given trait

is measured and the specification, or definition, of the corresponding variable (e.g. the unit, whether it is continuous or categorical, and more) is captured in the single, monolithic *PhenotypeMethod* class. This necessitates the creation of an instance of *PhenotypeMethod* even when no protocol information is available. A further side-effect of this design is that it is not possible to describe a protocol which measures multiple traits (e.g. a standard battery of biochemical blood tests).

4.2 Aligning with XGAP

In addition to the benefits to study metadata exchange outlined above, establishing mappings between the HGVbaseG2P model and XGAP-OM on the level of study data could be highly advantageous. This would facilitate, for example, gathering of G2P datasets from multiple systems into one place for integration, or bi-directional data exchange between the two or more federated systems. My analysis was therefore focused on these G2P domain-specific aspects of the two models.

Study organisation. Having emerged rather later on the scene than the HGVbaseG2P model, XGAP-OM developers have taken advantage of an existing domain standard and constructed the model as a set of extensions of several core elements of the more general FuGE-OM model. Because of this grounding in FuGE-OM, the cross-model mappings of high-level concepts from the previous section also apply here and thus need not be discussed.

A pragmatic approach to modelling G2P datasets. Despite close similarities at the organisational level and several common mappings to high-level FuGE-OM classes, the two models differ fundamentally in their approach to modelling study data. As illustrated in Fig. §4.2, XGAP-OM specifies a simple, generic representation of data elements as a two-dimensional data matrix which, like the higher-level organisation classes of the model, is based on generic FuGE-OM constructs. A pair of classes derived from

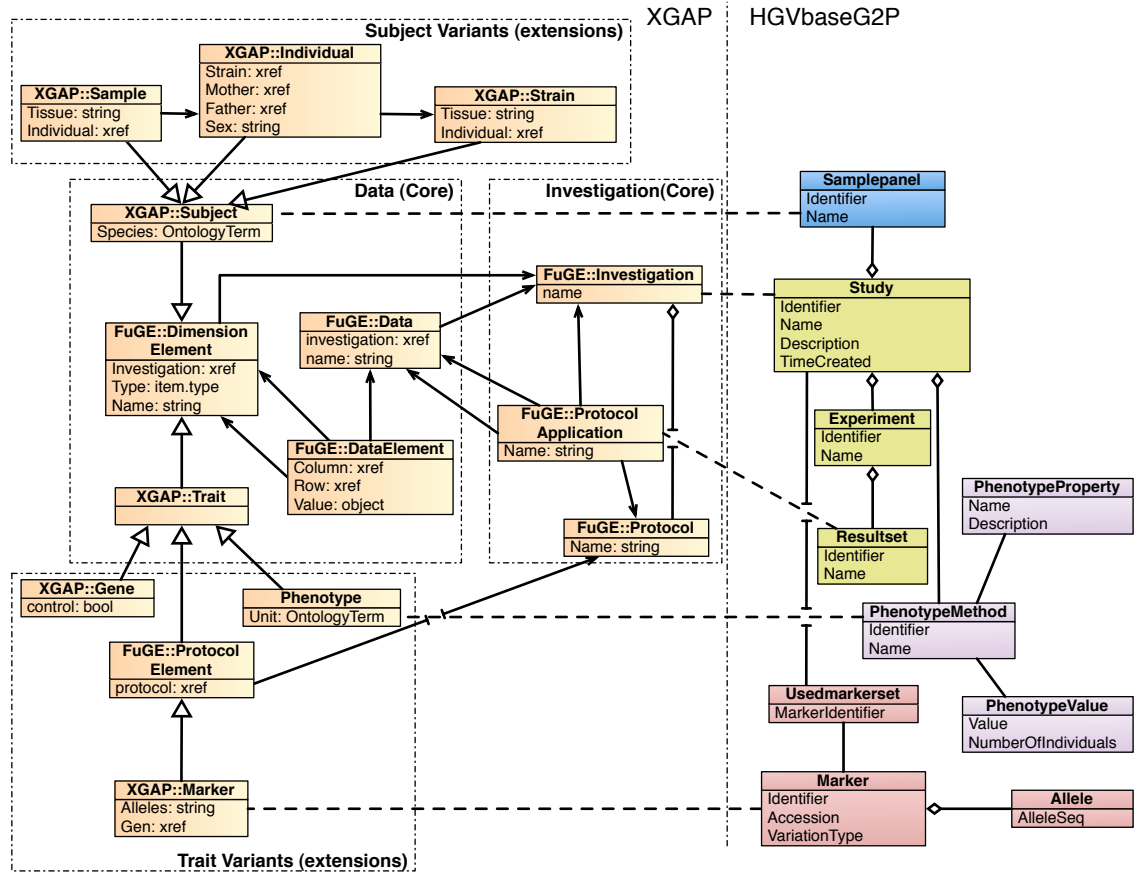


Fig. 4.2: Aligning the HGVBbaseG2P model with XGAP. The simplified representation of XGAP-OM (left) is derived from XGAP-OM model overview diagram (<http://www.xgap.org/wiki/XgapDataModel>) and XGAP-OM model documentation(<http://gbicserver1.biol.rug.nl:8080/xgap4exampledatasets/doc/objectmodel.html>). *XGAP::Investigation*, *itshapeXGAP::ProtocolApplication*, *itshapeXGAP::DataElement* and *itshapeXGAP::DimensionElement* extend *FuGE-OM* classes with the same names.

FuGE::DimensionElement - *XGAP::Subject* and *XGAP::Trait* - represent the subject of the measurements and the specific feature that was measured, respectively. These classes serve as pointers into the data matrix and facilitate annotation of individual data elements. For example, the data element “G/T” (genotype) can be annotated on one axis with rs1234 (SNP marker) and CEPH1341.02 (sample) on the other axis. This minimal core is then extended via subclassing to create custom datatypes, subjects and traits, in order to cater for different flavors of G2P experiments. This pragmatic modelling pattern, which

focuses on data and minimal annotation of data elements, contrasts with the HGVbaseG2P approach of specialised, unrelated classes to represent specific types of study data (e.g. *GenotypeFrequency*, *Significance*, *PhenotypeValue*).

Some parts of the minimal XGAP-OM model can nevertheless be aligned with the HGVbaseG2P model. *Samplepanel* aligns directly to *XGAP::Subject*, since the both are conceptually equivalent to *PaGE::Abstract_observation_target*. Similarly, *XGAP::Marker* and *XGAP::Phenotype* - both are a kind of trait in XGAP-OM - align with *Marker* and *PhenotypeMethod* in the HGVbaseG2P model, respectively. Where the minimal XGAP-OM model ends, however, the HGVbaseG2P model provides a much richer set of classes for describing markers, phenotypes and panels. For example, the minimal *XGAP::Phenotype* class does not provide a means to describe a “phenotype” in a fine-grained way in terms of its core sub-concepts, as discussed in §3.1.4.

Use of ontologies. Another contrasting feature of XGAP-OM compared to HGVbaseG2P is its extensive use of ontologies which, again, is based on generic facilities provided by FuGE-OM. Several of the XGAP-OM classes in Fig. §4.2 specify attribute values to be constrained to terms from a controlled vocabulary. When combined with a domain ontology or ontologies, this provides a way to standardise terminology where required, and to embed domain knowledge in data described with the model. Similar facilities are provided by PaGE-OM, whereas in the HGVbaseG2P model, controlled vocabularies are only used informally for two purposes; i) to specify variation type as SO terms, and ii) to annotate phenotype properties with terms from the Medical Subject Heading (MeSH) controlled vocabulary⁶ (see also Chapter 5).

Use of multiple inheritance. As with some aspects of PaGE-OM, XGAP-OM makes use of multiple inheritance to endow the *XGAP::Gene* class (subclass of *XGAP::Trait*) with properties of a *Locus* which can be located on a genomic sequence, for example. Multiple inheritance is also used liberally by various XGAP-OM classes to endow

⁶<http://www.nlm.nih.gov/mesh/>

them with standard FuGE-OM attributes derived from the *FuGE::Identifiable* and *FuGE::Describable* interface classes. These advanced modelling techniques have thus far been used sparingly in the HGVbaseG2P model, mainly through informal use of FuGE-derived attributes in several classes (see also Discussion and §5.1.2).

4.3 Pheno-OM: creating an improved object model for phenotypes

This section introduces a new phenotype model, provisionally named the GEN2PHEN Phenotype Object Model (Pheno-OM), which is a key outcome of collaborative modelling work involving our group, several other GEN2PHEN partners and non-partner collaborators from G2P domain. The primary objective of this work was to create a minimal, extensible object model for describing phenotypes from a range of sources and of varying levels of complexity, from minimal annotations in LSDBs to GWAS case/control investigations and rich clinical reports from longitudinal cohort studies. The resulting model draws on practical experiences with several domain models, including the HGVbaseG2P model, PaGE-OM and XGAP already discussed, and models from the GenomEUTwin project mentioned in Chapter 2, the OBiBa project and others.

My role in the creation of Pheno-OM was as follows. By participating in a number of GEN2PHEN-sponsored modelling workshops in 2008 and 2009, I contributed to the overall discussion and model design in early stages of development. Outside of these focused workshops, I also maintained a dialogue with GEN2PHEN partners who led Pheno-OM development (Morris Swertz, Juha Muiilu and Tomasz Adamusiak) regarding certain key features of the model.

A description of the Pheno-OM development process has been published online as GEN2PHEN deliverable report D3.5⁷. Additionally, detailed Pheno-OM documentation and logical diagrams are available online. This section will therefore not describe the full

⁷<http://www.gen2phen.org/document/d35-high-level-domain-model-version-2-samplephen>

model details, but instead focus on the overall organisation of the model and present a comparative analysis of several key features as these relate to the phenotype modules of the HGVbaseG2P and PaGE-OM models.

Reusing existing object models. On a technical level, a key characteristic of the Pheno-OM model is extensive incorporation of concepts from available domain models, much like the approach taken by XGAP as already discussed. Indeed, as is evident from class names in the logical diagram shown in Fig. §4.3, nearly all of the Pheno-OM classes originate from either FuGE-OM or PaGE-OM. This reflects an overall agreement on the core concepts defined by these standard reference models. The critical difference, however, lies in how several of the phenotype-related classes derived from PaGE-OM are connected. The impact of this reorganisation and overall model design will be explored in the sections to follow.

4.3.1 Describing protocols and provenance

Pheno-OM leverages the FuGE-derived classes *Pheno::Protocol* and *Pheno::ProtocolApplication* to describe how features are measured and the process of obtaining measurements, respectively. With regard to the limitations of the monolithic *PhenotypeMethod* class in the HGVbaseG2P model mentioned above, one of the benefits resulting from this design is that protocol information is factored out of *PhenotypeMethod* into its own class, thereby facilitating protocol reuse across studies. Also, a useful consequence of the way these two classes and *Pheno::ObservedValue* are associated (see Fig. §4.3) is that protocol information is not required. This is useful for applications such as LSDBs, where often only minimal structures for phenotype information are needed. In such cases, requiring the implementation of structures for holding protocol details, which then would be mostly populated by empty records, would seem an unnecessary complication.

Another benefit is that a single protocol can be associated with many features, thus

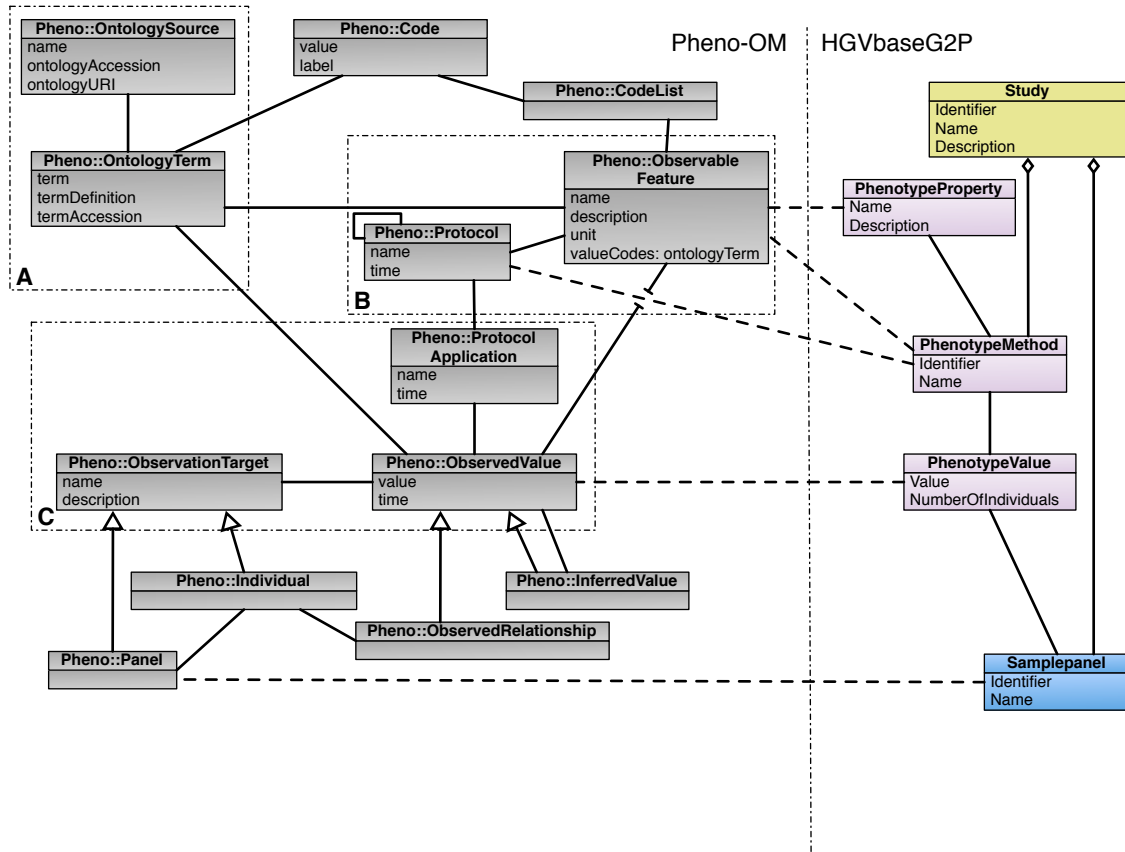


Fig. 4.3: Aligning the HGVbaseG2P model (right) with Pheno-OM (left). The simplified representation of Pheno-OM is derived from the current (as of May 2010) Pheno-OM specification (<http://www.gen2phen.org/system/files/private/GEN2PHEN%20Phenotype%20Model%202010-03-10.docx>). For simplicity the *Pheno::Investigation* class (with which most of the Pheno-OM classes shown are associated) is not shown.

addressing the example of biochemical blood tests above. The option of nesting protocols within protocol makes it possible to aggregate several protocols each measuring one or more features (e.g. series of eye exams) into a single super-protocol (e.g. clinical assessment of participants in a study of the AMD disease). Protocol nesting also enables straightforward modelling of a questionnaire composed of a series of questions and sub-questions (e.g. smoker yes/no? And if yes, for how long?).

Finally, the use of *Pheno::ProtocolApplication* also addresses the problem of aggregating and providing provenance for collections of *PhenotypeValues* in the HGVbaseG2P model.

Reflecting on the example from the previous section, in Pheno-OM a visit to the clinic would be modelled as an instance *Pheno::ProtocolApplication*, which captures the date, name of the nurse taking the measurements and other context, as well as a reference to the standard protocol used. The blood pressure measurements taken during this visit would be represented as series of *Pheno::ObservedValues* associated with the *Pheno::ProtocolApplication*. Measurements taken during the next visit will subsequently be associated with a different instance of *Pheno::ProtocolApplication* linked to the same protocol, thereby enabling observation data from the two sets to be distinguished.

4.3.2 Describing features

One of several important themes to emerge from the above modelling work was the importance of modelling phenotype features as variables. It soon became clear that, when it comes to the practicalities of dealing with phenotype data, the focus tends to be on the eventual statistical analysis that is to be undertaken, and therefore on the trait as a statistical variable.

PaGE-OM and FuGE-OM provide facilities for describing results from measurements with various kinds of values (e.g. string, boolean; see Fig. §B.12), where the unit and other attributes of the value are specified. However, these constructs have the following significant disadvantage; it is not possible to glean these important statistical attributes of the observed feature without the presence of actual data. In other words, when considered in isolation, this incomplete feature description does not allow objective comparison of phenotype measures to be made. For example, it would be impossible to determine whether two variables relating to body weight are equivalent, because the unit is not specified; data for one variable may be in pounds, whereas data for the other may be in kilograms.

These and related issues pertaining to statistical analysis highlight the need for a different way of modelling features in the new model. This led to the current design of *Pheno::ObservableFeature* class, which aligns to *PhenotypeProperty*, and also encapsulates the variable type properties of *PhenotypeMethod* as shown in Fig. §4.3.

An optional link to an instance of *Pheno::CodeList* allows a list of allowed values to be specified, thus addressing the issue raised in §3.2.4 regarding coded phenotype data. Finally, an association with *Pheno::OntologyTerm* provides a mechanism for annotating a feature with a term from a controlled vocabulary. A crucial advantage of this is an effective decoupling of the operational definition of a feature in the statistical sense on one hand, and the logical definition on the other. For example, the length of finger in cm versus the question of what is a finger, and how the finger entity relates to other anatomical parts.

4.3.3 Describing the outcome of observations

Since attributes relating to variable specification are declared at the level of the observable feature itself, actual results from observations are modelled in a far simpler way in Pheno-OM compared to PaGE-OM. The *Pheno::ObservedValue* class describes the observation result simply as value (with an optional timestamp), with associations to the classes discussed above providing the necessary context. Critically, Pheno-OM adds an optional reference to an ontology, which enables observation values to be specified as terms from a controlled vocabulary. This provides the means for standardising phenotype descriptions, either on a small scale via locally-defined vocabularies of standard terms, or more broadly via the use of concepts from rich, external domain ontologies.

Pheno-OM also provides a means to specify inferred observation values which are derived from one or more primary values. *Pheno::InferredValue* resembles the hierarchical, value-within-value constructs in PaGE-OM (see Fig. §B.8) and addresses many of the same modelling problems, but in a simpler way. For example, in the commonly-used example of blood pressure, it is straightforward to represent systolic and diastolic blood pressure measurements as a pair of *Pheno::ObservedValues*, and subsequently add a *Pheno::InferredValue* linked to the first two to represent the inferred blood pressure status of the individual (“normal” vs “high”).

The same method can be used to represent descriptive summary statistics of multitudes of individual datapoints. For example, a *Pheno::ProtocolApplication* can be used to aggregate

a series of primary weight measurements, each linked to the *Pheno::Individual* observed. An instance of *Pheno::InferredValue* can then hold the mean, median, standard deviation and other descriptive statistics of interest across the primary observation data. Or, as in the dbGaP example from §3.2.4, summary statistics could be stored by themselves, in the absence of individual level data, with a reference to a *Pheno::Panel* to represent the group of individuals. It is unclear, however, what the best way is to model example data as dbGaP provides for continuous variables.

4.4 Discussion

This chapter presented results from conceptual modelling work undertaken as a natural follow-on to the work reported in the previous chapter. In an effort to align my work to create the HGVbaseG2P model with other modelling work in the G2P domain, I compared it with FuGE-OM and XGAP-OM and identified key conceptual commonalities and differences between these related domain models. I also contributed significantly to the creation of a new, minimal object model for describing phenotypes which I have described above and aligned with the HGVbaseG2P model. The results from these two main activities are summarised and discussed below.

4.4.1 HGVbaseG2P versus FuGE-OM and XGAP

Study organisation and provenance. The comparison with FuGE-OM confirmed that the general high-level organisation of the HGVbaseG2P model is broadly consistent with models emerging from “omics” standardisation activities. This is not surprising, as the traditional journal paper template shared by many of these models has proved a useful one. A natural next step would be to aim for increased compatibility with the generic FuGE-OM model and investigate to which extent class attributes can be aligned with and/or adapted in the HGVbaseG2P model. This is already done in an informal way in the HGVbaseG2P implementation (see next chapter), and several other FuGE-OM constructs could be used

in a similar way to further enhance and standardise the HGVbaseG2P model. As argued in the previous chapter, model formalisation is a prerequisite for effective work in this area. The second issue pertains to whether or not additional support for provenance is needed in the HGVbaseG2P model. Though adequate for current use in the HGVbaseG2P project, the partial provenance support in the model is not sufficient if more complex study designs and more detailed phenotype data are to be represented. For the three categories of study data in HGVbaseG2P (genotype and allele frequencies, phenotype observations and outcomes from statistical analysis), only the last of these supports nesting and capture of provenance below the level of *Experiment*. This imbalance means that statistical methods employed in the study and the outputs from multiple analyses can be described in some detail, but by contrast the input dataset or datasets for said analyses can be described in only a very limited way. One way to address this is to employ the same provenance model across all three categories of data, so genotype and phenotype data would be nested under *Resultset*-like containers. Alternatively, a better strategy may be to instead adopt a single, core model for representing data, as discussed below.

Experimental data and analysis results. Another issue concerns the representation of *data*.⁸ The creators of XGAP-OM have approached this problem from an entirely different angle compared to the HGVbaseG2P and PaGE-OM models. The core representation of datapoints as elements in a two-dimensional matrix is the heart of the model. Interestingly, XGAP-OM/XGAP developers have been able to take several generic, highly-abstract (and, in my view, challenging to understand) FuGE-OM constructs and extend them in a way that makes the resulting derived model more suitable for G2P information, and far easier to comprehend. The simple abstractions, combined with generic, Molgenis-based data loading tools, are the basis of a data infrastructure into which any given tabular G2P dataset can be loaded with few or no modifications. Importantly, the base model is highly extensible, in that new categories of G2P information can be added without altering how

⁸I use the term “data” loosely here to refer to both raw experimental data (e.g. SNP genotypes or mass spectrometry measurements) and analysis results (e.g. p-values and other metrics from statistical analysis).

the actual data are represented in the core model and, consequently, manipulated in the software implementation.

This is quite different from the HGVbaseG2P model which specifies three categories of G2P data, each of which has its own distinct representations. To illustrate this, consider a panel of 56 individuals for which we have the following information: a relative frequency of 0.20 for the T allele of SNP rs49234 and an average height with a value of 156.3cm. In the HGVbaseG2P model, the two numeric values share nothing conceptually and are stored in completely different parts of the model. In the data-centric XGAP-OM model, on the other hand, both values would be represented by an instance of the same *XGAP::DataElement* class. The data elements would be annotated with the same subclass of *XGAP::Subject* on one axis (the group of individuals) but with different subclasses of *XGAP::Trait* on the other axis (average weight, allele frequency). Adding a new datatype would therefore merely require a new subclass to be created, whereas the HGVbaseG2P model this would in effect require an entirely new set of constructs, at least if new datatypes do not fit neatly into the three existing categories.

My conclusion from the above is that the HGVbaseG2P model does not lend itself to extending to handle new datatypes beyond its GWAS-centred scope. This is arguably the result of the singular focus of the HGVbaseG2P modelling work on describing GWAS investigations at the summary level, with relatively little thought given to broader use of the model for describing other kinds of G2P investigations. As noted in the previous chapter, in its current use the model fulfils this requirement adequately. Conversely, XGAP-OM and the XGAP system were designed from the ground up as a general solution for practical G2P-focused data management and analysis needs of bench scientists. The general modelling philosophy and technical approach employed thus better lends itself to creating extensible models.

Model “richness” for annotating data. Notwithstanding the above strengths of XGAP-OM for modelling data and provenance in a simple, extensible way, some other aspects of the model design are overly simple. The “Trait” part of XGAP-OM (see lower

third of Fig. §4.2) for annotating data elements is not rich enough to, for example, capture genetic variation information to the level of detail required in the HGVbaseG2P system. The GENOTYPE and SEQUENCE domains of HGVbaseG2P and PaGE-OM, by comparison, offer far richer constructs for capturing this information. Similarly, the single *XGAP::Phenotype* class is far less expressive than the multiple phenotype-related classes provided in the other two models. It is thus evident that the two sets of models each have their own strengths, and indeed complement one another in a number of ways. One way to exploit this would be for XGAP-OM to incorporate the richer GENOTYPE and SEQUENCE classes from PaGE-OM, possibly as an optional add-on module. The Pheno-OM model discussed below could perhaps be used as such a modular extension to XGAP-OM.

4.4.2 Pheno-OM and “pluggable” G2P information models

Though not part of my original workplan, the involvement of our group in the GEN2PHEN project provided a valuable opportunity for me to tackle the difficult challenge of modelling phenotypes in collaboration with several partners. By way of input from our group and other stakeholders, the resulting Pheno-OM design incorporates the salient features of each of the domain models considered during the design process. The new model also addresses several key issues with some of these models, including issues highlighted in this and the previous chapter. With minor deviations, the high-level organisation of Pheno-OM is equivalent to XGAP-OM and FuGE-OM, and so the points made above regarding general study organisation and provenance need not be repeated here. Several specific aspects of Pheno-OM merit further discussion, however, especially as this relates to the other models considered in this chapter and future development of the model in the context of a unified G2P modelling framework.

Modelling the elusive “phenotype” concept. The design of Pheno-OM reflects a general conclusion reached by several groups independently: namely, that unambiguously

specifying i) which trait was observed, ii) how the trait was observed and iii) the outcome of the observation, is crucial to interpretation and cross-study comparison of observational data. The proposed tripartite model of a “phenotype” in Pheno-OM is based on the same conceptual elements as used in the HGVbaseG2P model and PaGE-OM. Crucially, in Pheno-OM these elements are connected to one another in a slightly different way which addresses a range of issues in practical use, in particular concerning protocol reuse across studies and handling absence of protocol information, protocols which measure multiple traits, and provenance of observational data.

Pheno-OM also provides a simple, elegant method for dealing with the potentially very complex nature of the “trait” part of the tripartite model. A single entity encapsulates the abstract concept of a trait and, optionally, the operational definition of the trait as a variable. This design facilitates the storage and manipulation of observation data in a relatively simple, lightweight core framework, which may well be sufficient for many applications on its own.

However, it is the built-in support for semantics which is the key to the extensibility and flexibility of the model. The design enables the complex task of describing biological concepts to be delegated, or “outsourced”, to domain ontologies, by allowing features and values to be annotated by ontology terms. Similar ontology-driven data typing is at the core of the Chado database schema and is the key to interoperability between various GMOD software components (Mungall and Emmert, 2007). Other related work includes ontology-based representation of study variable metadata in the Data Schema and Harmonization Platform for Epidemiological Research project (DataSHaPER)⁹ Fortier *et al.* (2010).

Finally, Pheno-OM addresses what I consider to be a flaw in the way observations, or measurements, are represented in both FuGE-OM and PaGE-OM. Both reference models subscribe to the notion that individual values, or datapoints, from observations are to be richly annotated with respect to data type and semantics, and associated with a “slim” representation of the variable or variables being studied (or study factors, in FuGE-OM

⁹<http://www.datashaper.org>

terminology). This ostensibly overlooks the importance of associating such crucial details with the variable itself in the form of metadata, which is the approach taken in Pheno-OM (and also HGVbaseG2P). As a result, two instances of *Pheno::ObservableFeature* from different studies can be objectively compared in the *absence* of actual data, whereas this is difficult or impossible to do for the under-specified *FuGE::Factor* and *PaGE::Observed_value*. Such comparison is important in scenarios where, for example, only study metadata are publicly available and access to primary study data requires special privileges.

Future work: mixing and matching object models. Considering the above, the current version of Pheno-OM is relatively complete, and could in principle be used as is a basis for a simple application for managing phenotype data. Such an application has indeed been created by GEN2PHEN partners and is being used for model validation (see also below). This Molgenis-based reference implementation is publicly accessible at <http://wwwdev.ebi.ac.uk/microarray-srv/pheno/>. However, further work is needed to realise the full potential of the model. Increasing interoperability with other models will be particularly important, as this relates to the GEN2PHEN strategy of creating multiple small domain models which can be combined with one another to build standards-based, flexible G2P data infrastructure. A comparison between Pheno-OM and XGAP-OM illustrates this. The central backbone of Pheno-OM - *Pheno::ObservableFeature*, *Pheno::ObservedValue* and *Pheno::ObservationTarget* - maps directly to the XGAP-OM core constructs *XGAP::Trait*, *XGAP::DataElement* and *XGAP::Subject*, respectively. The Pheno-OM constructs can in fact be treated as a “drop-in”, richer replacement for the simpler XGAP-OM classes. The resulting composite G2P model would thus be endowed with the generic data management facilities of XGAP-OM on one hand, and the rich Pheno-OM facilities for describing observations on the other. Combining XGAP-OM and Pheno-OM (and possibly other models) in this way would not be an entirely smooth operation at present. One reason for this is slight inconsistencies in logical connections between higher-level organisational constructs. For example, observation values in Pheno-OM

are connected directly to the *Pheno::Investigation* class, and therefore values cannot be aggregated into discrete data collections (this is what *XGAP::Data* is for). However, this particular issue and others like it could be remedied by packaging and publishing the core of Pheno-OM as a standalone module, explicitly intended for use as a “bolt-on” extension to XGAP-OM or similar minimal model. This will probably require further coordination of the modelling initiatives involved, and perhaps the creation of a new core model in the XGAP-OM vein, specifically designed to be used with such extensions to add specific types of functionality. As noted in the previous subsection, the GENOTYPE and SEQUENCE domains of PaGE-OM could conceivably also be used in this way. Alternatively, a single monolithic model could be devised which merges all the features of the various models, but this would negate the advantages of smaller, more nimble object models when it comes to actually implementing software infrastructure.

Whatever the strategy followed, an important next step will be to formalise the relationship of Pheno-OM to other domain models. The current Pheno-OM specification (in the Molgenis domain-specific language) indicates class mappings to FuGE-OM, PaGE-OM and XGAP-OM, but in free-text comments only which has a number of disadvantages as noted in the previous chapter. Ideally, Pheno-OM should be published as a formal UML model, where such links to external, published models are stated unambiguously (e.g. *Pheno::Investigation* extends *FuGE::Investigation*). Work in this direction is already being undertaken by GEN2PHEN partners, who have devised software tools and processes for translating UML models published in the standard XML Metadata Interchange (XMI) format into the Molgenis DSL (Swertz *et al.*, 2010). Also, it may be beneficial in this context to take note of the guidelines for extending FuGE-OM provided by Jones *et al.* (2009) and investigate whether Pheno-OM (and possibly other GEN2PHEN models) should perhaps be published as formal FuGE-OM extensions, much like has been done for FuGEFlow¹⁰ for flow cytometry (Qian *et al.*, 2009) and several other standard models and formats in recent years.

¹⁰<http://flowcyt.sourceforge.net/fugeflow/>

5. Creating a federated database of genetic association studies

Several issues were raised in Chapter 2 concerning accessibility of data generated in genetic association studies, and the challenges faced by researchers who need to locate, compare, integrate and synthesise all available association study data for genes, genomic regions or diseases of interest. High-level G2P knowledgebases and GWAS catalogs provide a broad view of published GWAS findings in the scholarly literature. But due to the limited data resolution and bias towards “publishable” study findings, these resources do not provide a fully detailed and comprehensive comparison. By contrast, central G2P archives contain fully-detailed, primary study data, but association datasets are fragmented across several of these unconnected archives. This situation is further exacerbated by stringent access controls on identifiable individual-level data, dataset size, data complexity and other factors, all of which act as barriers to effective data integration and exploitation.

The HGVbaseG2P project, as originally envisioned, aimed to address these urgent issues as follows. First, via a combination of active data gathering and data submissions from the community, HGVbaseG2P would provide a more comprehensive, unified and unbiased collection of aggregate, non-identifiable experimental data and findings from published and unpublished human genetic association studies. Second, to facilitate optimal integration and use of this information, HGVbaseG2P would provide a suite of web-based software tools for searching, browsing, visualisation and mining of database content.

The topic of this chapter concerns a team effort to build the informatics infrastructure and undertake data gathering to achieve these goals. My role in this work was that of chief architect and creator of initial versions of the HGVbaseG2P system, and as overall coordinator and project lead during the first half of my PhD. In later stages of the project, day-to-day HGVbaseG2P development and other operations have been mostly undertaken

and coordinated by other bioinformatics developers who have since joined our group. In an attempt to clarify my contributions to this collaborative work, I will indicate key roles of group members as appropriate in the sections to follow.

The chapter starts with an in-depth technical description of the system, its architecture, database and software components and web-based tools. This will be followed by a series of use case scenarios which demonstrate key features of the system. Next, strategies for gathering data for the HGVbaseG2P study catalog during the course of the project will be discussed. The final section summarises the chapter, discusses some of the limitations of the system and suggests areas of future work.

5.1 Constructing the HGVbaseG2P system

Although system requirements have not remained constant through the lifetime of the HGVbaseG2P project, the following considerations were deemed most important during its design and implementation:

Standards. The database and software would be grounded in the conceptual models presented in the previous chapter, with a long-term aim of making HGVbaseG2P compatible with other systems based on common domain G2P information models.

Types of data supported. Initial emphasis would be placed on gathering data from SNP-based genome-wide case-control association studies which were becoming available at the start of the project. The key rationale for this was the relative ease of access to these useful data, compared to the many challenges that would be faced in gathering thousands of pre-GWAS candidate gene association studies. However, the data model and software would ultimately support all common association study designs and data outputs.

Website functionality. The HGVbaseG2P website would need to provide all the customary data browsing and searching tools users now expect from a modern online

biological database, for easy access to all database content. Graphical genome views of association study findings would play a central role, with a special emphasis on customisable aligned views of many study findings instead of limiting users to viewing only the significant associations reported in the original publication, one study at a time.

Programmatic access. The system would be made fully web service-enabled for programmatic access. Workflow construction tools such as Taverna were considered an initial primary target client of data retrieval services, with a move towards more general web service-based grid capabilities in the future.

Software development philosophy. HGVbaseG2P would be constructed as a modular, extensible system according to good software engineering practices (e.g. object-oriented programming and conceptual modelling, as discussed in the previous two chapters). Existing open-source software would be re-used and adapted wherever possible, both for pragmatic reasons (to save development effort and increase overall quality and standardisation of the system) and for philosophical reasons¹.

5.1.1 System overview

The HGVbaseG2P system is designed in highly modular fashion to ensure that the system can be easily maintained, extended and tested. The various system components are organised into a three-tier, layered architecture as shown in Fig. §5.1. Such a tiered architecture is often used for database-backed software systems, with the major benefit that high-level applications are effectively shielded by the middleware from database implementation details and changes in data storage over time. A modular, layered architecture is certainly a great deal more complicated than simpler web scripts and direct database access methods. But in the long run, advanced architecture such as this is invaluable and indeed essential for simplifying development of complex systems, and

¹I have long been a strong advocate of open-source, collaborative software development.

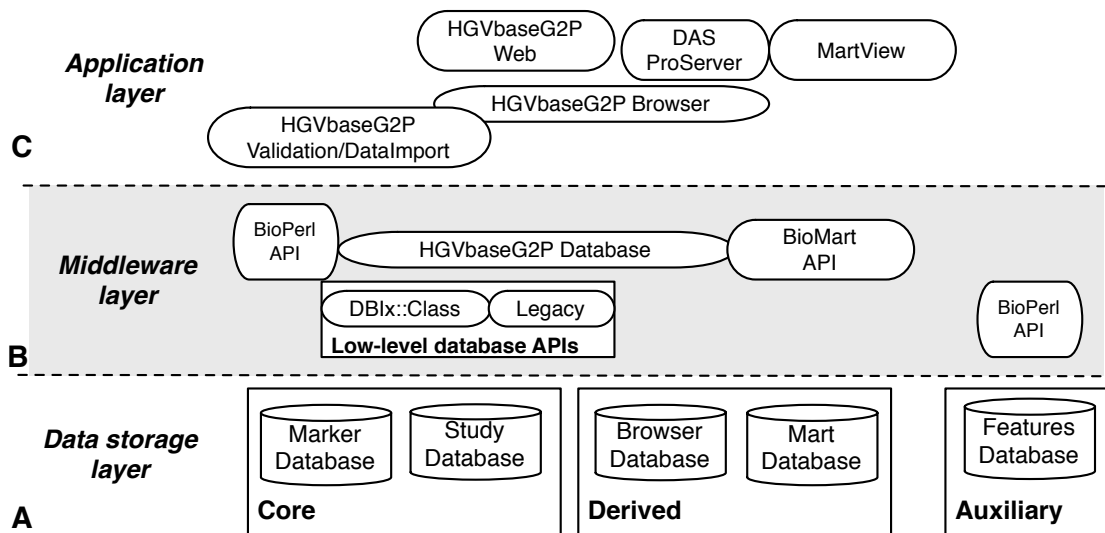


Fig. 5.1: An overview of the HGVbaseG2P system. The top-level application layer (C) contains all logic pertaining to data processing, loading, web-based access and other applications. All data querying, retrieval and storage operations are handled by the middleware layer (B) which interfaces with the databases in the storage layer (A).

making them easier to maintain and extend.

Based on project requirements and on my own experience from previous work in the field, I judged that the benefits from the advanced architecture would be worth the extra up-front design and development effort (compared to a more bespoke, less sophisticated approach). The HGVbaseG2P software components are a mixture of custom-built and freely available, open-source software. A summary of key third-party tools and software libraries used to construct the system and other technical specifications can be found in §A.1. These software components, the underlying databases and other key parts of the system are described in the subsections to follow, with full sourcecode and further documentation provided on the accompanying DVD (see §D).

5.1.2 Database design and content

HGVbaseG2P data storage comprises three types of relational databases (Fig. §5.1A). The primary Marker and Study databases contain core sequence variation data and association

study data, respectively. The derived Browser and Mart databases are generated from the core databases and contain simplified, de-normalised forms (i.e. allowing redundant storage of certain data to enhance query performance) of the core data, as required by certain software components. The auxiliary databases contain supporting data acquired from third-party sources. Table §5.1 lists the five databases which currently underpin the system. All databases were implemented in conventional fashion using the MySQL relational database management system, though in principle any of the major RDMS systems (e.g. PostgreSQL², Oracle³) could be deployed for this purpose.

Table 5.1: Summary of databases used by the HGVbaseG2P system.

Database	Used by	Database contents
Study	Core system	Metadata describing studies in the HGVbaseG2P catalog, and genotyping data and association analysis results from those studies.
Marker	Core system	Basal, summary information on genomic variation in the HGVbaseG2P catalog.
Browser	Genome browser	De-normalised data from the Study and Marker databases and pre-calculated aggregate data on number of significant markers per genomic interval.
Features	Genome browser	Known genes, repeats and other genome annotations from external data sources
Mart	HGVmart	De-normalised data from the Study and Marker database.

Technical presentation. All relational model diagrams shown below were created with the open-source MySQLWorkbench application⁴. All table columns are displayed, but not indexes or other relational schema details (full relational definitions can be found on the DVD, see §D). Many-to-one cardinality of entity relationships is indicated with the infinity symbol (∞). As in the previous chapter, names of classes in the object model are italicised, whereas table and column names are rendered in fixed-width font (e.g. `Study`).

²<http://www.postgresql.org>

³<http://www.oracle.com>

⁴<http://dev.mysql.com/downloads/workbench/>

Implementing the object model. As an extension of my role as chief designer and maintainer of the conceptual model described in Chapter 3, I created core relational databases based on the object model according to the following general principles. Each class in the object model is represented by a table with the same name in the relational schema, except where practical considerations (e.g. performance, complexity) demanded otherwise. Table relations represent associations between classes, with linking tables added as required to implement many-to-many associations. Table column names match object class attributes, and extra columns are added to several tables to support implementation-specific logic. For columns holding the same type of information, such as object identifiers or free-text description, uniform column names are used across the database (e.g. `Identifier`, `Description`). In most cases, these uniform column names are derived from class attributes adopted from FuGE-OM and this arrangement thus provides, in an informal way, a certain level of compatibility with the high-level FuGE-OM model discussed in §4.1.

Object identifiers and names. Several types of data object in the core databases require a stable, unique identifier which is assigned by HGVbaseG2P. These object IDs are stored in the `Identifier` column of the respective tables and are formatted as an alphanumeric string. It is good practice to make database IDs as opaque as possible, to avoid the “brittleness” that comes with semantic overloading - that is, the inclusion of semantic information into the identifier (directly or inadvertently) or if opportunity is created for people to “project” semantic meaning onto the identifier. Nevertheless, I deemed it important to embed a small amount of semantics in these HGVbaseG2P-assigned IDs. Each ID is prefixed with HGV to clearly indicate HGVbaseG2P as the source of the database entry (this is commonly known as “branding”). This is followed by a 1-2 letter indicator of the object type (e.g. marker, study) and finally an incrementing number which is unique across all objects of that type. For example, the ID HGV`M12345` refers to a *Marker* object, whereas the ID HGV`ST54` refers to a *Study* object.

Certain data objects are also assigned human-readable names which are only required to be

unique in a more restricted scope, rather than across all objects of that type in the database. For example, the *Assayedpanel* object HGVAP1675 has the short, convenient name “Italian cohort” which only needs to be unique within the parent study HGVST375, and therefore *Assayedpanels* in other studies can use the same short name without conflict.

The Marker database. This database holds core marker information and genome mappings represented by classes from the SEQUENCE part of the conceptual model. The relational structure can be considered as two separate, distinct parts born out of two different design approaches. The first part, shown in Fig. §5.2 follows the table-per-class convention, with the one significant exception that *Alleles* for a *Marker* are collapsed into a text field in the *Marker* table (e.g. `` (A) : (T) '' for a biallelic SNP), encoded according to the published HGVbaseG2P nomenclature rules⁵. Earlier designs utilised a separate *Allele* table linked to the *Marker* table for holding allele information, but for several reasons (notably the physical split between the Marker and Study databases) this was deemed a more pragmatic solution to accomplish the same goal.

The second part of the Marker database schema is a set of standard tables specified by the `Bio::DB::SeqFeature::Store` package from the BioPerl collection (Stajich *et al.*, 2002) (see Table §A.1 and also below). These tables hold genomic coordinates and related information for markers, modelled as feature annotations on the reference and other genome sequence assemblies. This standard schema and accompanying mature software tools are optimised for common queries over genome annotation data (e.g. retrieving features in a given genomic interval, retrieving feature by name). Therefore, choosing this approach over the class-per-table design provides a standardised, optimised mechanism for storing and retrieving feature information.

The relational schema for the feature database tables is shown in Fig. §5.3. The Marker database contains one set of these standard tables per assembly, each set distinguished by a table name prefix. This arrangement facilitates straightforward storage of marker mappings to multiple assemblies (such as the alternative Celera assembly) in addition to

⁵http://www.hgvbaseg2p.org/docs/hgvbaseg2p_nomenclature_system.pdf

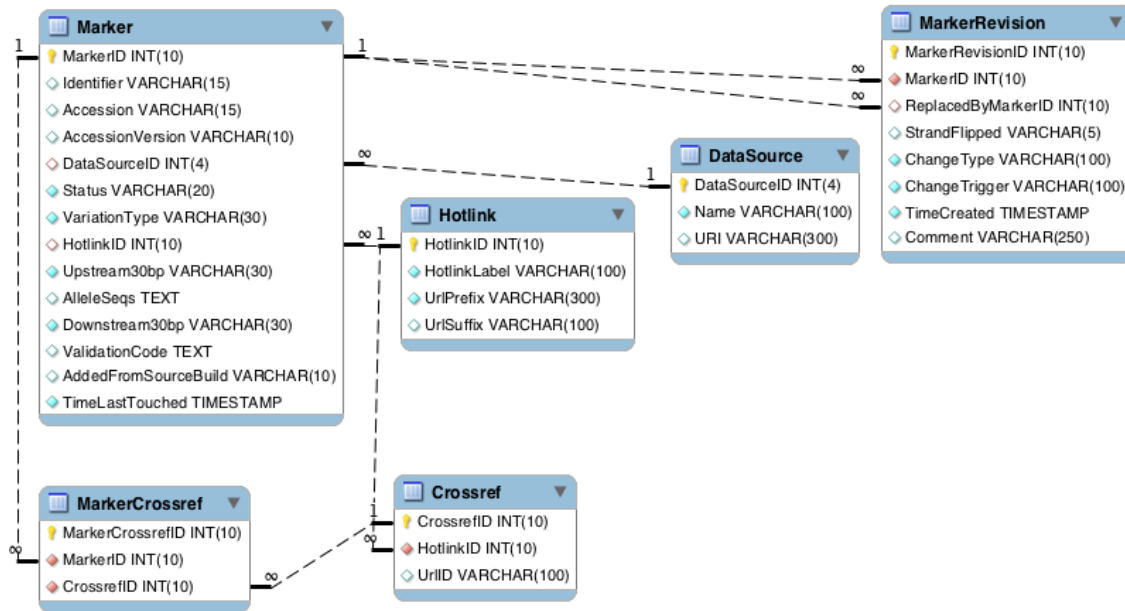


Fig. 5.2: Relational implementation of a part of the GENOTYPE domain of the object model. The MarkerRevision table tracks changes to marker entries in the source database between releases. Other tables map directly to classes in the object model

the NCBI reference sequence. The `reference_feature` table stores the coordinates of instances of the *Marker* class to a given sequence assembly which is stored in `reference_sequence`. The `reference_attribute` and supporting tables provide generic facilities for attaching feature annotations as attribute-value pairs. These are used to store a copy of a subset of core marker information, for convenient access via the BioPerl software described below.

The Study database. With few exceptions, the Study database is implemented according to the table-per-class convention. The net effect of this strategy is that nearly all of the relational model is identical to the conceptual model as presented in the previous chapter. For this reason, only significant deviations from the table-per-class convention will be discussed here. ER-diagrams illustrating other database tables, table columns and table relations are shown in Figs. §5.5 through §5.10 at the end of this subsection.

Fig. §5.4 shows tables which store aggregate genotype information described by the

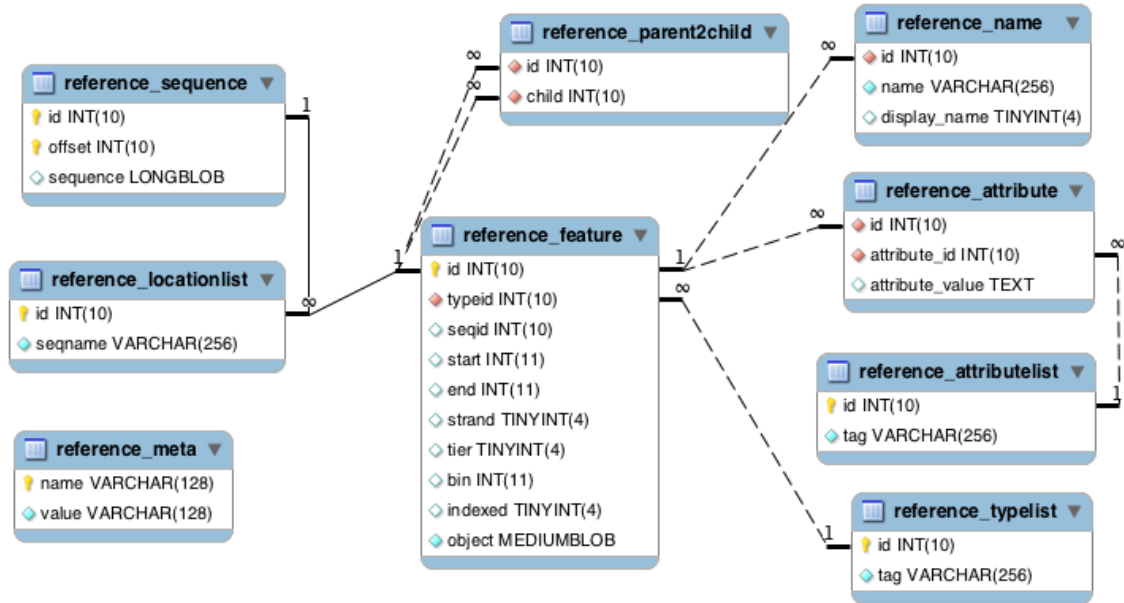


Fig. 5.3: The Bio::DB::SeqFeature::Store relational schema implementing the SEQUENCE domain of the object model. Only tables for reference sequence assembly mappings are shown.

GENOTYPE domain of the conceptual model, with the exception of the *Marker* and *Allele* classes already discussed. This part of the schema exhibits a substantial break from the straightforward table-per-class convention. Rather than implementing the *AlleleCombo*, *GenotypeCombo*, *HaplotypeCombo*, *Haplotype*, *Genotype* and *Allele* classes as tables with relations to the corresponding tables in the Marker database, instead these constructs are encoded as strings according to the nomenclature mentioned above, and stored in a text field in the corresponding frequency table. Compared to earlier implementations of the database done with the former approach, the latter method substantially reduces the number of tables required. This greatly simplifies database queries and software logic dealing with frequency information, while retaining the necessary querying ability.

One drawback of this scheme is that links from allele frequency data to marker alleles are no longer enforced on the relational level (e.g. alleles can be deleted without violating a foreign-key constraint). However, this is not considered a problem, as the software described in the next section implements numerous routines for validating these links

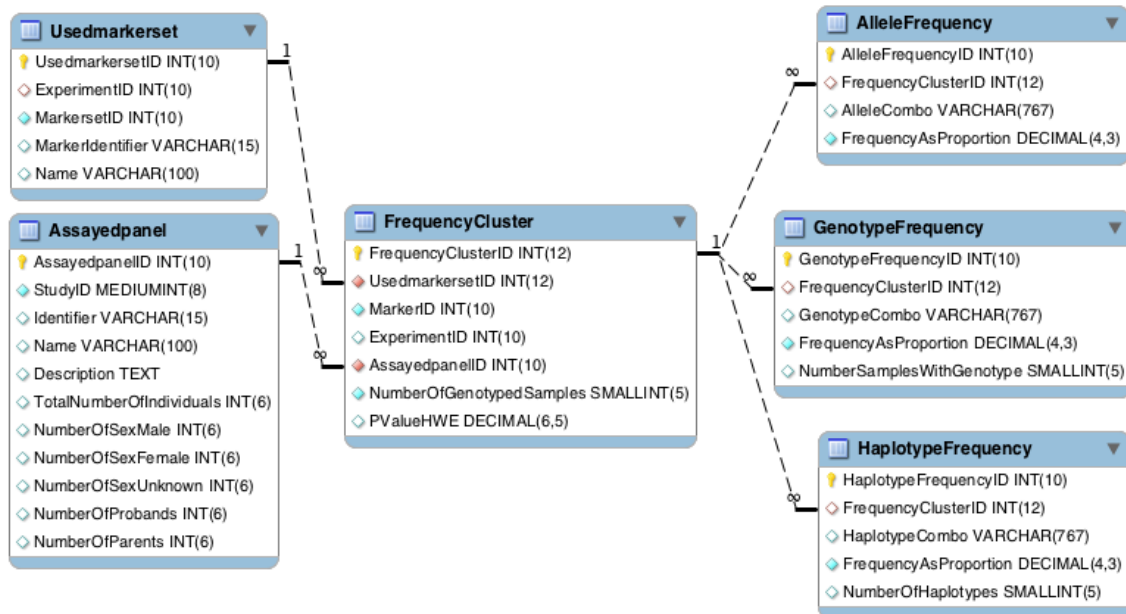


Fig. 5.4: Relational implementation of the GENOTYPE domain of the conceptual model. The *Usedmarkerset* table stores the identifier to the corresponding entry in the *Marker* table as a plain string, rather than as a true foreign-key reference. The link to the *Marker* database is handled in software logic.

between the *Marker* database and the *Study* database. Another drawback is that custom code is required to encode/decode the combo constructs (see the next section for details). But the advantages gained from the streamlined schema design outweigh both of these drawbacks. Overall, the lesson learned here is that slavishly adhering to the simple table-per-class approach when creating implementations of conceptual models will frequently lead to overly complex and inefficient relational schemas and/or software.

The Browser database. The denormalised structure of the Browser database (see Fig. §5.11) is specifically designed to facilitate a very restricted set of queries over marker and study data. These queries drive the web-based genome browser tools described below, and it is thus essential that these queries execute very fast, even as the number of entries in the tables increases to millions or tens of millions. To meet these requirements, Rob Free (who undertook this work) deemed it necessary to create such an *ad hoc*, non-standard schema,

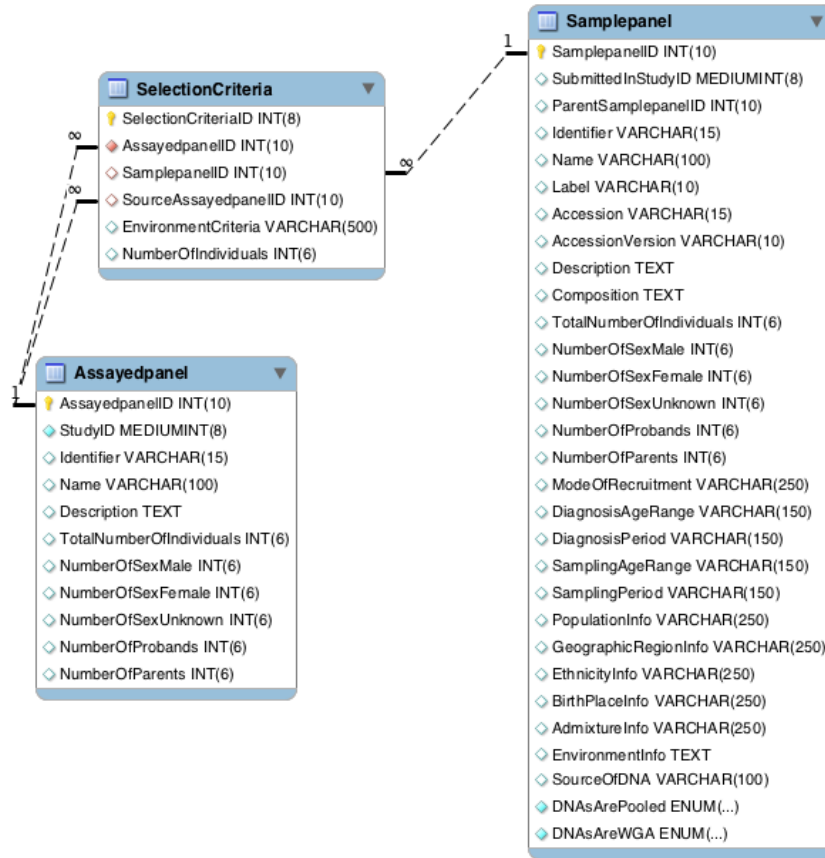


Fig. 5.5: Relational implementation of the SAMPLE domain of the conceptual model.

rather than deploying a standard feature database as described above.

The Features database. This is a standard BioPerl feature database containing publicly-available genome annotation data acquired from external sources. The relational structure of this database is identical to that shown for the feature part of Marker database.

The Mart database. The relational schema of the Mart database is specially designed to work with the standard BioMart data mining system⁶ (Smedley *et al.*, 2009). BioMart is built around a reverse-star schema optimised for rapid queries over large datasets, as described in (Kasprzyk *et al.*, 2004). The HGVbaseG2P Mart database schema was built

⁶<http://www.biomart.org>

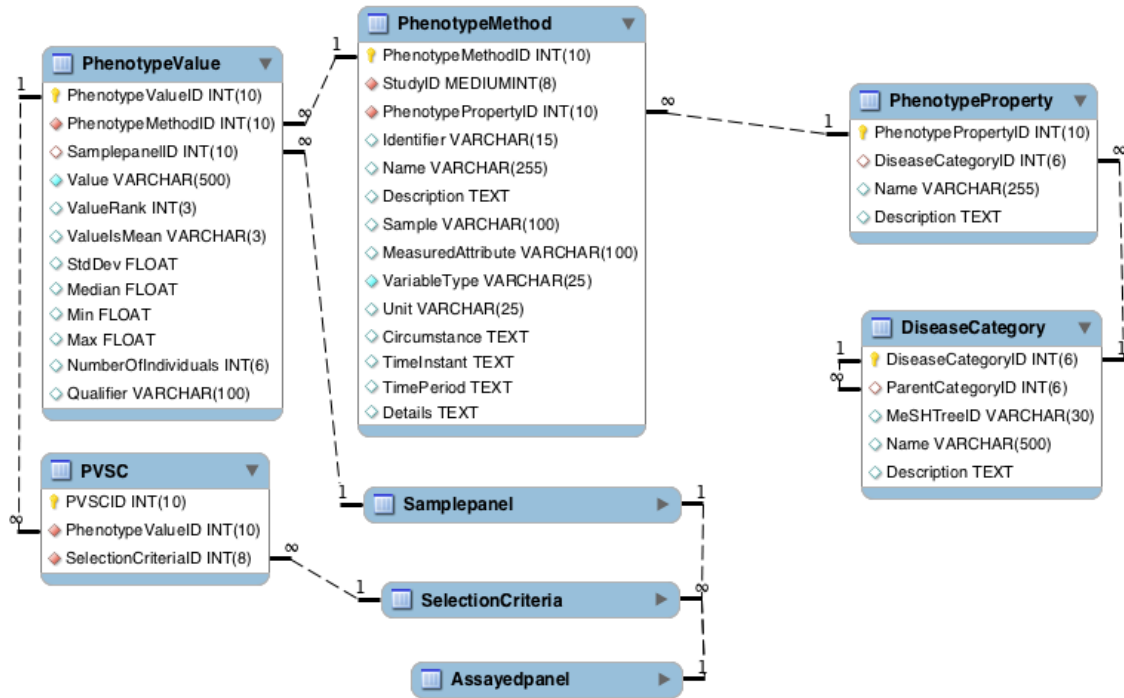


Fig. 5.6: Relational implementation of the PHENOTYPE domain of the conceptual model. An extra linking table enables many *SelectionCriteria* instances to be associated with many *PhenotypeValues*, for maximum flexibility in describing how *Assayedpanels* are created from *SamplePanels* or other *Assayedpanels*.

by Rob Free, with help from the MartBuilder tool (provided with the BioMart software distribution) which facilitates semi-automated schema construction.

5.1.3 Database APIs

The middleware layer of the HGVbaseG2P system shown in Figure §5.1B is built from several database application programming interface (API) libraries. As noted above, such modularisation is a key part of the tiered system architecture employed for this project. The strategy of factoring non-trivial logic into dedicated library modules follows general good practice in software design, and is used throughout to make the system easier to maintain and more robust.

The HGVbaseG2P API libraries are a mixture of custom-built modules which query one of

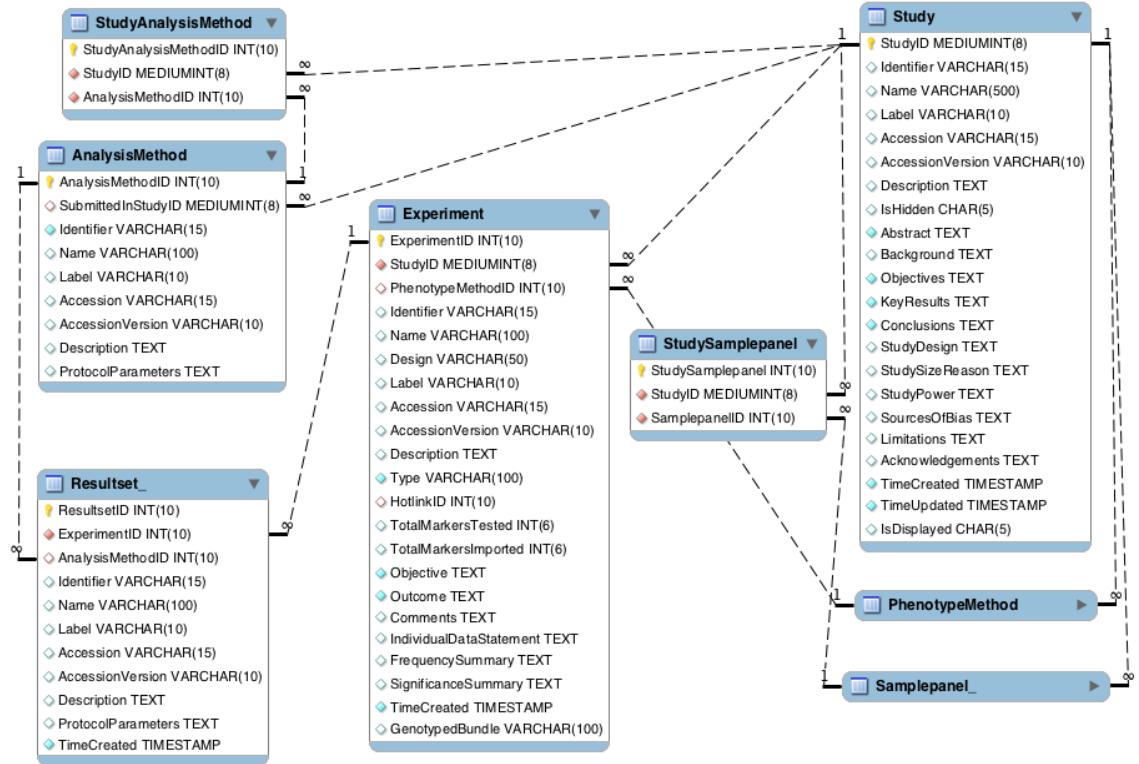


Fig. 5.7: Relational implementation of the EXPERIMENT domain of the conceptual model, part 1. A linking table provides the many-to-many relationship which enables an *AnalysisMethod* submitted in one particular *Study* to be re-used in multiple *Studies*, and similarly for *Samplepanel*.

the core or derived databases directly, custom modules which interact with the databases through a collection of lower-level database API modules, and third-party API toolkits which are used without modification alongside a standard database schema. The main features of these database APIs are briefly described below. Further details regarding individual library modules and their functions are provided on the DVD (see §D).

Custom ORM middleware. In order to programmatically access the core HGVbaseG2P marker and study databases, a custom API library was required. To this end, I used the Perl-based DBIx::Class object-to-relational mapping (ORM) framework, widely used in the open-source community, to generate a collection of custom software modules. I chose this methodology because it greatly minimises the programming effort involved in creating

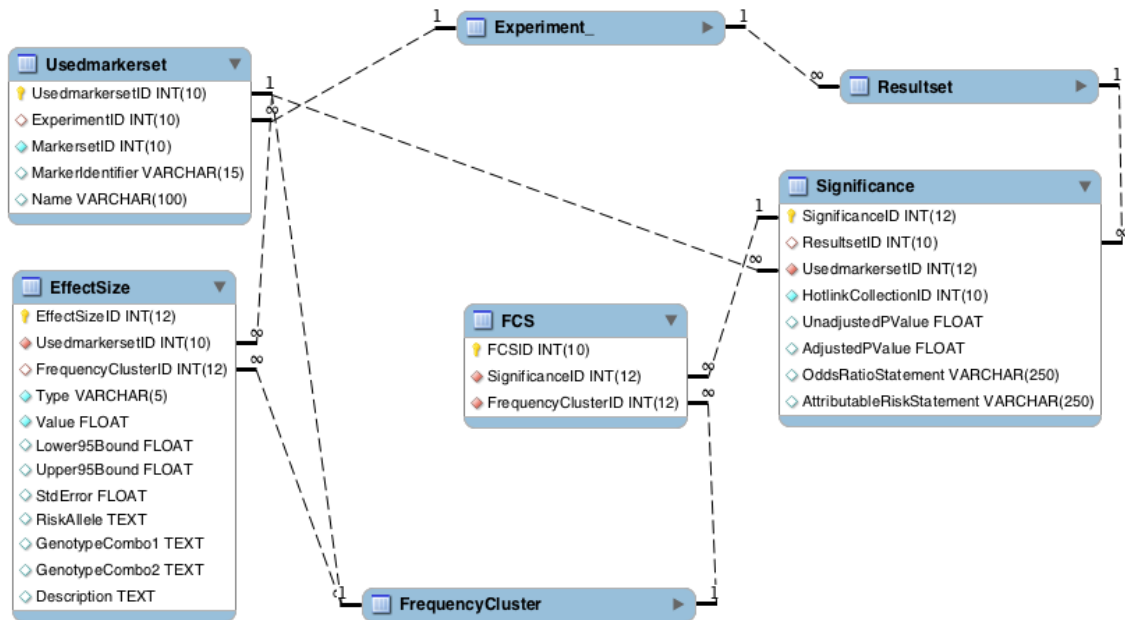


Fig. 5.8: Relational implementation of the EXPERIMENT domain of the conceptual model, part 2. A linking table enables the many-to-many relation from *Significance* to *FrequencyCluster* which is necessary to support multiple analyses with the same genotype data as input.

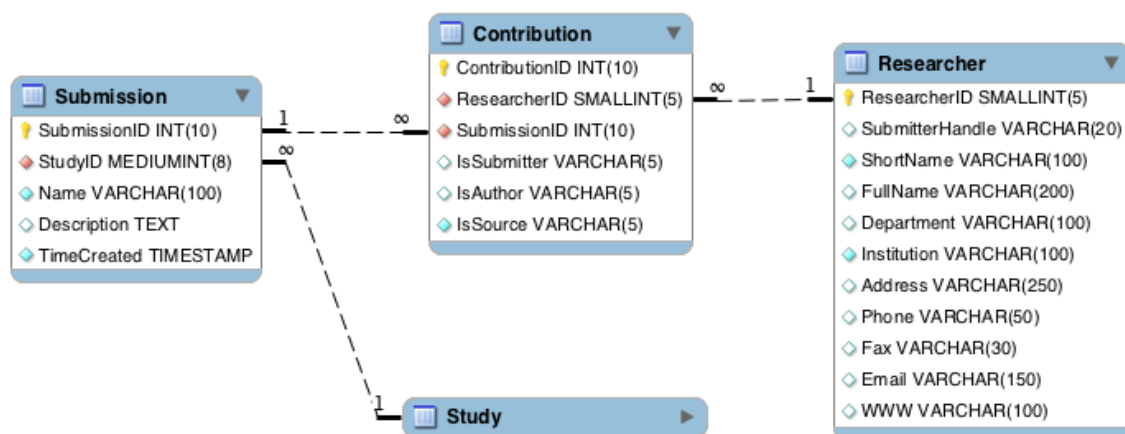


Fig. 5.9: Relational implementation of the EXPERIMENT domain of the conceptual model, part 3.

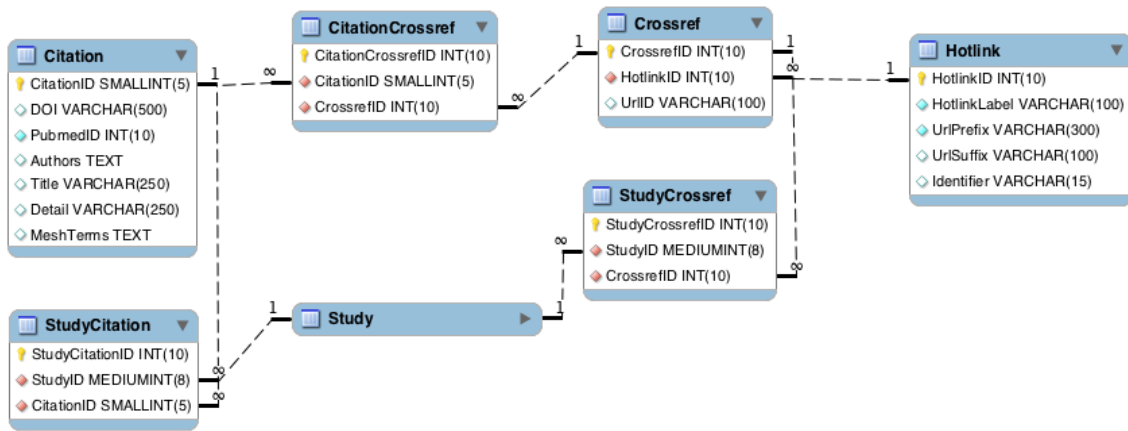


Fig. 5.10: Relational implementation of the generic cross-referencing facilities of the conceptual model. Only the many-to-many links from *CrossRef* to *Study* and *Citation* are displayed for brevity. Similar linking tables to *Crossref* exist for several other key classes in the model.

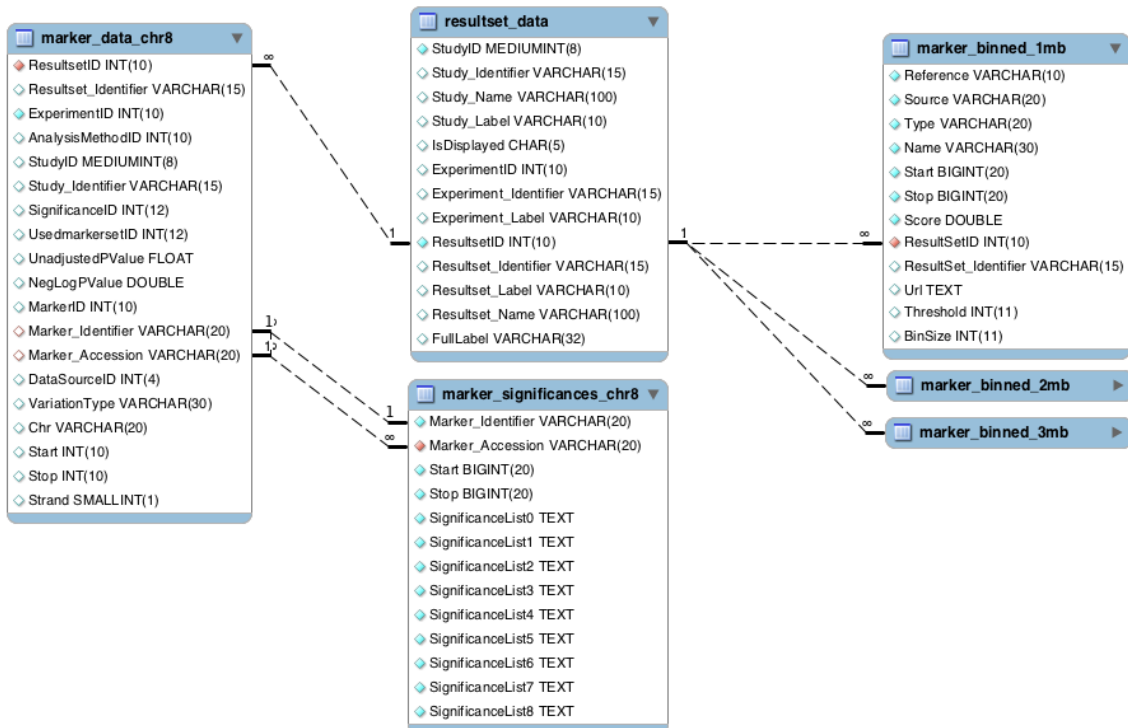


Fig. 5.11: The relational schema of the Browser database. Marker and marker significance tables for Chr8 are shown; tables with identical structure exist for the other chromosomes. All three binnedmarker_* tables have identical structure.

database-driven applications and eases maintenance, compared to simpler methods using relational SQL-queries “hardcoded” and spread throughout the application code.

The generated ORM modules match the relational table structures and facilitate manipulation of database contents using object-oriented (OO) programming techniques. For example, data retrieved from the `Study` table are represented by instances of the *Study* class in application code, and contents of table columns can be accessed and updated via method calls on that object.

Legacy middleware. The `DBIx::Class`-based API is used by the majority of HGVbaseG2P software components in the current version of the system, including all web application components. Certain components, however, still use a different database API which I constructed in early stages of the project. This legacy API, based on the `DBIx::DBStag` Perl package originally developed for the Gene Ontology project, facilitates manipulation of nested data structures and database storage/retrieval of those structures based on the relational structure of the database. Although it has some merit, notably the provision of XML-based import/export functionality (see also §5.1.4 and §A.3), as the project progressed this nested-structure approach was found to be unwieldy, and was thus replaced by the `DBIx::Class`-based API.

The HGVbaseG2P Database library. Most higher-level software components of the system do not utilise the ORM middleware directly, but do so indirectly via the HGVbaseG2P Database library. This library, created by Rob Free, provides a common API layer on top of the ORM modules and several other database modules, and thus effectively abstracts away the idiosyncrasies of the lower-level database APIs. The primary advantage of this arrangement is overall simplification (and therefore easier maintenance) of higher-level code, since components which use the common database API do not need to be concerned with from which database a particular piece of data is retrieved, nor the

mechanics of how this takes place. This also makes it easy to extend the system with additional data sources in the future.

The BioPerl feature database API. The third-party `Bio::DB::SeqFeature::Store` and various related BioPerl libraries provide a compendium of standard software tools for efficiently storing, querying and retrieving genome annotation data using the relational tables already described.

The BioMart API. The BioMart-compatible database is accessed via the third-party BioMart API library. This library is primarily used by the companion MartView web interface described below, but can also be used programmatically by other system components.

5.1.4 Software tools for data transformation and loading

A significant part of routine HGVbaseG2P database operations involves processing, transforming and loading marker and study data, both from external sources into the primary databases and from the primary databases into various derived databases (illustrated in Figs. §5.12 and §5.13). The following subsections describe the suite of command-line utilities and libraries created to support these activities.

GFF3-based marker data import. The marker import system, the first of the two steps shown in Figure §5.12A, is concerned with processing and importing marker data acquired from reference catalogs of genetic variation (see §2.2.2). A key feature of this system, developed by myself in collaboration with Pallavi Sarmah (from India-based GEN2PHEN partner CSIR), is comparison of data from the source database with data already present in the Marker database from previous loads, with the aim of detecting and tracking changes in allele orientation, flanking sequences and other marker information over time. Such detailed tracking was deemed necessary for HGVbaseG2P, in order to maintain consistent link from genotypes and association data in the Study database to

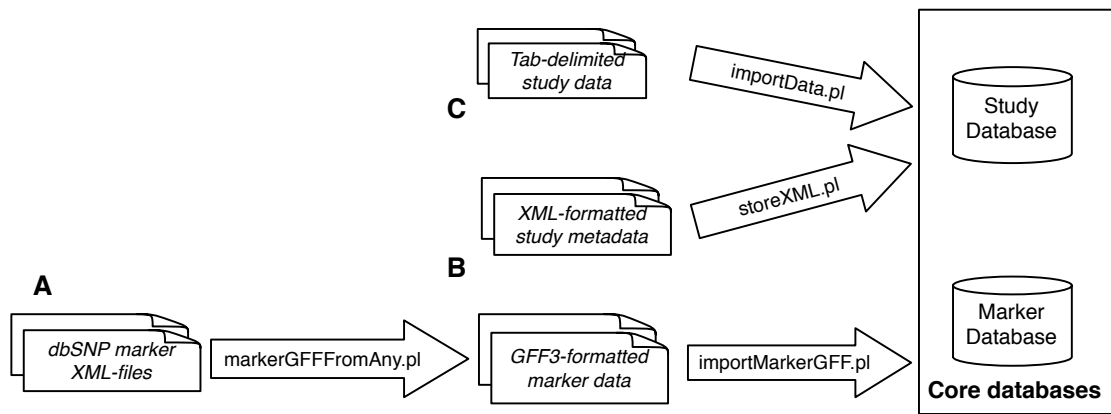


Fig. 5.12: Workflows for loading data into the core databases.

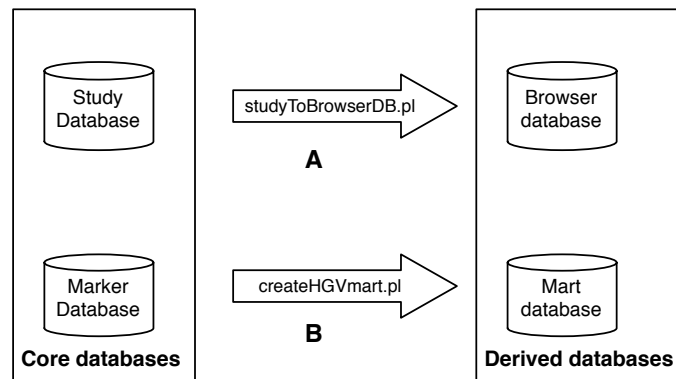


Fig. 5.13: Workflows for exporting data from the core databases to populate the derived databases.

reference marker information, as the content of the dbSNP database and other archives changes.

The overall architecture of the marker import system, known as “dbSNP-lite”, is illustrated in Fig. §5.14. The result from running dbSNP-lite on a target marker database (currently only dbSNP is supported) is a set of standard GFF3-formatted feature files. These files are subsequently loaded into the Marker database using a custom GFF3-loader tool (adapted from the `Bio::DB::SeqFeature::Store` package) to store new markers, update existing markers if required and flag markers which have been deleted from the source

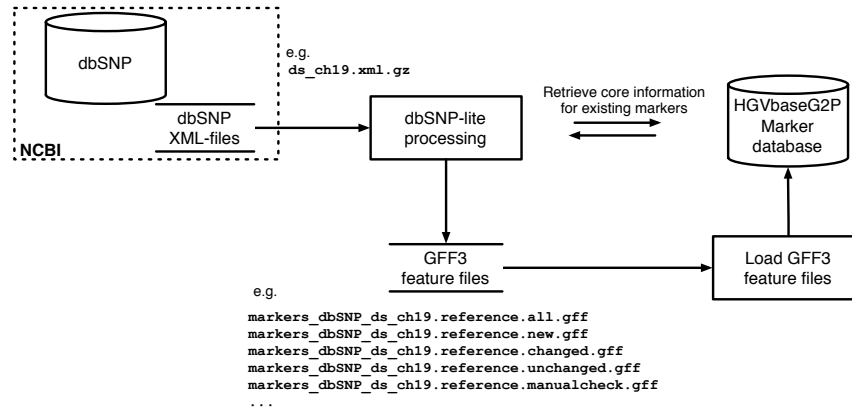


Fig. 5.14: dbSNP-lite architecture for processing marker data from dbSNP and importing into the HGVbaseG2P Marker database.

database. Further implementation details and rationale for dbSNP-lite are provided in §A.2. Results from running the tool to process a recent release of dbSNP are presented in §5.3.

XML-based study data and metadata import. One of the two main tools for importing data into the Study database relies on the aforementioned DBIx::DBStag-based legacy database API. This pipeline, which I built early in the project, utilises the hierarchical structure of XML to represent table relations in the database and does not require a collection of pre-built ORM modules as described above. Any XML data structure which matches the relational schema can be imported with little or no additional configuration to the core system. This provided certain benefits during early HGVbaseG2P development, when a flexible import mechanism was required for iterative testing of the rapidly evolving relational database schema. Further details and data examples are provided in §A.3.

Limitations of XML for large-scale data. In the first half of the project, the XML-based procedure was used for all HGVbaseG2P data import tasks. However, this technique was ultimately found to be unsuitable for processing and loading genome-wide aggregate GWAS datasets. One key issue concerned scalability of XML as a transport format for large-scale datasets. XML is very verbose and thus CPU-intensive to process, and we found

XML-formatted allele/genotype frequency and association datafiles to be unacceptably slow to load - and thus impractical for routine processing of GWAS data (see also §A.3). For smaller datasets, XML-loading performance is not an issue. The XML-based procedure is therefore still (as of May 2010) used for study metadata import in day-to-day HGVbaseG2P operations, albeit scheduled to be phased out and replaced with new tools in the near future (see Discussion).

Loading study genotype and association data. The issues outlined above prompted the design and creation of new tools for processing genome-wide study data in a more flexible and expedient manner. This work was undertaken by Rob Free and resulted in a new import pipeline based on the DBIx::Class-based database API. The main features of this system are i) re-usable templates specifying the layout and formatting of data in the source datafiles and ii) a “plugin” architecture for extending the system with custom processing and validation logic. The aforementioned HGVbaseG2P data validation/import library provides modules which implement the template and plugin functionality, as well as a variety of routines for checking marker identifiers, allele strand orientation and other data validation tasks. Importantly, unlike the XML-based method, in the new pipeline the data format is no longer directly linked with the database, and so the database structure can be altered without invalidating existing datafiles.

Populating derived databases. The workflow depicted in Figure §5.13A uses various facilities provided by the HGVbaseG2P Browser library to extract data from the core Study database and transform into datafiles suitable for loading into the derived Browser database. Similarly, the workflow in Figure §5.13B uses modules from the G2Pmart library to create datafiles suitable for loading into the BioMart database. These utilities were all created by Rob Free.

5.1.5 Web-based applications for online access to database contents

In order to facilitate browsing, searching, retrieval and analysis of HGVbaseG2P contents, a suite of Web-based applications was created. These applications and issues relating to their design and implementation will now be described, whilst section §5.2 demonstrates how the various analysis and reporting tools work together in practical usage.

Building an MVC-based web application. Traditional websites for biological databases have frequently been constructed in “throw away” or one-off fashion, often as a series of crude web scripts which are very difficult to maintain and extend as future needs demand. By contrast, the HGVbaseG2P website was designed and built from the ground up as a sophisticated web application using a wide range of third-party software components. The main HGVbaseG2P web application underpins the project website at <http://www.hgvbaseg2p.org> and drives many of the report displays, summaries and search interfaces described below. This application was initially designed and constructed by myself, and later developed further by Rob Free.

At the design stage, I considered a number of existing bioscience-centric toolkits with which to build the web application, including the aforementioned SYMBioMS system, GMODWeb⁷, the caBIG infrastructure and others. However, my overall assessment of these toolkits was that they were either unsuitable for this project, or potentially suitable but too complex to be of use considering the project scope and timeframe. Therefore, I elected to create a custom tool based on a generic web application framework, and combine this with various custom-built and off-the-shelf data retrieval and analysis tools.

In part due to my familiarity with the Perl programming language, I chose the Perl-based Catalyst framework⁸ as a foundation for the HGVbaseG2P web application. Catalyst follows the established Model/View/Controller (MVC) software design paradigm, which

⁷<http://gmod.org/wiki/GMODWeb>

⁸<http://www.catalystframework.org>

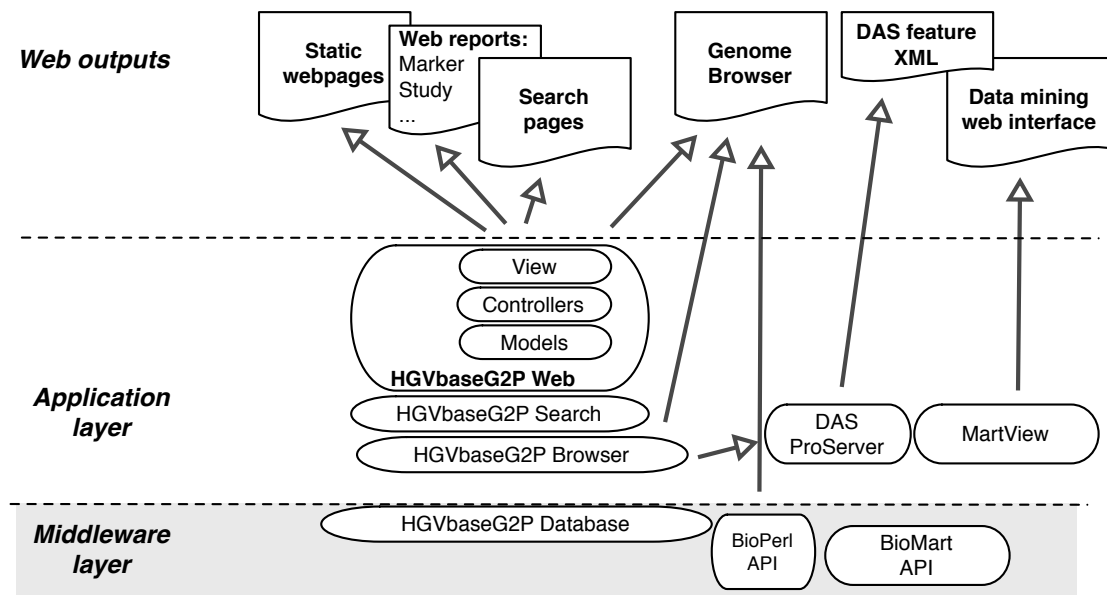


Fig. 5.15: Overview of the web-based application layer of the HGVbaseG2P system and the main outputs generated by various components of the system. For brevity, the data storage layer and low-level database APIs from the previous system overview diagram are not shown.

emphasises a clear separation of concerns between i) the underlying data model, ii) the “business” or domain logic of the application and iii) presentation to the user.

The software tools described below are either extensions of the primary Catalyst-based web application, or separate components that connect to it. Together these components comprise the web-based application layer of the HGVbaseG2P system illustrated in §5.15.

Stable, readable URLs for identifying web resources. As mentioned in §2.5, globally-unique, persistent identifiers for database objects on the Web are crucial for cross-database data integration, whether integration is performed via traditional means or via more sophisticated semantic approaches. From early stages of the project, URL readability was a key consideration in using URLs to identify HGVbaseG2P data objects. Care was taken to make all URLs exposed to the outside world as simple and human-readable as possible, according to established Web conventions (see e.g. “Cool URIs don’t change”⁹).

⁹<http://www.w3.org/Provider/Style/URI>

Examples of this are <http://www.hgvbaseg2p.org/study/HGVST1> to identify a study and <http://www.hgvbaseg2p.org/marker/HGVM16524389> to identify a marker.

A major aim of the URL scheme is to avoid as much as possible future alterations of any HGVbaseG2P URLs, so as to provide external parties with stable links to HGVbaseG2P content. Nevertheless, even if the various web applications in the HGVbaseG2P system maintain stable URLs locally, it is important to note that at present this GUID scheme is not immune to wholesale changes of the Internet domain from where the website is served. For example, if our group (as current operators of HGVbaseG2P) were to lose control of the [hgvbaseg2p.org](http://www.hgvbaseg2p.org) domain, all HGVbaseG2P URLs in use “in the wild” would become invalid. Indeed, this very thing happened after the predecessor of this project was relocated from its original host institution in Sweden; URLs for HGVbase webpages were tied to the hgvbase.cgb.ki.se domain which has since been deactivated.

These concerns relate to another important aspect of GUIDs - *persistence* - and highlight a fundamental weakness of URLs as GUIDs: namely a dependence on the Internet domain name. Several alternative GUID schemes have been devised to address the issue of persistent identification of online resources, including permanent URLs (PURLs)¹⁰, Life Science Identifiers (LSIDs)¹¹ (Martin *et al.*, 2005), ARKs¹² and Digital Object Identifiers (DOIs)¹³ (Paskin, 2000). Some of these schemes were considered for use in the HGVbaseG2P system, but the time and effort required for deployment was deemed too great considering project timeframe.

Tools for browsing and searching database content. Several web-based views and search tools were created to facilitate access to the main categories of HGVbaseG2P database content: studies, phenotypes and markers. Many of these views employ a user interface paradigm known as “faceted browsing” (Yee *et al.*, 2003): for a given category,

¹⁰<http://purl.oclc.org>

¹¹<http://lsids.sourceforge.net>

¹²<http://www.cdlib.org/inside/diglib/ark/>

¹³<http://www.doi.org>

the data items are listed on the main display panel, whilst a secondary panel (to the left or above the main panel) provides various filtering options that the user can use to narrow down the set of items shown according to certain criteria. When the total number of items is relatively small (studies, phenotypes), the option is given to list all items. Conversely, when the total number of items is very large (markers) and a global listing is thus not useful, the faceted view is used for presenting search results. The screenshots shown in §5.2 below illustrate how this works in practice.

Early versions of the search tools were created by myself, by way of extending the Catalyst-based web application and using the database middleware already described. Later, more sophisticated, versions created by Rob Free use the high-level HGVbaseG2P Search library. Regardless of the category of data to be queried, the Search library provides a common API for executing searches and manipulating search results in a uniform way. For marker database entries, queries are performed against the primary database constrained by values for specific fields, such as identifier and genomic location. For study and phenotype entries, a sophisticated text indexing and search facility, based on the third-party Xapian search engine library (see Table §A.1), facilitates boolean keyword searches (e.g. “cancer AND breast”) across all the various fields available, such as titles, descriptions and other free-text content. As with other parts of the system, the modular design brings benefits in terms of extensibility and ease of maintenance. The design also facilitated the construction of a search aggregator which provides a global search across all available data categories.

Web reports. All the content listings and search result displays described above are hyperlinked to various report pages, which contain further details on individual database entries. Table §5.2 summarises the range of reports available and section §5.2 below describes how some of these reports are used in practice.

Graphical displays of association study data. A number of open-source genome browser software tools suitable for local installation were considered for constructing browser-based views of association study data, including toolkits underpinning the well-

Table 5.2: Web-based reports for individual entries in the core HGVbaseG2P database.

Report type	Report content	Example screenshot
Study report	Study summary information and cross-references. List of panels and phenotypes used. List of G2P experiments. List of most significant disease-associated markers.	Figure §5.21
Marker report	Marker summary information and cross-references. Lists of available aggregate genotypes and association results from studies where marker has been tested.	Figure §5.19
Panel reports	Panel summary information and cross-references. Details on how Assayedpanels are composed of Samplepanels	Figure §5.16
Phenotype reports	Phenotype method summary information and cross-references	Figure §5.17

HOME STUDIES PHENOTYPES MARKERS HGVMART BROWSER

» Prostate cancer (HGVST1)

⊖ Data not available to Browser ➤ Go to Browser ➤ Go to Markers

Summary Panels Phenotypes Genotype Experiments Analysis Experiments

The following Sample Panels are defined in this Study. [What is a Sample Panel?](#)

Sample Panel name	No. individuals	Panel composition	Geographic region
PLCO cohort	155,000	Unrelated	NORTH AMERICA
CGEMS prostate cancer sub-cohort	28,521	Unrelated	NORTH AMERICA

The following Assayed Panels are defined in this Study. [What is an Assayed Panel?](#)

Assayed Panel name	No. individuals	Selected from Sample Panel(s)	Phenotype(s) Tested
Case aggressive	737	CGEMS prostate cancer sub-cohort	Prostate cancer
Case non-aggressive	624	CGEMS prostate cancer sub-cohort	Prostate cancer
Control	1,230	CGEMS prostate cancer sub-cohort	Prostate cancer
Case	1,361	Undefined	

Fig. 5.16: The HGVbaseG2P panel report page.

HOME

STUDIES

PHENOTYPES

MARKERS

HGMART

BROWSER

» HGVST1: Prostate cancer » Phenotype method *Prostate cancer status*

Phenotype Summary

HGVbaseG2P identifier	HGVPM1
Phenotype property assayed	Prostate cancer
Description	Classify subjects into one of three categories: without disease, with non-aggressive form of disease, or with aggressive disease
Variable type	Ordinal
Cross-references	Not supplied
Phenotype categories	<ul style="list-style-type: none"> Male Urogenital Diseases Neoplasms

Method Details

Circumstance	Not supplied
Time instant	Not supplied
Time period	Not supplied
Citations	none

Web Site Release 2.2 - Sept 2009

About | Disclaimer | Contact Us © HGVbaseG2P

Fig. 5.17: The HGVbaseG2P phenotype report page.

known Ensembl and UCSC browsers. Due to factors such as ease of installation and customisation, as well as my own familiarity with the toolkits from previous projects, the GBrowse package mentioned in §2.2.3 was chosen for use in the project. Through work undertaken primarily by Rob Free and with contributions from myself, GBrowse and the companion package GBrowse_karyotype (see Table §A.1) were configured and extended to create a series of novel genome-wide and region-level displays for association study findings. The capabilities of these graphical tools are further explained in the context of usage scenarios presented in section §5.2 below.

Web service APIs. Programmatic API access to HGVbaseG2P content is provided via several means. First, the MartView application provides API access to the BioMart data mining system, as described in Smedley *et al.* (2009). Second, a DAS server exports study association data as genome features via the DAS protocol mentioned in §2.5. This DAS server, created by Owen Lancaster, is a custom extension of the third-party ProServer

package (Finn *et al.*, 2007) which fetches study data as genome annotations from the Browser database and transmits in a standard format to a remote DAS client. Both of these APIs return genome-wide GWAS data, and due to data sensitivity concerns neither is currently active on the public website.

The third API is concerned with metadata only. The various browsing and search tools described above optionally return machine-readable XML in the standard Atom syndication format¹⁴. The potential for Atom “data feeds” and the standard Atom Publishing Protocol (AtomPub)¹⁵ as a generic framework for publishing G2P data on the Web, and for “funnelling” data from one database to another, has been described elsewhere (Thorisson, 2009b) (see also Cafe RouGE in §??).

5.2 Using the HGVbaseG2P website toolkit

The web-based software components described above work together in various configurations to enable exploration of database content from several different perspectives. The functional aspects of the system will now be described, by way of several use cases which showcase the capabilities of the toolkit. Unless specifically listed as a development website feature only, at the time of writing all these scenarios are supported on the live HGVbaseG2P website at <http://www.hgvbaseg2p.org>. A more in-depth description of how to use the system has been published as online help documentation¹⁶.

Finding markers of interest. Users are often interested in all available information for a particular variant site in the genome, typically because the marker has been reported as significantly associated with a disease phenotype in a GWAS publication. The marker search modality provides a way to explore HGVbaseG2P study content from a marker perspective. Fig. §5.18 shows the user interface for querying the database by entering one

¹⁴<http://tools.ietf.org/html/rfc4287>

¹⁵<http://tools.ietf.org/html/rfc5023>

¹⁶<http://www.hgvbaseg2p.org/info/help>

baseG/P Genotype-to-Phenotype

HOME STUDIES PHENOTYPES M

Markers

Use this page to find Markers in the database. [What is a Marker?](#)

Search: **Significance threshold:** **Display only markers with data:** ☐

Enter any HGVbase IDs, dbSNP IDs, HGNC gene symbol or genomic region.
(For multiple markers enter space-separated IDs).

Show Markers with association results where p-values are greater or equal this value

42 Markers in region 'chr4:100000..400000' (1-42 shown) (Reset All Filters)

Export these results as: **Report** ☐ **Table** | **Sort by:**

Marker: HGVM3565 : rs3946 Go to Association Results Genomic location: Chr4:357927..357927 Variation type: SNP
Marker: HGVM177798 : rs718429 Go to Association Results Genomic location: Chr4:398952..398952 Variation type: SNP
Marker: HGVM189286 : rs1443076 Go to Association Results Genomic location: Chr4:357956..357956 Variation type: SNP
Marker: HGVM1519188 : rs1986557 Go to Association Results Genomic location: Chr4:273118..273118 Variation type: SNP
Marker: HGVM1531430 : rs2353605 Go to Association Results Genomic location: Chr4:209538..209538 Variation type: SNP
Marker: HGVM1918691 : rs1822345 Go to Association Results Genomic location: Chr4:374533..374533 Variation type: SNP
Marker: HGVM2264286 : rs3749523 Go to Association Results

Fig. 5.18: The marker search page on the HGVbaseG2P website (<http://www.hgvbaseg2p.org/markers>), showing results from querying for all markers mapping to a 300Kbp region on Chr4.

or several marker identifiers, a genomic region or a gene symbol as landmark. The results page shows a list of markers found, hyperlinked to individual marker reports which the user can subsequently peruse (Figure §5.19).

Finding studies of interest. Another common scenario is when a user is interested in findings from a particular association study he has already discovered by other means. The study search modality caters for such basic study-list browsing. The study section of the website (Figure §5.20) lists all studies in the database, but the list can be easily filtered by entering a keyword or author name into the textfield that is common to all the search pages. The user can then follow each of the hyperlinked list entries to peruse the study report page shown in Figure §5.21.

» Marker report for HGVM1285772

Summary | Association Results | Genotype Frequencies

HGVbaseG2P Marker ID HGVM1285772

Source database info External marker accession [rs2420946](#), imported from dbSNP

Variation type SNP

Observed alleles [GATCCATAAG](#) [(C):(T)] [GCATCCACAG](#) (IUPAC ambiguity code: Y)

Genomic location

Marker status

Cross-references

» Marker report for HGVM1285772

Summary | Association Results | Genotype Frequencies

Use the control below to show association results above other significance thresholds.
Significance Threshold:

13 Studies contain Markers with P Values above the threshold
[Show association results as table](#)

- [4 kHz hearing threshold \(dB\)](#) (HGVST311)
- [Adult body mass index \(kg/m2\)](#) (HGVST308)
- [Breast cancer](#) (HGVST2)
- [Breast cancer](#) (HGVST63)
- [Systolic blood pressure \(mm Hg\)](#) (HGVST307)
- [Prostate cancer](#) (HGVST1)

[Experiment 1: Association analysis from WGAS Phase 1 \(Illumina 317K and 240K\).](#)

Analysis Method	Result Set	Unadjusted P value
Single selection sampling, Adjusted score test	Single selection sampling, Adjusted score test	Not available
Incidence density sampling, Unadjusted score test	Incidence density sampling, Unadjusted score test	Not available
Incidence density sampling, Adjusted score test	Incidence density sampling, Adjusted score test	Not available
Single selection sampling, Unadjusted score test	Single selection sampling, Unadjusted score test	Not available

[Cholesterol \(mmol/L\)](#) (HGVST312)

Fig. 5.19: HGVbaseG2P marker report page (<http://www.hgvbaseg2p.org/marker/HGVM1285772>). A) Summary information describing the exemplar rs2420946 marker imported from dbSNP is shown by default under the first tab. B) A full listing of all studies where this marker has been tested is under the second tab.

Finding phenotypes of interest. Users are frequently interested in a particular phenotype and wish to browse available studies in the catalog in more targeted fashion. The phenotype browser page in Figure §5.22 displays a list of all available phenotypes across all studies in the database. As with the study list above, hyperlinks are provided to the study report, phenotype report and other relevant parts of the website. The list can be filtered by choosing one or more categories on the left-hand menu, or by entering a keyword. Currently the phenotype categories presented for filtering comprise a flat, alphabetical list of high-level MeSH terms associated with each phenotype entry, but work is now underway in the group to create a far more useful tool for hierarchical category browsing as a means to navigate the collection of available phenotypes.

The screenshot displays the HGVbaseG2P website interface. At the top, there are navigation tabs: HOME, STUDIES, PHENOTYPES, MARKERS, and BROWSER. Below the tabs, a search bar contains the text 'yeager'. To the right of the search bar, there are buttons for 'Add all Studies to the Browser' and 'Remove all added Studies'. Below the search bar, there are filter options: 'Filter Query Results' with radio buttons for 'match everything (6)', 'match keywords (6)', 'matches gene feature or region (0)', and 'matches markers (0)'. Below the filter options, there is a section for '6 Studies matching 'yeager' (1-6 shown)'. Each study entry includes a plus icon, a study ID (e.g., HGVST387), a title, a brief description, and a link to 'Go to Browser'. The first study, HGVST387, is expanded, showing its abstract and related citations. The second study, HGVST357, is also expanded, showing its title and a link to 'Go to Browser'. The third study, HGVST155, is partially visible, showing its title and a link to 'Go to Browser'.

Fig. 5.20: HGVbaseG2P study listing page (<http://www.hgvbaseg2p.org/studies?q=yeager>). The full collection of several hundred studies in the catalog has been filtered down to 6 via a keyword search for an author's family name.

Graphical display and comparison of genome-wide study findings. The filtered lists in the two previous scenarios also feature as entry points to the graphical genome browser tools. Except for studies where data are not yet available for public viewing, for each study or phenotype listed the study can be added to the browser by clicking the “plus” symbol. In the resulting popup-window, the user can select which of up to several available sets of results to add to the browser. Up to sixteen datasets can be chosen for comparison in this way. The user can then click the “Go to Browser” link to switch to the browser view. The main purpose of the browser view is to facilitate comparison of findings across the chosen studies, or for comparing results within a single study (e.g. different analysis approaches, or different cohorts). Once interesting regions are identified in the genome view, the user can click each region to zoom in to a more detailed region-level view. This functionality is illustrated in Figure §5.23.

Study Report - GWAS of prostate cancer (HGVST1)

Data access not provided

Summary | Panels | Phenotypes | Analysis Experiments | Markers

Export Study as: --choose a format--

HGVbaseG2P Identifier: HGVST1

Study name: GWAS of prostate cancer

Phenotype(s) tested: Prostate cancer

Study design: Case and control

Genotype Platforms: Illumina HumanHap240S Illumina

Abstract: Summary-level data and analysis re found to be associated with prostate cancer. The most significant signal is 70 kb additional studies (total: 4,296 case ratio (OR): 1.26, 95% confidence in (rs1447295 P = 1.41 x 10⁻¹¹); rs indicate the presence of at least two the new locus, marked by rs698324

Submission information

Contributor	Date Submitted	Analysis
CGEMS	2009-03-30	
HGVbaseG2P	2009-03-30	

Cross-references: NCI CGEMS website¹⁷

Study Report - GWAS of prostate cancer (HGVST1)

Data access not provided

Summary | Panels | Phenotypes | Analysis Experiments | Markers

The Study contains 2 Analysis Experiments. [What is an Analysis Experiment?](#)

Experiment: HGVST1

Association analysis from WGAS Phase 1A (Illumina 317K).

Experiment: HGVST2

Association analysis from WGAS Phase 1 (Illumina 317K and 240K).

Phenotype: Prostate cancer

Total No. Markers Imported: 527,698

Analysis Summary

Analysis Method	Result Set
Incidence density sampling, Adjusted score test	Incidence density sampling, Adjusted score test (HGVST2)
Single selection sampling, Unadjusted score test	Single selection sampling, Unadjusted score test (HGVST5)
Single selection sampling, Adjusted score test	Single selection sampling, Adjusted score test (HGVST6)
Incidence density sampling, Unadjusted score test	Incidence density sampling, Unadjusted score test (HGVST8)

Assayed Panels Used In Experiment

Assayed Panel name	No. individuals	Selected from Sample Panel(s)
Case aggressive	737	CGEMS prostate cancer sub-cohort
Case non-aggressive	624	CGEMS prostate cancer sub-cohort

Fig. 5.21: HGVbaseG2P study report page. (<http://www.hgvbaseg2p.org/study/HGVST1>). A) Summary information describing the study is shown by default under the first tab. B) Available association analysis results are summarised under the fifth tab, with options to add the result sets to the browser if study data are available. Additional study details (not shown) are accessible via the remaining tabs.

Retrieving website data updates via feed reader. All the search functions described above can alternatively display results as web feeds in the machine-readable Atom XML-format. Users can click the ubiquitous feed icon highlighted in Fig. §5.24A to retrieve the feed URL for the current page (e.g. a list of studies found with a search for the keyword “Crohn’s”). This URL can subsequently be used to subscribe to the feed and monitor for updates using Google Reader¹⁷ as shown in Fig. §5.24B or other feed reader software. This feature is analogous to web feeds offered by PubMed, the NHGRI GWAS catalog and growing number of biological data sources. Web feeds offer a simple yet powerful way of monitoring many websites for updated content.

¹⁷<http://www.google.com/reader>

HOME STUDIES PHENOTYPES MARKERS HGV MART BROWSER

Browse all phenotypes

The following Phenotypes have been investigated by Studies in the database. [What is a Phenotype?](#)

3 Phenotypes added to Browser
3 Result Sets added to Browser
[Go to Browser](#) [Remove](#)

Phenotype List Filters

Use the filters below to refine the Phenotype displayed in the list on the right.

Search Phenotypes by keyword(s)

Select Phenotype categories
1/66 Categories Selected

[A-D](#) [E-H](#) [I-L](#) [M-P](#) [Q-T](#) [U-X](#) [Y-Z](#)

☐ Reproductive and Urinary Physiological Phenomena
☒ Respiratory Tract Diseases
☐ Skin and Connective Tissue Diseases
☐ Stomatognathic Diseases
☐ TODO

13 Phenotypes filtered from 631 (1-13 shown) (Reset All Filters)

Phenotype	Add to Browser	Phenotypic variable	Study Title	Study ID
Lung cancer	+	Lung cancer status	Lung cancer	HGVST72
Lung cancer, smokers with versus smokers without	+	Lung cancer, smokers with versus smokers without status	Lung cancer, smokers with versus smokers without	HGVST74
Childhood asthma	+	Childhood asthma status	Childhood asthma	HGVST84
Asthma (toluene diisocyanate-induced)	+	Asthma (toluene diisocyanate-induced) status	Asthma (toluene diisocyanate-induced)	HGVST110
Lung cancer	+	Lung cancer status	Lung cancer	HGVST119
Lung cancer	+	Lung cancer status	Lung cancer	HGVST136
Asthma	+	Asthma status	Asthma	HGVST137
Chronic Obstructive Pulmonary Disease	+	Chronic Obstructive Pulmonary Disease status	Chronic obstructive pulmonary disease	HGVST144
Lung cancer	+	Lung cancer status	Lung cancer	HGVST164
Idiopathic pulmonary fibrosis	+	Idiopathic pulmonary fibrosis status	Idiopathic pulmonary fibrosis	HGVST168
Lung cancer	+	Lung cancer status	Lung cancer	HGVST170
Lung cancer	+	Lung cancer status	Lung cancer	HGVST180
Lung cancer	+	Lung cancer status	Lung cancer	HGVST329

Fig. 5.22: HGVbaseG2P phenotype listing page. (<http://www.hgvbaseg2p.org/phenotypes>). The full collection of 631 phenotypic traits have been filtered via the dynamic category selector in the left-hand panel.

In its simplest form, the feed version of the main HGVbaseG2P study listing is nearly identical to the NHGRI GWAS catalog¹⁸. However, the true power of this feature is apparent once the user subscribes to a feed for a filtered study list (e.g. results from a keyword search), a list of all phenotypes in a broader category (e.g. all cardiovascular diseases) or one of the other search pages on the website. In this way, web feeds enables the user to create customised listings of HGVbaseG2P catalog contents and, importantly, to monitor these listings in an automated fashion.

5.3 Gathering reference variation data for HGVbaseG2P

In order to provide context for association study data in the Study database, the HGVbaseG2P Marker database contains a copy of core dbSNP data which serves as a reference information layer of known genetic variation. The dbSNP-lite tool, which facilitates routine transformation and import of a dbSNP release into HGVbaseG2P, was

¹⁸<http://feeds.feedburner.com/NhgriGWASCatalogAdditions>

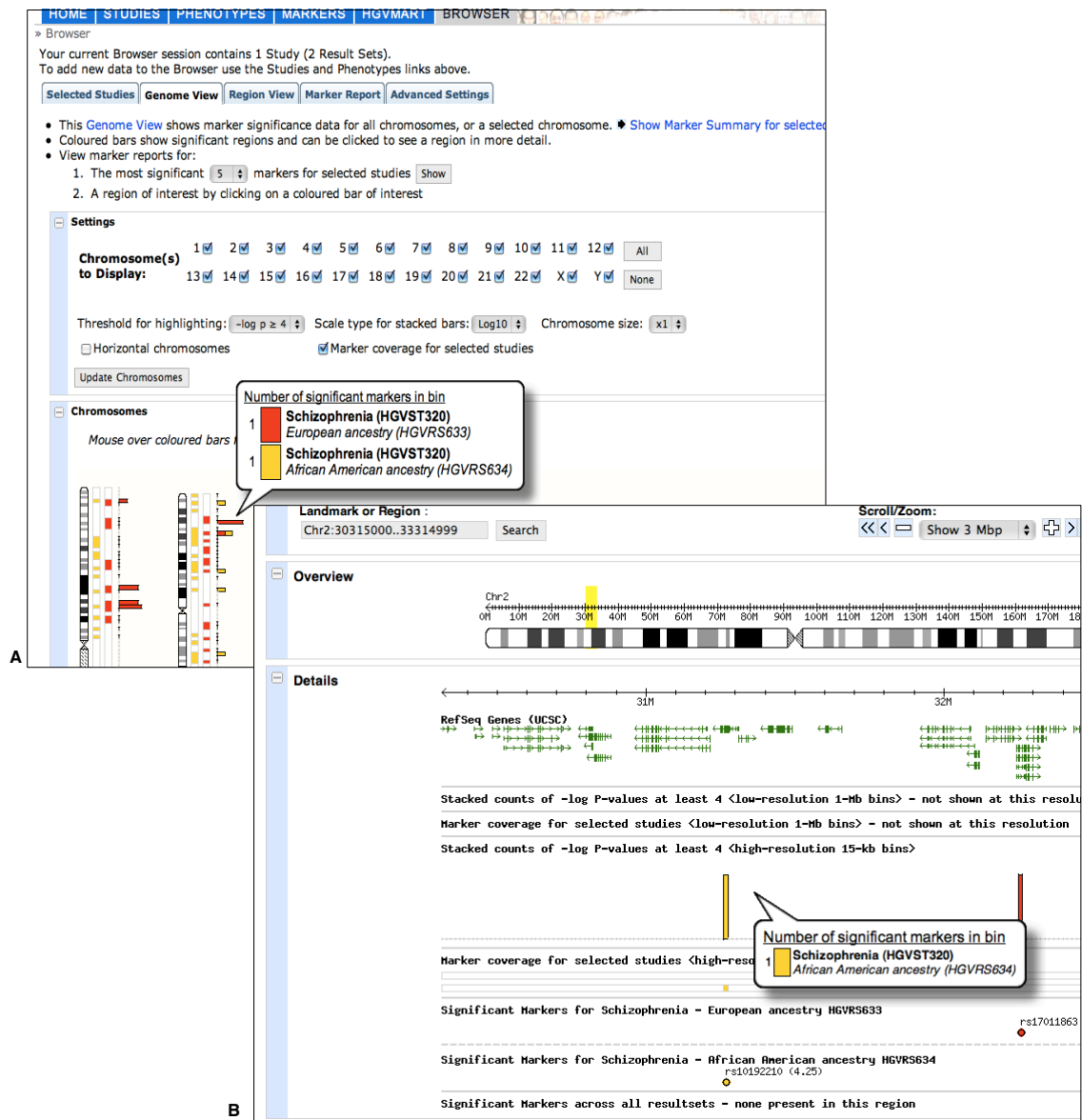


Fig. 5.23: HGVbaseG2P graphical display capabilities. A) Two association datasets are being compared in the genome-wide view, both from the same schizophrenia study (<http://www.hgvbaseg2p.org/study/HGVST320>) but using different cohorts. The histogram alongside each chromosome indicates the number of markers per 3Mbp bin with a P-value from the association test that passes a tuneable significance threshold. The highlighted bin on the p-arm of chromosome 2 contains significant association signals in both scans. B) Clicking on regions of interest zooms in to a detailed, customisable view in the region-level browser. A variety of annotation tracks are displayed in this view to facilitate further exploration of study findings. In this particular study comparison, the region-level browser reveals that the two association signals originate from two SNPs spaced approx. 1Mbp apart. Known genes and other common annotations are optionally displayed in separate tracks.

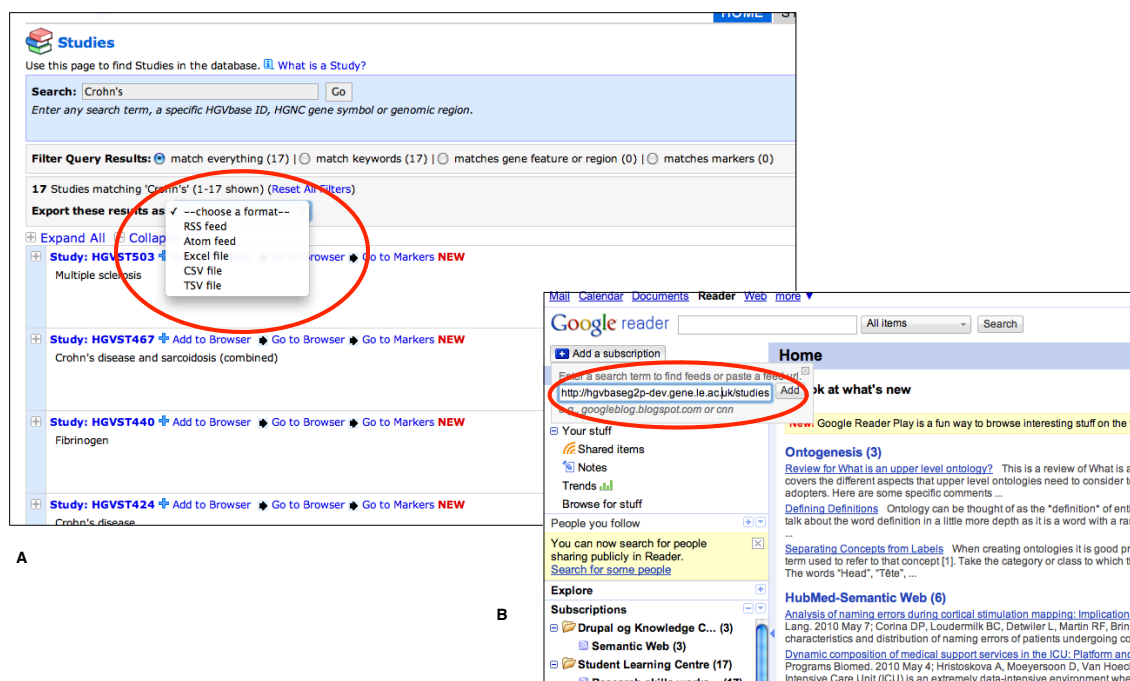


Fig. 5.24: HGVbaseG2P web feeds. The search results can be exported as Atom or RSS feed XML (A). The feed URL can be added to Google Reader (B).

already introduced in §5.1.4. This section presents results from the latest dbSNP import cycle.

Processing dbSNP b130 against HGVbaseG2P content. The full XML-release of dbSNP b130 content was processed against the contents of the Marker database containing data from dbSNP b129, as described in §A.2. The results are summarised in Fig. §5.25. Approximately 3.5M new rs# entries (an increase of ~20%) have been created in dbSNP from data submitted by variation discovery projects since the previous release, primarily the 1,000 Genomes Project.

Changes in core dbSNP marker information. Out of the ~14.2M dbSNP markers already present in HGVbaseG2P before this latest update cycle, the vast majority were found to have unchanged core information. Only ~35,000 (~0.2%) exhibited changes compared to the previous b129. This is a substantial reduction from a similar analysis

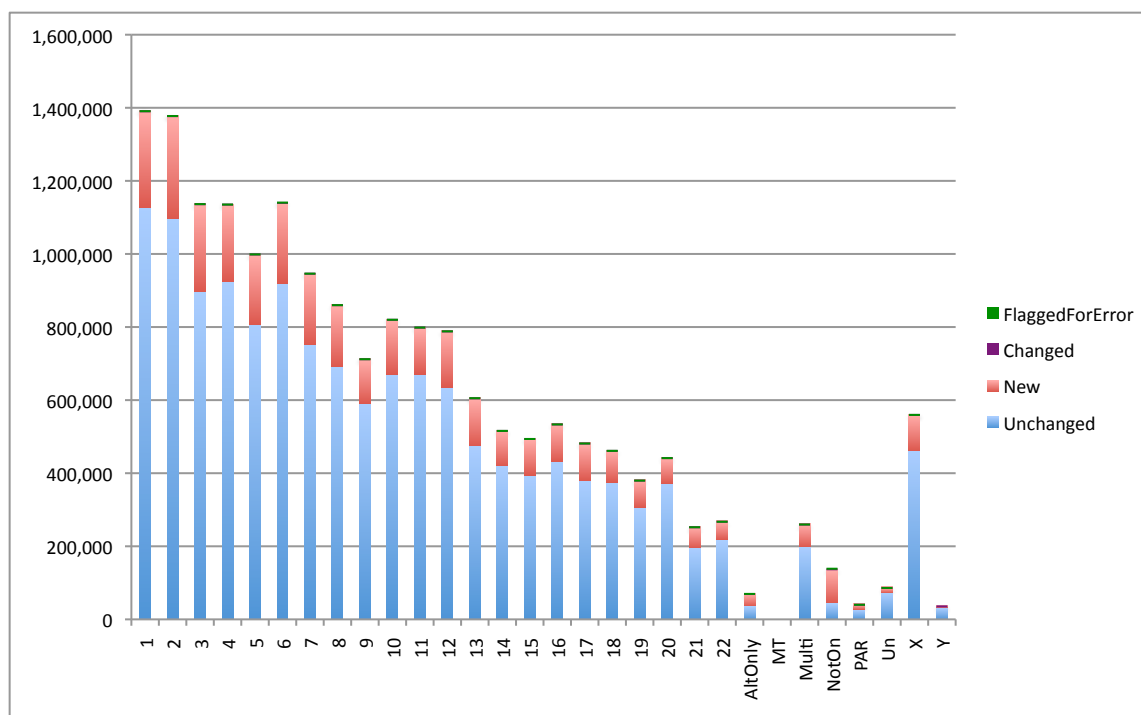


Fig. 5.25: Summary of changes in dbSNP b130 core marker content compared to b129 as stored HGVBbaseG2P (see Table §5.3 for the underlying data). Y-axis shows no. rs# entries. X-axis is sorted by dbSNP XML-file labels. Abbreviations: AltOnly=rs#'s mapped to a non-reference, alternative genome assembly only; Multi=rs#'s mapped to multiple chromosomes in the reference assembly; MT=rs#'s mapped to the mitochondrion; NotOn=rs#'s not mapped to any chromosome in any assembly; Un=mapped to a chromosome in the reference assembly but no definite sequence coordinates; PAR=rs#'s mapped to the pseudo-autosomal region on ChrY.

undertaken previously in the project comparing b129 with the previous b128 release, where ~1% of rs#'s were found to have changed between releases.

A total of 5,885 rs#'s exhibited inconsistencies in flanking sequences, allele assignments or other changes which could not be reconciled in an automated fashion by the dbSNP-lite tool. These markers were therefore provisionally excluded from further processing. A full analysis of these and the rest of the total ~35,000 markers with changes has not yet been undertaken, but a preliminary analysis based on cross-checking rs# lists with the HGVBbaseG2P Study database indicates that 10-20 GWAS-tested markers may be affected.

Table 5.3: Summary of changes in core marker information in dbSNP build b130 vs b129.

Chromosome	Unchanged	New	Changed	FlaggedForError	Total
1	1,127,418	264,067	1,646	287	1,393,418
2	1,096,022	282,148	1,792	204	1,380,166
3	896,737	240,595	1,543	239	1,139,114
4	924,960	210,716	2,141	120	1,137,937
5	807,799	191,746	1,770	183	1,001,498
6	920,476	220,498	1,460	813	1,143,247
7	751,543	195,878	1,299	137	948,857
8	692,306	168,800	1,442	93	862,641
9	592,009	121,581	1,011	103	714,704
10	671,171	150,151	1,163	178	822,663
11	671,445	128,654	1,075	134	801,308
12	635,917	153,812	1,069	117	790,915
13	475,310	131,302	841	566	608,019
14	420,251	97,329	870	58	518,508
15	395,041	100,636	681	70	496,428
16	432,096	103,288	780	68	536,232
17	380,174	103,051	715	132	484,072
18	375,274	87,759	816	62	463,911
19	307,025	74,973	523	47	382,568
20	371,729	71,473	855	34	444,091
21	195,622	58,803	400	30	254,855
22	218,680	50,817	405	23	269,925
AltOnly	36,692	35,381	33	344	72,450
MT	659	11	0	0	670
Multi	198,856	63,297	82	121	262,356
NotOn	44,907	91,402	3,409	1,652	141,370
PAR	27,678	14,995	39	13	42,725
Un	73,470	15,274	6	8	88,758
X	461,885	99,725	629	49	562,288
Y	32,875	5,456	9	0	38,340
Total	14,236,027	3,533,618	28,504	5,885	17,804,034

Deleted and merged dbSNP markers. All marker GFF3 feature files, except those labelled “manualcheck”, were loaded into the Marker database using the GFF3-loader utility described in §A.2.5. dbSNP merge history information was subsequently added to the database, followed by the “legacy” check, as described in §A.2.6. This resulted in a total 3,744,437 rs# entries flagged as merged with another rs#. This large number of mergers (compared to previous processing jobs) reflect large-scale fixes in the b130 release which addressed an error in the dbSNP build procedure for b129 (see the dbSNP mailing

list announcement¹⁹ for details). Another 37,129 rs# entries which featured in b129 were found to be deleted from b130 altogether.

Other sources of reference variation data. At present, a limitation of the HGVbaseG2P Marker database is that only simple sequence variants as cataloged by dbSNP are included. This fulfills the principal purpose of supporting association studies held in the Study database, all of which are currently SNP-based. However, as discussed in §2.2.2, reference variation data are available from various other sources as well. If variation data from DGV, UniSTS and other sources were also incorporated, HGVbaseG2P would become a truly comprehensive catalog of genetic variation, and indeed both of these are currently being considered for routine import into HGVbaseG2P. Reference information on structural variants, whether acquired from DGV, dbVar or elsewhere, will be particularly important going forward as CNVs are included in GWAS scans. But as noted in §3.2.3, handling CNVs and other structural variants adequately may require fundamental modifications to the HGVbaseG2P conceptual model and parts of the HGVbaseG2P system itself.

5.4 Gathering association study data for HGVbaseG2P

The major initial aim of the project is to create a global catalog of association study findings. Central to this aim is active gathering of study data from a variety of sources, combined with direct submissions from the community. However, the findings by Homer *et al.* (2008) and subsequent reactions from G2P data providers as discussed in §2.3.5 radically changed the G2P data sharing landscape and necessitated a significant departure from this initial strategy. This section discusses the data collection methodology employed early in the project, provides a brief historical perspective on how these events have affected the project, and then discusses the data collection methodology currently followed.

¹⁹<http://www.ncbi.nlm.nih.gov/mailman/pipermail/dbsnp-announce/2008q2/000082.html>

Published aggregate datasets in the early “GWAS era”. In the early stages of the project, one of the first publicly-accessible GWAS datasets were aggregate genotype data and association analysis results from the CGEMS project (see also §3.10). Soon after, several other aggregate GWAS datasets became available from a number of sources (see Table §5.4). This initial collection of complete, aggregate GWAS datasets were downloaded and subsequently imported into the Study database according to the procedure described in §A.3. The datasets were thereafter made available via the HGVbaseG2P website on its launch in July 2008. This work was carried out primarily by Rob Hastings, with contributions from myself.

Table 5.4: Summary of GWAS datasets gathered and imported into HGVbaseG2P prior to Homer *et al.* (2008). Abbreviations: WTCCC=Wellcome Trust Case-Control Consortium, CGEMS=Cancer Genetic Markers of Susceptibility, DGI=Diabetes Genetics Initiative. a) A common set of 3,000 controls were compared with 2,000 cases in each of the seven WTCCC disease studies.

Datasource	Disease phenotype	No. individuals genotyped	No. markers tested
CGEMS	Prostate cancer	2,700	317K/550K
	Breast cancer	2,400	550K
dbGaP	Age-related macular degeneration	600	100K
	Type I diabetes	2,300	300K
	Ischemic stroke	550	400K
WTCCC ^a	Type I diabetes	5,000	550K
	Type II diabetes	5,000	550K
	Crohn’s disease	5,000	550K
	Coronary heart disease	5,000	550K
	Hypertension	5,000	550K
	Bipolar disorder	5,000	550K
	Rheumatoid arthritis	5,000	550K
DGI	Type 2 diabetes	5,000	550K

A changing data sharing landscape. The HGVbaseG2P catalog initially covered only a minority of the total number of association studies published at the time of website launch. However, this was nevertheless considered a good first step towards a comprehensive study catalog, especially given that GWAS datasets from many published studies were not available online at all at the time. The two primary G2P archives - dbGaP in the

US and the EGA in Europe (see §2.2.7) - were being established around that time, and it seemed reasonable to expect that the majority of primary data from rapidly rising numbers of GWAS publications would in time be submitted to these archives. By monitoring dbGaP, EGA and other key GWAS project websites and incorporating aggregate datasets as soon as pre-publication data release embargoes were lifted, a significant body of study data could be routinely gathered with relatively little effort. By “priming the pump” in this way, the usefulness of the HGVbaseG2P catalog and website tools could be demonstrated to the research community, and there was optimism that this would in time encourage more widespread data sharing, including via submissions directly to HGVbaseG2P.

However, this strategy ultimately proved ineffective for several reasons. First, the GWAS projects we approached directly regarding submissions to HGVbaseG2P either rejected requests for providing aggregate representations of their data (some cited intra-consortium agreements which precluded data sharing outside a closed group of partners), or else chose not to respond to our advances at all. Second, for numerous reasons routine exchange of aggregate data could not be established with the EGA, with the result that only dbGaP remained as a major datasource. Finally, after the publication of Homer *et al.* (2008), dbGaP, NCI and other data providers revised their data sharing policies and relocated all aggregate GWAS datasets from their public download sites to secure, controlled-access locations. Data providers then proceeded to strongly urge HGVbaseG2P (and other secondary data distributors) to block access to the same datasets, and this was summarily done.

Switching tactics: gathering partial GWAS datasets. The net effect of the events chronicled above was that the HGVbaseG2P catalog contained at the end of 2008 only a single complete GWAS dataset which could be browsed via the website, with little near-term prospect of acquiring additional genome-wide datasets as originally planned. A decision was thus taken to change data gathering tactics and commence with incorporating data from all published GWAS investigations, including those where only highly-significant associations were reported in the journal manuscript and no aggregate

data were available. Though far from the original aspirations of the project, considering the alternatives this strategy was considered a reasonable compromise with respect to ensuring some level of usefulness for the HGVbaseG2P catalog, and indeed to ensure viability of the project. Concurrently, work was undertaken to lobby for a less restrictive approach to sharing of safe components of GWAS datasets, based on statistical guidance regarding the validity and extents of the risks suggested by Homer *et al.*.

It was acknowledged that, given the manpower and expertise available in the HGVbaseG2P group, comprehensive literature extraction was not feasible as a means to achieve this goal. Instead, secondary sources of GWAS study information were utilised where possible. To this end, a collaboration was struck with the research group who maintain the NHGRI GWAS catalog. NHGRI provided a version of their catalog containing more detailed study information than provided on the NHGRI website. In return, the HGVbaseG2P website displays hyperlinks back to the NHGRI website where applicable. Furthermore, since the NHGRI catalog is not comprehensive, additional study data were acquired from the Open Access Database of Genome-wide Association Results (OADGAR) created by Johnson and O'donnell (2009). Though overlapping in study coverage with the NHGRI catalog, this datasource provided a substantial number of new studies, including several studies where more complete aggregate data were available, either as downloads from the project website or as supplementary data on the journal publisher's website.

To supplement the study data gathered via the two GWAS catalogs, an additional set of GWAS publications and study data not included in either catalog was identified via PubMed literature searches, in supplementary data tables, and other means. A collaboration with the 1958 Birth Cohort project²⁰ also provided GWAS data for several non-disease phenotypes, such as birth weight and cholesterol levels. Fig. §5.26 summarises the proportion of studies in the Study database acquired via each of the three approaches. Data gathering work was carried out by Rob Hastings, with minor assistance from others in our group (not including myself), using the data import tools described above.

²⁰<http://www.b58cgene.sgul.ac.uk>

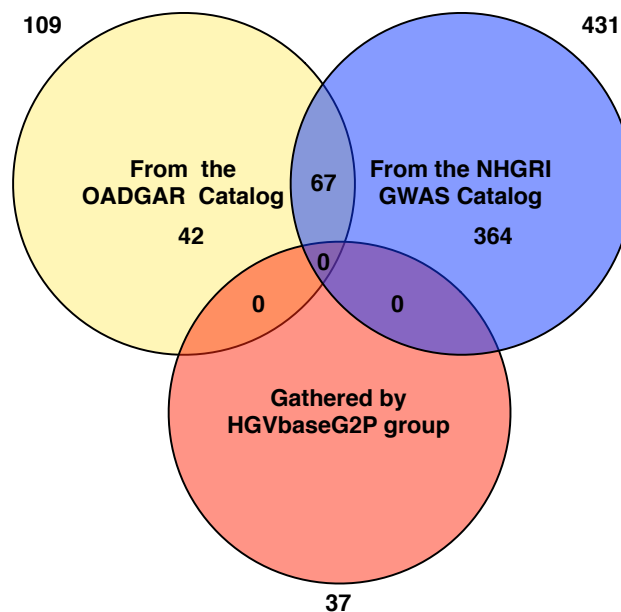


Fig. 5.26: Venn diagram showing the the number of studies contributed by each data gathering method to the grand total of 510 GWAS studies in HGvbaseG2P as of May 2010. 67 studies appear in both GWAS catalogs. Numbers outside the circles indicate total no. studies per source.

An up to date listing of all studies currently in the HGvbaseG2P catalog can be found on the Study page on the live website at <http://www.hgvbaseg2p.org/studies>. The total of ~500 studies available in HGvbaseG2P (as of May 2010) includes the studies listed in Table §5.4, albeit with the caveat that browsing is limited to metadata only, as access to the full datasets from these studies remains blocked.

5.5 Discussion

The work presented in this chapter aimed to address several shortcomings of existing online database resources for genetic association study findings. Via a collaborative effort - to which I contributed significantly - the HGvbaseG2P database and software systems were constructed, with a grounding in the conceptual data model presented in Chapter 3. The central aim of the HGvbaseG2P system, and of the GWAS data gathering efforts

undertaken chiefly by others in the group, was to provide unified access to all published and unpublished findings and complete aggregate datasets from association studies. This final section summarises the work and includes a critique of main system shortcomings and several suggestions for future work to improve and extend the system.

5.5.1 Core informatics infrastructure

Core database implementation. The conceptual data model presented in the previous chapter was used as a basis for a relational database implementation for storing reference genetic variation data, association study metadata, aggregate genotype data and association analysis results. With few exceptions, the table structures and table relations in these core databases map directly to the conceptual object model presented in Chapter 3. Notable departures from a conventional table-per-class strategy include alternative table structures to better handle high-volume, aggregate genotype datasets and deployment of a standard BioPerl feature database schema to hold genomic mapping information for markers.

These exceptions aside, the overall core database implementation is firmly grounded in the HGVbaseG2P model and, by extension, the PaGE-OM reference model. However, as noted in the previous chapter, several aspects of the HGVbaseG2P data model could be simplified, with corresponding streamlining of the HGVbaseG2P database and software. For example, the number of columns in the `Study` and `Samplepanel` tables is excessive and most of these are seldom or never used. Here, a generic attribute-value (also known as entity-attribute-value, or EAV) approach could be used for flexible annotation of data objects, possibly using a particular restricted set of attributes defined on a per-class basis. A shortcoming of the current core database implementation is the lack of a formalised connection to the object model, as explained in Chapter 3. It is non-trivial for HGVbaseG2P developers (not all of whom are necessarily well-versed in the conceptual model and all of the system intricacies) to comprehend, on a practical level, how elements of the conceptual model relate to corresponding database constructs and various software components, and how these evolve over time. Granted, comprehensively documenting of

all aspects of the system is an important project activity, and indeed existing documentation could be much improved and would certainly help in this regard. However, given that system documentation needs to be maintained separately from the relational specification proper and not linked directly to it, keeping every documentation detail synchronised with the actual implementation would be an arduous task for any group of developers. It would therefore seem that high-level, written system documentation, coupled with the formal model notation and auto-generated documentation facilities previously described, would be a superior arrangement with respect to future development of the system. Other advantages and specific suggestions for future development strategies are discussed below.

Derived databases. In addition to the core databases, two derived databases were created to support specific web-based software tools: a BioMart-compatible database intended for data mining, and a custom Browser database for storing marker and association data as genomic annotations. Though deemed necessary at the time, the *ad hoc*, non-standard design of the Browser database and corresponding software library remains a sub-optimal solution. Thus, the prospects of replacing this database with a standard BioPerl feature database or purpose-built file formats and software tools, such as SAMtools²¹ (Li *et al.*, 2009) or the indexed binary “bigWig” format²², should be investigated.

Software infrastructure and tools. A modular, extensible system comprising a number of core software components for database access, data processing and validation was constructed. Extensive use was made of third-party generic and domain-specific software packages and other tools, with corresponding savings in development time and increased standardisation (compared with a system built completely from scratch). Using this core infrastructure, a series of applications were created to process marker and study data gathered from external sources, loading these data into the core databases and other key tasks.

²¹<http://samtools.sourceforge.net>

²²<http://genome.ucsc.edu/goldenPath/help/bigWig.html>

Although the core infrastructure and tools in their present state are considered well-engineered and feature-rich, there is room for improvement in several areas. Better use could be made of the BioPerl feature database tools for manipulation and storage of genome annotation data in the core HGVbaseG2P databases. A tool needs to be implemented for checking and updating study data for changed reference marker data identified by the dbSNP-lite tool. The study data import/validation toolkit has proved very powerful for importing genotype and association data in various formats, but could be improved by decoupling data transformation from data import, by way of a two-step process involving a standardised intermediary data format (similar to that used for marker processing). Adopting emerging standard formats and tools for capturing, processing and exporting study metadata would be highly beneficial in this respect, in particular the MAGE-TAB and ISA-TAB spreadsheet-based formats mentioned in the previous chapter.

5.5.2 Website features and usability

Building on the core database and software infrastructure, a web portal and a suite of software tools for data browsing, searching and analysis were constructed. The HGVbaseG2P portal has been online and open for all to use for approximately 1 1/2 years. User traffic has increased steadily to approximately 5,000 page views per month (see Fig. §5.27), and anecdotal reports and interviews with local users indicate that website functionality and usability is generally good. Nevertheless, in order to further improve the usability of the site, ideally a broader segment of the user population should be approached regarding new and improved website features, design flaws in existing features and general usability, possibly via one-to-one interviews or via online surveys. Furthermore, a detailed analysis of user traffic data (collected into webserver logs as part of routine operation of the website) would provide a more fine-grained view of site usage patterns, further informing work to improve usability.

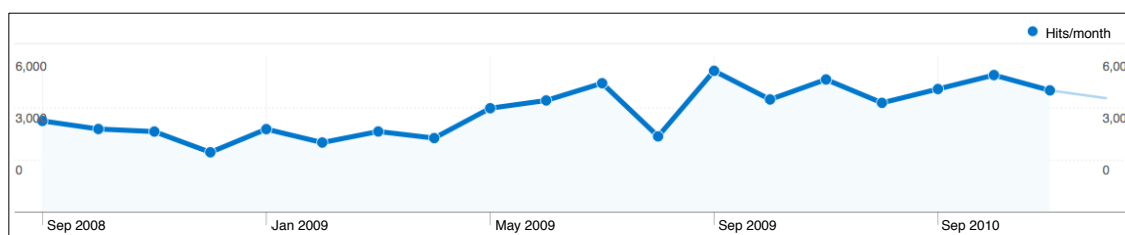


Fig. 5.27: Monthly web traffic to <http://www.hgvbaseg2p.org> over the period September 2008 to May 2010. Y-axis shows number of page hits per month. Generated with the online Google Analytics service (<http://www.google.com/analytics/>).

Web application design. As with the core infrastructure of the system, the extensive use of third-party software packages has brought significant dividends with respect to web tool development. A sophisticated, standard MVC-architecture provides a solid foundation on which the main HGVbaseG2P web application was built. The increased flexibility, extensibility and other features of this architecture compared to simpler designs have proved to be worth the increased complexity, and so the current base web application infrastructure is generally considered a good platform for future development. However, other platforms may be more suitable for certain parts of the infrastructure, as further discussed below.

Text-based searching and browsing. Several web-based tools were created to enable users to search and browse HGVbaseG2P content, with search modalities centered on each of the three main categories of data: studies, phenotypes and markers. Major emphasis was placed on creating uncomplicated, streamlined user interface to these search tools, with a special focus on powerful keyword-based searches. The current design can be considered a success with regard to these initial, general design considerations. However, a limitation of these search modalities as originally designed is that indexing and searching is entirely based on the limited content available in study, phenotype and other types of database records. Though this no different from the way Google and other search engine treat regular Web content, the limited (and largely unstructured) content does not lend itself to

the kind of “smart” searches that are needed to navigate complex biomedical information. For example, a study search for “malignant neoplasm” will only identify records where this exact search term occurs, but not records containing common synonyms such as “cancer”, nor more specific terms such as “prostate cancer”.

An obvious strategy for improvement in this regard is to extend the current search system and database content to enable semantic query expansion. This strategy utilises the knowledge formalised in domain ontologies to perform additional steps in order to enrich the search results and return more useful information to the user. Term synonyms and definitions provide additional content to search against, and term relationships facilitate basic inferencing over class hierarchies. For example, a study search for “cardiovascular disease” would return not only records where this phrase occurs, but also all studies annotated with “stroke”, “arrhythmia” and other more specific terms. A recent paper by Malone *et al.* (2010) illustrates how this technique was used to semantically enhance queries to the ArrayExpress Gene Expression Atlas²³, by way of a minimal, application-focused Experimental Factor Ontology (EFO)²⁴ cross-linked to many other domain ontologies. Current work by others in the group aims to add such functionality to HGVbaseG2P.

Graphical genome viewing tools. A series of novel, genome-wide and region-level views of association study findings were created by reusing and modifying the open-source GBrowse and other GMOD software packages. The primary purpose of these tools is to facilitate cross-study or within-study comparison of study findings, and at present this functionality is not matched by other available databases. Several related projects with some degree of overlap in functionality have been published, including standalone applications specialised for analysis and browsing of GWAS findings. A recent example is the GWAS GUI²⁵ created by Chen *et al.* (2008) which facilitates many of the same cross-

²³<http://www.ebi.ac.uk/gxa/>

²⁴<http://www.ebi.ac.uk/efo/>

²⁵<http://www.sph.umich.edu/csg/weich/>

study comparisons. GWAS GUI and similar applications, which are installed on the user's computer, can offer a level of interactivity and sophisticated graphical features that cannot be matched by web-based applications.

Though powerful as such, these applications typically have the drawback of requiring local download of all relevant data from multiple sources, and that the datasets be no bigger than the user's computer can handle. These and a number of other factors relating to ease of implementation (e.g. the need for supporting multiple operating systems) have led this project to focus on creating web applications and on the Web browser as the target platform. Nevertheless, standalone applications for G2P-focused analysis would ideally be interoperable with the web application, so users could discover and peruse study data on the HGVbaseG2P website, and subsequently download data in a common format for loading into the GWAS GUI and similar tools for more detailed analysis.

The major genome browsers all offer views of association study findings. For example, the Ensembl and UCSC browsers provides the option of switching on a feature track with diseases-associated SNPs acquired from the NHGRI GWAS catalog. However, as a rule these displays are very generic and do not offer anything close to the specialised features of the HGVbaseG2P tools. This is understandable, given that the central browsers have to cater to a wide user audience, and that association study findings are only one of several hundreds of genome annotation datasets they provide access to. Conversely, the highly-specialised HGVbaseG2P browser tools are not intended to serve as a global browser for all available genome annotations, but rather a strategically-chosen subset of the most relevant annotations (e.g. known genes, but not all EST alignments) that can be reasonably managed on the HGVbaseG2P side. Nevertheless, it is recognised that viewing study findings in the context of arbitrary genome annotations can be a powerful analysis tool, thus underlining the need for flexible export and import of feature data between HGVbaseG2P and the major browsers. This would ideally be done via standard web service protocols such as DAS as already discussed (see also below).

For future development, it is worth noting that new features are continually added to GBrowse as the software continues to be developed by the GMOD community, and these

improvements will directly benefit this project. For example, recent additions to GBrowse include facilities for visualising next-generation sequencing data, which will no doubt be useful for future sequencing-based association studies.

Federation capabilities. Two standard software packages were originally intended to provide programmatic access to database content, as first steps towards a grid-enabled HGVbaseG2P system. The web service component of the BioMart data mining tool would facilitate fine-grained API access to all marker and study data, whereas the ProServer package would enable study data to be exported via the DAS protocol and added as custom tracks to external genome browsers. However, given that the aggregate genotype data exposed by these APIs have individual identifiability implications, it has not proved possible to utilise these tools as intended and they remain deactivated until appropriate access controls have been implemented.

At present, API access to HGVbaseG2P content is limited to web feed representations of search results. The standard Atom XML-data optionally returned from website searches facilitate automated notifications of updated marker and study data content, thus providing increased utility for website users. Atom-based feed access to HGVbaseG2P may have utility beyond this limited context, however. Efforts are already underway by various GEN2PHEN partners to utilise the RESTful Atom-based web services for “lightweight” federation of LOVD-based mutation databases, as a simpler alternative to conventional “heavyweight” SOAP-based grid architecture (Thorisson, 2009c). For example, G2P database search services could be augmented with a standard OpenSearch²⁶ description file and OpenSearch-enhanced Atom XML-feed query results, thus enabling distributed search across many databases from a central portal.

²⁶<http://www.opensearch.org>

5.5.3 Data content and usefulness

An important issue to reflect on is whether HGVbaseG2P as a whole is a useful G2P information resource or not. As with other online biological databases, there is no single answer to this question, given that “usefulness” is a subjective criteria based on each user’s personal assessment of website features, balanced against the coverage and richness of the data in the database and their relationships to other available data. Given that existing resources already provide high-level GWAS cataloging services, it is pertinent to try and assess what HGVbaseG2P provides in terms of added value compared to these resources, and to also contemplate where future work should be focused.

Study coverage and data depth. The other main GWAS resource comparable to HGVbaseG2P - the NHGRI catalog - looks at first glance like a more complete (and therefore more useful) study listing. Study tallies do not tell the whole story, however. A number of entries in the NHGRI catalog are provisional and do not contain any information pertaining to study findings, raising questions about the purpose of listing them in the first place. Moreover, the NHGRI catalog is not complete, in that it does not include a substantial number of studies classified as “genome wide” by the HuGE Watch. Several of these studies have already been added to HGVbaseG2P, and current data gathering work in the group aims to continue in this direction and make HGVbaseG2P into a truly complete catalog of association studies and study findings.

Another issue is that the NHGRI listing is updated quite often, sometimes several times each week. By comparison, HGVbaseG2P is currently updated every few months, and this detracts from its utility as a way to keep abreast of the GWAS literature. A planned switch to a more frequent update cycle for HGVbaseG2P will address this issue.

In addition to overall study coverage, or catalog “breadth”, it is also useful to compare the per-study data “depth”. The extent of reported associations that are included from each study differs markedly between the two catalogs. HGVbaseG2P is in large part populated by study metadata and partial findings acquired from the NHGRI catalog and other sources,

which only includes the most significant associations reported in the primary GWAS paper. Meanwhile, further literature mining work by our group has brought in additional study metadata and reported associations from papers and supplementary materials for many of these same studies. Although this only partially redresses the problem of data accessibility and reporting bias discussed in previous chapters (see also below), in our view the increased depth of information on reported disease associations is a marked improvement over other catalogs.

Tools for exploiting catalog contents. The scientific value of gathering published GWAS findings is not disputed. For example, systematic analysis undertaken by Hindorff *et al.* (2009b) and Zhang *et al.* (2010) using data content from the NHGRI catalog and GAD, respectively, combined with G2P data from other sources, demonstrate their usefulness in preparing and publishing synoptic “snapshots” of research in the field. The simple online presentation mode employed - typically limited to hyperlinked text table views and basic filtering options - has some merit, as noted in §2.2.4. But one consequence of this simplicity is that these catalogs are far less useful for providing such integrative views on an ongoing basis, given that new findings are constantly being generated, and so comprehensive synopses and reviews become rapidly outdated.

Cross-linking provided by other parties offers a partial remedy to these limitations. For example, the GWAS Integrator²⁷ integrates various HuGE Navigator resources with the NHGRI catalog, and also provides a utility for displaying selected GWAS associations in the UCSC browser. The UCSC browser also provides basic annotation tracks for both the GWAS catalog and GAD (though the latter has not been updated since early 2008). Nonetheless, the general lack of accessible online tools for exploiting the valuable information contained in these catalogs substantially limits their overall utility. HGVbaseG2P addresses this by bringing the power of the web-based toolkit to bear on the problem. For example, though originally intended for visualising complete, genome-wide association findings, we were able to easily repurpose the specialised genome and region-

²⁷<http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do>

level graphical tools to display partial study findings. From a functional perspective, this adds substantial utility to information acquired from the NHGRI catalog and other sources.

5.5.4 Towards a model-driven, semantics-enabled architecture

As noted above, the IT technology choices made at the start of the HGVbaseG2P project have served well overall. Given the overall complexity of the system, there is now a substantial investment in Perl-based software components (notably genome browser tools) and a general long-term commitment to the Perl platform. Even so, other programming languages and platforms are now also being investigated for use in certain parts of the system. In particular, the Java²⁸-based Molgenis platform is an attractive means for migrating to a more sophisticated model-driven architecture.

Model-driven generative software. The advantages of domain-specific language (DSL) based software engineering methods were already discussed (see §2.6.1). The value of Molgenis as a tool for iterative model development and rapid, model-driven software prototyping has been demonstrated by several GEN2PHEN partners in the past 18 months, in particular for the modelling work presented in Chapter 5. A collaboration is now underway between our group and the creators of Molgenis and XGAP to jointly develop the platform itself further, and to create a modular, “pluggable” object model and software extensions of broader utility in G2P databasing. The future prospects for this work are further discussed in the final chapter.

Web-based tools for data entry and management. The DSL-based, model-driven Molgenis approach will be particularly useful for addressing the aforementioned lack of facilities for entering and editing study metadata into the HGVbaseG2P Study database (and for data management in general). Other approaches, notably the option of building such facilities into our existing Catalyst-based web application, were considered previously

²⁸<http://java.sun.com>

in the project and have some merit. For the reasons outlined above, a Molgenis-based approach was instead chosen to construct these tools.

As further discussed in the final chapter, a crucial advantage of this strategy is that any improvements to the base Molgenis system and add-on components created by this work can be contributed to the G2P community and reused in other Molgenis installations if others find them useful. Conversely, specialised components created by others, such as custom user interface widgets for looking up terms in bio-ontologies, can be reused for this project (e.g. for ontological annotations of phenotypes). Text mining of study metadata in order to identify putative ontology terms, such as the method employed by Malone *et al.* (2010), may be a particularly useful complementary strategy to such semantic annotation facilities.

Publishing on the Semantic Web. As noted in previous chapters, semantic technologies and standards are emerging as important tools to address data integration problems which are hard or impossible to solve via traditional means. An important area of future work in the HGVbaseG2P project is to explore ways of exposing database content as Linked Data (see §2.6.2). Rather than aiming for a fully-fledged RDF-based triplestore solution, a less radical strategy would be to leverage available tools, such as D2R Server²⁹ or Virtuoso³⁰, which serve as RDF “front ends” or views to traditional relational databases. Such relational-to-RDF technologies could be deployed on their own, but of particular interest in this regard is the potential for leveraging them indirectly via Molgenis as the facilitator. For example, rather than manually generating relational-to-RDF mapping files for the D2R Server, the DSL-based code generation framework can in principle be used to auto-generate these mappings, thus enabling RDF-publishing without any extra effort. An important issue for consideration in this regard will be how to transform the object model (as specified in the Molgenis DSL) into an OWL/RDFS ontology, for lending structure and semantics to the RDF-ized data (see also §3.3.2).

²⁹<http://www4.wiwiw.fu-berlin.de/bizer/d2rmap/D2Rmap.htm>

³⁰<http://virtuoso.openlinksw.com>

A dual-platform architecture. The preceding paragraphs may seem to argue for a wholesale change from the Perl platform to a Java-based one, given the plethora of powerful technologies available for the latter which are becoming important for future development. Indeed, many of the in-progress or suggested areas of future work discussed here, whether it be iterative model development or Linked Data publishing, would be far harder to undertake using only the existing Perl-based infrastructure. However, due to the aforementioned investment in construction (and reuse of) Perl-based software components, it is impractical to migrate completely to a Java-based architecture. But this does not at all preclude a Perl-based web application, browsers and other tools and a Molgenis-based application running in parallel, both interfacing with the same database using a common data model. In fact, such a dual-platform strategy would enable HGVbaseG2P to continue to take advantage of sophisticated genome browser packages and other tools which are not available for the Java platform, and combine them with the aforementioned Java-based technologies.

6. Exploring the role of digital identity in data publication

The data gathering challenges described in the previous chapter highlighted the need for access controls in the HGVbaseG2P system, a feature not considered in the original design which assumed completely open access to all association study data. This prompted me to investigate authentication and authorisation techniques and open-source software components that would be suitable for implementing the required functionality. From this initial analysis came the realisation that the notion of *identity* on the Internet would not only be crucial to addressing GWAS data sharing challenges, but could potentially have important implications for G2P databasing and dissemination of research data in a wider context. I deemed these findings important enough to warrant an expansion of the initial project plan (centred on data modelling and HGVbaseG2P development) to include an exploration of the the role of identity in data publication.

The two first sections of this chapter provide a summary of my analysis of key concepts relating to identity on the Internet and the main relevant technologies. The next section introduces some of the ideas thereafter developed - both independently and via engagement with the research community - regarding the application of identity in two main problem areas: i) reliable attribution of published datasets and other digital contributions and ii) practical access management for sensitive G2P data. The final section summarises work undertaken to develop formal use cases for a variety of identity-enabled online data access and data submission scenarios, as a precursor for future software implementations.

6.1 Identity on the Internet

The term *identity* is notoriously ambiguous, with many interpretations depending on the context in which it is applied and what it refers to. For example, a person's notion of his/her own ethnic, religious or national identity is quite different from a government's administrative definition of that person's identity. Philosophical speculations into personal identity and the existence of self invoke a different notion of identity still. The following definition of identity is useful as a starting point in the context of identity on the Internet:

The collective aspect of the set of characteristics by which a thing is definitively recognizable or known.¹

This pragmatic definition is broad enough to also apply to non-person entities: for example, Chapter 2 briefly discussed identification of biological objects. However, the topic of this chapter is identification of persons - specifically researchers - on the Internet, and so we can usefully paraphrase the above and say that identity is information on the Internet that is associated with individuals and can be used for identification.

As Internet users, or Net citizens (the term "netizens" is also sometimes used), we make a substantial amount information about ourselves freely available on the Web, via personal websites, professional networking services such as LinkedIn² and various social networking sites such as Facebook³ and MySpace⁴. To maintain our privacy, we release certain types of information in controlled way, such as our full Facebook profile and activity stream to only our close circle of friends, or our name, home address and credit card information to online retailers when purchasing goods. Yet other kinds of personal information are not provided or controlled by ourselves, but rather by third parties. Some of this information is publicly accessible; for example, a scientist's publication record can be found in bibliographic databases. Other information is tightly access-controlled, usually

¹<http://www.thefreedictionary.com/identity>

²<http://www.linkedin.com>

³<http://www.facebook.com>

⁴<http://www.myspace.com>

for good reason; for example, our driver's license details are held by government agencies and are not available to the general public, but are by necessity accessible online by certain parties, such as law enforcement officers. Irrespective of the source or level of access, when considered in its totality, all this information collectively amounts to a digital profile which describes us - a *digital identity*.

The rise of Web 2.0 and social networking. If we disregard for the moment the various types of personal information which we do not control directly, we can define online identity more narrowly as a Web presence - what we choose to reveal about ourselves online and to whom. The notion of an online presence has been at the core of the Web itself from its beginning, starting from simple personal homepages, but only in the last few years has it gone mainstream and been adopted by the masses of people with Internet access. According to Internet World Stats⁵, the total number of Internet users is approximately 2 billion. At least 500 million of these visit Facebook, MySpace, FriendFeed⁶, Yahoo⁷ and other social networking websites every day, to interact with their fellow netizens in self-organising online communities. These community websites and the companies operating them are the vanguard of the so-called Web 2.0 revolution⁸, which has transformed the Web from principally static content intended for passive viewing to a highly interactive experience characterised by online collaboration, information sharing and user-generated content.

Reputation systems and trust. Another way to think about online identity is in terms of *reputation*. Online reputation systems have existed in one form or another since the early Web era and now underpin major e-commerce enterprises, discussion forums and related websites. The best known online reputation system was created in the mid-1990s by eBay⁹,

⁵<http://www.internetworldstats.com/stats.htm>

⁶<http://friendfeed.com>

⁷<http://www.yahoo.com>

⁸<http://oreilly.com/web2/archive/what-is-web-20.html>

⁹<http://www.ebay.com>

which uses buyer/seller feedback and ratings in their online auction forum and marketplace to encourage good behaviour in transactions and to weed out “bad actors”. In his analysis, Ubois (2003) considered eBay and several other reputation-based services, concluding that systems in which reputation approximates reality is the key to establishing trust in online interactions between individuals.

Online reputation is also at the core of Wikipedia, where authors can build a reputation as page editors on their own merit, irrespective of academic credentials or other real-world qualifications. As an example of measures for quantifying reputation, the system proposed by Adler and de Alfaro (2007) is based on an objective measure of reputation which incorporates information about a Wikipedia author’s text addition and how long- or short-lived these revisions are.

The problem with names. Reputation, however, is meaningless without a way to verify the identities (real or pseudonymous) of the parties involved in the trust relationship. For example, Frishauf (2009) points out that if reputation-based systems for scientific peer-review are to be established in the future, identification is essential for reliably associating editors and reviewers with their expertise and with their reputation or track record. As in the real world, the non-uniqueness of person names makes them unsuitable as a means for identification, which a simple Google search with a common name as keywords quickly demonstrates: determining which of the myriad search results returned are associated with the specific person of interest is often far from straightforward. In a global online community of hundreds of millions Internet users, better methods for identification are therefore required for inter-individual transactions where reputation, trust and personal accountability are important.

Fragmented identity. eBay and other community websites deal with name ambiguities by assigning their users personal identifiers which are unique within the system, and used *in lieu* of names for looking up profile information, tracking reputation and when dealing with other local users. These identifiers serve as pointers to the particular subset of our

personal information stored on the system, and enable us to communicate with other local users about our identity.

Whilst useful within the confines of each individual site, this does not address the global identification challenge. One problem is that the plethora of community (and other) websites have historically not been connected with one another. From the business perspective of the website operator, this makes perfect sense; the more time the user spends on the site and the more personal information he or she supplies the vendor with, the more potential for advertising revenue (or whichever revenue model is employed by the vendor). But from the perspective of users, many of which make use of several of these services, this lack of connectivity has numerous disadvantages. The first, and most obvious, is the inconvenience of having to create and maintain a separate user account on each site one wishes to utilise. “Password fatigue” - the tendency to use the same username and password on multiple websites - is a serious security concern; if a malicious person gets a hold of a person’s login credentials for one site, he can use the same credentials to access the victim’s accounts on the other sites as well. Finally, because various pieces of our personal information are stored in separate user profiles on each of these unconnected sites, the net effect is that our online identity is spread out, or fragmented, across a multitude of *identity silos*. This makes it challenging to consolidate all information about us in a way that accurately and comprehensively represents us on the Web, to the extent that we wish to do this.

Privacy and security. An inevitable consequence of increased use of the Web for personal and professional interactions with other netizens is the increased amount of personal information we place on the Web, and with it the opportunity for theft and misuse of this information. Bilge *et al.* (2009) recently demonstrated how identity theft attacks (i.e. account cloning) and illicit access to personal details on various social networking sites could be relatively easily automated via a combination of open-source software and custom scripts. As an example of less sophisticated identity theft involving so-called impersonation attacks, in 2008 a network of fake Facebook profiles for over 100 stem-cell scientists was

created for an as-yet unknown purpose (Laursen, 2009).

Another problem is that our personal information is in many cases controlled by (and often the property of) the for-profit companies who operate the various social networking sites. We have limited power to influence changes in these companies' data release and data use policies we do not agree with, except, of course, to stop using (and thus stop reaping the benefits of) the service. As an example of this, Facebook recently changed their policy to make available for public view which websites their users have indicated they "like" - a piece of information previously visible only to a user's designated friends. This seemingly innocuous piece of information is nevertheless potentially damaging to the individual if seen by an employer or the government, for example if a user "likes" the website published by a controversial political activist movement. The company's decision to treat this as an "opt-out" feature - that is, users have to explicitly hide this information via Facebook's complex privacy management configuration¹⁰ - has resulted in a significant backlash in the online community, with many users choosing to delete their Facebook accounts (see recent Wired article published online¹¹).

The above illustrates several key issues. First, it seems inevitable that we must give up some of our privacy in exchange for what are often substantial benefits and convenience from using Facebook and related tools. Second, there will always be some risks involved in using such tools and we can never be 100% sure that privacy breaches will not occur. Third, a major challenge in the domain is to devise privacy controls and policies that are i) sufficiently flexible to allow fine-grained control of how personal information is to be shared with which third parties, and ii) easy enough for users to comprehend and use effectively. Otherwise, a "Hobson's choice"¹² situation is created, whereby confused users have to choose between the lesser of two evils and mistakenly divulge more personal information than they intended. All of the above becomes increasingly important as we

¹⁰<http://www.nytimes.com/interactive/2010/05/12/business/facebook-privacy.html>

¹¹<http://www.wired.com/epicenter/2010/05/facebook-rogue/>

¹²<http://www.phrases.org.uk/meanings/hobsons-choice.html>

move beyond the traditional single-site online community context and into a distributed, Web-wide environment, where the use of federated identity is likely to be commonplace.

6.2 Federated identity technologies

On a practical level, digital identity is facilitated by a collection of Internet standards and technologies, the primary function of which is to provide a reliable means of identification. In this section I outline the most important of these technologies and key related concepts.

Distributed identity systems. The most basic means of identification over the network is signing into a website using a username and password in order to prove one's ownership of the user account on that site. An important aspect of this which I will not go into here (but see the ICGC case study in §6.3.2) relates to how a user account is created in the first place, involving in some cases (e.g. online banking) verifying one's real, "offline" identity with a letter or telephone call. As already noted, local account credentials (hereafter referred to as ID credentials) are only useful in the restricted context of that particular identity silo. A key aim of distributed identity systems is to link together disparate sites by means of *federated authentication* - the process of signing on to an Internet site using ID credentials from another site, the ID provider. Briefly, this typically involves the following steps illustrated in Fig. §6.1:

- i) On the first visit to the target website (the relying party), the user indicates that he wishes to use his digital ID to sign in, rather than creating a new account in the conventional way.
- ii) The user's web browser is redirected to his ID provider website, where he is asked to prove his identity with a username and password (or some other, stronger form of authentication such as a smart card, see below).
- iii) If the authentication step is successful, the user is returned to the target site, already signed on.

As long as he remains signed in to the ID provider site in the browser session, the user does not have to re-authenticate on subsequent visits to the target site. This is much the same as

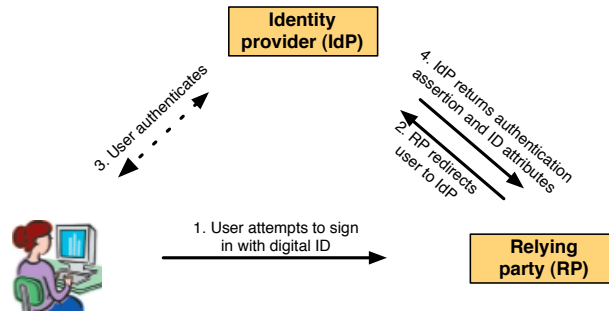


Fig. 6.1: Simplified illustration of federated authentication. The user attempts to sign into a relying party (RP) website (often referred to as service or resource provider). The user is redirected to his identity provider (IdP) where he proves his identity, and is subsequently redirected to the original site. Importantly, the RP never sees the user's ID credentials, but receives only an assertion from the IdP stating that the user to whom a particular identifier refers has authenticated successfully. In Step 4, the IdP will typically also send additional profile information, if requested by the RP and if the user consents to share this information.

happens with regular websites which use browser cookies to maintain a persistent, signed-in session between site visits. The crucial difference is that the user remains authenticated with the same ID at all the other sites where he has signed in. Such *single sign-on* (SSO) across many websites with the same ID, and the provision of a portable Internet identity profile via a central provider, are among the chief benefits of federated authentication. The importance of this will be further illustrated in the use case discussion in §6.4.

Organisation-centric identity. Most federated identity systems have been deployed in industry and in the higher-education (HE) and research domain. The main purpose of these systems for managing and using organisation-issued IDs is typically within- or cross-institutional SSO and managing access to shared, distributed resources within an organisation, or amongst multiple members of a virtual organisation (VO). Many of these systems are based on public-key infrastructure (PKI) and related technologies¹³ created and standardised in the 1980s and 1990s. PKI is based on public-key encryption and digital certificates issued by central authorities and is used widely to implement security

¹³<http://www.sun.com/blueprints/0801/publickey.pdf>

in distributed network environments. Related, decentralised approaches such as Pretty Good Privacy (PGP)¹⁴ based on “webs of trust” and exchange of encryption keys between individuals have also been developed.

An emphasis on handling a wide range of “enterprise-centric” use cases, with a strong focus on security, has tended to make these traditional systems complex and costly to implement and maintain, and difficult to use. A major challenge in this domain therefore concerns how to deploy federated identity technologies in such a way that is sufficiently secure and flexible for a given purpose, and at the same time user-friendly and not unreasonably difficult to manage (Harding *et al.*, 2008).

An example of a large identity federation is InCommon¹⁵, a consortium of over 100 US higher education and research institutions, scholarly publishers and government entities in the US. In the UK, the Athens¹⁶ system is widely used in the higher education and research domain and the National Health Service (NHS)¹⁷. Athens is a suite of software products and services, provided by the non-profit Eduserv organisation¹⁸ since 1995, which enables organisations to leverage federated access and identity management with minimal local investments in informatics infrastructure and expertise.

A common use case for the Athens service is provision of off-campus access to scientific journals with institution-based subscription; university students and staff click the “Athens login” button on a journal website and are directed to the institution website for Athens-based authentication, as per the SSO scenario outlined in Fig. §6.1. As of 2007, Eduserv has extended the Athens framework to create the OpenAthens service¹⁹, which provides interoperability with other federated identity standards such as the Shibboleth open-source

¹⁴<http://www.openpgp.org>

¹⁵<http://www.incommonfederation.org>

¹⁶<http://www.athensams.net>

¹⁷<http://www.nhs.uk>

¹⁸<http://www.eduserv.org.uk>

¹⁹<http://www.athensams.net/products-services/OpenAthens.aspx>

platform²⁰, the Security Assertion Markup Language (SAML)²¹ and related protocols. The key advantage of this “outsourcing” of federated identity is that the organisation and, importantly, the user need not be concerned of the intricacies of the various underlying, complex technologies.

Despite advances in usability, in particular the web-based SSO scenarios like the one described above, a disadvantage of these traditional identity systems is their organisation-centricity and limited scope. Upon joining an organisation, the individual (e.g. student, employee) is assigned a set of ID credentials which are valid while he or she is with the organisation. When the individual departs (e.g. graduates, switches jobs), the ID - analogous to an organisational E-mail address - ceases to be valid and the link between the person and his/her organisational identity is broken. This is a major drawback for people who frequently move between organisations, are members of multiple organisations, or are perhaps not affiliated with any organisations. Organisation-issued IDs also only work on certain Internet sites; for example, a student at the University of Leicester can use his institutional ID to access a journal website via the Athens system, but not for signing in to the NCBI PubMed website for accessing his literature search history. Though effective and widely used for identification across sometimes very large federations of hundreds of ID providers, such as the UK Access Management Federation for Education and Research²², such IDs are therefore of limited utility as a universal means of identification.

Decentralised identity systems. Efforts aimed at addressing this challenge of global online identification are based on assigning identifiers to persons which are unique across the entire Web, as opposed to within local systems or federations only. Early attempts to establish such systems with centrally-assigned IDs were not successful, largely because of concerns over any one entity - especially a for-profit company - storing and controlling personal information for all Internet users. An example of this is Microsoft’s failed

²⁰<http://shibboleth.internet2.edu>

²¹<http://saml.xml.org>

²²<http://www.ukfederation.org.uk>

attempt in the late 1990s to dominate the mainstream online identity space with their centralized Passport SSO service²³. Microsoft's service did not engender trust in the wider online community, neither with end users nor website developers, due to factors such as the proprietary, closed nature of the technology and security concerns (see e.g. Oppliger (2004)). As a result, Passport did not see widespread adoption beyond Microsoft's own family of websites.

Later efforts aimed at building open authentication platforms, such as the Liberty Alliance²⁴, have addressed centralisation concerns by way of a distributed identity architecture. A key feature of this architecture is multiple, rather than just one, identity providers, from which users can choose to host their identity. But as Weitzner (2006) notes in his review, none these systems have been suitable (or are yet mature enough) to serve as a general-purpose, global *identity infrastructure* that i) scales with the Web, and ii) is user-friendly enough for the general public to use for secure online communications.

A recent flavour of decentralised identity systems based on globally-unique URIs as person identifiers is enjoying a better run of success. In a more recent review, Weitzner (2007) outlines some of the key features of these systems and their significance in facilitating rapid, widespread adoption in the Web 2.0 space in the past 3-4 years. A major driver is the utility of the new systems in bridging the identity silos and enabling *user-centric* identity (Maler, 2009), a new concept which is currently attracting a great deal of interest in the online community. User-centric identity refers to the empowerment of Internet users to link together their identity profiles across the various websites they use and, critically, control which personal identifier(s) and potentially-identifiable information represents them online and how this information is shared with other parties. This focus on the needs and wishes of the individual, rather than the needs of an organisation, is one of the defining characteristics of this new brand of online identity.

Key to this movement towards user-centric, federated identity is widespread adoption

²³<http://www.passport.com>

²⁴<http://www.projectliberty.org>

of new Internet standards, notably OpenID²⁵ (Recordon and Reed, 2006). OpenID is a federated authentication protocol (see Fig. §6.1) designed to be simple to implement and secure enough for many common use cases. The philosophy behind OpenID (and indeed many other successful Internet standards) is focused on creating a multitude of small, interoperable specifications which are deployed on an as-needed basis. This modular approach - focused on scalability rather than security - greatly eases the task of building user-friendly, identity-enabled applications with modest security requirements, compared to using traditional “heavyweight” identity technologies. The main use case for OpenID - transfer of the user’s own personal data across various social networking websites - is an example of a key use case scenario where such a simple trust model is sufficient. The simplicity of the protocol leads to ease of implementation, and as a result this and related use cases in the social networking arena are currently driving adoption of OpenID and related technologies, helped by backing from major Web 2.0 companies and organisations such as Google, Yahoo, Facebook and others.

SSO and privacy/security concerns. The numerous advantages aside, SSO introduces new - and amplifies existing - Internet security and privacy concerns. One of the most important of these relates to the inherent risk in using a single set of credentials to access a multitude of online services. If a user’s account credentials with the central ID provider are compromised, then an unscrupulous hacker would be able to access all the sites where this ID has previously been used to authenticate. Moreover, the user’s profile with the master ID provider would typically hold a rich collection of personal information. All this information, which as already noted is currently fragmented across numerous identity silos, would now be conveniently accessible to the hacker in one place.

Clearly, putting all one’s eggs in the single proverbial basket can potentially have serious consequences for the user. Indeed, federated identity or not, the above are but a few of many issues relating to centralisation of personal information in general (as illustrated by the aforementioned controversy over privacy on Facebook). However, it should be

²⁵<http://openid.net>

emphasised that the present Internet security situation is far from optimal. A general lack of awareness of security issues in the broader user population is a huge problem, exemplified by password fatigue as already mentioned. Security practices on popular mainstream website also leave much to be desired. For example, a typical Internet user's E-mail inbox contains password reset instructions for the various sites where he has registered. A hacker would therefore only need to break into the master E-mail account in order to gain control of the user's accounts on those sites.

Considering the above, it would seem that a centralised ID is no less secure than the present situation. Consider also that moving user registration and sign-on away from individual websites to dedicated services specifically designed for managing online identity can actually lead to overall increased security. The element of choice is important in this regard: a user can, for example, better protect his central identity by choosing a provider which offers stronger forms of authentication than just a username and password (see LoA below). Verisign²⁶ and TrustBearer²⁷ are two examples of OpenID providers which offer such enhanced services. Providers will also often provide other useful identity management services, such as auditing tools which track websites where the user has authenticated with his ID and allow him to remove a website from his list of trusted sites if he later finds that site to be untrustworthy. All of these features are improvements over the present situation, where account credentials are stored in the various identity silos with varying levels of protection, over which users have no control.

Authorisation and trust. Leveraging digital identity and federated authentication for access control involves at its core three distinct processes, which often take place in physically different locations on the Internet:

- i) Verifying the identity of the user, by way of federated authentication as described above.
- ii) Storage and retrieval of access privilege attributes, or assertions, relating to the identity of the user.

²⁶<http://pip.verisignlabs.com>

²⁷<http://www.trustbearer.com>

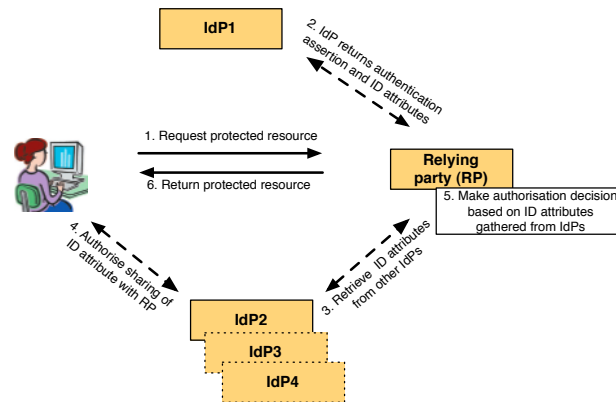


Fig. 6.2: Simplified illustration of federated authorisation. User is already authenticated via her primary identity provider (IdP1) as shown in Fig. §6.2. The ID attributes available to the relying party (RP) from IdP1 are insufficient to authorise her for access to the resource. Once additional attributes have been retrieved from secondary sources (e.g. an institutional IdP confirming the person's Principal Investigator status), the user is authorised. Attribute gathering would typically involve each IdP asking user to approve sharing specific information with the RP, and possibly additional authentication if the user is not already signed into the secondary IdPs.

iii) Authorization, or determining based on available ID attributes whether or not to grant the user access to the resource.

This is illustrated in Fig. §6.2. Importantly, attribute assertions - often referred to simply as identity credentials - on which the access control decision is based may be sourced from up to several Internet locations. Trust and identity assurance are critical this process: the resource provider needs some way of assessing the certainty that the user is who he says he is, and that the credentials presented are trustworthy. This is important not only in the initial authentication step, but also in subsequent steps when credentials are retrieved from various other identity providers where the user has an account. A particular distinction is made between credentials issued by the user, or self-asserted, and those issued by a trusted authority (e.g. the government or a research organisation). This is based on the common-sense premise that, for certain kinds of information, users should not be trusted to provide accurate information about themselves (e.g. professional qualifications and criminal records).

As with authentication protocols, a variety of standards and technologies have been

developed for managing access to resources in distributed network environments. Traditional role-based access control (RBAC) based systems such as PERMIS²⁸ Chadwick *et al.* (2008) can be leveraged with other grid technologies and standard software components, such as Globus Toolkit²⁹ and GridShib³⁰, to create flexible, fine-grained authorisation schemes. However, these sophisticated technologies tend to require expert knowledge to install, configure and manage, making them difficult to deploy. Watt *et al.* (2009) discuss some of these issue in their paper describing efforts to create identity-based security solutions for federated data services for two major UK-based e-Science projects. More generally, shortcomings of RBAC-based approaches in complex, distributed scenarios have motivated development of alternative federated authorisation approaches. Of particular interest to my work is the Advanced Grid Authorisation through Semantic Technologies (AGAST) framework recently developed by Sinnott *et al.* (2009). AGAST features prominently in the use cases presented later in this chapter, and will be introduced in more detail there.

Identity linking and attribute aggregation. A particular challenge in complex authorisation scenarios is how to gather all the required ID credential information. To support multiple sets of credentials, grid security solutions often require users to upload digital certificate files from their local computers, but this introduces various usability and security issues. The identity linking service proposed by Chadwick and Inman (2009) is an interesting approach to address this problem. In this model, implemented in the Shib-Grid Integrated Authorization (Shintau) project³¹, the user links together his various IDs at different providers to create a sort of “composite” identity. This composite identity, containing so-called attribute referrals to the various ID providers, is then presented to the resource provider for authorisation. The resource provider follows the digitally-signed

²⁸<http://www.permis.org>

²⁹<http://www.globus.org>

³⁰<http://gridshib.globus.org>

³¹<http://sec.cs.kent.ac.uk/shintau/>

referrals to aggregate the required attributes from each ID provider, and finally makes the access control decision. Though the method is still at the developmental stage, this concept of linked identities may have utility in some of the access control scenarios discussed later in this chapter.

Authentication Level of Assurance (LoA). An important factor in many authorisation scenarios is the strength of the authentication method employed. So far I have only mentioned simple authentication with a username and password - a method which is easy to use and simple for non-technical Internet users to understand, but at the same time easily compromised. If an unscrupulous party discovers the password, he can then use this information to sign into a user's account and, in, effect hijack his identity. This is particularly a concern with centralised accounts on ID providers, as noted above.

Stronger authentication methods address these limitations and thus enhance security. For example, multi-factor authentication involves at least two factors (a factor is something the user has or knows or is), often a password combined with a digital PKI certificate physically located on the user's computer. An even stronger form of authentication involves a physical factor, such as a password combined with a PIN-protected smart card or a random key generator. As an example of mainstream use of the latter method, all online banking websites in Iceland have for several years required (in addition to a username and password) a one-time password from a random key generator linked to the person³².

A scheme for authentication LoA has been devised to represent the degree of confidence in the identity of the user (Nenadic *et al.*, 2007). In this scheme, an LoA of 1 (username and password) is the weakest and an LoA of 4 (multi-factor, hardware-based) is the strongest. When standardised and codified in this way, authentication LoA can be usefully incorporated as a parameter in access control decisions. For example, a data provider would be able to specify that access to identifiable individual-level G2P data should only be

³²<http://www.landsbanki.is/english/Personal/PersonalInternetBanking/AuthenticationKey/>

granted to users who have authenticated with an LoA of 2 or greater, as well as possessing additional attributes (e.g. a data access permit).

OpenID and security plug&play. Given their emphasis on trust and identity assurance, Internet security technologies and protocols such as Shibboleth operate under a federated trust model: the resource provider trusts only a limited set of identity providers (i.e. those in the federation, as described above). Conversely, OpenID assumes a much simpler open trust model, where the user can authenticate using any identity provider, and it provides no mechanism for exchanging credentials other than self-asserted. For example, if an OpenID user self-asserts that he is affiliated with the Royal Bank of Scotland, the resource provider has no means of ascertaining that this is indeed true, and therefore this assertion does not comprise reliable information on which to base an access control decision. This lack of a trust fabric in the OpenID protocol is intentional, and a direct result of the philosophy that underlies its design as noted above. However, this means that for use cases where trust and identity assurance are important, OpenID by itself is not sufficient, and thus enterprise-level features provided by “heavyweight” technologies will be required.

As Maler and Reed (2008) discuss in their review, user-centric identity can be implemented in a number of ways and with a variety of technologies, depending on the trust and security requirement of the problem at hand. Optimal solutions for many use cases may comprise a hybrid approach using OpenID for authentication (easy to use and implement) combined with other frameworks, such as SAML, for exchange of attribute assertions, digitally signed if required (to verify the source of the information and ensure that the information has not been tampered with). When used in this limited way, OpenID may be feature-rich enough to offer sufficient security for many use cases. For example, the Provider Authentication Policy Extension (PAPE)³³ enables the relying party to specify that only IDs authenticated with certain methods (e.g. multi-factor) are accepted, thus providing a way to enforce a specific LoA for increased security.

³³http://openid.net/specs/openid-provider-authentication-policy-extension-1_0.html

6.3 A digital identity for researchers

A presence on the Internet is perhaps even more important to professional scientists than to the general public. The primary traditional measure of scientific reputation - the publication record - is already digitised and available online via bibliographic databases, though the aforementioned name ambiguity problem causes significant difficulties in determining which authors have contributed to a published work. These problems are further amplified in Web searches; distinguishing between scholars and non-scholars sharing the same or similar common name in Web content returned from Google searches is often far from straightforward.

Further to the issue of name ambiguity and attribution, researchers' scientific outputs increasingly comprise more than just traditional journal publications. Researchers working in large teams or as individuals generate datasets which are submitted to central databases, and the work of data curators is important to ensuring and enhancing the quality of published datasets (see §2.3.3). New forms of digital contributions are also emerging, including Web 2.0 activities such as science blogging and scientific wiki authoring. All of these contributions can impart significant scientific impact and thus contribute to a scientist's professional online reputation. But as they do not confer publication credit in the traditional sense, such important contributions go largely unrecognized when it comes to assessment for career promotion or grant funding. Cultural momentum and overall slowness to adapt centuries-old conventions of the scholarly literature to the modern, digital world are some of the reasons for this state of affairs. But a fundamental hindrance to meaningful progress towards recognising digital contributions to science is the lack of robust means for verifying the identity of each researcher, and for comprehensively attributing all his scientific contributions to this identity.

It is undisputed that access to identifiable biomedical data must be carefully controlled, but as noted in previous chapters current measures to achieve this are generally at a local level (i.e. independently at various central archives and/or projects), typically involve bureaucratic procedures to gain access to the data, and are non-interoperable. While there

are good reasons for these measures, they already impede the rate of research progress by curtailing broad data integration and meta-analysis, whether by individual scientists or by secondary data providers which add value to, repackage and redistribute the data. Global access permit systems, as well as unified means for identification are key to streamlining data access. Therefore, here too online identity for researchers is likely to be important to progress beyond the status quo.

Community outreach. Considering the above issues relating to name ambiguity, attribution and access control, I concluded from my analysis that online identity and online reputation should be a central concern to researchers as a community going forward. Based on this conclusion, from early 2009 to summer 2009 I made significant efforts reach out to and engage with the community regarding this important topic. As a first step, I co-authored a whitepaper (of which the preceding paragraphs are a condensed summary) which was published online in February 2009. This whitepaper was also distributed to and discussed with various partner organisations via E-mail.

Responses from this initial round of community engagement were encouraging, so I proceeded to co-organise an international workshop titled Identifying Researchers on the Biomedical Web (IRBW2009). Details on the workshop programme, meeting minutes, an executive summary and other materials are available on the GEN2PHEN Knowledge Centre website³⁴.

As well as providing additional background, the subsections to follow discuss the main outcomes of the IRBW2009 workshop, and also incorporate further developments and thinking on this topic that have since been formed. Much of this material has also been published online in a compact and less developed form as a series of blog posts and articles, mainly via the Researcher Identification group on the GEN2PHEN website³⁵.

³⁴<http://www.gen2phen.org/event/irbw2009-workshop-may-13-14-toronto>

³⁵<http://www.gen2phen.org/groups/researcher-identification>

6.3.1 Attribution and accreditation for scientific contributions

In order to understand how digital identity and unambiguous attribution can help to incentivise data publication, it is helpful to briefly consider the “publish or perish” culture in science. Scholarly communication has become deeply intertwined with the process of measuring or assessing scientists’ contributions, and the abstract notion of *scientific impact* and ways to quantify it is therefore of great interest to many stakeholders in research. The centuries-old tradition of acknowledging contributions by others by citing their work is central to measuring impact. Citation cross-linking data are the raw material for calculating commonly-used metrics such as the Journal Impact Factor (JIF)³⁶, originally devised by E. Garfield over five decades ago (Garfield, 1955, 1999), which measures per-journal average citation rate based on articles published in the previous two years. The JIF, and over-reliance on it as the principal means for measuring impact, has long been criticised and is considered a major contributor to the present situation where the pressure is on scientists to publish in leading journals (Cockerill, 2004; Editors, 2006). Lawrence (2003) argues that the resulting emphasis on *where* the research is published, rather than the quality of the research itself, has detrimental results on the presentation and quality of the published research. In an attempt to address this, other measures for measuring impact are being developed (Bollen *et al.*, 2009a,b). But as Neylon and Wu (2009) and others argue, it is paramount that, whichever metric or metrics are used, they be centered on individual publications, on the basis that a journal-based impact factor is not a good indicator of the impact of articles published in a given journal.

The data citing problem. Given the importance of citations and impact measures as a motivating factor in science, Costello (2009) and others have argued that treating online data releases as publications is key to tackling numerous issues surrounding research data availability. The general idea is as follows: if a data publication is likely to increase the overall impact assessment for a researcher in the same or similar way that a journal article

³⁶<http://science.thomsonreuters.com/citationimpactcenter/>

does (assuming a publication-level, rather than journal-level, citation metric), he would have an incentive to make the extra effort to prepare, package and publish/share his data. However, a significant obstacle to this strategy has been a lack of a standard mechanism for citing datasets, and subsequent difficulties in systematically tracking data citations and entering them into the permanent scholarly record. Citing database identifiers (e.g. sequence accessions) in the article text is an established convention and required by many journals, but this method is only applicable where reference archives have already been established, and furthermore does not facilitate giving credit to the original data contributors in a structured way.

One method for crediting the data creators is to cite the main journal publication associated with a dataset; for example, users of HapMap data typically cite the primary HapMap paper (The International HapMap Consortium, 2005). But this method has many shortcomings. First, citing a specific dataset version (if data have been updated post-publication) is not possible, nor is there a way to cite a specific dataset out of several datasets described in the paper. For example, the HapMap project has released several dozen versions, or ‘freezes’, of data generated in the project, and information about the specific data freeze used in an analysis is typically provides as free text in the manuscript and is thus not captured in the scholarly record. Another problem results from the common practice in “Big Science” projects to attribute publications to large consortia (e.g. The International HapMap Consortium), rather attributing dozens or hundreds of persons as authors, resulting in obscured individual contributions to a potentially high-impact data publication³⁷. A third, and perhaps rather obvious, downside is that datasets not associated with a publication cannot be cited in this way, although so-called “marker” papers which are now being piloted will hopefully help with this Peterson and Campbell (2010).

Internet references address some of the issues above. Given that research data are increasingly accessible via the Internet, a URL reference is a straightforward method to

³⁷An interesting, and extreme, example of this from particle physics research is a recent paper on analysis of LHC data with over 2,000 contributors and 170 institutions from around the world (see [http://dx.doi.org/10.1007/JHEP02\(2010\)041](http://dx.doi.org/10.1007/JHEP02(2010)041))

cite particular online dataset. URL references, however, come with several caveats: a dataset hyperlink is indistinguishable from any other web reference; a URL citation does not credit the data contributors in a structured way; and web hyperlink decay or “link rot” in the literature is a well-documented problem (Dellavalle *et al.*, 2003).

Several persistent identifier schemes have been devised to address these problems relating to Internet references. One of these schemes - the DOI system - is now near-ubiquitous in the scholarly and professional publishing domain. CrossRef³⁸, a not-for-profit association comprising over 700 publishers, libraries and other stakeholders, provides technological and social infrastructure which supports identifying, locating and citing digital resources via DOIs. Nearly all scientific journals now assign DOIs to published articles, and most (but not all) publication-to-publication citation cross-links are deposited in the system, as are publication metadata (e.g. title, authors, date published). This facilitates “forward” citation linking - i.e. given a particular paper as a starting point, retrieve a list of publications citing that paper - which underpins citation metric calculations.

Data DOIs. Persistence, citation cross-linking and several other properties of DOIs and the DOI infrastructure makes them ideal for identifying, locating and citing datasets published online (Paskin, 2005). Data DOIs have not yet been adopted to any meaningful degree in the biosciences, but efforts in this direction have been ongoing in other disciplines for several years. For example, the DOI name 10.1594/PANGAEA.119754 identifies a set of surface sediment observations published as a supplement to the paper by Stein *et al.* (2004). Resolving the DOI via the central DOI resolver using a web browser (direct link: <http://dx.doi.org/10.1594/PANGAEA.119754>) takes the user to the Publishing Network for Geoscientific & Environmental Data (PANGAEA) open-access library³⁹, where the dataset is archived.

As with DOIs for article and journal websites, if the location of an archived dataset changes, the data manager updates the metadata for the DOI with the new URL. Because of the

³⁸<http://www.crossref.org>

³⁹<http://www.pangaea.de>

resolution service provided by the central system, citations using the DOI identifier remain intact even as the dataset location changes (as long as the resolver is operational).

The system also addresses dataset versioning. Datasets are cited by way of the unique, persistent DOI, rather than the potentially ambiguous author-year combination, with the DOI referring to specific version of the dataset. Updated versions of a given dataset would each be assigned a different DOI, and metadata held in the central system would indicate that each of up to several DOIs refers to a different *manifestation* of the same published work (much like different editions of a book).

Building on the success of PANGEA and associated pilot projects in the Earth and Environmental Sciences (Brase, 2004; Brase and Schindler, 2006), the international DataCite Consortium⁴⁰ was recently formed to promote the use of DOIs for scientific data. Despite the glaring absence of major bioinformatics centres such as the EBI and NCBI on the list of partners, this initiative is a positive step towards a unified system for data citation, not the least because it involves extending and adapting established infrastructure and so technological hurdles should be minimal.

The author name problem. In addition to a robust infrastructure for registering and citing published datasets, several other factors need to be considered, not least issues relating to the identity of researchers (Thorisson, 2009a). At present, authorship information held by the DOI system has the fundamental limitation that authors are identified by name only. As noted above, this introduces troublesome ambiguity. Furthermore, individuals sometimes change their name (e.g. women marrying and taking their husband's family name). Increasing numbers of scholarly contributions from non-Western countries in recent years have exacerbated the problem, with a limited supply of family names and transliteration of similar-sounding last names into Western character sets causing further confusion (Qiu, 2008).

Ambiguous author names may not necessarily pose a serious problem in a small discipline.

⁴⁰<http://www.datacite.org>

But in a broader context the situation is more serious, as illustrated by Garfield (1969) in his correspondence to Nature:

[...] While A. Kantrowitz is “obviously” a heart specialist (to the cardiologist) and obviously a physicist to the physicists, it is not obvious to many others interested in their work which one is the author of a paper appearing in a multi-disciplinary journal.[...]

Over 40 years on, the volume of published literature and the total number of academic authors has grown by several orders of magnitude. A recent analysis by Torvik and Smalheiser (2009) found that $\sim 2/3$ of the ~ 6 million authors in the MEDLINE bibliographic database share a last name and first initial with at least one other author, and that an ambiguous name refers to ~ 8 persons on average. The consequences of this are far-reaching, ranging from inaccuracy in literature searches to the wrong person being to be asked to peer-review a paper.

Crucially - and central to the present discussion - non-unique person names are a fundamental obstacle to unambiguously attributing published works to individuals, be it journal articles, books or research data. Consequently, the problem of author name disambiguation, recently documented in detail by Smalheiser and Torvik (2009), has been the subject of considerable research in computer science. A plethora of automated methods have been developed for disambiguating author-publication associations using publicly available information in the literature, such as co-authorship, author affiliation and author E-mail address. For example, the “Author-ity” method by Torvik and Smalheiser (2009) estimates the probability that two publications, tagged with the same author name, were actually written by the same individual. However, although these automated data mining approaches can identify the great majority of authors, manual intervention is always required to resolve the remaining ambiguities.

Identifiers for authors and other contributors. The adoption of unique author identifiers instead of person names for identification is generally recognised as central to solving the author name problem. Like identifiers for scholarly publications, author identifiers need to fulfill certain key requirements critical to the long-term scholarly record,

in particular that they be i) persistent and ii) never recycled (i.e. an identifier never re-used for another person in the future). For these and other reasons, a centralised system specialised for assigning and managing identifiers for authors and contributors in general (analogous to the DOI system) would seem to be an optimal solution in this regard.

Two of the better known efforts in this area are commercially-provided services: ResearcherID⁴¹ by Thomson-Reuters and the Scopus Author Identifier⁴² by Elsevier. The general approach taken by these and similar projects is to use a disambiguation algorithm to pre-compute a set of author identifiers based on available bibliographic data, and use this to populate the ID system with provisional profiles for authors and lists of their publications. This information is then presented to authors who have registered with the system, with the expectation that these users will then proceed to help with resolving errors in their profiles. For example, users can alert system administrators if their information appears to be split across multiple profiles (e.g. “G. Thorisson” vs “G. A. Thorisson”), or if their profile has been incorrectly merged with another author with the same name (e.g. the numerous authors named “J. Smith”).

The underlying assumption in these schemes is that it is in the best interest of authors that their bibliographic information is accurate, thus giving them an incentive to participate. This kind of engagement with authors helps (and is indeed essential) to address the problem retrospectively, though with the obvious downside that incorrect information for authors who do not participate will remain incorrect. If the unique identifier for a given author were linked to his future publications, this would address the ambiguity problem in a prospective manner. More generally, a central profile linked to a contributor ID, pre-populated with existing author-publication data, and basic tools to manage this profile would be a major asset to scientists and publishers alike and help to streamline various processes, such as manuscript tracking and peer review⁴³.

⁴¹<http://www.researcherid.com>

⁴²http://help.scopus.com/robo/projects/schelp/h_authsrch_intro.htm

⁴³http://www.crossref.org/CrossTech/2007/02/crossref_author_id_meeting.html

However, these perceived benefits of author identifiers have so far not been realised. One problem is that the two commercial services and smaller-scale discipline- or country-specific author ID systems are identity silos, analogous to the Web 2.0 websites and organisation-centric ID systems referred to above. In other words, they are entirely separate from one another and not interoperable, so authors scholarly identity is fragmented across several systems.

Another obstacle is an overall lack of uptake in the community, in part due to uncertainty as to who should run a global author ID system (Enserink, 2009). Understandably, many researchers, funding agencies and other organizations are wary of commercial, for-profit companies operating and controlling access to such crucial infrastructure and vast stores of valuable information, and thus a global contributor ID system should ideally be operated by a neutral, non-profit international organisation.

The ORCID initiative. Acknowledging that community support and trust are crucial to the success of the enterprise, Thomson-Reuters, Elsevier, Nature Publishing Group and other parties represented by CrossRef, as well as numerous other stakeholders in the publishing and research domains, have for the past several years been working towards a common, pre-competitive solution. The outcome of these deliberations was the December 2009 launch of the Open Researcher Contributor Identification Initiative (ORCID)⁴⁴. As stated on the project website, ORCID aims to:

establish an open, independent registry that is adopted and embraced as the industry's de facto standard. Our mission is to resolve the systemic name ambiguity, by means of assigning unique identifiers linkable to an individual's research output, to enhance the scientific discovery process and improve the efficiency of funding and collaboration.

On a political and strategic level, the unprecedented level of backing from the publishing industry, academia and other stakeholders means that an ORCID-operated contributor ID system stands a very good chance of succeeding where other global efforts (such as the

⁴⁴<http://www.orcid.org>

proprietary services discussed above) have failed. On a technical level, the contribution by Thomson-Reuters of software and data underpinning the ResearcherID service to “jump start” the ORCID system further bolsters the likelihood of success. It would therefore seem safe to expect that within a few years, the infrastructure will be in place to provide researchers with a home for the portion of their digital identity that matters the most - information relating to their publication record, and therefore their academic reputation. Scientists will thus have numerous incentives to join the system. Moreover, if ORCID also offers identity provision as an opt-in service for users of the system, this could drive awareness and adoption of digital IDs in the scientific community. All this will create a fertile ground on which to grow the new systems needed for reliable attribution of digital contributions.

6.3.2 Managing access to G2P data

One of the key outcomes of the IRBW2009 workshop relates to an apparent dichotomy of the researcher identity challenge. Leveraging identity for managing access to sensitive data and other protected resources on the Internet is primarily a security problem wherein authentication and trust requirements feature prominently. Conversely, the various issues and strategies elaborated in the previous section are chiefly concerned with *knowledge discovery* using for the most part information which is in the public domain (i.e. part of the scholarly record), for which security requirements are a relatively small part of the overall picture.

The opinion was shared amongst many workshop attendees that a single, universal identity system aiming to cover the broad range of use cases from both of these problem domains would be a daunting, if not impossible, task. Importantly, the key to success going forward will be to not only distinguish between these two main problem domains, but also to further “segment” the problem space and focus on creating or adapting standards-based, interoperable identity-enabled systems which address specific problems, or small sets of

related problems. The creation of the ORCID system principally for solving the author name problem is an example of this approach.

The G2P data accessibility problem. The G2P databasing challenges discussed in Chapter 2 bring into focus the issue of data accessibility. It is not disputed here that access to detailed, identifiable data on participants in biomedical studies needs to be carefully controlled for personal privacy reasons (see §2.2.7 and §2.3.5), and this involves many stakeholders (see e.g. Foster and Sharp (2007)). However, it is clear that access management on a case by case basis is becoming unfeasible, given increases in; the number of such studies; the number of groups/consortia generating such datasets; the number of databases wishing to integrate and disseminate the information; and the number of researchers wishing to access these data. With some exceptions (e.g. BIRN, caBIG), current disseminating practices focus on secure local access only and thus hamper effective data reuse, as previously discussed. The analogy with the Web 2.0 domain is readily apparent here; for example, the EGA and dbGaP are unconnected silos, with access managed entirely independently and with no means of cross-site communication about the identity and authorisation level of users.

This is no doubt the result of G2P archives being mandated to provide a certain minimum level of secure access according to data release policies, with distributed access models lying outside their remit. However, another factor - technological overhead - may also be a significant barrier at present. The technologies described in the previous section are available and can be used to build the necessary infrastructure for secure G2P data dissemination, but are difficult to deploy. This may be especially true for smaller projects or organisations with fewer resources (Harding *et al.*, 2008). The proverbial “chicken and egg” problem may be yet another factor: the true benefit of distributed data access would not be apparent until a certain critical mass of data providers participate, and until then there may be little incentive (and even risk) for providers to join as early adopters. Finally, a related challenge is making identity-enabled distributed systems sufficiently easy to use, so non-expert researchers can locate, authenticate and securely retrieve the relevant datasets.

A registry for users of biomedical data. It would be prudent to follow the advice given at the beginning of the section and not attempt to solve the entire G2P data accessibility problem at once. Instead, my main focus concerns the process of determining data access privileges for a user who wishes to retrieve a dataset from a G2P data provider, in a way that scales beyond the simple single-site use case. This would be greatly streamlined if one or more services (possibly operated by major regional data centres such as WTSI and NCBI) were to store information on access privileges for each user, linked to his digital ID. This registry (or registries) could then be used by various primary and secondary data providers (whether or not part of the WTSI and NCBI) to check whether or not a user should be granted access or not. Furthermore, the same registry could also be used to “blacklist” the IDs of individuals found guilty of inappropriate use of data (though the complex issue of sanctions needs much further consideration, whatever mechanism for access approval is in operation).

The registry and participating data providers could have different levels of granularity for access permissions. For example, in the simplest scenario a person who is listed in the registry (thereby confirming his status as a researcher) could be given “blanket” access to quasi-sensitive data (such as aggregate genotypes). In a more complex scenario involving individual-level data, a researcher could be granted access to all datasets from a particular archive (e.g. dbGaP), or all data from a particular consortium which has submitted several datasets to primary archive (e.g. all WTCCC data). Finally, a researcher could be given access to only a particular dataset (e.g. WTCCC bipolar study).

A crucial advantages of such a distributed system would be that participating data providers would not have to deploy the full arsenal of complex grid security technologies to enable distributed access. They would only have to implement a minimal federated identity solution similar to the Web 2.0 social networking applications previously discussed. Further details on how this could work in practice are provided in the next section. Another potential advantage is that data generators - or other parties who govern data access, such as ethics committees - could manage access permissions (add, remove and modify permits) centrally, and the changes would have immediate effect across all participating primary

or secondary data providers. The following two case studies serve to illustrate this idea further.

Case study: The caBIG grid security infrastructure

The caBIG security infrastructure is an example of how federated identity technologies are used for access control within the confines of a special-purpose grid. An interesting aspect of the caBIG Grid Authentication and Authorization with Reliably Distributed Services (GAARDS) as described by Langella *et al.* (2008) (see Fig. §6.3) is the use of strong PKI-based authentication in conjunction with users' local, institution-supplied identity. Recognising that centralised identity provisioning for the entire US cancer grid would not be practical, GAARDS architects have devised tools which enable users to sign into the cancer grid using their institution ID credentials and exchange these for grid credentials (in the form of a X509 certificate), which are subsequently used across the grid for authorisation purposes. This support for externally-supplied IDs is coupled with facilities for specifying and enforcing local authorisation policies, i.e. empowering individual caBIG data providers to control who has access to their data, an important consideration in the caBIG project (Manion *et al.*, 2009).

The GAARDS system is effective and appropriate for the kind of strictly-controlled distributed environment mandated by US cancer research, where individual-level data shared by dozens of institutions over the grid include tumour tissue annotations, genotype data, resequencing data, imaging data and more. Also, via a proxy service, the system enables caBIG users to interact securely with computing clusters connected to the grid and launch CPU-intensive batch analysis jobs without having to deal with certificates directly. Like other caBIG software components, GAARDS is available as open-source software⁴⁵ and can be used with other caBIG tools to construct identity-enabled, caBIG-like grids (Buetow, 2009). This effort to produce reusable software and make it available to the community is commendable and will no doubt be immensely useful for new grid

⁴⁵<http://www.cagrid.org/display/gaards/>

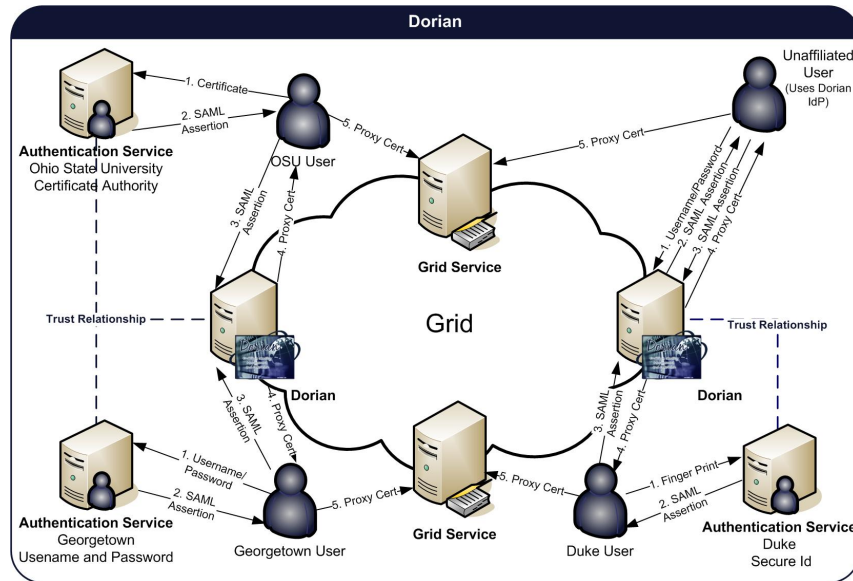


Fig. 6.3: Overview of a part of the caBIG security infrastructure and example usage scenarios. From <http://cagrid.org/display/dorian/Overview>.

projects with similar technical requirements to caBIG. However, from an implementation perspective, even the availability of such open-source security toolkits is of limited help in bringing identity-based security to smaller projects and software packages, such as HGVbaseG2P or LOVD. This is because security requirements are often far simpler and developers often do not have the required technical expertise to install and effectively operate such “heavyweight” grid security solutions.

Case study: OpenID-based controlled-access data sharing in the ICGC

A contrasting example to the above is the access control scheme that will be used for sharing individual-level data generated by the International Cancer Genome Consortium (see §1.2.6). The central data portal for the project at <http://dcc.icgc.org> is operated by the ICGC Data Coordination Center (DCC) which will provide two-tiered access to project data: aggregate genotype data and minimal cancer patient information (e.g. cancer histology and gender) will be publicly accessible, whereas access to genotypes, gene expression and other potentially identifiable data on individuals will be controlled.

The system being devised by the DCC group and scheduled to go online mid-2010, is illustrated in Fig. §6.4, based on a presentation given by Lincoln D. Stein at the IRBW2009 workshop and a report from the 2nd ICGC scientific workshop published online⁴⁶. A key feature of the system is the OpenID-based authentication scheme. This removes the need for managing local user accounts in the DCC system and, importantly, enables single sign-on across the various participating systems. The paper-based interaction between the DCC and the ICGC Data Access Compliance Office (DACO)⁴⁷ is crucial to establishing a link between the digital identity (as represented by the OpenID identifier) and the real-life persona and credentials of the investigator. The DACO also fulfills other important regulatory requirements, such as maintaining a paper trail for audit purposes. After this initial process, the user's OpenID serves to streamline the interaction with the DCC and with the various ICGC partners' franchise databases, thus doing away with the need for separate user accounts and certification/approval processes with each partner on a case by case basis.

Though built as a custom solution specific to the needs of the ICGC project, the overall design is generalisable to many other data sharing scenarios. The most important aspect of the above is the planned collaboration with the EGA. ICGC partners will deposit primary sequence data in EGA for archiving, and users already registered with the DCC will be able to securely retrieve these data without having to create a separate account with the EGA and going through a separate approval process. Extending this scenario beyond the set of actors shown in Fig. §6.4, various manifestations of protected ICGC data could in principle reside in many other locations and be served in the same secure way by a variety of secondary data providers. These providers would naturally need to similarly establish an initial trust relationship with the DCC with respect to the identities of authorised users.

⁴⁶http://www.icgc.org/files/icgc/ICGC%20Scientific%20Workshop%20Report%20June%2022-24,%202009_en.pdf

⁴⁷<http://www.icgc.org/daco>

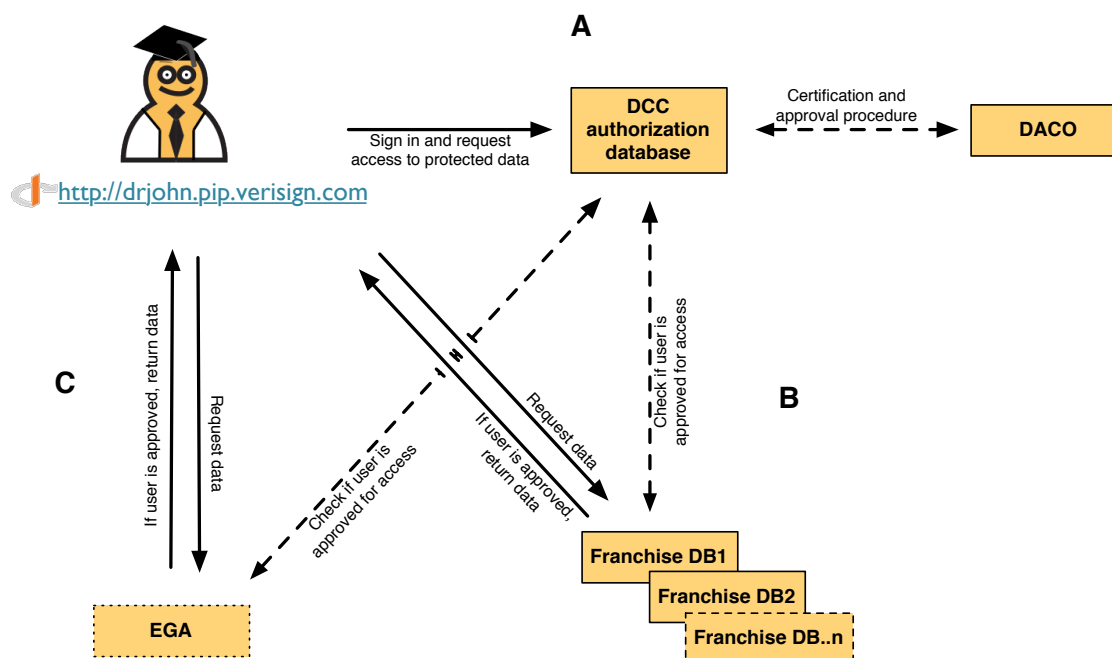


Fig. 6.4: Overview of the ICGC federated controlled-access scheme. A) The first phase involves the user signing in and initiating the access request process, during which paperwork (e.g. material transfer agreements, certifications and other paperwork) is sent to the DACO for processing. This is done in batch rather than on a case-by-case basis, with the DACO sending lists of approved requestors' OpenIDs back to the DCC. B) By way of federated search tools provided by the central data portal (this step is not shown), in the second phase the user locates data of interest in up to several of the ICGC partner databases. As the user proceeds to retrieve the data, each franchise database can look up the user's OpenID in the DCC database and verify that the user has been approved for access. C) Although not an ICGC franchise database, the EGA will use the same federated authorisation scheme to facilitate primary data retrieval by users pre-approved with the DCC.

6.4 Use cases for identity-enabled access management and attribution

Having undertaken a field analysis and engaged with the community, I concluded that the next logical step forward was a pilot software development project to demonstrate the technical feasibility of digital IDs for identifying researchers in the two main contexts already discussed: i) as data consumers accessing non-public datasets, and ii) as data

generators contributing data to central repositories. Since it was clear that the timeframe for my PhD would not allow the completion of such demonstrator, I focused my efforts on developing a series of use cases which would serve as precursors to later implementation work. These use cases are centred on a number of scenarios involving exchange of genetic mutation data via the Café for Routine Genetic data Exchange (Cafe RouGE). introduced below.

The next four sections contain brief summaries of a total of ten use cases, each referring to a subsection in §C.1 which provides a detailed step-by-step description, use case diagram and additional notes. Appendix §C.1 also provides a summary of the Cafe RouGE platform, as well as a brief description of the GEN2PHEN Knowledge Centre which also plays a role in some of documented scenarios.

6.4.1 Basic access management with local user IDs

This first collection of Cafe RouGE use cases does not involve digital IDs, but is presented primarily to provide context for later sections. This section therefore does not present an exhaustive set of all Cafe RouGE use cases, but rather only a limited set of use cases on which the more advanced, identity-enabled scenarios build and extend. For example, various data management-related tasks (e.g. deleting submissions) are important from an overall Cafe RouGE operations perspective, but use cases for such scenarios will not be presented here.

All these use cases assume that the actors (data submitter/owner and data consumer) have already registered for a local user account on the Cafe RouGE site, and therefore possess local ID credentials in the form of a conventional username and password.

Use Case #1: data submission from diagnostic laboratory. A diagnostic lab operator (the data owner) finishes a series of genetic tests on a de-identified patient sample and wishes to publish the results. He uses the Gensearch® analysis software⁴⁸ to make a secure

⁴⁸<http://www.phenosystems.com>

connection to the central Cafe RouGE depot and upload a mutation report for publication (§C.1.2).

Use Case #2: data consumer pre-authorised to access Cafe data. A data consumer wishes to retrieve a mutation report from the Cafe RouGE depot. The diagnostic lab operator (the data owner) has elected to only share his data with specific persons, and has already added the ID of the data consumer to the list of IDs for users who are permitted to access all data generated by his lab (§C.1.3).

Use Case #3: data consumer requesting access to Cafe data. A data consumer wishes to retrieve a mutation report from the Cafe RouGE depot, after discovering the dataset via a feed syndication service on the Knowledge Centre. He does not yet have permission to access the protected data, so he first needs to request access from the diagnostic lab operator who generated the data (the data owner) (§C.1.4).

6.4.2 Access management with digital IDs

The next set of use cases utilises digital IDs to enhance various aspects of the Cafe RouGE website. As noted in §6.2, a key advantage of federated authentication is that the user does not have to go through the trouble of creating an account on yet another website, and can instead reuse an existing profile from his ID provider. Cafe RouGE users - data submitters and data consumers alike - can thus sign in to the website and have a local user account created automatically with minimal effort. This lowers the threshold for users to join, and is thus expected to increase the proportion of users who participate in the scheme. The other main benefit is single sign-on across multiple websites which further enhances the user experience.

From the perspective of the Cafe, an important benefit of “outsourcing” authentication to the ID provider is reduced administration overhead relating to user management. Given that users’ ID credentials would not be stored locally, the all too common problem of users forgetting their passwords and otherwise needing help with their accounts would be greatly minimised.

Use Case #4: authenticating via OpenID. A user wishes to register on the Cafe Rouge website. Instead of a local user account and authentication on the Cafe system, the user opts to identify himself with his OpenID credentials. The Cafe website therefore redirects him to his OpenID provider where he authenticates, and his OpenID is subsequently linked to his local ID in the Cafe system (§C.1.5).

Use Case #5: browsing and accessing Cafe data from the Knowledge Centre portal. A data consumer is interested in reported variants in and around the BRCA2 gene. He uses the Knowledge Centre mutation feed browser to retrieve a list of mutation reports tagged with the BRCA2 gene symbol, and now wishes to access the full report details (§C.1.6).

Use Case #6: manage access based on ID provider whitelisting. A diagnostic lab operator (the data owner) decides to adopt a liberal data release policy: automatically approve all data access requests from data consumers with an identity from a certain subset or “whitelist” of OpenID providers (§C.1.7).

6.4.3 Access management with digital IDs and semantic authorisation

The use cases considered in the previous section illustrate how data access management in the Cafe RouGE system can be usefully augmented by digital IDs and single sign-on. However, whether the data consumer is identified with a local ID or with a digital ID from a remote ID provider, and whether ID attributes from remote sources are considered or not, one disadvantage is that the final authorisation decision ultimately takes place locally, in the Cafe system itself.

In principle, this need not to be a problem if access control requirements are relatively basic. But if more sophisticated access control strategies are required in the future, the system would need to be extended with increasingly complex software logic and data storage capabilities. As already noted in §6.3.2, conventional role-based security in distributed systems can quickly become very difficult to manage and use. This

section introduces a newly-developed semantic authorisation technique which addresses the problem, and then summarises several use cases which demonstrates its utility in the Cafe RouGE setting.

Semantic authorisation with the AGAST framework. Sinnott *et al.* (2009) elaborate further on the various shortcomings of conventional grid security solutions as they describe the Advanced Grid Authorisation through Semantic Technologies (AGAST) project ⁴⁹. Briefly, the AGAST approach is built around the concept of describing hierarchies of roles and access privileges as ontologies expressed in the RDF/OWL language. Instance data (i.e. ID attributes, such as assigned role(s) or group membership) are also converted into RDF. With access policy, ID attributes and additional knowledge of administrative or organisational structures all expressed in a semantically-computable format, this information can be gathered and merged into a single RDF graph. A SPARQL query is subsequently executed over this graph by a semantic reasoner, or *decider*, to answer the question: “*does this user belong to the class of users who are permitted access?*”. This method thus leverages i) the “universal solvent” properties of the primitive RDF model to represent and integrate potentially very complex information from different sources, and ii) the power of ontologies to expressing classes and logical relationships in a way that can be reasoned over to infer relationships not explicitly declared.

Use Case #7: access based on inferred virtual organisation membership. A data consumer wishes to access a dataset published in Cafe RouGE. The data consumer and data owner are affiliated with different research institutions, both of which are part of the same virtual organisation (VO). An intra-organisational data exchange agreement is in place, such that any member of a partner institution can access mutation data generated by the other partners. This access policy is expressed as an ontology which describes access permissions and organisational structure of each partner in the VO (§C.1.9).

Use Case #8: access based on status as *bona fide* researcher. The data consumer

⁴⁹<http://www.nesc.ac.uk/hub/projects/agast>

wishes to access a dataset published in Cafe Rouge. The data owner has chosen a liberal data release policy stating that anyone who has an active ORCID contributor profile can access the data. But since the Cafe does not have this profile information in hand, the data consumer is prompted for additional ID information proving that he is in the ORCID system (§C.1.10).

Use Case #9: access based on permission from an external registry. A data consumer wishes to access a dataset published via Cafe Rouge. The data include identifiable information on patients, and access is therefore regulated on a case-by-case basis by a Data Access Committee (DAC). The data consumer has already obtained permission to access the dataset by sending a request to the DAC (see e.g. §2.2.7). Information on his authorisation level for this and other related datasets is held in a central registry of data access privileges, in which the data consumer's ID has been linked to his OpenID (§C.1.11).

6.4.4 Attribution with ORCID IDs and DOIs

As already noted, my primary motivation for exploring digital identity came from the need for controlled access to GWAS data in the HGVbaseG2P database, and so the majority of my use cases are focused on data access scenarios. However, I reserve one final use case to illustrate a potential key role for digital IDs in data registration and attribution schemes discussed in §6.3.1.

Use Case #10: attributing data publication to submitter. A diagnostic lab operator (the data owner) publishes a dataset via Cafe RouGE. He wishes to associate this dataset with his ORCID profile in order for this digital contribution to be attributed to him. He has previously authenticated with the Cafe using his OpenID, acquired submitter privileges and associated the Gensearch® application with his Cafe ID as per previous use cases (§C.1.12).

6.5 Discussion

This chapter reported on my forays into the realm of digital identity. This work, initially prompted by a technical problem in the HGVbaseG2P project, has grown into a standalone project in its own right, with implications far beyond the original scope of the HGVbaseG2P project. From my initial informal investigation, I concluded that a global Internet identity infrastructure could potentially be a key enabling technology in scientific research. In particular, identity could potentially be crucial to successfully tackling several of the G2P databasing challenges outlined in Chapter 2. This exciting prospect prompted me to undertake further work in this area, including a more extensive follow-up analysis of key concepts and technologies that I had not previously considered, community outreach efforts which included organising a workshop on the topic, and use case development in preparation for pilot software implementations. I will now summarise key results and conclusions from these three main identity-related aspects of my work.

6.5.1 Federated identity and research

The exponential growth in adoption of OpenID over the past 5 years underlines the importance of user-centric identity in the evolution of the Internet. At present, an estimated⁵⁰ 1 billion website user accounts are OpenID-enabled - a substantial fraction of the entire online community - and several million websites where these IDs are accepted for registration and sign-on. Stakeholders include major companies such as the Paypal online payment service⁵¹ and Amazon. Institutions within the NIH (itself already using federated identity⁵²) are involved in various OpenID pilot projects. User-centric identity also features prominently in so-called Government 2.0 initiatives in the US and elsewhere. All of these are signs that user-centric identity is not solely a niche concept in the Web 2.0 and social networking, but is gradually permeating the mainstream Internet.

⁵⁰<http://openid.net/2009/12/16/openid-2009-year-in-review/>

⁵¹<https://www.paypal.com>

⁵²<http://federatedidentity.nih.gov>

If the current trend continues, it seems reasonable to expect that, within a few years, signing into a website with a 3rd party account of choice will be a process familiar to the average Internet user (including scientists), and not be merely a niche activity of interest to a tech-savvy minority. Therefore, I believe the time is now ripe to seriously consider how federated identity can be applied much more broadly in scientific databasing than is currently the case.

A user-centric, scholarly identity on the Internet. By fortunate coincidence, at the same time as adoption of user-centric identity is soaring, there has been a growing interest in the community in finding a practical solution to the perennial author name problem in the literature. None of the commercial author identifier systems developed previously had reached broad adoption, a major obstacle being the sentiment shared by many researchers that such a system should be operated by a not-for-profit, independent organisation. The creation of the ORCID initiative was the culmination of several years of negotiations between publishers, funders, universities and numerous other stakeholders find a common solution. If ORCID is successful - and backing by an international consortium of over 100 companies and organisations increases the chance of success - the prospect of a global registry of unique identifiers for contributors will be key to solving a myriad problems and increase efficiency in academic publishing.

One key question concerns the future role of ORCID-hosted identity profiles. Will contributor identifiers be mostly used in the background, largely invisible to end users? Or will these identifiers and ORCID profile pages be used prominently by researchers as a focal point of their scholarly activities? There may be substantial advantages to centralised, authoritative provision of identity management services for professional researchers, a prominent one being convenience for scholars who have no desire to manually maintain a publication listing on their departmental homepage. This is indeed a prominent use case of the proprietary ResearcherID service, on which ORCID will be based.

However, the mantra of the user-centric identity movement is decentralisation and freedom of choice as to where one's identity is hosted. Ideally, it should be possible to treat their

ORCID-hosted identity as one of several identities, and then link these together as desired. Enabling scholars to display publication listings and other ORCID-hosted details on their own webpages will be key to this (see e.g. Warner (2010) for a discussion of how this can be done with the arXiv preprint archive⁵³). Encouragingly, this view is shared by many ORCID stakeholders, and the prototype system now being constructed will support programmatic, web serviced-based interactions between the central system and various other Internet sites. In fact, one major function of the ORCID system will be to serve as a “hub” which links together existing person identifiers in various other systems.

ORCID will without a doubt be immensely important in the publishing domain and to advancement of digital scholarly identity in general. But the key relevance of the above to my work relates to the utility of scholarly IDs outside the specific realm of traditional scholarly publishing.

Identity-based attribution and data sharing. The various attribution scenarios I have considered are built around the general principle of creating a persistent link between a digital contribution on the one hand, and the identifier of the person on the other. Centrally-sourced identifiers are not required for this, and new attribution tracking systems could in principle be built based on whichever IDs are presented by data submitters as they upload data to central repositories. However, given the unique properties of the new ORCID identifiers, the infrastructure being built around them and the crucial link to the scholarly profile of a submitter, repurposing ORCID IDs for attributing digital contributions would be a superior strategy.

The infrastructure being created by DataCite (see §6.3.1) for identifying and citing datasets via DOIs provides the other vital technological piece in this puzzle. Once persons and the datasets they contribute to can be identified with persistent, non-recyclable identifiers, the attribution link (i.e. publication credit) can finally enter the permanent scholarly record. Having solved the identification and tracking problem, the community will then be faced with challenge of how best use the data citation links, online access and other data collected

⁵³<http://arxiv.org>

by the new system. But this is an altogether more attractive problem to be faced with, compared to having no data at all.

6.5.2 Identity-based use cases for Cafe RouGE

The third main part of my work was a collaboration with others in our group in an effort to create a proof-of-concept software implementation. The aim of this was to demonstrate certain practical aspects of federated identity for solving problems relating to data sharing and data accessibility. Due to time constraints, this development could unfortunately not be completed as planned. My preliminary work, presented in this chapter in the form of ten documented use cases, is nonetheless a useful resource on which to base future implementation work.

The use cases as documented describe scenarios involving the exchange of genetic mutation data, specifically the flow of data from diagnostic laboratories through the central Cafe RouGE depot to third party data consumers. However, the main motivation behind this work was to explore ways to enhance data access and publication of G2P data in general. Therefore, most of the use cases are deliberately vague on many details, with the intent that they be extended, refined and adapted to other categories of G2P data, and used as a basis for software implementations.

Use cases for access management. All the use cases but one entailed access control scenarios. This simply reflects the importance placed on such scenarios in the HGVbaseG2P project, and my conviction that tools for practical data access management will be instrumental to persuading researchers - many of whom are concerned about data re-identifiability following the findings of Homer *et al.* (2008) - to deposit data in HGVbaseG2P. In many of the Cafe RouGE use cases, the term “mutation data” could be easily be replaced with “GWAS data”, with the general sequence of interactions between data consumers, data owners and the data provider remaining largely unaltered. In particular, future systems which support Use Case #9 and variations thereof may be vital to

addressing the GWAS data sharing problem, and in the long run help to streamline secure sharing of research data in general.

Another important issue pertains to the potential role of semantic authorisation in addressing access management scenarios where access policies are complex, or where identity profile information resides at multiple locations on the Internet, or both. One downside is that adopting the semantic approach can be expected to incur significant technological overhead. The various semantic technologies are not yet fully established in the field and therefore tend to be unfamiliar to many developers. For example, the SPARQL query language is far less likely to be familiar to developers than SQL. Also, constructing the necessary ontologies can be a non-trivial task to undertake for complex access policies (J. Lusted, personal communication). However, one counter-argument is that the various grid security solutions present their own challenges and tend to require expert knowledge to install and operate. Semantic authorisation may therefore on balance be a more suitable, scalable strategy for many applications, once ontologies are in place.

Newly developed frameworks for specifying semantic access policies in controlled natural language may also be worth of further exploration. Policy language platforms such as Protune⁵⁴ and user-friendly tools for creating and edit policies (Abel *et al.*, 2009) can be used to generate a set of semantic rules, which could then be integrated with other data and consumed by a semantic decision-making engine. Such techniques address the inherent difficulties in controlled-access systems, whereby a policy (or a set of policies) are first devised in natural language (by lawyers, ethicists and others) and subsequently need to be implemented in software logic.

Use case for attribution. The single data submission use case represents a realistic scenario, highly analogous to the process of submitting a manuscript to a journal, where identity can be leveraged for attribution. A key consideration for refinement of this use cases and eventual implementation will be to play on researchers' familiarity with manuscript submissions to journals. This may be critical to widespread adoption of the

⁵⁴<http://policy.l3s.uni-hannover.de:9080/policyFramework/protune/>

data publishing process proposed here. Although the proposed system does not require this, the presence of ORCID as an authoritative supplier of scholarly IDs could be major help in simplifying the user experience for data submitters. To further explore this, I will in coming months commence follow-up work to further develop the data submission use case into a set of user interface mockups, and later a fully-fledged implementation within Cafe RouGE which will interact with the prototype ORCID system. This will be done in collaboration with the ORCID Technical Working Group, with a major aim of keeping the data submission workflow as similar as possible to the ORCID-enabled manuscript submission workflow now being developed by ORCID partners (G. Bilder, personal communication).

Usability. A key to successful adoption of Cafe RouGE in the community will be usability, as neither data submitters nor users are likely to be receptive to a system that is overly complicated to use. Throughout the use cases developed, emphasis is placed on keeping the interactions between the user and the Cafe RouGE system as simple as possible. The process of initial registering and subsequent interactions should be as seamless a user experience as possible, with users ideally not needing to know advanced concepts like certificates (where higher level of security is required). But the process cannot be entirely automated, as cues will be required to prompt users to authenticate with other ID providers if required, and/or to supply additional attribute assertions. To leverage users' familiarity with privacy controls in social networking applications such as Facebook, a key consideration for future work is to (where possible and appropriate) model privacy and authorisation tools in Café RouGE on existing user interface designs and processes. The online photo sharing paradigm, on which use case #1 is modelled, is an example of this. In general, it would seem that in order for identity-based solutions to gain acceptance in the community, developers must ensure that existing online behaviour and needs of users is taken into account (Maler, 2009).

Community engagement and evaluation. The overall Cafe RouGE strategy was conceived to respond to a growing data sharing problem, with input from several members of the community. However, as of yet most of the use cases as documented and discussed here have not been presented and discussed in depth outside this small group of collaborators. An important next step is to present these ideas and future plans to the broader community (e.g. workshops, or focused sessions at scientific meetings). In general, engagement with the community will help to refine use cases and to evaluate prototype implementations. Producing iteratively more sophisticated, preconfigured and standalone software packages which can be demonstrated at workshops, or downloaded and installed by users as a “sandbox” to try themselves, could be a useful strategy in this regard.

6.5.3 Cafe RouGE as an experimental platform for digital IDs

Cafe RouGE may have further utility beyond serving as a practical data publication platform. Many of the technical features already described make the system an ideal testbed for investigating and experimenting with sophisticated authentication and authorisation solutions. The standard, RESTful AtomPub API simplifies development of external clients. The feed-based data model is conceptually simple, and standardized. Overall, the system provides a foundation that appears simple enough for non-experts in grid security to contribute to its development, and for end users (who will be used to little or no security) to understand.

There are two main areas of opportunities in this regard. First, Cafe RouGE can serve as a sandbox for deploying various existing federated identity technologies. The system can play the role of a “training ground” to build developer experience and test software components destined for use in other projects in our group. As mentioned, this was a key motivational factor for formulating the various use cases for this particular platform (as opposed to some other, or an unspecified, platform). Second, Cafe RouGE is an ideal platform for experimenting with new, sophisticated authentication and authorisation solutions and ideas. Again, because of the overall simplicity of the system, it will

be straightforward to experiment with relatively simple access control scenarios and “layer” on sophisticated security as required. Indeed, the potential of Cafe RouGE as an experimental platform has led to it featuring in the proposal for a new project called Distributed AuthN/Z and Semantics in Networked Genetics (DANZSiNG). Our group and others (including the authors of AGAST) are now seeking funding to the Joint Information Systems Committee (JISC)⁵⁵, a UK-based funding agency. The goals of the DANZSiNG project include a Cafe RouGE demonstrator implementing variants of the semantic authorisation use cases from this chapter, as well as several other use cases where, for example, authentication level of assurance is one of the parameters used in authorisation decision-making.

⁵⁵<http://www.jisc.ac.uk>

7. Final summary, conclusions and future work

This thesis is set against the backdrop of a number of challenges to effective management, exchange and integration of G2P research data. The most prominent of these are increasing size and complexity of data generated in large-scale biomedical research projects. The rapidly expanding scope of research involving human subjects has also brought difficult social, ethical and legal issues to the forefront of the discussion. Addressing these issues, whilst providing access to the research data for secondary reuse, is becoming a huge problem in a field where unrestricted access to large-scale genomic data was previously the norm. Other important challenges include data discovery, data quality assurance, and effective representation of the scientific knowledge generated from the data. The purpose of the work reported in the preceding four chapters was to investigate a subset of these challenges.

The HGVbaseG2P project, together with a related spin-off activity focused on G2P data models, aims to address problems related to data complexity, discoverability and accessibility in the context of disease genetics, and these activities are firmly rooted in the initial problem domain. A second spin-off activity from the HGVbaseG2P project, centred on applications of digital identity in research, has evolved into a standalone project with potential implications for the broader G2P domain and beyond. The first three sections in this final chapter briefly summarise results and main conclusions from these three projects, and suggest future work in several key areas. In the final section I proffer a brief future perspective on the broader field of G2P databasing.

7.1 The HGVbaseG2P project

7.1.1 Informatics infrastructure

At the core of the HGVbaseG2P project is a database and software infrastructure which we implemented following a conventional centralised databasing strategy. The majority of features originally envisioned as key to project success have been implemented. In its present state, the overall infrastructure is considered relatively complete, given the initial requirements and design considerations. The central HGVbaseG2P web portal has been operational at <http://www.hgvbaseg2p.org> since July 2008 as a service to the scientific community. Based on anecdotal evidence, overall usability of the web-based tools created for searching, browsing and visualising database contents is considered high. In particular, the HGVbaseG2P graphical genome views for comparing and contrasting findings within and across studies are novel and not provided by other online resources in the domain. However, several planned features of the website remain underdeveloped or deactivated due potential re-identification of research subjects from aggregate GWAS genotype data, including web service API access to full datasets, data export options and bulk downloads.

Future work. One area of ongoing infrastructure work is further development of the web-based tools, including enhancement of browser-based displays and ontology-enabled search functions. A particular challenge will be storage and visualisation of whole-genome sequencing data, given that sequencing has already started to replace array-based GWAS genotyping. Dealing with data sensitivity issues is also a priority, and work to implement access management in the web application is already underway.

The other main area of work is centred on a recently established collaboration with the creators of the Molgenis platform. One objective in the medium term is to create a Molgenis-based application which implements several important features - notably facilities for entering and managing study metadata - which were not considered initially in project design, but later deemed to be essential. This work is part of a longer-term strategy to jointly develop the Molgenis platform further into a core component of an open-source,

grid-enabled G2P data management and analysis toolkit “in-a-box”. One aim of this is to substantially lower the technological barrier to sharing data over the grid. Built-in support for domain standards will provide semantic interoperability amongst nodes on the grid. The platform will supply G2P data managers with an easy-to-use set of tools for constructing bespoke “mini-grids” of multiple installations of the same platform in other locations (e.g. by supplying a list of URLs for trusted partner databases).

Such *ad hoc* decentralised G2P database federation would potentially help to address scenarios where, for example, a consortium has a need for restricted sharing of project data amongst partners for a certain period (e.g. a pre-publication embargo), after which the data are to be shared more broadly. Importantly, the federation facilities could be used to connect a remote installation to the master HGVbaseG2P “hub” in order to, for example, open up local datasets for searching and browsing via the central HGVbaseG2P portal.

7.1.2 A global, semantically enhanced GWAS catalog

A core project objective was to gather information for as many association studies as possible, in order to provide comprehensive overview of the GWAS field. At the time of writing (May 2010), work undertaken by others in the group to this end has resulted in over 500 study entries in HGVbaseG2P, each containing at least basic study metadata and significant SNP associations as presented in the primary journal publication.

Comparison with other GWAS catalogs. HGVbaseG2P represents an improvement over existing resources in two main areas. First, in comparison with the NHGRI catalog (currently the most comprehensive, continually updated resource comparable to HGVbaseG2P), the proportion of available primary GWAS publications and amount of information provided per study is substantially greater. The current HGVbaseG2P data gathering plan calls for a continual updating of the catalog as a service to the community, and a much more frequent update cycle than has hitherto been the case in the project.

In addition to catalog breadth and depth, the other main improvement is the advanced

search and visualisation tools on the HGVbaseG2P website. These provide users with a much more powerful means for exploiting database contents than the simple tabular listing offered by the NHGRI catalog and similar resources.

Semantic enhancement of study metadata. A core curation activity in the project has been annotation of disease phenotypes with terms from existing controlled vocabularies. This work is necessary to address several issues relating to terminology inconsistencies and lack of structure in the source materials (including the NHGRI catalog). The resulting enhanced study metadata have already enabled the creation of improved searching and browsing facilities on the HGVbaseG2P website, which return more meaningful results to the user than tools created previously in the project. The longer term aim of this work (led by Tim Beck) is to facilitate annotation using terms sourced from biomedical ontologies such as Human Phenotype Ontology (HPO), which is based on medical knowledge and is therefore more suitable for describing disease phenotypes (Robinson *et al.*, 2008) than the Medical Subject Heading (MeSH) controlled vocabulary used earlier in the project.

A major driver for further development in this direction is the potential future impact of HGVbaseG2P as an integration platform for bridging the two worlds of genome-scale human disease genetics and model organism research. Semantic integration of GWAS findings with datasets generated by high-throughput mouse and rat phenotyping programmes could be immensely useful. For example, cross-species semantic phenotype similarity searching (see e.g. (Washington *et al.*, 2009)) might identify - independently of GWAS evidence - an evolutionary conserved biological pathway not previously linked to a given human genetic disorder.

Underlining the importance of the connection to model organisms, the International Mouse Phenotyping Consortium (IMPC) is currently seeking funding to extend previous work done in EUMODIC and other international pilot projects to whole-genome scale (Abbott, 2010). If funded, the IMPC will undertake extended phenotypic screening of mutant strains for each of the ~20,000 genes in the mouse genome. The standardised phenotype data and knockout mouse strains generated in this project will be a tremendously important

resource for human disease research. HGVbaseG2P would be a natural home for some of the sophisticated analysis and semantic integration tools that will be required to fully exploit such a resource.

7.1.3 Centralised GWAS data gathering and unpublished studies

The other main data gathering objective in the project was to bring together complete aggregate genotype datasets and analysis findings from various GWAS data sources into the central HGVbaseG2P database. This strategy ultimately proved unsuccessful, and indeed this part of the project serves as an excellent case study in the ineffectiveness of centralised databasing to solve this particular problem. The main factor critical to this outcome was concerns over re-identification from aggregate GWAS data. Following the findings by Homer *et al.* (2008) and subsequent changes in G2P data release policies, it is now clear that full aggregate genotype data (frequencies, alleles, odds ratios and more) from large-scale G2P investigations will, at least for some time, be subject to the same access restrictions as individual-level primary data. In the present data sharing landscape, therefore, secondary distribution of complete aggregate GWAS data via the HGVbaseG2P portal as originally planned is not possible. Encouragingly, recent developments on this front indicate that marker identifiers and p-values from association testing present no risk, and that primary data providers will soon agree to share this subset of GWAS results without restrictions.

The data sharing dilemma. The ongoing debate over whether less restrictive release policies should be applied for reduced-resolution, ostensibly safe components of GWAS datasets is likely to take some years to resolve. Given this circumstance and the limited timeframe in which array-based genotyping will be a relevant technology in this field, the current project strategy - i.e. emphasising the cataloging and study metadata aspect of HGVbaseG2P, as well as p-values from association testing - seems justified in the present context of SNP-based association studies. It is, however, vital that the debate is settled and solutions found, because the same data sharing challenges will be posed, on a

much larger scale, by sequencing-based G2P studies. The sheer volume of whole-genome sequencing data for thousands of individuals will make it entirely impractical to transfer the primary data over the Internet for secondary analysis. Therefore, facilitating controlled dissemination of whole-genome sequence data at various levels of resolution (from base call quality to aggregate statistics) will be important going forward.

Other ways of dealing with data volume and sensitivity challenges involve making a sensitive dataset available for certain kinds of analyses by external users, using locally-provided computing resources. The main advantage of such methods for “bringing the analysis to the data” is that sensitive data never need leave the boundaries of the hosting site. One such approach - Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-levEL Databases (DataSHIELD)- tailored to meta-analysis of individual-level data from multiple studies has recently been proposed by Wolfson *et al.* (2010). However, these schemes bring their own IT infrastructural challenges, and their use for datasets of substantial size is therefore likely to be limited to major international or regional bioinformatics centres.

A data journal for unpublished studies. The above data gathering challenges have brought into focus another project objective which had received little attention earlier in our work: that is, facilitating publication of data from unpublished association studies. A promising strategy in this regard may be to partner the HGVbaseG2P catalog and web-based toolkit with a G2P data registration, reviewing, validation and archiving service operating in similar fashion to the PLoS ONE open-access journal¹. Analogous to journal-like online data archives already established in other scientific disciplines (e.g. PANGAEA mentioned in Chapter 6 and Earth System Science Data (ESSD)²), such a hybrid database/journal could be a powerful platform for investigators to publish G2P datasets from scientifically sound research that would otherwise not be published in traditional journals. More generally, the prospect of receiving publication credit for

¹<http://www.plosone.org>

²<http://earth-system-science-data.net>

releasing data online could be instrumental in dragging unpublished, non-significant findings from researchers' "file drawers" (Rosenthal, 1979) and out into the open, thus helping to address publication bias.

7.2 G2P data models

A significant proportion of my work involved conceptual data modelling undertaken in two overlapping phases, with two distinct goals in mind.

7.2.1 An implementation model for genetic associations studies

The first phase of modelling involved the creation of a working implementation model, to serve as a conceptual foundation and guide to the design and construction of the HGVbaseG2P informatics infrastructure. The resulting HGVbaseG2P model incorporates numerous established conventions and domain standards for genotype and sequence information, and adds several key features which are critical to describing association studies. The most important of these is a tripartite representation of a "phenotype": i) the concept of the trait that was measured, ii) how the trait was measured, and iii) the result from measuring the trait with the specified method. This and many other features are shared with the more generic and detailed PaGE-OM reference model, the development of which was influenced by this work (and *vice versa*).

Through real-world use in the HGVbaseG2P project and informal validation exercises, the model has been tested and refined, and in its current form we consider it to be feature-complete and adequate for representing SNP-based association studies. However, further work will be needed to add support for CNVs and other forms of structural variation which are now beginning to be routinely interrogated in GWAS investigations.

Comparison to related domain models. Two related object models were published during the course of the work summarised above. A comparison of these models with the HGVbaseG2P model revealed several important commonalities and differences.

The generic, technology-agnostic FuGE-OM reference model was found to be broadly compatible with the HGVbaseG2P at the study metadata level, but provides better support for representing study provenance and protocols. XGAP-OM - a G2P-focused implementation model based on FuGE-OM - is not as expressive as the HGVbaseG2P model when it comes to describing markers, genotypes and phenotypes and other study metadata, but has at its core a far more flexible and extensible scheme for representing study data. Finally, both XGAP-OM and FuGE-OM provide standard facilities for associating most types of objects with ontology terms, whereas in the HGVbaseG2P model the use of ontologies is limited to sequence feature types.

The overall conclusion from this exercise is that, compared to these newly-emerged models, the HGVbaseG2P model is lacking in several important features. This does not detract from its utility in current use, as noted above, but does have certain implications relating to future use and development of the model and interoperability with other models (see more below).

7.2.2 A minimal data model for phenotypes

As work on the HGVbaseG2P project progressed, we entered into a collaboration with a number of groups with similar interests to form the GEN2PHEN Consortium. Several project activities are focused on developing domain data standards, with an overall goal of increasing standardisation in the G2P domain and to enable intra-domain data exchange and integration. Through GEN2PHEN, I became involved in these activities, and the previous work summarised above proved to be valuable input into collaborative modelling work led by others in GEN2PHEN.

Pheno-OM: an extensible model for observations. The main outcome of the above was a new object model for phenotypes and other observations. The modelling strategy followed is well-aligned with a recent trend in the field, towards reuse of core concepts from existing standard models and only creating new classes where absolutely necessary. As a

result of this strategy, the majority of Pheno-OM comprises reused or derived classes from FuGE-OM and PaGE-OM. The tripartite representation of “phenotype” is the core of the model, but with the critical difference that the three phenotype sub-concepts are connected to one another differently compared to PaGE-OM. This core is augmented by connections to several high-level organisational and utility classes from the generic FuGE-OM model. The final result from this “reshuffling” and extensive class reuse is a compact, composite object model which combines key features from key relevant models. Moreover, when compared to the HGVbaseG2P model, the new model addresses several of the limitations highlighted in the cross-model exercise summarised above.

A key significance of this work is not the creation of a new object model where there were no models before, but rather that the result is a model that is *practical*. Pheno-OM was deliberately designed to be minimal, easy to understand and, ultimately, to be extended and customised to fit specific implementation requirements (such as exchange formats and software infrastructure). This contrasts with some other related information models in the domain, such those developed by caBIG, mentioned in previous chapters, which contain hundreds of classes and are non-trivial to comprehend, and therefore challenging to reuse outside of the caBIG infrastructure proper. The minimal model, combined with facilities for describing complex biological concepts using domain knowledge captured by external ontologies, will be a powerful way of addressing syntactic and semantic interoperability challenges relating to phenotype data complexity in the G2P domain.

Future work. The Pheno-OM specification has now been published and a reference implementation has been created by lead partners on this project in order to test the model. However, further work is required to make the model fully usable. There is not currently a way to aggregate sets of observations into discrete datasets within a study, so the addition of a “Data” concept may be required. Representation of individual observations is simplistic and could be improved by adopting the flexible scheme employed by XGAP-OM. Finally, devising and publishing a formal UML representation should be a priority, so the model can be combined with other reference models, as further discussed below.

7.2.3 A modular G2P object model architecture

Given the advances in G2P domain models summarised above, it is pertinent to examine how these can now be applied in the HGVbaseG2P project. It is clear that several important features of the three other models considered here will be needed in HGVbaseG2P system in the future. For example, ontology support will be required for effective semantic annotation of GWAS traits, and better provenance for genotype and phenotype data is needed if more complex study designs (including GWAS meta-analyses) are to be captured in greater detail. New types of data will also need to be supported, including facilities for managing individual-level genotype and sequencing data which will be needed by the aforementioned in-a-box federated platform.

Re-engineering the HGVbaseG2P model as a composite. One strategy is to develop the existing model further and adopt key features from other models if they are needed. However, for the reasons discussed in Chapter 4, continuing a parallel line of development in this way would have numerous disadvantages, not the least from a maintenance perspective. In my view, it would be better to instead use the modelling strategy from above and re-specify the HGVbaseG2P model formally as a derivation of the models already discussed, by extending and adding classes and implementation-specific features where required. Such “re-engineering” would be a substantial undertaking, but would bring about a transformation of what is effectively a standalone modelling effort with informal ties to other modelling projects, into a GWAS-focused branch of broader standardisation activities linked together in a common modelling framework. This would bring benefits in the longer term for HGVbaseG2P development, as work could then be focused on the sub-domain specific aspects of the model, and would also help to ensure that model enhancements are shared with the community.

Assembling a standards-based data infrastructure. The proposed HGVbaseG2P model reworking above would start with the creation of an amalgam of at least 3 different models sub-model modules/domains: Pheno-OM and the SEQUENCE and GENOTYPE

domains from PaGE-OM (or derivations thereof). Conceptually, the composite model would cover the core of the current HGVbaseG2P model, but would then need numerous enhancements, such as addition of various “housekeeping” classes specific to the GWAS-focused HGVbaseG2P implementation. Importantly, unlike the current incarnation of the model, the final outcome of the above would be formally compatible with the source reference models, and with other implementation models built from them in the same way. Although useful on its own from a pure modelling perspective, the real value of such “mix and match” object model reuse relates to the broader context of standards-based, reusable and customisable software infrastructure for biology. Consider the biologist who needs to manage G2P data generated in his lab. At present, an feasible option is to download XGAP to serve as a base platform, then alter the underlying model and/or software configuration parameters to generate a customised system to suit his needs. Extending this concept further, one can envision a future scenario where the user elects to extend the base system by installing one or more add-on packages or bundles, each of which supplies a certain set of features. One package may comprise a series of data model enhancements and a suite of specialised software to support storage, retrieval and manipulation of sequence annotations. Another package based on Pheno-OM could add support for storing phenotype data, along with various ontology-based user interface components to aid with semantic annotation of data stored in the system. A third package would add support for genotype data, and so on. Importantly, an individual user or research lab would need to deploy exactly the set of add-ons required and would then customise the installation further if needed. This would result in a system that meets local requirements, but which is also standards-based and (at least in principle) provides for easy future integration with other systems.

The above proposal is based on the generative software strategy advocated by Swertz and Jansen (2007) previously discussed, and indeed numerous conversations and E-mails with the authors (M. Swertz, personal communication) have proved fruitful in developing many of these ideas. Similarities with (and inspiration from) the GMOD project and its philosophy are also acknowledged. However, the loose conglomerate of heterogeneous tools in the GMOD collection differs markedly from the strategy proposed

here, which is based on a common, formal conceptual framework and advanced model-driven architecture.

7.3 The role of digital identity in research

The other main offshoot of the HGVbaseG2P project originally arose from the need for user registration and access management facilities in the HGVbaseG2P web-based toolkit. A rather mundane search for suitable software tools for implementing this functionality resulted in a quite unexpected finding: that identity on the Internet could be pivotal in solving a variety of problems relating to data sharing in the biosciences, including those encountered in the HGVbaseG2P project.

7.3.1 A scholarly identity and contributor recognition

A key conclusion from my initial investigation was that digital identity would be key to incentive/reward-based schemes to encourage data sharing. This was the primary motivation for my work to develop a data submission use case involving DOI-based data registration and ORCID ID-based attribution. Although the use case is set in the specific context of mutation data exchange via the Cafe RouGE platform, the general concept can be applied to other data publication and attribution scenarios, including GWAS data sharing. Therefore, I expect that key technical aspects of this proof-of-principle software implementation will be useful to others as well. In particular, the implementation will serve as an important test case for the prototype system now being developed by the ORCID Technical Working Group.

Future work. If the Cafe RouGE pilot project is successful, a next step will be to apply the same concept in other projects in our group, such as HGVbaseG2P, and to seek collaborations with other research groups. We recently opened a dialogue with the research

team which develops the Dataverse data sharing platform³ (King, 2007). Dataverse, which is widely used in the social sciences for publishing and archiving data, provides facilities for persistent citations of data (akin to DOIs), but not does at present support federated identity. Interestingly, DataVerse has a number of other features which make it well suited to serving as the underlying platform for the aforementioned G2P data journal.

Another interesting area of future work is to extend the concept of identity-based attribution to address some of data quality challenges discussed in §2.3.3. Contributions of professional bio-curators could be tracked and attributed to their scholarly identity, thereby giving them publication credit for this important work. Community curation projects could also benefit in a similar way; for example, WikiProteins users could link their accounts to their ORCID IDs in order to be accredited for their contributions to the knowledge “commons”. However, as such “nano-publications” (to use the terminology of Mons and Velterop (2009)) would probably not be assigned persistent DOIs, specialised micro-attribution tracker systems would have to be developed for collecting citation links and other data, to be later mined for various purposes.

Assessing impact of data publications. There is a clear need for a transition to more appropriate and meaningful metrics which relate to the quality and impact of individual articles, rather than that of the journal they appear in (Campbell, 2008), and the same applies to data publication. However, the fundamental obstacle to accrediting researchers for online data releases has not only been the lack of an appropriate, accepted metric, but also a lack of data. Science metrics are data driven, and efforts to create, evaluate and use measures of academic productivity are dependent on the provision of infrastructure for gathering the necessary information on which to base them.

Now, as DataCite and ORCID gradually gain momentum, the required support systems for identifying, citing and attributing data publications will hopefully soon fall into place, and so the time is right to start investigating which metrics are appropriate for this form of scholarly output. Bourne and Fink (2008) have suggested a single Scholar Factor (SF),

³<http://thedata.org>

which incorporates a number of citations to papers authored, grant and paper reviews undertaken, submissions to databases, and more. Another view is that the more metrics that are provided, the better, and that different metrics should be used in different contexts. However, as Lane (2010) points out, choosing the right measure(s) for the right context requires some thought and should involve expert advice.

Privacy concerns. I expect that the majority of scholars will see a central scholarly ID as an opportunity to properly organise - and make more accurate - the publication-related information about them online that is in the public domain already. Many will also wish to use this ID for various other purposes as well. However, any centralised systems for tracking individuals on the Internet raises concerns about potential for abuse by governmental agencies or other parties. Some will be sceptical and view the whole scheme as a potential threat to privacy. However, as Wolinsky (2008) points out, there is “a careful balance to be struck between giving credit where credit is due and knowing everything about everyone”. Individual recognition and reputation are key drivers in science, and clearly some degree of tracking is certainly necessary to ensure scientists are properly credited for their contributions.

Provision of choice will perhaps be key to addressing such concerns: sceptics could simply choose not to use their scholarly identity for non-publishing related online activities, and so avoid the risk of being tracked. In this respect, an opt-in scheme based on the marriage of user-centric identity and a central scholarly profile contrasts starkly with top-down global identifier schemes with “Big Brother” connotations, such as the controversial UK national ID card scheme which has now been abolished⁴.

7.3.2 Identity-based security in distributed G2P data sharing

Another principal conclusion concerns the various data gathering challenges we faced in the HGVbaseG2P project and how digital identity could be applied to address those

⁴<http://www.telegraph.co.uk/news/newstopics/politics/7757720/ID-card-scheme-will-be-scrapped-with-no-refund-to-holders.html>

challenges. It seemed technically feasible to utilise OpenID and related simple protocols and open-source software tools to create practical solutions for managing access to potentially-identifiable aggregate GWAS data. This led me to develop a set of additional use cases for such controlled-access data sharing scenarios, again in the specific context of mutation data exchange via Cafe RouGE but with the general case in mind. Two approaches explored in these use cases are especially promising.

OpenID-based federated authorisation. The first scenario involved using remotely-located registry of access privileges and simple OpenID-based authentication to solve a specific problem - controlled access to mutation data - in a simple, clean way. The ICGC has come to similar conclusions and is currently implementing a system much like that I have proposed, for facilitating sharing of data generated in a major international project. Such a simple federated authorisation infrastructure, if adopted by other data providers in the domain, can greatly streamline sharing of GWAS and other sensitive G2P data. This would avoid the substantial overhead required for deploying sophisticated security infrastructure, such as that employed in caBIG and other large-scale biomedical grid projects. Importantly, such a strategy would enable secondary data providers (such as HGVbaseG2P) to re-distribute data to authorised users. Future work on this front should ideally include pilot projects with ICGC and EGA, focused on both devising software implementations and establishing a set of best practices and guidelines.

Semantic authorisation. The other main scenario involves the use of semantic reasoning for authorisation decision-making. One key benefit of semantic authorisation concerns scalability in a future distributed network of multiple data providers, access registries and heterogeneous ID profile information from many sources. When combined with lightweight federation infrastructure, this technique may be useful to help with minimising complexity of individual systems in such a network. I suggest as future work a proof-of-principle pilot project to create one or more software implementations to assess the feasibility of this approach.

Trust and identity assurance. If the federated authorisation scheme above is to be applied broadly in the domain, a key challenge will be to establish trust between data providers and identity providers concerning the identities of users. The scope of this challenge should not be underestimated. According to Manion *et al.* (2009), reaching agreement over a plethora of social and ethico-legal (rather than technical) issues concerning sharing of research data was one of the more difficult obstacles to overcome early in the caBIG project. With potentially thousands of data providers and identity providers all over the world, the task of establish pairwise trust between all of these is formidable.

Reassuringly, like many other non domain-specific issues surrounding security and privacy, this challenge is being tackled globally by various stakeholders. The recent creation of the Open Identity Exchange (OIX)⁵ - an independent, neutral certification body for digital identity providers supported by, amongst others, the NIH - is an interesting development on this front, albeit US-focused at present and mainly aimed at online interactions between citizens and government entities. However, if expanded to other domains as well as internationally, OIX and similar initiatives aiming to create a trust framework may in the longer term be key to increasing trust in Internet identities in general, and therefore ameliorate some the trust-related difficulties in research data sharing.

7.4 Future perspective

To conclude this thesis, I wish to finish with a short list of observations and predictions concerning future progress in the broader field of G2P databasing.

Simplifying grid building. There has been great progress towards sophisticated distributed infrastructure tailored to the needs of biomedical research. Open-source toolkits created in caBIG and BIRN will become increasingly important in large-scale translational medicine (Buetow, 2009). Though certainly of great value to BRICCS (see §2.7.1) and

⁵<http://openidentityexchange.org>

similar projects of that ilk, with multi-year funding and adequate in-house informatics expertise, the substantial technological overheads puts such an infrastructure strategy beyond the reach of many smaller projects.

Sharing and integration of genome annotation data via the DAS protocol is an excellent counter-example, which illustrates what can be achieved by simple, well-designed standards based on the REST style of web service architecture (Fielding, 2000). The REST approach offers numerous advantages due to its simplicity for both data consumers and tool builders (Stockinger *et al.*, 2008), whereas the prevailing mainstream SOAP-based grid architecture style tends to make the resulting web service APIs and overall infrastructure unnecessarily complex. Clearly, large-scale projects such as caBIG will continue to mandate such enterprise-level grid computing solutions. But for smaller databasing projects that require grid capability, a simpler, bespoke approach will often be sufficient, and a better fit for the scale of their operations and available resources and expertise (Pautasso *et al.*, 2008).

I predict that we will see these simpler federation techniques applied much more widely in the “long tail” of G2P databasing in coming years. A particular benefit of this will be to make easier the task of creating off-the-shelf, grid-enabled software packages such as those described above, which drastically reduce the level of competence required of users to create and populate G2P databases and connecting them to the grid.

Smarter search engines: the “killer app”⁶? As more and more databases find their way onto the grid, the next challenge is to make the most effective use of the available information and putting the grid to work. Most impact in the short term is likely to come from central search portals, which will enable non-technical users to undertake distributed searches across a segment of the total information space that is most relevant to their discipline. A community of users accustomed to routine “Googling” for Web content and searching the literature via PubMed will find it easy to adapt to a data access and

⁶The term was popularised by the VisiCalc spreadsheet software for the Apple II personal computer, see <http://www.dssresources.com/history/sshistory.html>.

retrieval modality which works in much the same way as the tools they are already familiar with, but with added capabilities. Moreover, recently developed methods for incorporating ontologies into search engines allow scientific knowledge to be taken into account when performing searches and presenting results (see also §5.5.2). Future grid-enabled search engines will therefore not only be able to search across multiple repositories, but they will also be a great deal smarter.

An integrated Web of G2P knowledge. In the same way that a Google search enables users to find Web content of interest, and subsequently locate and browse related content by clicking on hyperlinks, the new generation of specialised search portals will help researchers to find a suitable entry point into an integrated G2P knowledge environment. Initial iterations of this environment will likely be an evolutionary progression from present paradigms, such as the familiar report-style display of individual database entries or a key biological concept, cross-linked to related Web resources. As more online data resources become grid-aware, such reports will be greatly enhanced by dynamic discovery, retrieval and display of data from external sources on the grid.

As with the search modality, users' familiarity with the report mode of presentation means that they will be able to take advantage of federated resources without learning new tools. Genome browsers, already federated via DAS (and other means of overlaying external annotation tracks), can similarly be further enhanced. This general tactic - that is, to grid-enable web-based tools that researchers are already familiar with - will be key to bringing the benefits from federation to mainstream users in the research community.

Web 3.0 / the Semantic Web. The Web will continue to be an important platform for integration and analysis, not least because of the relative ease with which novel, lightweight bioinformatics tools can be created and quickly adapted to changing requirements. A key element in the Web's continued importance is its ongoing transition from a Web of documents to a Web of Data. New datasets are constantly added to the Linked Data

Cloud⁷, currently estimated to contain approximately 13 billion triples⁸. Semantic Web applications are emerging in disciplines as diverse as mobile phone industry and health care (Feigenbaum *et al.*, 2007).

Pilot projects in the life sciences, such as the NeuroCommons⁹ (Ruttenberg *et al.*, 2009), have been developing tools and best practices for a number of years, and these will now start to be widely applied in the G2P domain. Much of the relevant biomedical information is already published as Linked Data, some natively via the primary data providers, others via the Bio2RDF portal¹⁰ (Belleau *et al.*, 2008). However, various important datasets (including GWAS findings) presently are not, and many of the required ontologies are incomplete or nonexistent. Filling these gaps will be critical to progress.

“Rich” clients of the grid. The Web of data will provide fodder for a diverse collection of semantic mashup tools (Cheung *et al.*, 2008), which mediate semantic integration, or “smashups” of data from heterogeneous sources whilst hiding the complexities of semantic technologies from the user. But in order to support more advanced methods for utilising larger amounts of information accessible over the grid, web-based tools will need to be supplemented with specialised, standalone software tools running on users’ local computers. Such rich clients can provide far more sophisticated functionality than is possible in a browser, such as construction of web service-based workflows or visualisation of protein structures and interaction networks.

Given the relatively high level of tool sophistication, many of these standalone tools (notably those for creating complex workflows) will likely remain the domain of bioinformatics experts, although wet-lab biologists will need to be able to execute, reuse and adapt such workflows as part of their research activities. Rich clients will increasingly

⁷<http://richard.cyganiak.de/2007/10/lod/>

⁸[http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/
DataSets/Statistics](http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics)

⁹<http://neurocommons.org>

¹⁰<http://bio2rdf.org>

have semantic capabilities built in, enabling them to take full advantage of biomedical information available as Linked Data. Such knowledge management platforms will enable researchers to construct what Post *et al.* (2007) refer to as a “personal semantic framework” from a mix of external and local data and knowledge from diverse sources, and subsequently query and mine their custom knowledgebase to answer specific biological questions.

A. Research methods

A.1 Technical specification

Development work for the HGVbaseG2P project was carried out on a Hewlett-Packard ProLiant DL380 G5 server with 2x quad-core Intel Xeon 2.33GHz processors and 16GB RAM, running the open-source Debian Linux operating system v5 (Lenny)¹. Unless otherwise specified, the programming language used was Perl v5.10². Table §A.1 summarizes key third-party open-source software libraries and tools that were used to construct the various parts of the HGVbaseG2P system.

¹<http://www.debian.org>

²<http://www.perl.org>

Table A.1: Key third-party software libraries and tools used.

Library/tool and version	Source	Description
Apache v2.2.10	http://httpd.apache.org	Web server software providing the environment in which all web-based software components run.
BioMart v0.6	http://www.biomart.org	Query-optimized data mining platform used to create HGV mart.
BioPerl v1.6	http://www.bioperl.org	Library of modules for managing and manipulating biological data.
Catalyst v5.8	http://www.catalystframework.org	Provides Bio::DB::SeqFeature::Store and other modules employed to handle feature data.
DBIx::Class 0.08102	http://search.cpan.org/dist/DBIx-Class/	MVC framework. Underpins the main MVC web application.
DBIx::DBStag v0.09	http://search.cpan.org/dist/DBIx-DBStag/	Object-relational mapping framework used for core database access.
GBrowse v1.69	http://gmod.org/ggb	Older, legacy database API, still used by some applications for core database access.
GBrowse karyotype	http://gmod.org/GBrowse_karyotype	Region-level genome browser which underpins region-level views.
Getopt::Euclid v0.2.1	http://search.cpan.org/dist/Getopt-Euclid/	Genome-wide toolkit based on GBrowse which underpins the genome-wide views.
jQuery v1.2.6	http://jquery.com	Module which enables formal definition of a command-line interface as part of program documentation.
jQuery UI v1.5.3	http://jqueryui.com	Standard libraries used for dynamic client-side web page functionality.
MySQL v5.0.3.2	http://www.mysql.com	Relational database system used for data storage.
ProServer v2.0	https://www.sanger.ac.uk/Software/analysis/proserver/	Lightweight DAS server used to publish study association data as genome annotations.
Template Toolkit v2.19	http://www.template-toolkit.org	Template-based rendering of web application output in HTML or other formats.
Xapian v1.0.6	http://xapian.org	Search engine library, provides full-text indexing and searching facilities for certain types of database entries.

A.2 Processing and loading marker data with dbSNP-lite

In order to provide context for association study data in the Study database, the HGVbaseG2P Marker database contains a copy of core dbSNP data which serves as a reference information layer of known genetic variation. As with other reference sources of biological data, storing local copies of dbSNP data in HGVbaseG2P requires dealing with changes in the source database over time, as new data are gathered and existing records are altered. Ongoing variation discovery projects such as the 1,000 Genomes Project identify new and confirm suspected variant sites in the genome, and thus extend and enrich the dbSNP catalog. The reference genome sequence improves in quality, so genome mapping information in dbSNP changes with each new genome build. Erroneous dbSNP submissions are corrected or deleted. As a result of these and other changes, association study findings can be expected to gradually become inconsistent with data in the reference dbSNP archive. For example, some types of changes may invalidate earlier assumptions regarding genotyping assay designs or allele calling algorithms in generation of primary GWAS genotype data. Other changes may affect downstream analysis in subtle ways, for example changes in reported allele strand orientation. This complicates comparison and integration of association study findings from different points in time, generated in the context of different releases of dbSNP.

Motivation for dbSNP-lite. A key concern of HGVbaseG2P is to maintain a consistent link between association study data and the basal layer of reference marker data, such that references to marker information in the former (such as reported allele frequencies) can be properly updated to match changes in the latter. However, marker provenance information provided by dbSNP is not adequate for the detailed level of marker revision tracking deemed necessary for this task. Simply replacing a local copy of dbSNP data from a previous release with the contents of a new release was therefore deemed an unfeasible strategy for HGVbaseG2P, and maintaining complete copies of multiple previous dbSNP releases was also deemed impractical. Instead, the system for tracking changes in dbSNP

content was developed. The result from this work is the dbSNP-lite tool introduced in §5.1.4.

The output of dbSNP-lite is a “slim” version - or abstraction of core elements – from a complete dbSNP release in a standard format, enhanced with revision information for each marker describing the changes, if any, that were found. These data may have broader utility for others who wish to utilise dbSNP data in a similar way to HGVbaseG2P. For this reason, dbSNP-lite was developed as a standalone, reusable package which has previously been published as part of a report on GEN2PHEN deliverable D7.1. This report is accessible online ³ and also on the DVD (see §D). The subsections to follow contain a condensed, updated version of the D7.1 report and provide background, rationale and implementation details for the dbSNP-lite marker processing and import pipeline. Full sourcecode and further documentation are also provided on the DVD.

A.2.1 Overall design and architecture

A two step approach, illustrated in Fig. §5.14, was chosen for synchronising dbSNP with HGVbaseG2P. In the first step, the dbSNP-lite application processes dbSNP bulk datafiles and generates a set of standard GFF3-formatted output files. In the second step, the GFF3 files are loaded into the HGVbaseG2P Marker database. The rationale for this strategy was as follows. First, each of the two main software component could be focused on one task, and thus could be made simpler and easier to maintain (compared to a monolithic, more complex application). Second, a separate data loading step carries less risk of partial or incorrect data being imported into the database (since the full set of intermediary files from the processing stage can be checked beforehand). Third, the intermediary files in a standard format may by themselves be useful to others, and there was a desire to make the dbSNP-lite tool and its output more broadly useful, even without the presence of the full HGVbaseG2P system which is required for the import step.

³<http://www.gen2phen.org/document/d71-dbsnp-lite-established>

A.2.2 Input data for analysis

Each release of the dbSNP database is made available in bulk in several formats, including flat files, a variety of gene- and chromosome-oriented reports, full database table dumps and as a set of XML-files ordered by chromosome. For HGVbaseG2P purposes, the XML-dumps were considered sufficiently comprehensive, and more consistent and convenient to process than the other bulk download options. A partial example of an XML entry is provided in Listing §A.1. To prepare for dbSNP-lite processing, all available XMLs for the current dbSNP build (b130) were downloaded from the NCBI FTP-site⁴ onto the HGVbaseG2P Unix server.

Listing A.1: Partial rs# XML from dbSNP b130.

```

1  [...]
2    <Rs bitField="040108080001000000000100" molType="genomic" rsId="714"
   snpClass="snp" snpType="notwithdrawn">
3    <Validation byCluster="true"/>
4    <Create build="36" date="2000-09-19 17:02"/>
5    <Update build="130" date="2009-02-06 04:19"/>
6    <Sequence exemplarSs="11332849">
7      <Seq5>TTTCTTTATCCAGTTCTTCTACGGCTAT [... ]</Seq5>
8      <Observed>A/G</Observed>
9      <Seq3>ggaggcagagccttgagtgagccaagatcacaccactgcactacagcctggg [... ]</Seq3>
10   </Sequence>
11   <Ss batchId="485" buildId="36" handle="WIAF" locSnpId="WIAF-4147"
   methodClass="sequence" molType="genomic" orient="reverse"
   ssId="846" strand="top" subSnpClass="snp" validated="by-submitter">
12     <Sequence>
13       <Seq5>AAGCAGTAAATCTTCCATCATGCCA [... ]</Seq5>
14       <Observed>C/T</Observed>
15       <Seq3>GTTGGTTATAGCAGTCAACGACATCATCAATGA [... ]</Seq3>
16     </Sequence>
17   </Ss>
18  [...]
```

⁴ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/XML/

```

19     <Assembly current="true" dbSnpBuild="130" genomeBuild="36_3"
      groupLabel="Celera">
20     <Component accession="NW_926940.1" chromosome="18"
      componentType="contig" contigLabel="Celera" ctgId="980"
      end="14243427" gi="89047334" groupTerm="alt_assembly_1"
      name="HsCraAADB02.603" orientation="fwd" start="0">
21     <MapLoc alnQuality="1.00" asnFrom="10371899" asnTo="10371899"
      leftContigNeighborPos="10371898" leftFlankNeighborPos="99"
      locType="exact" numberOfDeletions="0"
      numberOfInsertions="0" numberOfMismatches="1"
      orient="reverse" physMapInt="10371899"
      rightContigNeighborPos="10371900"
      rightFlankNeighborPos="101"/>
22     </Component>
23     <SnpStat chromCount="1" hapCount="0" mapWeight="unique-in-contig"
      placedContigCount="1" seqlocCount="1" unplacedContigCount="0"/>
24 </Assembly>
25 [...]
26 <PrimarySequence accession="NM_153000" dbSnpBuild="130" gi="189409110"
      source="remap">
27     <MapLoc alnQuality="1.00" asnFrom="2478" asnTo="2478"
      leftContigNeighborPos="2477" leftFlankNeighborPos="0"
      locType="exact" orient="reverse" rightContigNeighborPos="2479"
      rightFlankNeighborPos="0"/>
28 </PrimarySequence>
29 <RsLinkout linkValue="2557" resourceId="1"/>
30 <RsLinkout linkValue="7811" resourceId="4"/>
31 <MergeHistory buildId="106" orientFlip="true" rsId="3170024"/>
32 <MergeHistory buildId="108" rsId="3748409"/>
33 <hgvs>NM_153000.3:c.*580A>G</hgvs>
34 <hgvs>NT_010859.14:g.10478615A>G</hgvs>
35 </Rs>
36 [...]

```

dbSNP organisation and build procedure. The central unit of organisation in dbSNP is the reference SNP cluster, usually referred to as a refSNP or rs#. An rs# represents a site in the genome, defined by a pair of 5' upstream and 3' downstream flanking sequences, which has been shown to vary between individuals. The stable, accessioned rs# record is created from one or more submitter SNP entries, or ss#'s, submitted by variation discovery projects large and small, which are clustered together based on sequence similarity and common mapping to a genomic contig. The final, reference list of alleles for the reference SNP entry is the combined set of alleles reported in each ss#. The number and source of the constituent ss#'s in a given rs#, as well as population data and other information, are used to assign validation status to the rs#, indicating the likelihood of the polymorphism being a *bona fide* variation and not an experimental artifact.

dbSNP reclustering. The NCBI prepares a new dbSNP build approximately 1-2 times per year by re-clustering all available ss#'s to create a new reference set of non-redundant rs#'s. During this build procedure, new ss#'s submitted to dbSNP since the last release are either clustered together with existing rs#'s, or else are used to seed new rs# clusters at sites in the genome not previously known to vary. Occasionally this results in changes at the rs# level (see below), but unlike GenBank and many other primary archives dbSNP does not employ a versioning and archiving for tracking such changes and making previous versions of changed rs#'s accessible.

rs# mergers and deletions. A side-product of evolution and improvements of the dbSNP re-clustering procedure and changing genome assemblies is that sometimes two or more rs# are found to represent the same polymorphic site in the genome. This results in a so-called refSNP 'merge' event between the co-located rs#'s, whereby the rs# with the higher number is merged into the rs# with the lower number (e.g. rs58061040 => rs626358) and subsequently deleted. rs#'s are also removed from the database if the underlying submitter-provided ss# entries are withdrawn for some reason. As with changes at the rs# level,

deleted markers in dbSNP are likewise not archived and kept accessible; instead they are removed completely from database releases (but see Discussion for exceptions to this).

A.2.3 Core processing of dbSNP entries

For each reference SNP entry found in the input XML-file, a core subset of available marker information is extracted:

- dbSNP rs# accession
- Variation class (e.g. “SNP”, “indel”)
- Validation code (e.g. “byFrequency”, “byHapMap”)
- List of reported alleles (e.g. “A/C”, “-/TTG”)
- 5’ upstream and 3’ downstream flanking sequences (30bp on either side)
- Cross-references to other NCBI resources, such as dbSTS and LocusLink
- Mapping information for all available genome sequence assemblies

Detecting and validating changes in core marker information. In addition to the data extraction step, for each rs# entry the workflow depicted in Fig. §A.1 is executed. The target HGVbaseG2P Marker database is queried using the dbSNP rs# accession for the marker. If the marker is not found in HGVbaseG2P, it is assumed to be a new marker. Otherwise the information from the XML-file is compared with information from the database and a series of checks are undertaken in order to identify which of a specific set of changes, if any, have occurred since the last dbSNP release was incorporated into HGVbaseG2P. Once this procedure has been completed for all markers in the input XML-files, each marker has been placed into one of the following categories:

New: Marker is not present in HGVbaseG2P.

Unchanged: Marker is present in HGVbaseG2P and no changes were detected.

Changed: Marker is present in HGVbaseG2P. One or more changes have occurred since the last source database build was processed and these could be reconciled automatically.

ManualCheck: Marker is present in HGVbaseG2P. One or more changes have occurred since the last source database build and these could NOT be reconciled automatically.

Fig. §A.1 depicts the workflow followed for each rs# to identify the following changes:

Change in variation class: a simple string comparison is used to detect changes in dbSNP-assigned variation class, such as a change from ‘snp’ to ‘mixed’ if an ss# reporting an insertion allele has been added to a rs# cluster where previously only single-nucleotide alleles had been reported.

Change in flanking sequences: occasionally the flanking sequences for established rs# entries is altered. This is typically due to changes in the set of ss#’s underlying the cluster, such as when new ss#’s have been clustered to the rs#. The ss# with the longest flanks is used as the cluster exemplar, and the master rs# flanks are changed to match the exemplar ss# if needed. Depending on the strand orientation of the exemplar ss# relative to the rs#, this process has in the past sometimes resulted in reverse-complementation of flanking sequences, or ‘strand flip’, for the rs# cluster. dbSNP-lite checks for flank changes and strand flips by aligning rs# flanking sequences in the current dbSNP build with those from the previous build.

Change in reported alleles: as new ss#’s are clustered into existing rs# clusters, previously known alleles are confirmed with independent observations or new alleles are added to the rs#. dbSNP-lite identifies these changes and checks that they are valid, taking into account potential reverse-complementation of the flanking sequences.

A.2.4 dbSNP-lite output

The output of dbSNP-lite is a set of standard GFF3-formatted feature files which contain the “lite” representation of dbSNP. A feature file labelled “all” is created for each input file

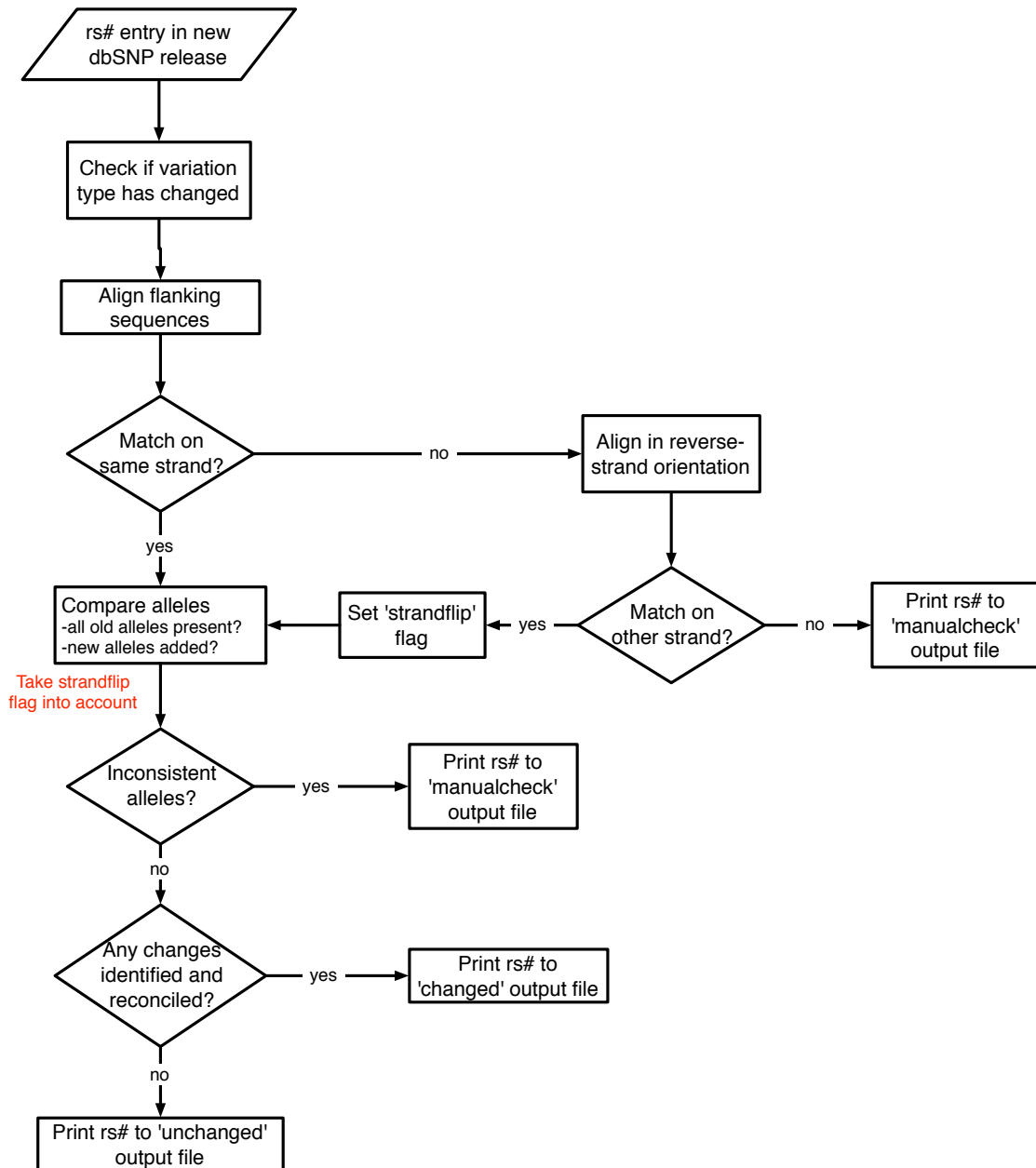


Fig. A.1: Workflow for comparing and validating marker entries in new dbSNP build against existing entries in the HGVbaseG2P Marker database.

per sequence assembly, containing all marker information and mappings but no change information. This file is provided as a convenience for users who only require a lite version of the current dbSNP and are not concerned with changes since the last build. A further set of feature files are created for each change category listed above. For example, new markers which are mapped to Chr18 in the reference sequence assembly are printed to the output file `markers_dbSNP_ds_ch18.reference.new.gff`, and existing markers with no changes mapping to Chr1 in the Celera assembly are printed to `markers_dbSNP_ds_ch1.Celera.unchanged.gff`.

As per the standard GFF3 specification⁵, the chromosome, genomic coordinates, source, type and strand are placed in the appropriate columns in the tab-delimited feature file. The variation type is specified as a standard Sequence Ontology terms, which are mapped to the non-standard dbSNP classification as shown in Table §A.2. Other marker information is encoded as attribute-value pairs in the 9th column, as per the specification. If a marker has multiple mappings within an assembly or to alternative assemblies, this core marker information is duplicated across the respective chromosome feature files as required. An example illustrating the organisation of marker data in a GFF3-formatted feature file is provided in Listing §A.2.

A.2.5 Loading marker data into the target database

After running the dbSNP-lite tool to produce the GFF3-formatted feature files, the feature files can be loaded into the target marker database or utilised in other ways. A variety of software tools are available for processing GFF3 files, such as those provided in the BioPerl toolkit. By reusing and extending one of these tools - the `Bio::DB::SeqFeature::Store` feature database and software library - a GFF3-loader tool for processing the marker feature files was created for importing both the core marker information and standard feature data into the HGVbaseG2P Marker database.

⁵<http://www.sequenceontology.org/gff3.shtml>

Table A.2: Mapping of dbSNP variation class to standard Sequence Ontology terms. Note that some dbSNP variation classes map to the same SO term.

dbSNP variation class	SO term name	SO term URL
snp	SNP	http://www.sequenceontology.org/miso/current_release/term/SO:0000694
in-del	indel	http://www.sequenceontology.org/miso/current_release/term/SO:1000032
heterozygous	complex_substitution	http://www.sequenceontology.org/miso/current_release/term/SO:1000005
microsatellite	tandem_repeat	http://www.sequenceontology.org/miso/current_release/term/SO:0000705
named-locus	complex_substitution	http://www.sequenceontology.org/miso/current_release/term/SO:1000005
mixed	complex_substitution	http://www.sequenceontology.org/miso/current_release/term/SO:1000005
CopyNumber	CNV	http://www.sequenceontology.org/miso/current_release/term/SO:0001019
Inversion	chromosomal_inversion	http://www.sequenceontology.org/miso/current_release/term/SO:1000030
InversionBreakpoint	inversion_breakpoint	http://www.sequenceontology.org/miso/current_release/term/SO:0001022

A.2.6 Checking for deleted markers in source database

As noted above, rs# entries do not appear in bulk dbSNP downloads after having been removed or merged with other rs#'s. This lack of archiving of deleted markers means that a separate procedure is needed to identify which rs#'s now present in HGVbaseG2P are no longer in dbSNP for either of these two reasons:

Merged markers: dbSNP provides a running log of merge events in their database table dump. The file RsMergeArch.bcp.gz containing this information was downloaded from the dbSNP FTP-site and processed with a custom script in the HGVbaseG2P toolkit. This adds

Listing A.2: Marker GFF3 example.

```

1 ##gff-version 3
2 ##genome-build NCBI B36
3 ##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=Homo+sapiens
4 # Timestamp=2010-01-11 17:35:31
5 # Label=none
6 # MarkerCategory=new
7 # AssemblyName=reference
8 Chr18 dbSNP indel 39617269 39617270 . - . ID=rs62659061;Name=rs62659061;upstream30bp=
AGAGCAAGAACTGTCTCAAAAAAAAAAAAA;downstream30bp=AAAAAAAAAGAAAGAAAAAGATTAT;alleles=(A):(C);
mapweight=unique-in-contig
9 Chr18 dbSNP SNP 33224768 33224768 . - . ID=rs62660160;Name=rs62660160;upstream30bp=
GGGAAGGCTGGTGGTGGTGGGGGGG;downstream30bp=GGGGGGGCAGTGTCTGGCCGGCACAGCA;alleles=(A):(G);
mapweight=unique-in-contig
10 Chr18 dbSNP indel 47613689 47613690 . - . ID=rs62668563;Name=rs62668563;upstream30bp=
AAAAAAAAAAAAAAAAAAAAAAAAAAAA;downstream30bp=AAAGAAAAAGAAAGAAAAAGAAATGC;alleles=(G):(C);
mapweight=unique-in-contig
11 Chr18 dbSNP SNP 47172821 47172821 . - . ID=rs62669164;Name=rs62669164;upstream30bp=
GTGGCACTGGGGAAATGTTATCAGCTGGG;downstream30bp=TTTTTTTTTTGTTTGACAGTTTGTTT;alleles=(G):(T);
mapweight=unique-in-contig
12 Chr18 dbSNP indel 53613034 53613035 . - . ID=rs62675162;Name=rs62675162;upstream30bp=
AGCGAAACTCCGCTCAAAAGAAAGAAAG;downstream30bp=AAAGAAAAAGAAAGAAATGGCATAT;alleles=(GAAAGAAA):(C);
mapweight=unique-in-contig
13 Chr18 dbSNP indel 11243564 11243565 . - . ID=rs62682662;Name=rs62682662;upstream30bp=
=CCTCCAAGTGAACAATCTAATTGAAAAAAT;downstream30bp=GCTAAAAAACAGTGATAACCAAAATTGTTGC;alleles=(A):(C);
mapweight=unique-in-contig
14 Chr18 dbSNP indel 23794009 23794010 . - . ID=rs62685163;Name=rs62685163;upstream30bp=
AAACACACACACACACACACACACA;downstream30bp=AACACACACTCGTGATTAGTTGCCATTT;alleles=(CA):(C);
mapweight=unique-in-contig
15 Chr18 dbSNP SNP 5579629 5579629 . - . ID=rs62691162;Name=rs62691162;upstream30bp=
AGAAACCTTCCATTGTTATATATACCCC;downstream30bp=GTCTTCCCATACCTATGCTTTAATTATCAA;alleles=(A):(C);
mapweight=unique-in-contig
16 Chr18 dbSNP SNP 22126911 22126911 . - . ID=rs62693060;Name=rs62693060;upstream30bp=
TTGATAAAACAACCTATTAGTAAATAAGGA;downstream30bp=TTATTAACTTATAAAATTTGCTATAGATT;alleles=(C):(T);
mapweight=unique-in-contig

```

to the HGVbaseG2P Marker database a revision history entry for every rs# which has been merged with another rs# (and subsequently deleted in dbSNP), and creates link between the two entries, effectively duplicating the dbSNP merge history in HGVbaseG2P.

Deleted markers: Following the loading of marker feature files into the HGVbaseG2P Marker database, a final “legacy” check is undertaken. This is a simple procedure which retrieves all markers which, according to a timestamp-flag, did not feature in the current dbSNP release. For each of these markers, if the rs# is not already logged as deleted due to a merge event, it is flagged as “dead”.

Importantly, in either of the two scenarios above the affected rs# entry is *not* deleted from the Marker database but archived. The net effect of this is that an archive of deleted rs# entries is maintained in HGVbaseG2P, even if they no longer appear in dbSNP itself. This information can then be used for rs# lookups and validation, for example when association study data loaded into HGVbaseG2P refers to rs# identifiers for deleted markers.

A.2.7 Limitations and future work

In its present form, the dbSNP-lite tool has some limitations. This first version of the software is capable of processing only dbSNP-data, but ideally other reference sources of variation data should be supported as well by adding new subclass modules. Some main sources of interest include the DGV and dbVar databases for structural variation, as noted in Chapter 5. Genome mapping information is presently handled in a simplified way, with existing mapping data in HGVbaseG2P simply replaced with mapping data from the new dbSNP release. A more intelligent procedure for verifying that marker mappings to the same genome assembly remain unchanged between dbSNP builds would be useful here.

The software implementation, albeit already modular and extensible, could be enhanced with a simple “plugin” architecture which would make it much easier to extend the package, in particular for the benefit of external users without access to the main

HGVbaseG2P database. A plugin architecture would help with the above and other future extensions, such as extracting additional fields from the source datafiles.

A.3 XML-based data loading tools

The HGVbaseG2P legacy XML-based import tools are based on the DBIx::Stag package. DBIx::Stag and supporting software packages were originally created in the Berkeley Drosophila Genome Project⁶ for manipulating Gene Ontology data as hierarchical structures, and also underpin XML-based import and export tools currently used for the GMOD Chado relational database⁷ (Mungall and Emmert, 2007). The main purpose of the DBIx::Stag middleware is to handle the mapping of data held in a hierarchical XML-structure to the multitude of SQL-statements needed for inserting or updating the relational database. DBIx::Stag inspects the table structure (including foreign-key relationships between tables) of the target database and automatically infers which XML-structures are valid. As a source XML document is parsed and the data extracted, the necessary SQL-statements are generated and sent to the database. This is best illustrated by an example. When the XML-structure shown in in Fig. §A.2 is processed by the import pipeline, the data elements are inserted into the `Samplepanel` table and several related tables. If the XML-structure does not match the structure of the target database, an error is reported and the import task aborted.

Advantages of XML to SQL mapping. Traditional solutions to the database-interaction part of the above would typically involve a substantial number of SQL-statements for querying, inserting and updating the various tables, nearly all of which would be specific to this particular type of data import. The main advantage of XML-to-SQL mapping is that all of the low-level database operations are handled by the DBIx::Stag middleware, including non-trivial lookups and inserts/updates for many-to-many linking tables. Importantly, any XML-data with a hierarchical structure matching the relational database structure (e.g. data destined for the `Study` and `Experiment` tables) can in principle be imported directly with no extra coding or custom SQL-statements. This flexible import facility was extremely

⁶<http://www.fruitfly.org>

⁷<http://gmod.org/wiki/Chado>

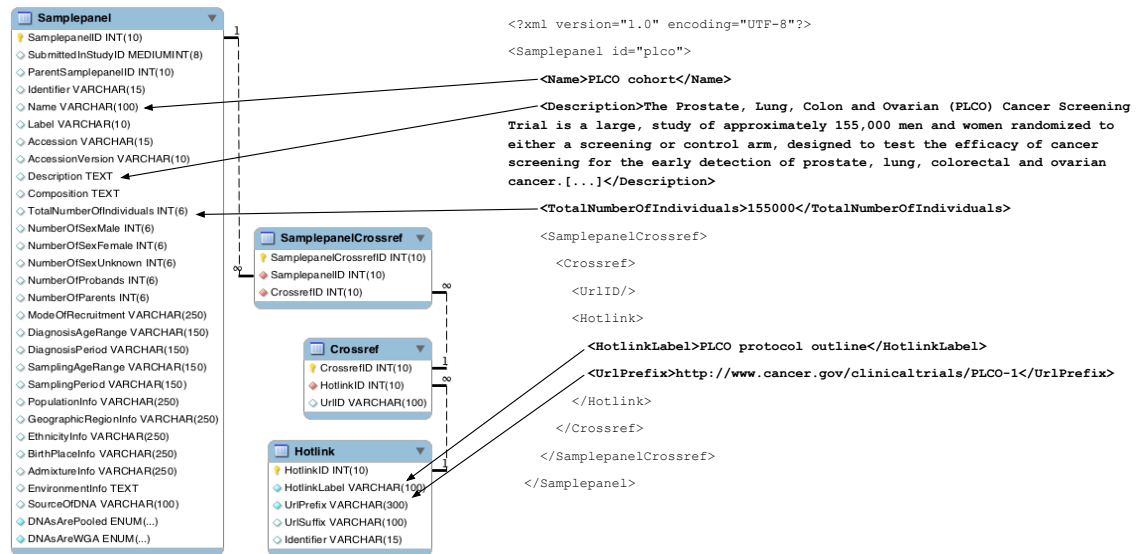


Fig. A.2: Example study metadata as XML and matching relational structures. The Samplepanel, SamplepanelCrossref, Crossref and Hotlink container elements in the XML all correspond to database tables with the same name. The four elements highlighted in bold, which contain textual or numeric data only, correspond to table columns into which the data will be inserted or updated. The two linking tables are populated automatically by the XML-to-SQL middleware.

useful in early stages of the project, at a time when the relational schema, data sources and various software components were under heavy development and therefore constantly changing.

HGVbaseG2P-specific tools for data loading. In order for the XML to SQL middleware to be fully usable for HGVbaseG2P-specific tasks, several enhancements were required. The most significant of these was the creation of a standalone command-line application and several Perl modules which tie together a low-level XML-parser, custom data handlers and the XML-to-SQL middleware into a common XML-loading framework. Full sourcecode and further documentation for these tools is provided on the DVD (see §D).

Disadvantages of XML to SQL mapping. Though useful earlier in the project, the XML-based data loading technique proved not to be a feasible strategy in the long run. One key factor was unacceptably slow performance with genome-wide aggregate GWAS

datasets in the early data gathering phase of the project (see §5.4). After conversion from the source data format into HGVbaseG2P XML, typical datasets for several hundreds thousands SNPs required up to several days to complete loading.

As noted in §5.1.4, one reason for the slow performance was the inherent overhead of XML as a transport format. However, another crucial factor was excessive verbosity of XML-data, mandated by the complex database structure to which it mapped. To illustrate this, consider the HGVbaseG2P XML example in Listing §A.3. The XML contains SNP marker information and genotype frequencies extracted from the tab-delimited representation shown in §A.4 and matches an earlier version of the HGVbaseG2P relational schema. This example illustrates the complex, multi-level nested XML-structures that are required for populating the database with relatively simple data. As a result, developing the necessary tools for converting source datafiles to XML which matched the database often became a more difficult task than expected. Another disadvantage of the scheme was that the generated XML-files were inherently “hardwired” to a specific relational schema, and so could typically not be imported into future versions of the database which were subtly different from previous ones.

Listing A.3: Aggregate GWAS genotype data as HGVbaseG2P XML, as used in early stages of the project. Only genotype frequencies for the case group are shown.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <Study>
3   [...]
4 <Experiment>
5   <Usedmarkerset>
6     <Name>rs8105536</Name>
7     <MarkersetMarker>
8       <Marker op="lookup" id="marker_rs8105536">
9         <Identifier>HGVM8366234</Identifier>
10        <Accession>rs8105536</Accession>
11      </Marker>
12    </MarkersetMarker>
13    <GenotypeFrequencyCluster>

```

```

14      <NumberOfGenotypedSamples>1960</NumberOfGenotypedSamples>
15      <PValueHWE>0.1</PValueHWE>
16      <AssayedpanelID>apanel_CASE</AssayedpanelID>
17      <GenotypeFrequency op="insert">
18          <FrequencyAsProportion>0.187</FrequencyAsProportion>
19          <NumberSamplesWithGenotype>366</NumberSamplesWithGenotype>
20          <GenotypeCombo op="insert">
21              <Genotype>
22                  <AutoMarkerID>marker_rs8105536</AutoMarkerID>
23                  <GenotypeLabel>(A)</GenotypeLabel>
24              </Genotype>
25          </GenotypeCombo>
26      </GenotypeFrequency>
27      <GenotypeFrequency op="insert">
28          <FrequencyAsProportion>0.350</FrequencyAsProportion>
29          <NumberSamplesWithGenotype>686</NumberSamplesWithGenotype>
30          <GenotypeCombo op="insert">
31              <Genotype>
32                  <AutoMarkerID>marker_rs8105536</AutoMarkerID>
33                  <GenotypeLabel>(G)</GenotypeLabel>
34              </Genotype>
35          </GenotypeCombo>
36      </GenotypeFrequency>
37      <GenotypeFrequency op="insert">
38          <FrequencyAsProportion>0.463</FrequencyAsProportion>
39          <NumberSamplesWithGenotype>908</NumberSamplesWithGenotype>
40          <GenotypeCombo op="insert">
41              <Genotype>
42                  <AutoMarkerID>marker_rs8105536</AutoMarkerID>
43                  <GenotypeLabel>(A)+(G)</GenotypeLabel>
44              </Genotype>
45          </GenotypeCombo>
46      </GenotypeFrequency>
47  </GenotypeFrequencyCluster>
48  [...]

```

Listing A.4: Aggregate GWAS genotype data and analysis results from the WTCCC study. The first two lines from the tab-delimited Chr19 file for bipolar are shown.

[illegible]

Ultimately, these various issues were deemed to outweigh the advantages gained by an XML-based data loading strategy. This led to the development of the template-based import pipeline described in §5.1.4, which processes source GWAS datafiles and inserts data directly into the database in a single step. This new pipeline, orders of magnitude faster than the XML-based pipeline, is currently used in HGVbaseG2P routine data loading.

The need for a more efficient relational implementation. Although effective performance-wise, the new data loading pipeline does not in itself address another core issue: the overall complexity of the HGVbaseG2P relational schema. As discussed in §5.1.2, the table-per-class strategy employed for creating a relational implementation from the conceptual model has resulted in excessively complex tables and table relations in several places in the relational schema (the most critical of which have been addressed by later work, it should be noted). In hindsight, a more pragmatic approach, centred on efficient, flexible relational structures for the core types of GWAS data that need to be stored, would have been more effective (see also discussion in Chapter 5).

It is clear from the experiences outlined above that XML as a data transport format can be useful in some settings, but is not an appropriate technique for representing large-scale datasets. My overall conclusion is that XML creates more problems than it solves, at least in the majority of scenarios of relevance in this project. In some settings, where dataset size and loading performance are not a concern, an XML-based solution has some merit; for example, XML is still used in the project for loading GWAS study metadata into the database. However, the current procedure of creating and editing the metadata

XML by hand is very cumbersome and error-prone. In hindsight, applications such as PEDRo⁸ (Garwood *et al.*, 2004b), originally created for the proteomics community, could perhaps have been adapted and used for study metadata capture and editing, and subsequent generation of the required XML.

Availability of tools for XML data entry does not, however, address a certain fundamental issue, namely that specialised tools are inevitably needed for manipulating XML data. For example, the aforementioned PEDRo tool was developed as the data entry component of broader data storage, search and presentation framework for experimental proteomics data (Garwood *et al.*, 2004a; Taylor *et al.*, 2003). After a period in the early 2000s when XML was the preferred solution for such projects, in recent years there has been a marked trend in the field towards simpler, tab-delimited data exchange formats which can be manipulated with spreadsheet software. A key advantage of this is that end users need not learn to use a specialised tool to enter data and, importantly, such tool therefore need not necessarily be developed (and maintained) by bioinformatics groups. Prominent examples of this trend are the MAGE-TAB format for microarray data and the more generic ISA-TAB format, both of which were mentioned in previous chapters. These formats are currently being piloted in various G2P databasing projects in GEN2PHEN and are also being evaluated for use in this project.

⁸<http://pedrodownload.man.ac.uk>

B. Supplementary materials for modelling chapters

B.1 PaGE-OM logical model diagrams

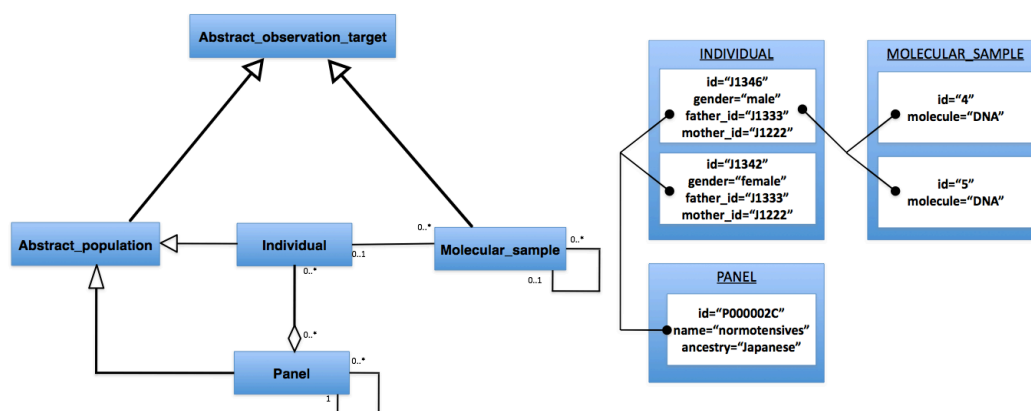
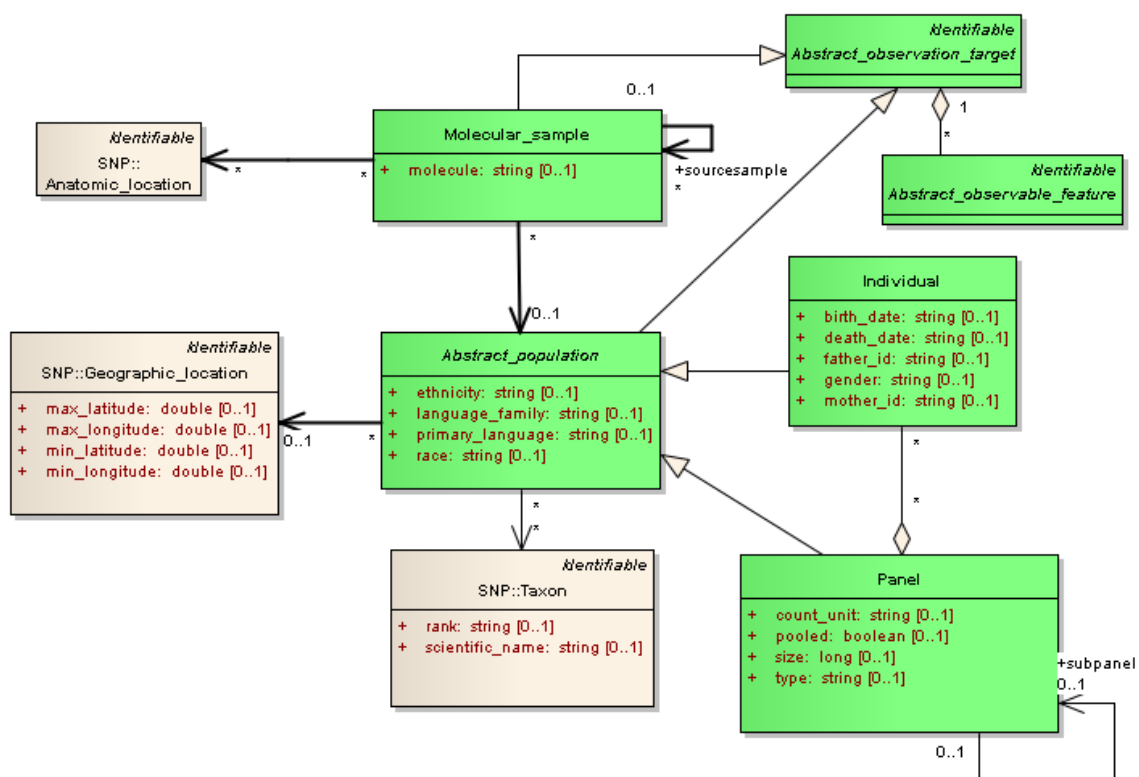


Fig. B.1: The PaGE-OM SAMPLE domain, simplified logical diagram and data example. From Brookes *et al.* (2009).



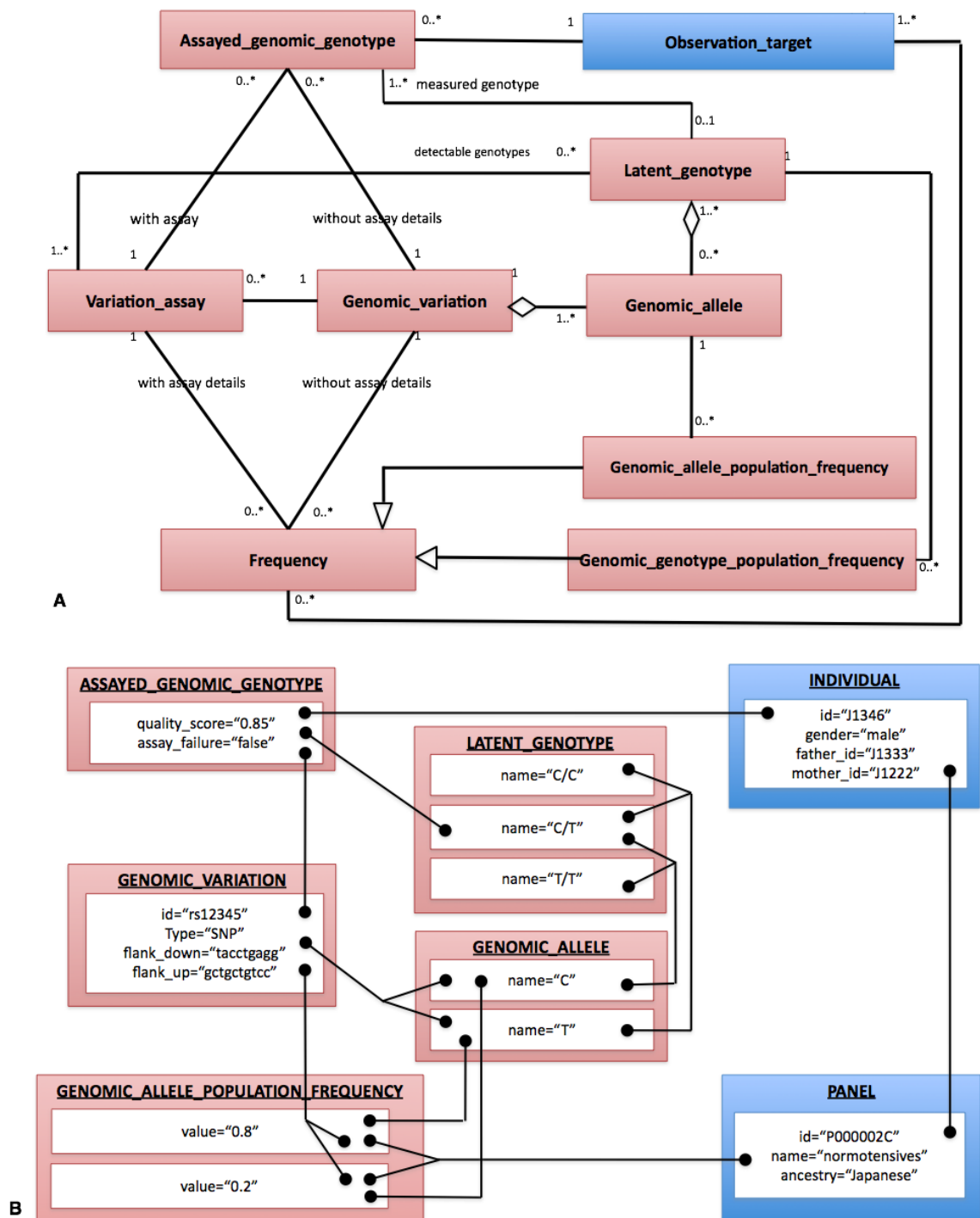


Fig. B.3: The PaGE-OM GENOTYPE domain, simplified logical diagram and data example. From Brookes *et al.* (2009).

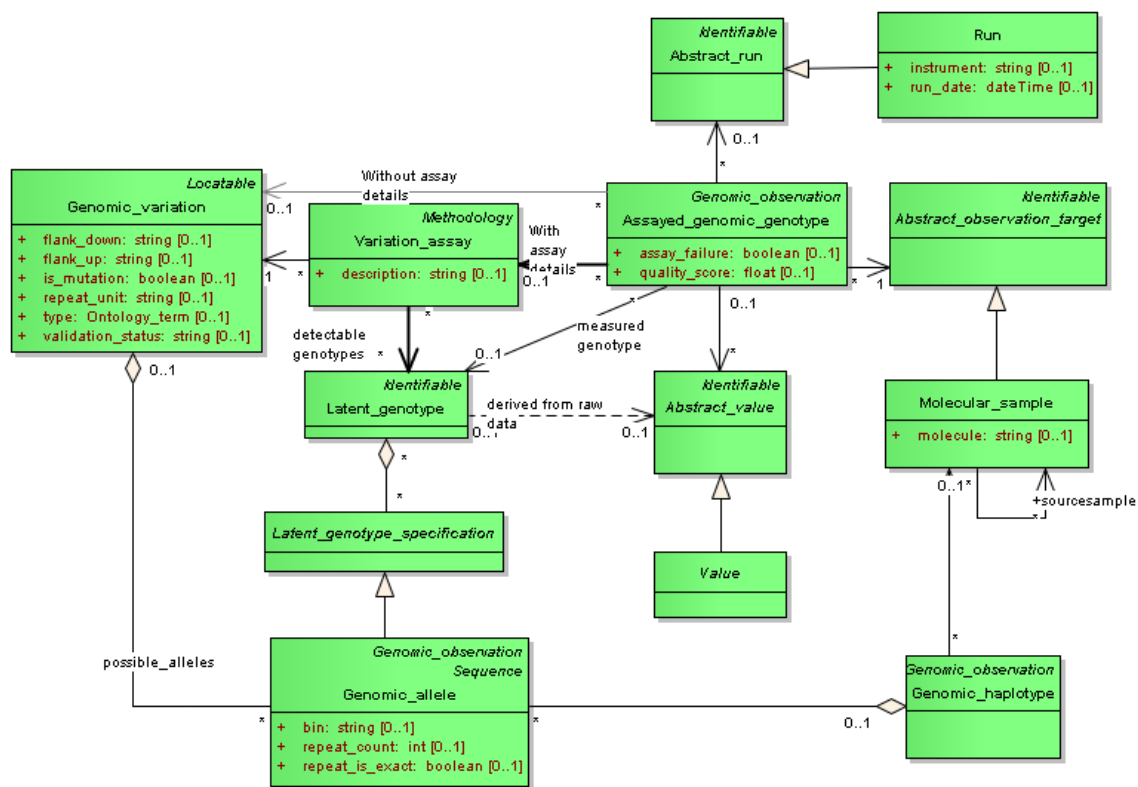


Fig. B.4: The PaGE-OM GENOTYPE domain, overview. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA5.htm).

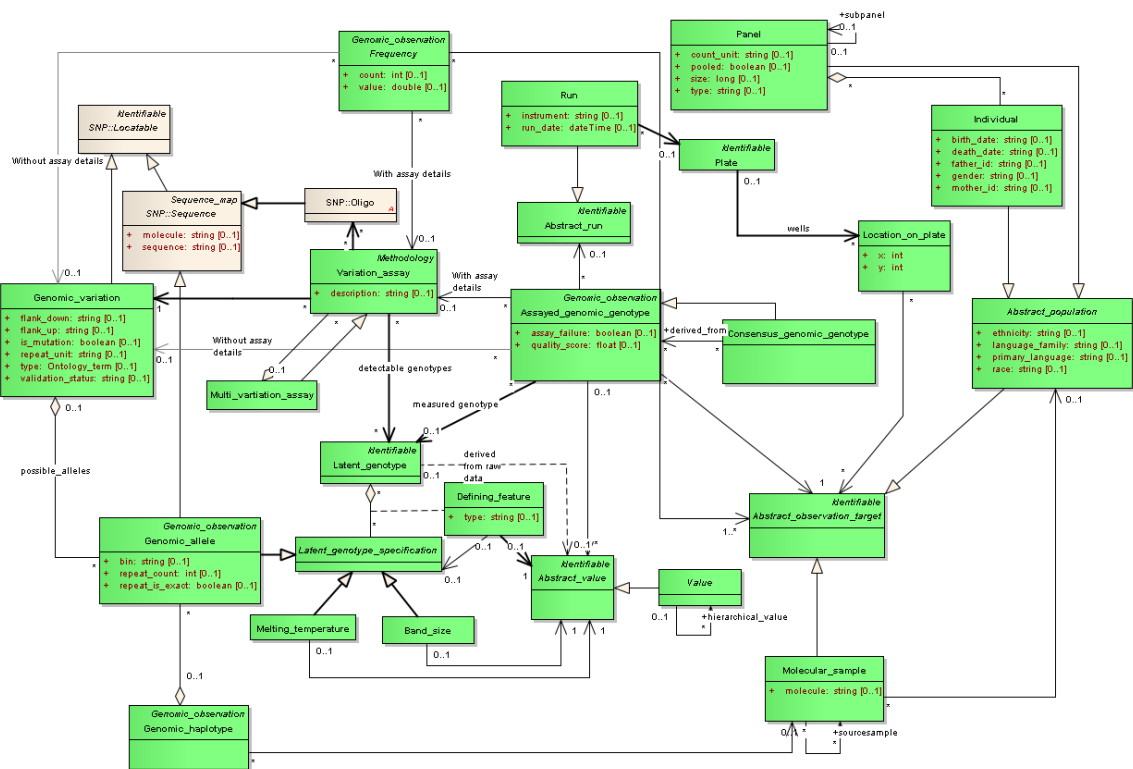
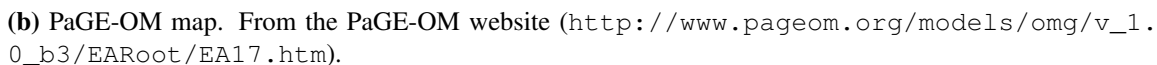
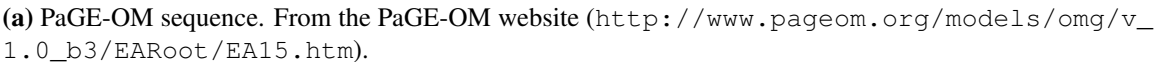


Fig. B.5: The PaGE-OM GENOTYPE domain, details. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA9.htm).



291

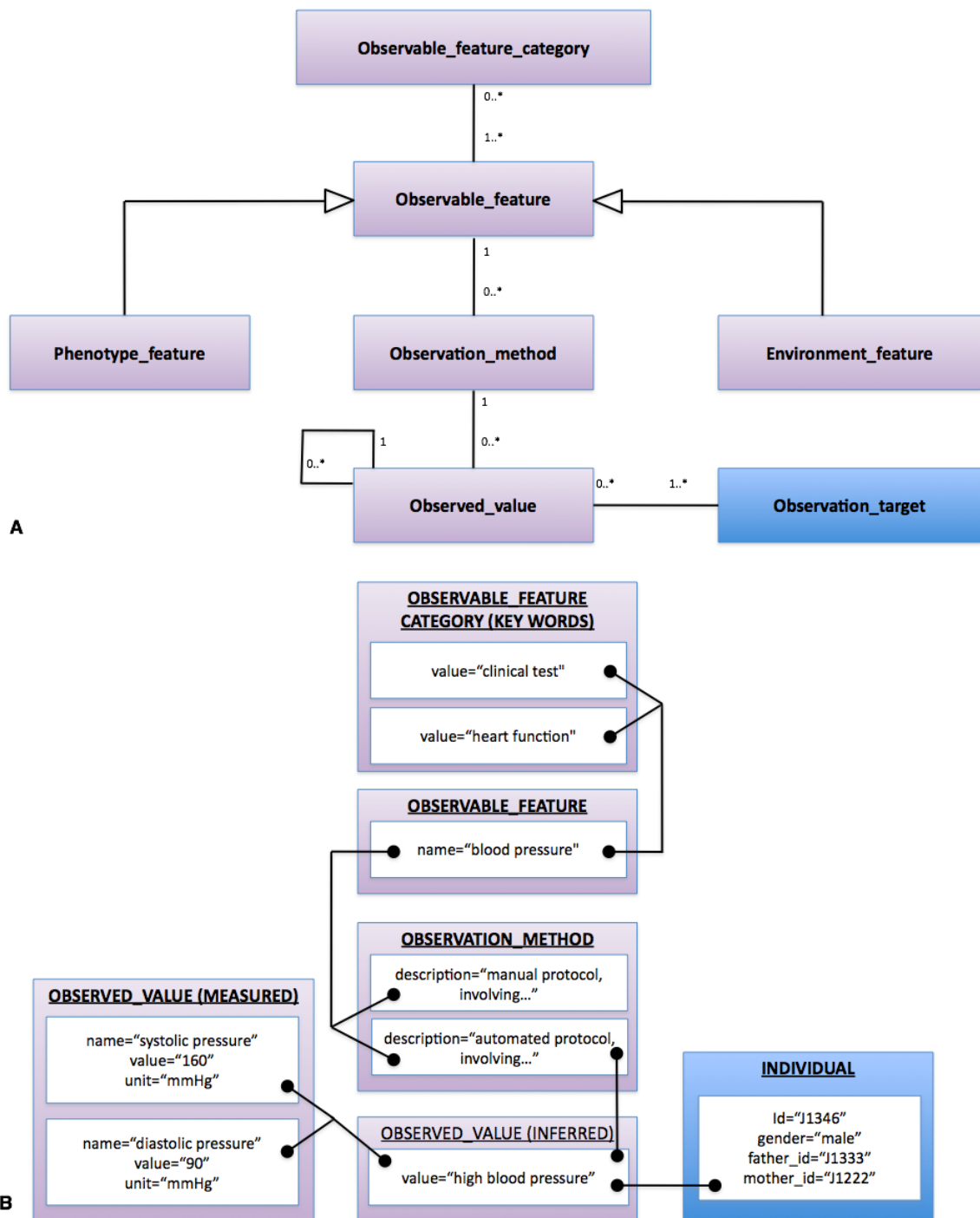


Fig. B.7: The PaGE-OM PHENOTYPE domain. Simplified logical diagram and data example. From Brookes *et al.* (2009).”

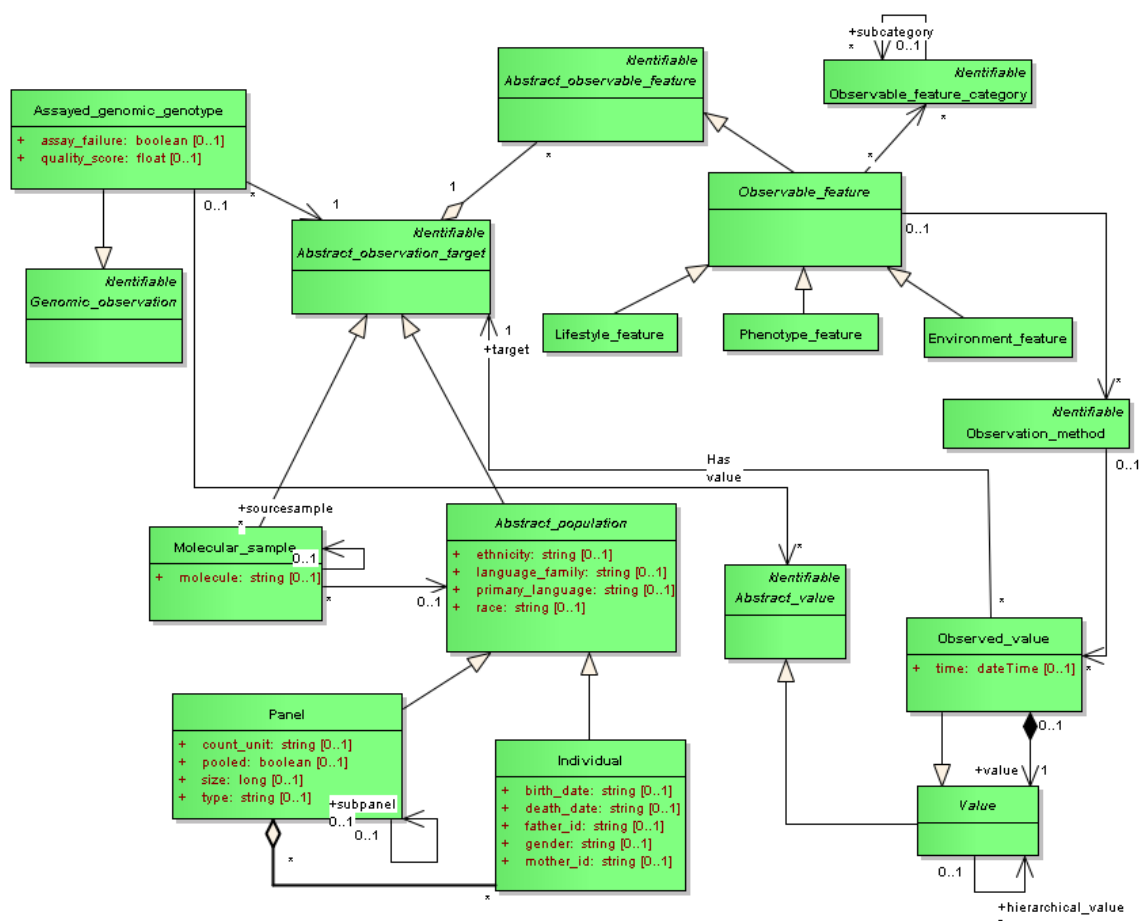


Fig. B.8: The PaGE-OM PHENOTYPE domain, details. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA7.htm).

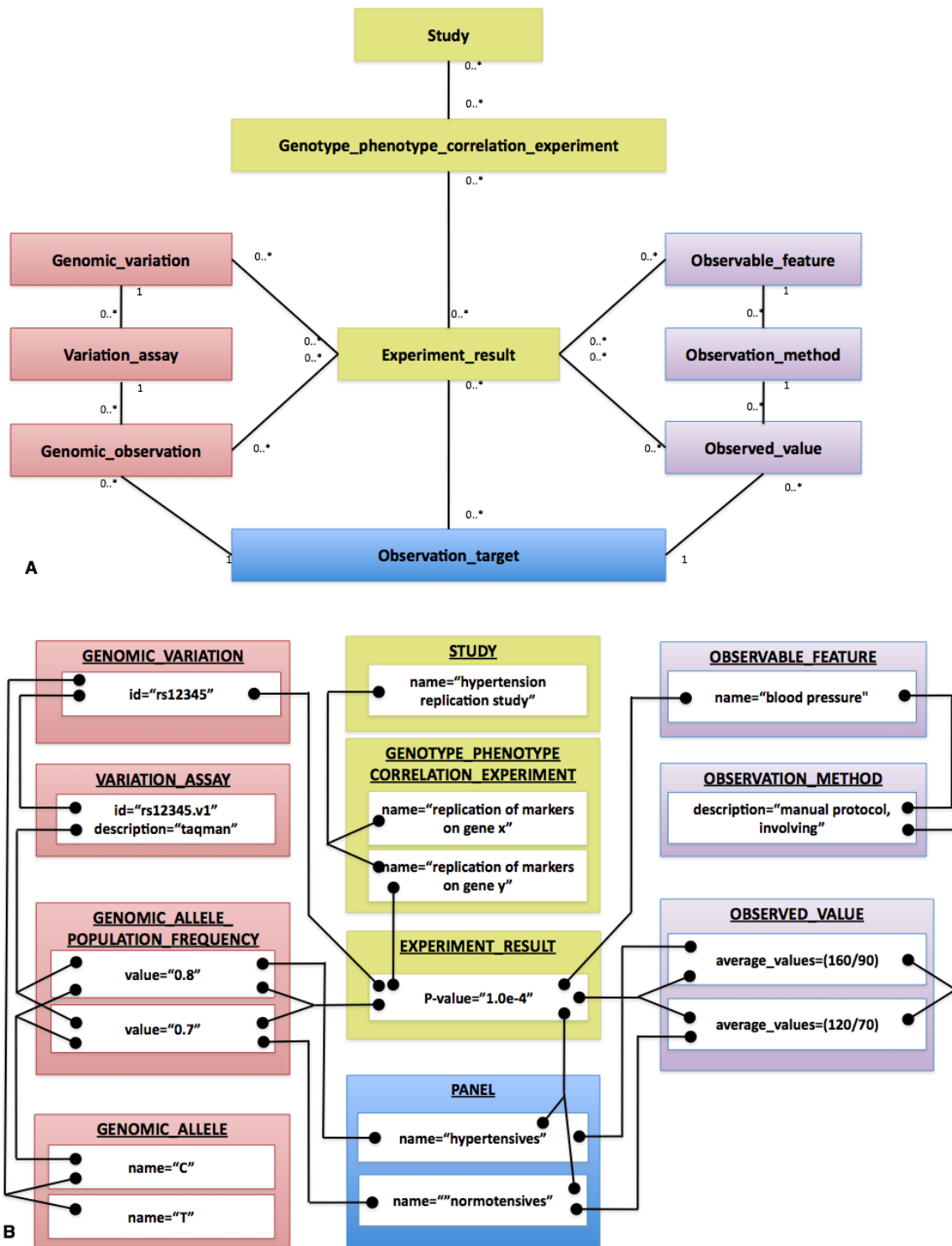


Fig. B.9: The PaGE-OM EXPERIMENT domain, simplified logical diagram and data example. From Brookes *et al.* (2009).

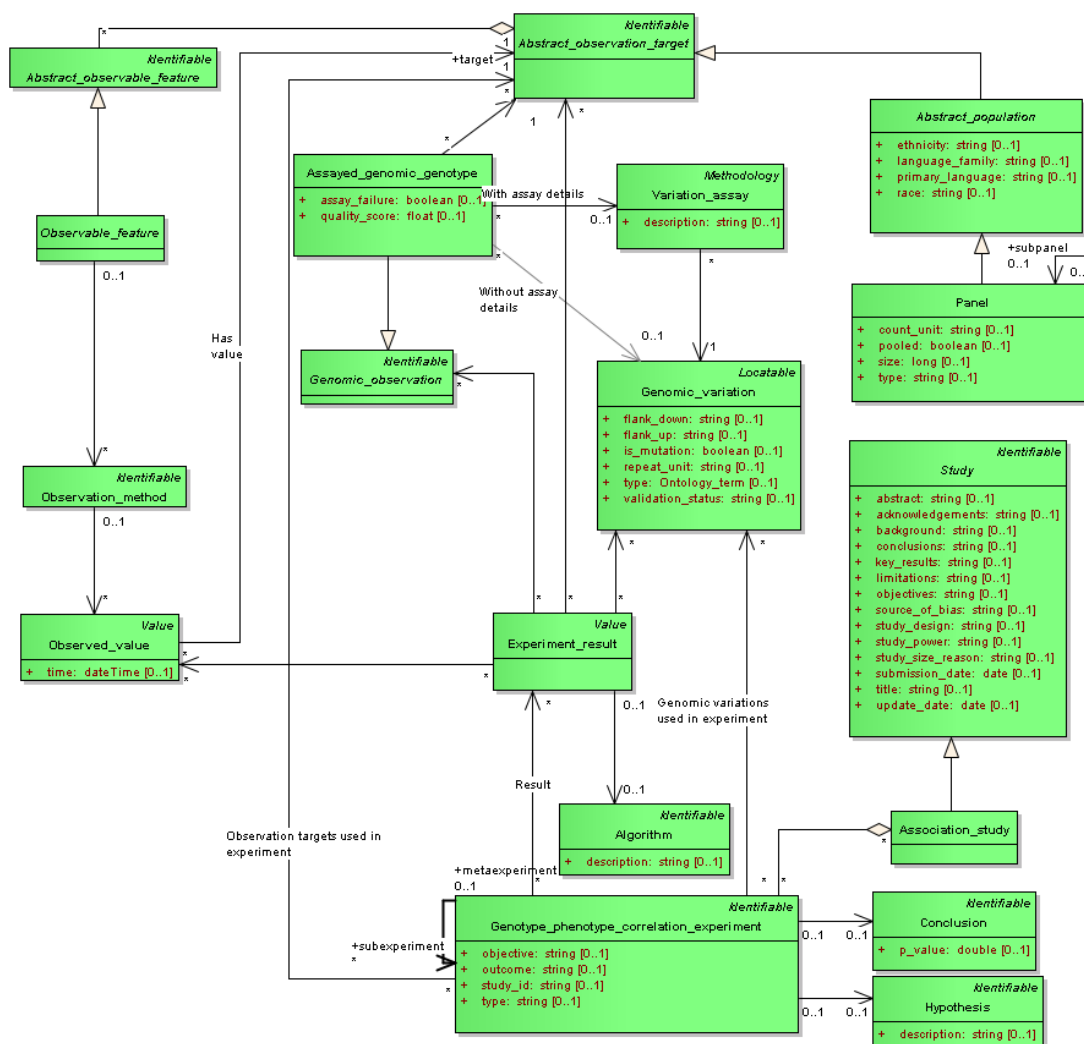


Fig. B.10: The PaGE-OM EXPERIMENT domain, details. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA1.htm).

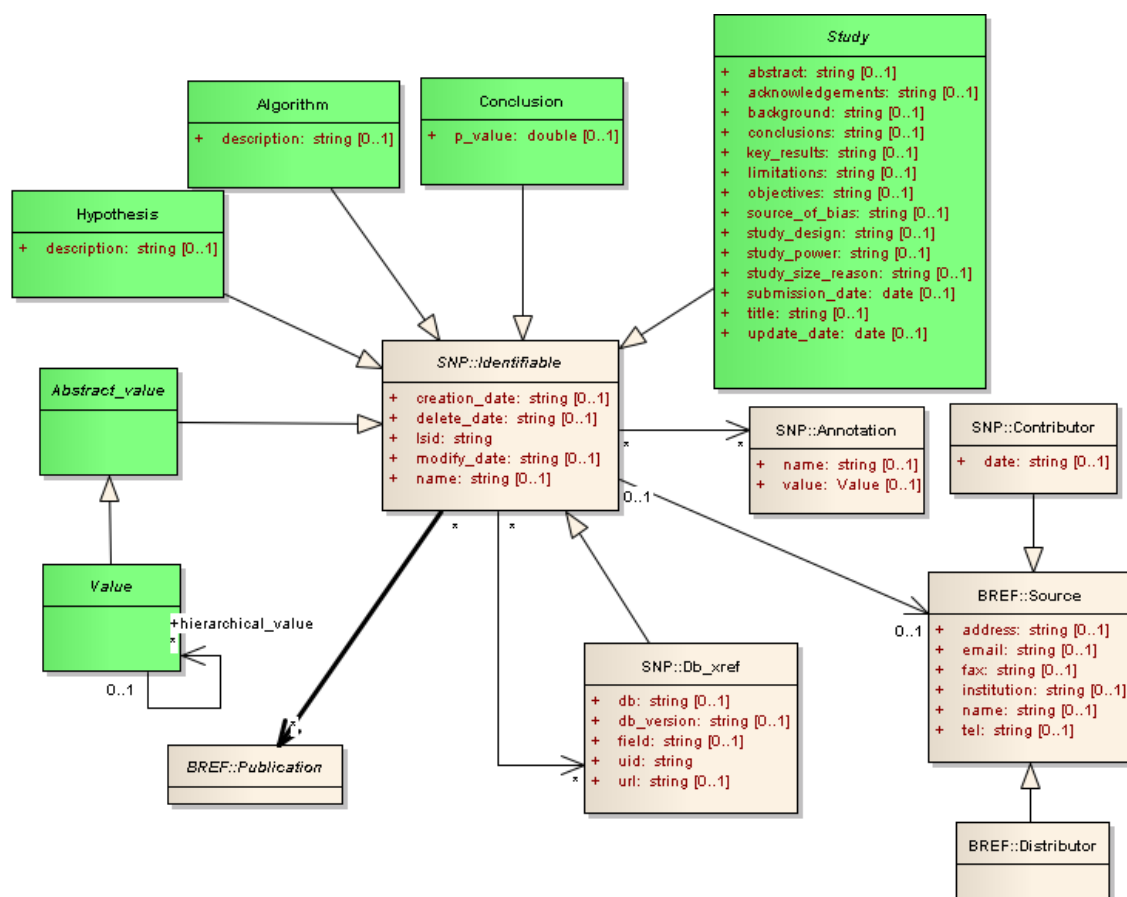


Fig. B.11: The PaGE-OM Identifiable abstract class and associated constructs. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA19.htm).

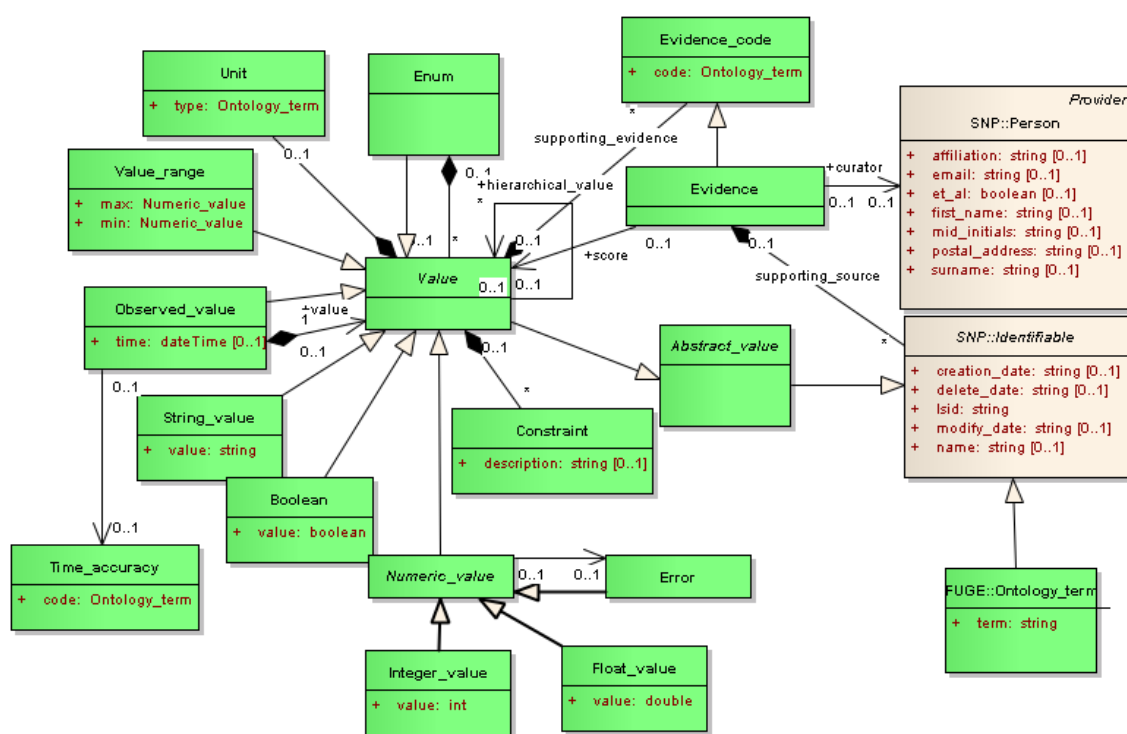


Fig. B.12: The PaGE-OM evidence and value, details. From the PaGE-OM website (http://www.pageom.org/models/omg/v_1.0_b3/EARoot/EA21.htm).

B.2 HGVbaseG2P object class definitions

Table B.1: HGVbaseG2P model class definitions.

Allele	A specific version of a set of different sequence alternatives of a Marker or DNA region resident at one or more locations in a genome
AlleleFrequency	Relative frequency of the given marker allele or allele-combo in a panel of individuals. All frequencies reported for other alleles/allele-combos for the marker in a study (normally adding up to 1) are grouped together into a FrequencyCluster
AnalysisMethod	A protocol for analyzing a set of input data (normally genotypes and data for one or more phenotypic variables) and outputting a set of statistical-test results (normally p-values from tests of association for each marker)
Assay	Describes the parameters for experimentally assaying a biological sample for a specific genomic marker, in order to obtain a genotype signal
Assayedpanel	A set of test subjects that are grouped into a named compilation, and used as the basis for examining and reporting Experiment data. Each Assayed Panel is derived from one or more Sample Panels (by splitting them into subsets and/or merging across Sample Panels) on the basis of some explicit phenotype criterion (such as presence/absence of a Phenotype, or a Phenotype value beyond some inclusion threshold)
Citation	Bibliographic reference
Crossref	Cross-reference to a website or citation, usually used in conjunction with a Hotlink, either directly to a website or a particular web report via a URL template
DataSource	Database or other source of data we have acquired data from directly (i.e. not via submissions)
DiseaseCategory	Broad disease categories for phenotype properties
Experiment	Represents either A) a genotyping experiment with frequency data only, or B) a limited set of statistical analyses which address ONE discrete researcher question, pertaining to at most ONE phenotype and only ONE set of Assayedpanels. B) will often contain results from up to several 'runs' of analysis with different statistical models and/or parameters, all using the same input data

FrequencyCluster	Set of allele and/or genotype and/or haplotype frequencies for a particular marker on a particular Assayedpanel, as reported by a study. Marker can have multiple freq-clusters for a panel, as many studies may have typed the same group of people for that marker
GenotypeFrequency	Relative frequency of the given marker genotype or genotype-combo in a panel of individuals
GenotypedLoci	List of all loci genotyped in a study bundled into a single string. Intended as a way of providing a complete list of markers in a compact way, w/o the actual results
HaplotypeFrequency	Relative frequency of the given haplotype or haplotype-combo in a panel of individuals
Hotlink	A labelled URL web address which can be used either on its own to point to e.g. a project website, or as a template combined with a database identifier to form a web report URL (e.g. dbSNP rs# report)
Marker	A DNA sequence for which identical or highly similar instances exist at one or more locations in a genome. Markers are typically used as the basis for designing an experimental assay for detection of those instances of that sequence
MarkerCoord	Genomic coordinates for a marker in a particular assembly. A marker can be mapped to multiple assemblies (e.g. human reference vs Celera) and can also be mapped to multiple sites within an assembly
MarkerRevision	Tracks changes to the marker information we store as the source data alters; e.g. dbSNP may report a new allele 'A' for a SNP marker that previously as reported as biallelic C/T
Markerset	If markers X,Y,and Z are tested as a set in four different studies, each study will get its separate Usedmarkerset (possibly with an intra-study local name. It may be useful to create a single Markerset to 'tag' those three markers as having been tested together in at least one study
PhenotypeMethod	Describes how a particular Phenotype Property is measured. Can be anything from manually counting or measuring some physical feature to experimental protocols for taking physiological measurements (e.g. protein serum levels)

PhenotypeProperty	The concept of the observable trait/character under study, irrespective of how it is ultimately measured, possibly in different ways by different studies
PhenotypeValue	The outcome or result from measuring/ascertaining a phenotype property by applying a phenotype method. Is usually associated with a Samplepanel as e.g. no. individuals with a certain value (red eyes) for a qualitative trait, or a mean for a quantitative trait
Researcher	A person involved with a study in one way or another, whether as an author or submitter or something else
Resultset	Represents the results from applying a particular analysis method to a discrete set of input data. One such 'run' for a GWA study will normally contain up to several hundred thousand individual test results
Samplepanel	A set of test subjects that are collected together and grouped into a named compilation to address some phenotype of interest. Typically, all the individuals in a Sample Panel are annotated in terms of one or more related Phenotypes, or share some commonality of another key metric (e.g., age, gender, ethnicity). Sample Panels may or may not be equivalent to the eventual groupings that are used as the basis for examining and reporting Experiment data, i.e., the Assayed Panels
SelectionCriteria	Describes the phenotype (or other) criteria used to select individuals from one or more Samplepanels to create an Assayedpanel. Typical use is to select disease-affected individuals to a 'cases' panel and non-affecteds to a 'controls' panel, for a straight-up case-control study design
Significance	Represents the outcome from applying a statistical test of G2P association to a single marker, or a set of markers (for multi-marker tests). Is normally a non-adjusted p-value and additional information such as which allele is the risk allele (for an allelic test) and log-odds ratio
Study	Similar in scope to a journal article, comprising information relevant to a given research question or set of related questions. Data and analysis results from a study are grouped into one or more Experiments
Usedmarkerset	A set of markers tested together for association with a phenotype in a particular study

C. Supplementary materials for identity chapter

C.1 Use cases for an identity-enabled Cafe RouGE system

C.1.1 Background

The Cafe RouGE platform for mutation data exchange

The recently launched Cafe RouGE project, led by Owen Lancaster in our group, aims to facilitate automated dissemination of genetic data generated by diagnostic laboratories which routinely test samples from patients suspected of suffering from inherited disorders. Considered in isolation, a mutation report produced by a diagnostic laboratory is of little relevance but for the person tested. But once the data are de-identified and combined with data from other patients and other laboratories, analysis can yield important insights into the underlying genetic basis of disease.

However, diagnostic laboratories do not usually share their data, not because they are reluctant to do so, but rather for practical reasons. The little data exchange taking place is typically arranged on a one-to-one basis, does not follow standard protocols or use standard data formats, and involves laboratory staff manually submitting data to LSDBs (the primary data consumers). This is inefficient and time-consuming, placing a burden on staff whose remit does not normally include data publication. Furthermore, diagnostic laboratory staff do not receive recognition or reward for this extra effort, giving them little incentive to release their data.

The Cafe RouGE strategy. The Café RouGE project aims to address these data flow problems by i) minimizing the effort required to publish mutation data, and ii) ensuring attribution for data creators working in diagnostic laboratories. Firstly, data publication will be automated by endowing standard analysis tools used by laboratories with a “data submission” function. Submissions will be received by a central depot or “clearinghouse” which will serve as a place where published datasets are advertised, and subsequently discovered and retrieved by third parties.

Secondly, datasets will be unambiguously linked with data submitters’ identities, and systems devised to facilitate citation of published mutation datasets so they can be cited in the literature. Data creators will thus be credited for their contributions. Overall, the project aims to lower the barriers for a willing community to share data, and thereby facilitate the broader exploitation of diagnostic laboratory data.

Cafe RouGE architecture. The system is grounded in the AtomPub protocol already mentioned (see §5.1.5). All interactions between the central Cafe Atom-store, the lightweight admin web application and any third- party applications take place via standard, RESTful AtomPub HTTP-requests. The standard Atom XML syndication format serves as a content-neutral “wrapper” to transmit metadata and data to and from the central depot, including Atom feeds for publishing dataset metadata for consumption and republishing by third parties. A prototype of this system based on the AtomServer off-the-shelf open-source framework¹ has now been constructed.

The GEN2PHEN Knowledge Centre. Some of the Cafe RouGE scenarios described below involve data exchange with the GEN2PHEN Knowledge Centre², a community web portal also being developed by our group. In addition to supporting online community activities, a major aim of this portal is to provide unified, holistic access to G2P data. A

¹<http://atomserver.codehaus.org>

²<http://www.gen2phen.org>

key part of project is to develop tools which facilitate federated searching and browsing across a wide range of G2P data resources on the Internet, including Cafe RouGE.

Technical presentation of Cafe RouGE use cases

The rest of this appendix presents a series of semi-formal use cases, a commonly used technique in software engineering. Use cases are presented as step-by-step textual descriptions, one for each Cafe RouGE scenario which I considered in my analysis. Most of these are accompanied by explanatory figures based on the UML sequence diagram notation³. Actors are represented by rectangles at the top of the diagram. Vertical, broken lines extending down from each actor box overlaid with smaller rectangles, represent the involvement of that actor in the sequence; if an actor partakes in two distinct phases, two smaller rectangles are shown. The sequence of interactions between the various actors proceeds downwards, with each interaction represented by a horizontal arrow. Some sequences are not shown in full, and are instead rendered more compactly as broken lines with arrows on either end or as text boxes.

C.1.2 Use Case #1: data submission from diagnostic laboratory

A diagnostic lab operator (the data owner) finishes a series of genetic tests on a de-identified patient sample and wishes to publish the results. He uses the Gensearch® analysis software⁴ to make a secure connection to the central Cafe RouGE depot and upload a mutation report for publication.

Scenario (Fig. §C.1):

- 1.** The data owner signs in to the Cafe RouGE website and requests submission privileges.
- 2.** The Cafe RouGE administrator assigns the “submitter” role to the ID of the data owner.
- 3.** In the Gensearch® application, the data owner clicks a button to start the authorisation process. This opens up a web browser window pointing to the the Cafe RouGE website.

³<http://www.ibm.com/developerworks/rational/library/3101.html>

⁴<http://www.phenosystems.com>

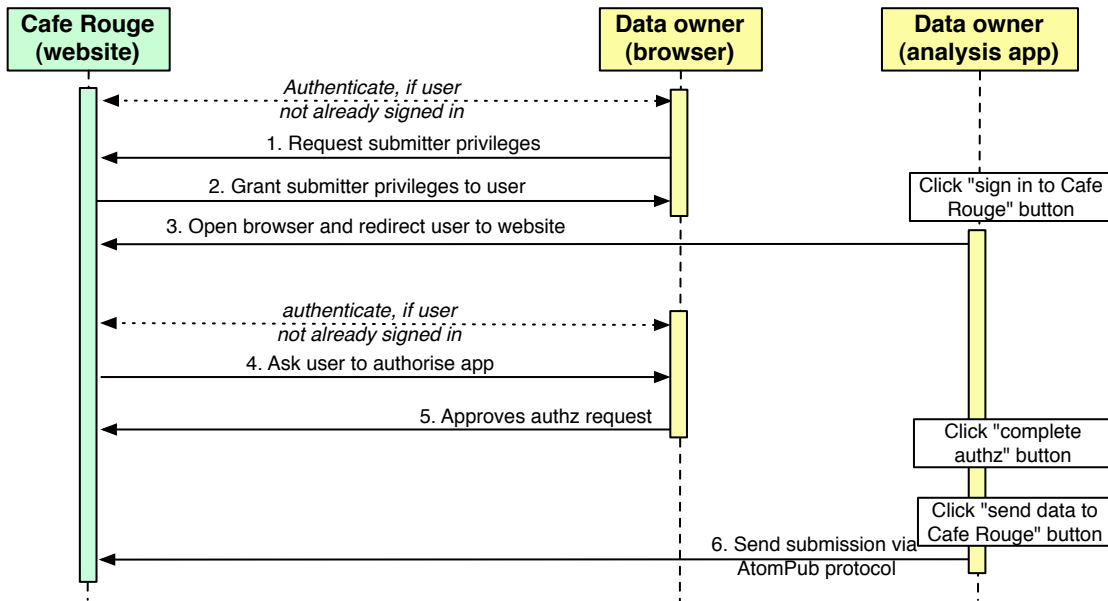


Fig. C.1: Sequence diagram for Use Case #1: data submission from diagnostic lab. The commonly-used 'authz' abbreviation is used for the term 'authorisation'.

4. The Cafe RouGE website displays a message asking the user to confirm that he wishes to link this instance of the Gensearch® application with his account and allow it to upload data.

5. The data owner approves the request, thus authorising the application to perform the prescribed actions on his behalf.

Note: The application is now linked to the data owner's account until he de-authorises it via the Cafe RouGE website, and so steps 1-5 need to be executed only once.

6. Returning to the Gensearch® application, the user now clicks the "Send data to Cafe RouGE" button. The application makes a secure connection to the Cafe RouGE website and uploads the data. The Cafe confirms that the data owner ID has submission privileges and subsequently accepts the upload.

Notes: This proposed mechanism for uploading submissions to the Cafe RouGE depot is modelled on the Flickr Uploadr photo sharing application⁵. A key advantage of this is that many Internet users are already familiar with the online photo sharing, and thus are more likely to be receptive to a data publication workflow based on the same paradigm. Another advantage to this approach - i.e. web-based authentication and delegated authorisation - is that it is relatively simple to implement in the Gensearch® software (and other applications), compared to implementing authentication in the standalone application itself.

C.1.3 Use Case #2: data consumer pre-authorised to access Cafe data

A data consumer wishes to retrieve a mutation report from the Cafe RouGE depot. The diagnostic lab operator (the data owner) has elected to only share his data with specific persons, and has already added the ID of the data consumer to the list of IDs for users who are permitted to access all data generated by his lab.

Scenario (Fig. §C.2):

- 1.** The data consumer attempts to retrieve a protected mutation report by following the URL which identifies it in the Cafe RouGE depot.
- 2a.** The Cafe checks that the ID of the signed-in data consumer is on the list of users approved for data access by the data owner. If the data consumer is authorised for data access, the Cafe returns the mutation report.
- 2b.** If data consumer ID is not authorised, the Cafe returns a standard ‘access denied’ error code and an informative message as guidance on how to request permission.

Notes: This scenario is an example of basic access control list (ACL) based access management which typically involves user IDs being assigned one or more roles (role-based access control, as previously mentioned). In this case, the data owner gives the data consumer “blanket” approval to access all his data, but the same approach can be used to control access on an individual dataset basis if required.

⁵<http://www.flickr.com/tools/>

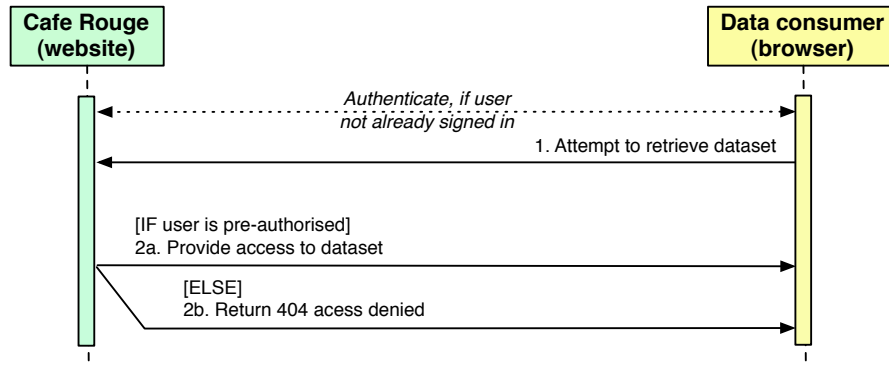


Fig. C.2: Sequence diagram for Use Case #2: data consumer pre-authorized to access Cafe data.

The same general concept can be extended and made more flexible (and easier to manage) by introducing groups which can be assigned roles. The overall sequence as described above would remain the same; the Cafe RouGE system would check if the user ID is in the appropriate group before providing access to the protected data. The data owner could thus define his own “buddy lists” comprising IDs for authorised data consumers who frequently need to access his data.

C.1.4 Use Case #3: data consumer requesting access to Cafe data

A data consumer wishes to retrieve a mutation report from the Cafe RouGE depot, after discovering the dataset via a feed syndication service on the Knowledge Centre. He does not yet have permission to access the protected data, so he first needs to request access from the diagnostic lab operator who generated the data (the data owner).

Scenario (Fig. §C.3):

1. The data consumer attempts to retrieve a protected mutation report by following the URL which identifies it in the Cafe RouGE depot.
2. The Cafe finds that the ID of the data consumer is not authorised for access and returns a standard ‘access denied’ error code, along with a message instructing the data consumer click a button to request access.

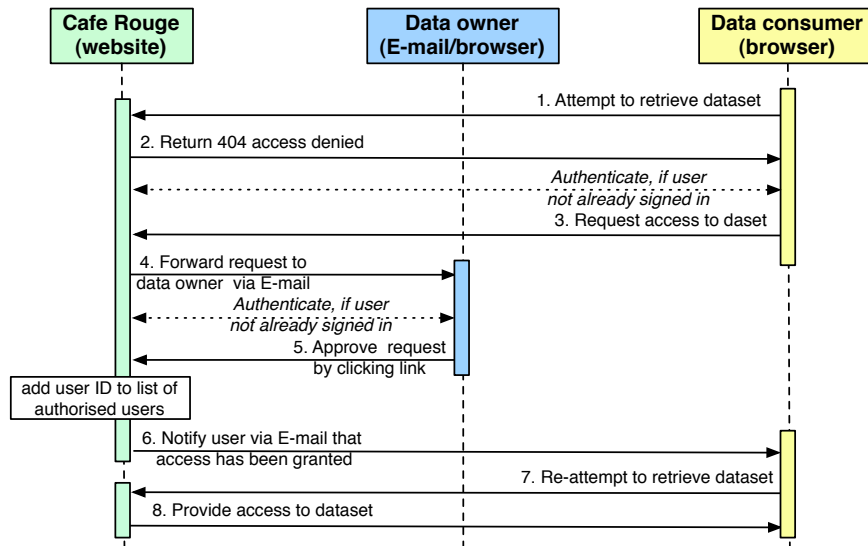


Fig. C.3: Sequence diagram for Use Case #3: consumer requesting access to Cafe data.

3. The data consumer clicks button to send access request.
4. The Cafe forwards the request to the data owner by sending a notification E-mail.
5. The data owner approves the access request by clicking on a hyperlink in the notification E-mail. The link takes the data owner to the Cafe website, where he confirms the authorisation and grants the data consumer access to all his data.
6. The Cafe adds the data consumer's local ID to the list of authorised users and sends a E-mail to the data consumer notifying him that the access request has been approved.
7. The data consumer returns to the Cafe website by following the link from the notification E-mail and re-attempts to retrieve the mutation report.
8. The Cafe confirms that the data consumer ID is on the list of users authorised for data access and returns the requested mutation report.

Notes: Like the photo-sharing paradigm in Use Case #1, this scenario is modelled on well-established usage patterns employed for mailing list subscriptions and similar scenarios. The simple system of E-mail notifications and a hyperlink-based workflow to approve or deny access is designed to mimic processes which are familiar to many Internet

users from online retailers and social networking websites.

The main distinction between this scenario and the previous one is that here the data consumer has not had prior dealings with the data owner. Therefore, the data owner needs to judge whether or not he trusts the data consumer, based on the personal's profile (e.g. E-mail address, name, affiliation). The simple approval process illustrated here assumes that the data owner will take whichever steps he feels are necessary to confirm the identity and credentials of the data consumer. If necessary, the process could be augmented with any number of formal confirmation and certification procedures. Conversely, the data owner may prefer a much more open data release policy, whereby any access request with a valid E-mail address is approved automatically.

C.1.5 Use Case #4: authenticating via OpenID

A user wishes to register on the Cafe Rouge website. Instead of a local user account and authentication on the Cafe system, the user opts to identify himself with his OpenID credentials. The Cafe website therefore redirects him to his OpenID provider where he authenticates, and his OpenID is subsequently linked to his local ID in the Cafe system.

Scenario (Fig. §C.4):

- 1.** On the Cafe website, the user opts to sign in with his OpenID and passes information on his identity to the Cafe. Depending on the OpenID authentication mode, this information comprises either the OpenID URL identifier itself or (in the case of so-called “directed” identity) the URL for the OpenID provider.
- 2.** The Cafe website either redirects the user's browser or displays a popup window to the OpenID provider.
- 3.** The OpenID provider displays a page asking the user to confirm that he wants to sign into the 3rd party Cafe website, and which personal profile information (if any) he wishes to share with the Cafe.
- 4.** The user approves the authentication request and release of his E-mail address, his full name and affiliation.

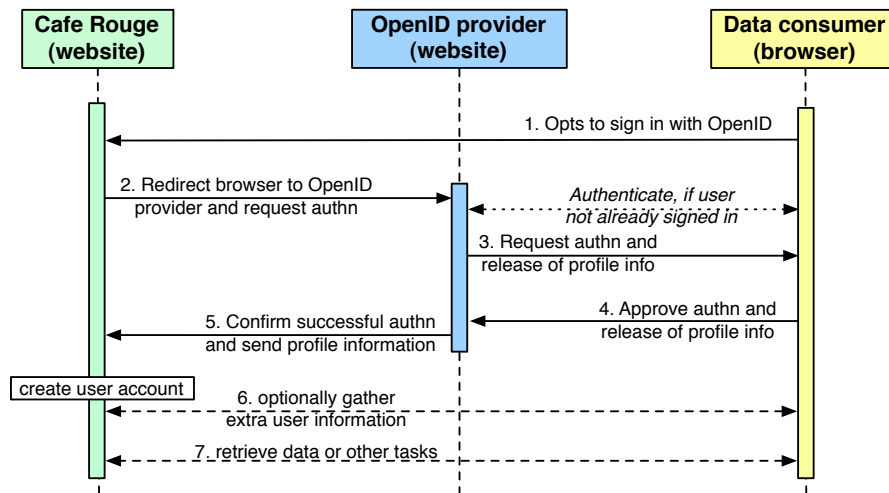


Fig. C.4: Sequence diagram for Use Case #4: authenticating via OpenID. The commonly-used abbreviation 'authn' is used for the term 'authentication'.

5. The OpenID provider redirects the user's browser back to the Cafe website, confirming that authentication was successful (i.e. the user has proved that he is in control of the ID that he presented) and also passing along the profile information the user chose to share with the Cafe.

6. The Cafe system transparently creates a local user account linked to the user's OpenID identifier and populates with the profile information provided. Optionally, the Cafe website may prompt the user for additional profile information if required.

7. The user is now signed in and can proceed to use the Cafe website to retrieve data, request submitter privileges or for other tasks (as per previous use cases).

Notes: The OpenID authentication process presented here is simplified and serves only to provide a general illustration of federated authentication in the Cafe Rouge context. For example, in reality the OpenID provider discovery process involves several exchanges between the Cafe relying party and the OpenID provider (see Recordon and Reed (2006) for details). But these exchanges do not affect the sequence of interactions from the perspective of the user, which is the focus here. Furthermore, although the OpenID protocol is used

here for authentication and profile data exchange, other federated identity protocols (e.g. Shibboleth for authentication, SAML for ID attribute exchange) could be substituted, with little or no change in the overall flow of interactions from the perspective of the user.

C.1.6 Use Case #5: browsing and accessing Cafe data from the Knowledge Centre portal

A data consumer is interested in reported variants in and around the BRCA2 gene. He uses the Knowledge Centre mutation feed browser to retrieve a list of mutation reports tagged with the BRCA2 gene symbol, and now wishes to access the full report details.

Scenario (Fig. §C.5):

1. On the Knowledge Centre website, the data consumer requests a listing of all available mutation reports for the BRCA2 gene.
2. The Knowledge Centre website requests an Atom category feed for the BRCA2 tag.
3. The Cafe website returns a feed containing a list of Atom entries annotated with metadata, one entry per mutation report.
4. The Knowledge Centre website displays the Atom entries for the data consumer to peruse.

Note: interactions between the Knowledge Centre and the Cafe in steps 2 and 3 do not require a secure connection, because only publicly-available mutation report metadata are being exchanged.

5. The data consumer wishes to access full details for one of the reports and follows the URL to the Cafe Rouge website to retrieve the protected data.
6. The Cafe website - aware of the identity of the user because he is signed into both sites via cross-domain SSO - can now immediately proceed to check if the data consumer ID is authorised for data access, as per Use Case #3.

Notes: This scenario assumes that the Knowledge Centre serves simply as a user-friendly front end or “showroom” for Cafe Rouge contents, and as such does not have a role beyond

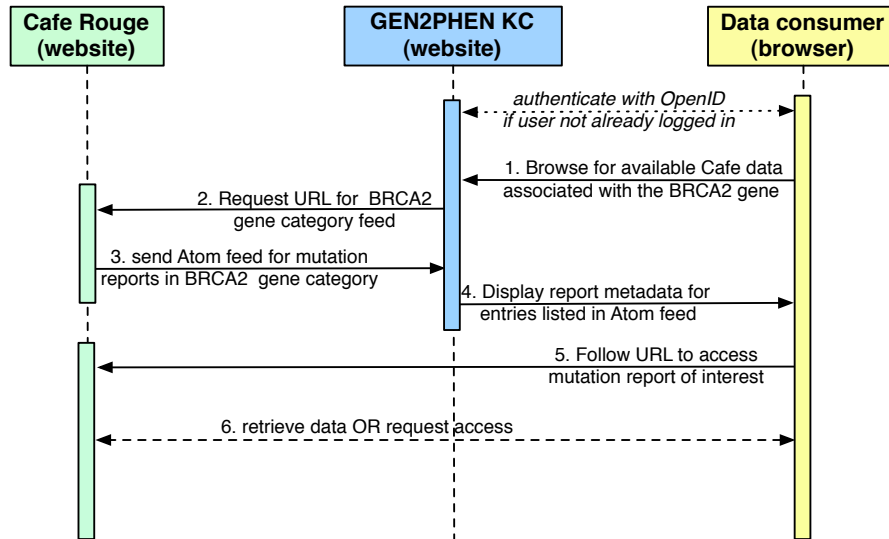


Fig. C.5: Sequence diagram for Use Case #5: browsing and accessing Cafe data from the GEN2PHEN Knowledge Centre.

retrieving and displaying public mutation report metadata. In this setting, the main benefit of cross-domain SSO - facilitated by OpenID authentication - is increased convenience for the user who is able to navigate freely between the websites without having to authenticate repeatedly. Similar results could be obtained, albeit less conveniently, if the user has already signed in to the two websites in the same browser session.

However, the full power of SSO becomes apparent if one considers richer functionality that the Knowledge Centre could provide. For example, instead of simply redirecting the data consumer to the Cafe Rouge website for a direct data retrieval, the Knowledge Centre could securely retrieve mutation reports and present them to the user in a custom viewer. This would require the same kind of delegated authorisation as illustrated in Use Case #1, whereby the data consumer would permit the Knowledge Centre website to connect on his behalf to the Cafe and retrieve data. As another example, the Knowledge Centre could connect to the Cafe and retrieve the user's authorisation level for each feed displayed, so the user could see at a glance whether or not he needs to request access from the data owner. More generally, using users' digital IDs as the "glue" to tie the web applications together, the

Knowledge Centre could provide a rich, fully-featured front end for the base Cafe depot, and indeed other applications could do the same and create specialised views over the data.

C.1.7 Use Case #6: manage access based on ID provider whitelisting

A diagnostic lab operator (the data owner) decides to adopt a liberal data release policy: automatically approve all data access requests from data consumers with an identity from a certain subset or “whitelist” of OpenID providers.

Scenario (no figure):

1. The data consumer identifies a mutation report of interest and attempts to retrieve the detailed report, as per Use Case #5.
2. As per Use Case #3, the Cafe website finds that the data consumer ID is not yet authorised and provides the option to request access.
3. The data consumer requests access to the dataset.
4. The Cafe website cross-checks the data consumer’s OpenID provider against the data owner’s whitelist. If the provider is on the whitelist, the data consumer’s ID is automatically authorised for access to all data generated by the data owner.
5. The Cafe website sends the data consumer a notification E-mail, and the sequence resumes from step 6 in Use Case #3.

Notes: This sequence represents a more clear-cut demonstration of the benefits of federated identity. Whitelisting, or its reverse, blacklisting - that is, accept IDs from all but a defined subset of providers - is a relatively simple means of leveraging federated identity to streamline data access, whilst still retaining a measure of control over the process. Blacklisting is deemed a minimal security strategy on social networking websites, as a means to avoid fraudulent user registrations from spammer-operated OpenID providers. For more security-focused scenarios, more stringent criteria involving an explicit ID provider whitelist known to be reliable and secure may be a useful strategy. For example,

customers of Microsoft's HealthVault service⁶ can sign in via OpenID, but only if their OpenID provider is on the short list of providers Microsoft has judged as reliable.

C.1.8 Cafe RouGE and AGAST

To set the stage for the next three use cases, Fig. §C.6 illustrates how the Cafe RouGE system could interact with the AGAST framework in authorisation decision-making involving a data consumer. In this setting, the Cafe RouGE system would be extended with a so-called policy enforcement point (PEP). The PEP is the software component responsible for gathering, and converting into RDF if necessary, ID attributes describing the data consumer. The Cafe PEP subsequently presents this information, along with the access policy expressed as an ontology, to the remote AGAST policy decision point (PDP) web service, known as Quadi. To obtain a yes/no authorisation decision, the Cafe PEP sends a SPARQL query to interrogate the Quadi decider. The overall access policy, then, is expressed as the combination of the SPARQL query and the ontology. As complexity of instance data and ontologies increases, and as this information is retrieved from an increasing number of sources, the overall architecture of the system and the authorisation process remains unchanged. The system thus scales to handle very complex, highly distributed scenarios.

C.1.9 Use Case #7: access based on inferred virtual organisation membership

A data consumer wishes to access a dataset published in Cafe RouGE. The data consumer and data owner are affiliated with different research institutions, both of which are part of the same virtual organisation (VO). An intra-organisational data exchange agreement is in place, such that any member of a partner institution can access mutation data generated by

⁶<http://www.healthvault.com>

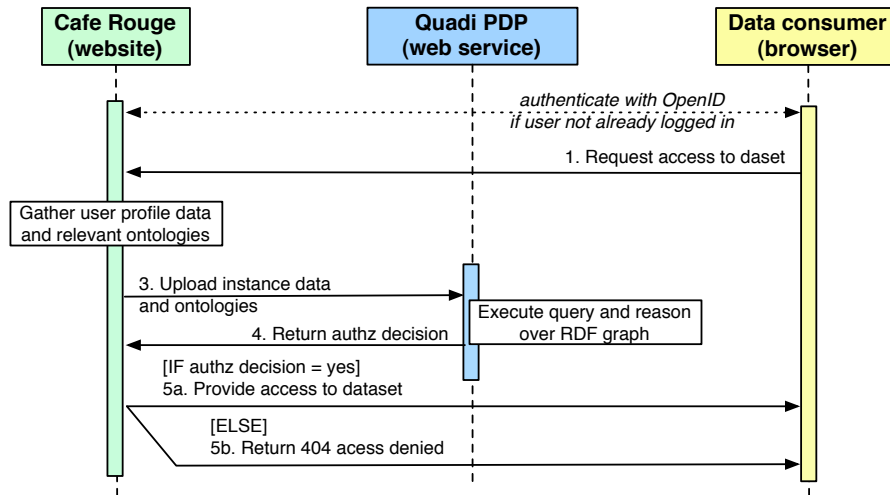


Fig. C.6: Sequence diagram for interaction between Cafe RouGE, the Quadi semantic authorisation service and the data consumer.

the other partners. This access policy is expressed as an ontology which describes access permissions and organisational structure of each partner in the VO.

Scenario (Fig. §C.6):

1. The data consumer has signed into the Cafe website with his institutional ID and attempts to retrieve a dataset of interest.
2. The Cafe PEP gathers the necessary profile information describing the data consumer. This information, available via the institutional ID he used to sign in to the Cafe, includes an assertion stating that he is a member of the Genetics department of his institution. The PEP then uploads the instance data, the access policy ontology and the SPARQL query to the Quadi service.
3. The Quadi decider executes the query, determines that the data consumer is authorised to access the data, and returns the result.
4. The Cafe website acts on the result of the authorisation decision and provides access to the data.

Notes: It is vital to stress the importance of what happens in step 3 above. Based on the provided information - the data consumer is a member of the Genetics department, the department is a part of the institution, and the institution is a part of the VO - the semantic reasoner infers that the data consumer is a member of the VO, even though this is not explicitly stated anywhere. The general principle would apply to larger, more complex VOs with thousands of members with various roles, all of which may change over time as individuals and partners institutions leave and join the VO. Furthermore, the information describing the organisational structure may be expressed in multiple ontologies retrieved from each VO partner website, rather than a single master ontology located at the Cafe.

C.1.10 Use Case #8: access based on status as *bona fide* researcher

The data consumer wishes to access a dataset published in Cafe Rouge. The data owner has chosen a liberal data release policy stating that anyone who has an active ORCID contributor profile can access the data. But since the Cafe does not have this profile information in hand, the data consumer is prompted for additional ID information proving that he is in the ORCID system.

Scenario:

1. The data consumer signs in to the Cafe website with his Google OpenID and attempts to retrieve a mutation report.
2. As per the previous use case, the Cafe PEP gathers the necessary information, the SPARQL query which expresses the policy, and sends to the Quadi service. Unlike the previous use case, no ontology is involved this time.
3. The Quadi service returns a negative result from the semantic query, along with further information indicating why the data consumer is not authorised.
4. The Cafe website informs the data consumer about the outcome and asks for his permission to connect to the ORCID system in order to retrieve his profile information.
5. The data consumer agrees to let the Cafe to connect to the ORCID system.
6. The Cafe redirects the data consumer's browser to the ORCID website.

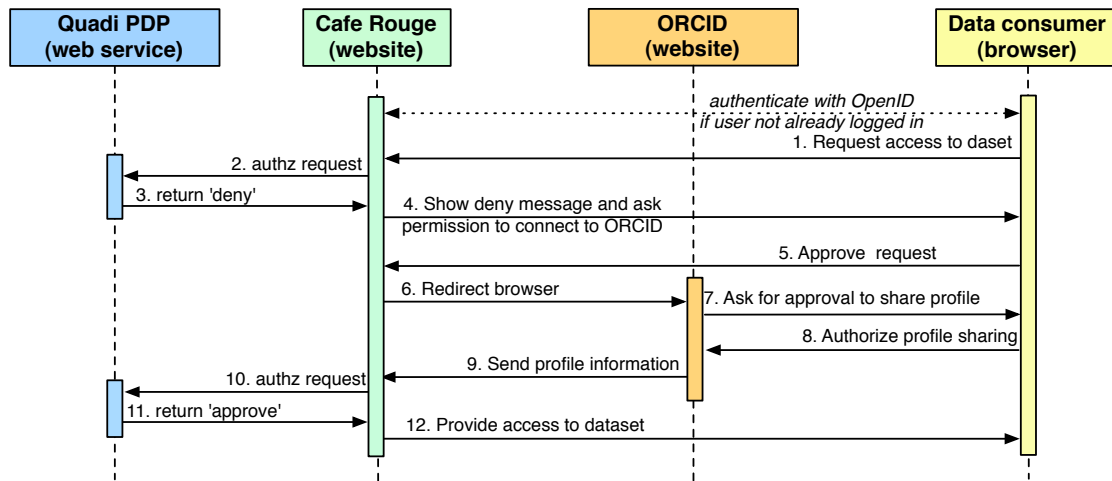


Fig. C.7: Sequence diagram for Use Case #8: access based on verified ORCID profile.

7. Having already linked his ORCID ID with his Google OpenID, the data consumer is already authenticated on the ORCID website, and is thus taken straight to a page where he is asked if he wants to share the required attributes from this profile with the Cafe website.

8. The data consumer authorises the profile sharing request.

9. Having obtained authorisation from the user, the ORCID website sends profile information to the Cafe.

10. The Cafe PEP re-sends ID profile information and SPARQL query to the Quadi services, but this time including an attribute assertion stating that the data consumer has a valid ORCID profile.

11. The Quadi decider now concludes, on the basis of updated ID attributions including profile information retrieved from ORCID, that the data consumer is authorised for data access and therefore returns a positive result from the semantic query.

12. The Cafe website acts on the result of the authorisation decision and provides access to the data.

Notes: The assertion that the data consumer controls an ORCID contributor profile is used here as a convenient proxy for *bona fide* researcher status. This is based on the

assumption that an author who claims an ORCID profile will be required to offer some tangible proof that the disambiguated author identifier in question actually refers to him. Therefore, an ORCID ID would provide a reliable link to a real person which would be difficult to forge. It should be noted, however, that various issues relating to authentication, certification of an author's affiliation and authorship and related matters are currently being worked out by ORCID partners. Those issues aside, the eventual ORCID implementation can be reasonably assumed to support a scenario very similar to the above.

From a privacy standpoint, this use case demonstrates how a user can grant fine-grained access to his personal identity federated across multiple identifiers - in this case his Google OpenID identifier which is securely linked to his ORCID identifier - in order to gain access to a protected resource. Importantly, the user shared only sufficient personal information with the Cafe to fulfill the access policy requirements, nothing more. More advanced variants of this scenario might involve ID linking services (see §6.2).

C.1.11 Use Case #9: access based on permission from an external registry

A data consumer wishes to access a dataset published via Cafe Rouge. The data include identifiable information on patients, and access is therefore regulated on a case-by-case basis by a Data Access Committee (DAC). The data consumer has already obtained permission to access the dataset by sending a request to the DAC (see e.g. §2.2.7). Information on his authorisation level for this and other related datasets is held in a central registry of data access privileges, in which the data consumer's ID has been linked to his OpenID.

Scenario:

1. The data consumer signs in to the Cafe website with his OpenID. Since he is already aware of the authorisation requirements for this dataset, he navigates to a page where additional ID attributes can be supplied.
2. The Cafe website presents a page asking the user to present the ID he wishes to proffer

additional information for.

3. The data consumer indicates that he wishes to proffer information from his profile with the central access registry.
4. The Cafe website redirects the data consumer's browser to the central registry website.
5. The data consumer authenticates at the registry website (if not already signed in via SSO) and authorises the registry website to share his access privilege information with the Cafe.
5. As in the previous use case, the Cafe now has enough profile information in hand to proceed with the semantic authorisation process and grant (or deny) access to the data.

Notes: The salient feature of the above is that the system where the dataset is archived is decoupled from the system holding the privileges. This is important for the following reasons. First, the same dataset could be disseminated from multiple data providers, each of which would implement certain core authentication and authorisation facilities. This would enable, for example, a pre-approved user to download raw GWAS data from the EGA, or alternatively view high-level integrated views of the data via a specialised secondary provider. Moreover, a secondary data provider could disseminate sensitive data originating from multiple primary archives, retrieving access permissions remotely on a dataset-by-dataset basis⁷. For example, dbGaP could conceivably serve primary GWAS data originating from the EGA, and vice versa. However, it should be clear from previous chapters that the secondary data provider of interest to this discussion is, of course, HGVbaseG2P.

Another useful feature of this scheme is that it would support the tracking of IDs for persons found guilty of inappropriate use of sensitive data, and the sharing such “blacklists” amongst data providers. A prominent example of this is the first (and, at the time of writing, the only) breach of a publication embargo agreement, whereby authors who accessed GWAS data from dbGaP submitted a journal manuscript before the embargo date (see Guttmacher *et al.* (2009) and PNAS editorial by Schekman (2009)).

⁷In this case, the primary data provider and access registry might be one and the same.

C.1.12 Use Case #10: attributing data publication to submitter

A diagnostic lab operator (the data owner) publishes a dataset via Cafe RouGE. He wishes to associate this dataset with his ORCID profile in order for this digital contribution to be attributed to him. He has previously authenticated with the Cafe using his OpenID, acquired submitter privileges and associated the Gensearch® application with his Cafe ID as per previous use cases.

Scenario:

1. The data owner uploads a mutation report from the Gensearch® application to the Cafe website, as per step 6 in use case #1.
2. The Cafe website, in its role as a DOI publication agent, registers a DOI name for the mutation report.
3. The Cafe sends an upload confirmation E-mail to the data owner, also inviting him to follow a one-time hyperlink to link his data publication with his ORCID identifier.
4. The data owner follows the link to the Cafe website.
5. The Cafe website page asks for the data owner's approval to retrieve his ORCID profile information, and to associate this and future data publications with his ORCID identifier.
6. The data owner approves the request.
7. The data owner's browser is redirected to the ORCID website, where he is already authenticated via SSO.
8. The ORCID website asks the data owner to confirm that he wishes to share profile data with the Cafe website, and that the Cafe website should be allowed to associate contributions with his ORCID identifier.
9. The data owner confirms that he wants the submission to be attributed to his ORCID ID and approves sharing of his profile information with the Cafe website.
10. The ORCID website redirects the user's browser back to the Cafe website and passes along his profile information, including the ORCID contributor identifier.

Note: the link with the ORCID site would only need to be configured once. In future submissions, steps 5 through 10 would thus not have to be repeated.

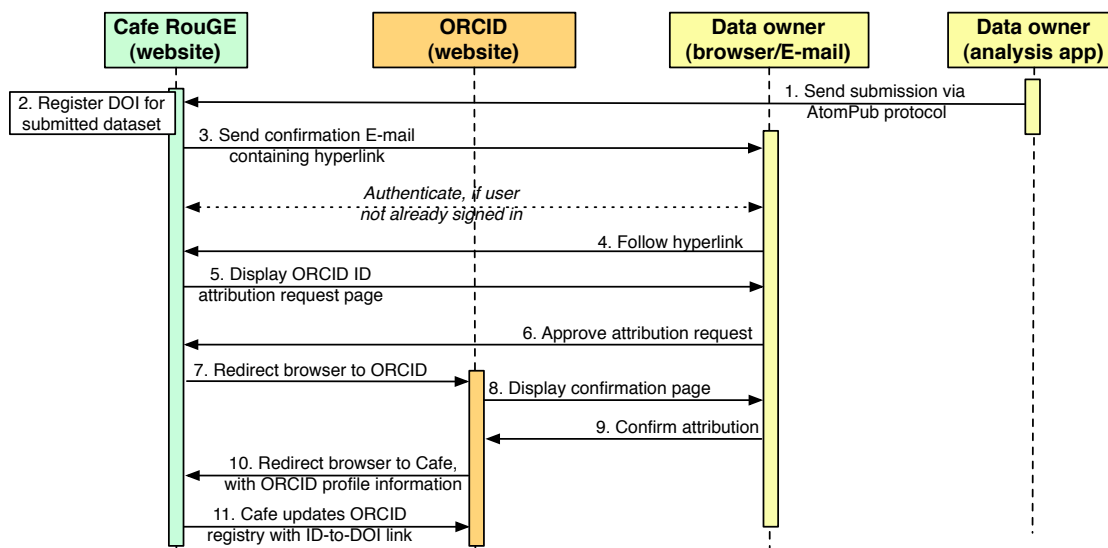


Fig. C.8: Sequence diagram for Use Case #10: attributing a Cafe RouGE data publication to a submitter's ORCID ID. The user does not interact with the DOI system directly, so to simplify the diagram the DOI system is not shown as a separate actor in the sequence.

11. The Cafe website updates the ORCID registry to include a link from the data owner's ORCID identifier to the dataset DOI, thereby making the data publication a part of his public ORCID record.

Notes: This use case has two key outcomes. First, the mutation dataset can now be accessed and cited via its persistent, unique DOI name, as described in §6.3.1. Citations, online access statistics and other information relating to this data publication can therefore be tracked from now on, in much the same way as is done for journal articles and using the same infrastructure, and potentially be used to assess impact of the dataset over time. Second, the data owner has been unambiguously identified, and through his ID the data publication has been attributed to his identity and is, consequently, part of his publication record. A variant of this sequence could include a step for associating additional persons with a data submission, whereby the submitter would specify up to several ORCID IDs for laboratory technicians, analysts or other laboratory personnel who should be credited for their contributions to the work.

D. DVD-ROM contents

The printed version of this thesis is accompanied by a DVD-ROM disc inside the back cover. The DVD contains complete program sourcecode for HGVbaseG2P, as well as various other supplementary materials. Details on the organisation of this content and brief descriptions of key software tools can found in the file `00README.txt` on the disc.

References

- Abbott, A. (2010). Mouse project to find each gene's role. *Nature*, 465(7297), 410. doi:10.1038/465410a 250
- Abel, F., Coi, J. L. D., Henze, N. *et al.* (2009). A User Interface to Define and Adjust Policies for Dynamic User Models. In *Fifth International Conference on Web Information Systems and Technologies (WEBIST)*, pages 184–191. INSTICC Press. Available online at <http://www.l3s.de/~olmedilla/pub/2009/2009-WEBIST-interfaces.pdf>. 242
- Adler, B. T. and de Alfaro, L. (2007). A Content-Driven Reputation System for the Wikipedia. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 261–270. doi:10.1145/1242572.1242608 203
- Allen, N. C., Bagade, S., McQueen, M. B. *et al.* (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics*, 40(7), 827–834. doi:10.1038/ng.171 47
- Amberger, J., Bocchini, C. A., Scott, A. F. *et al.* (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Research*, 37(Database issue), D793–D796. doi:10.1093/nar/gkn665 4
- Andrade, M. A. and Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Current Opinion in Biotechnology*, 8(6), 675–683. doi:10.1016/S0958-1669(97)80118-8 21
- Arnaiz, O., Cain, S., Cohen, J. *et al.* (2007). ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Research*, 35(Database issue), D439–D444. doi:10.1093/nar/gkl777 35
- Aulchenko, Y. S., Ripatti, S., Lindqvist, I. *et al.* (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genetics*, 41(1), 47–55. doi:10.1038/ng.269 50
- Aurrecoechea, C., Brestelli, J., Brunk, B. *et al.* (2009). PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research*, 37(Database issue), D539–D543. doi:10.1093/nar/gkn814 31
- Aurrecoechea, C., Heiges, M., Wang, H. *et al.* (2007). ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Research*, 35(Database issue), D427–D430. doi:10.1093/nar/gkl880 31

- Bard, J. (2003). Ontologies: Formalising biological knowledge for bioinformatics. *BioEssays*, 25(5), 501–6. doi:10.1002/bies.10260 36
- Barrell, D., Dimmer, E., Huntley, R. P. *et al.* (2009). The GOA database in 2009 — an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(Database issue), D396–D403. doi:10.1093/nar/gkn803 36
- Bateman, A. and Wood, M. (2009). Cloud computing. *Bioinformatics*, 25(12), 1475. doi:10.1093/bioinformatics/btp274 53
- Beck, T., Morgan, H., Blake, A. *et al.* (2009). Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. *BMC Bioinformatics*, 10(Suppl 5), S2. doi:10.1186/1471-2105-10-S5-S2 36
- Becker, K. G., Barnes, K. C., Bright, T. J. *et al.* (2004). The Genetic Association Database. *Nature Genetics*, 36(5), 431–432. doi:10.1038/ng0504-431 39
- Belleau, F., Nolin, M., Tourigny, N. *et al.* (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5), 706–716. doi:10.1016/j.jbi.2008.03.004 264
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J. *et al.* (2008). GenBank. *Nucleic Acids Research*, 36(Database issue), D25–D30. doi:10.1093/nar/gkm929 20
- Berners-Lee, T. and Hendler, J. (2001). Publishing on the semantic web. *Nature*, 410(6832), 1023–1024. doi:10.1038/35074206 78
- Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284, 34–43. doi:10.1038/scientificamerican0501-34 78
- Beroud, C., Hamroun, D., Collod-Beroud, G. *et al.* (2005). UMD (Universal Mutation Database): 2005 update. *Human Mutation*, 26(3), 184–191. doi:10.1002/humu.20210 46
- Bertram, L., McQueen, M. B., Mullin, K. *et al.* (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1), 17–23. doi:10.1038/ng1934 47
- Bhagat, J., Tanoh, F., Nzuobontane, E. *et al.* (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research*, 38(suppl 2), W689–W694. doi:10.1093/nar/gkq394 85
- Biesecker, L. G., Mullikin, J. C., Facio, F. M. *et al.* (2009). The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Research*, 19(9), 1665–1674. doi:10.1101/gr.092841.109 13

- Bilge, L., Strufe, T., Balzarotti, D. *et al.* (2009). All your contacts are belong to us: automated identity theft attacks on social networks. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 551–560. ACM, New York, USA. doi:10.1145/1526709.1526784 204
- Blake, J. A., Bult, C. J., Eppig, J. T. *et al.* (2009). The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Research*, 37(Database issue), D712–D719. doi:10.1093/nar/gkn886 32
- Blomqvist, M. E.-L., Reynolds, C., Katzov, H. *et al.* (2006). Towards compendia of negative genetic association studies: an example for Alzheimer disease. *Human Genetics*, 119(1-2), 29–37. doi:10.1007/s00439-005-0078-9 40
- Bollen, J., de Sompel, H. V., Hagberg, A. *et al.* (2009a). Clickstream data yields high-resolution maps of science. *PLoS ONE*, 4(3), e4803. doi:10.1371/journal.pone.0004803 219
- Bollen, J., de Sompel, H. V., Hagberg, A. *et al.* (2009b). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4(6), e6022. doi:10.1371/journal.pone.0006022 219
- Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33 Suppl, 228–237. doi:10.1038/ng1090 6
- Botstein, D., White, R. L., Skolnick, M. *et al.* (1980). Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *American Journal of Human Genetics*, 32(3), 314–331. 5
- Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology*, 1(3), 179–181. doi:10.1371/journal.pcbi.0010034 64
- Bourne, P. E. and Fink, J. L. (2008). I am not a scientist, I am a number. *PLoS Computational Biology*, 4(12), e1000247. doi:10.1371/journal.pcbi.1000247 259
- Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.
- Brase, J. (2004). Using Digital Library Techniques – Registration of Scientific Primary Data. *Research and Advanced Technology for Digital Libraries*, pages 488–494. doi:10.1007/b100389 222
- Brase, J. and Schindler, U. (2006). The publication of scientific data by World Data Centers and the National Library of Science and Technology in Germany. *Data Science Journal*, 5, 205–208. doi:10.2481/dsj.5.205 222

- Braun, R., Rowe, W., Schaefer, C. *et al.* (2009). Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genetics*, 5(10), e1000668. doi:10.1371/journal.pgen.1000668 60
- Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4), 365–371. doi:10.1038/ng1201-365 74
- Brazma, A., Krestyaninova, M. and Sarkans, U. (2006). Standards for systems biology. *Nature Reviews Genetics*, 7(8), 593–605. doi:10.1038/nrg1922 74, 76
- Brookes, A., Lehvaslaiho, H., Muilu, J. *et al.* (2009). The phenotype and genotype experiment object model (PaGE-OM): a robust data structure for information related to DNA variation. *Human Mutation*, 30(6), 968–977. doi:10.1002/humu.20973 71, 90, 91, 92, 94, 287, 289, 293, 295
- Brookes, A. J. and Prince, J. A. (2005). Genetic association analysis: lessons from the study of Alzheimers disease. *Mutation Research*, 573(1-2), 152–159. doi:10.1016/j.mrfmmm.2004.08.017 40
- Buetow, K. H. (2005). Cyberinfrastructure: empowering a "third way" in biomedical research. *Science*, 308(5723), 821–824. doi:10.1126/science.1112120 68
- Buetow, K. H. (2009). An infrastructure for interconnecting research institutions. *Drug Discovery Today*, 14(11-12), 605–610. doi:10.1016/j.drudis.2009.03.011 85, 230, 262
- Burton, P. R., Tobin, M. D. and Hopper, J. L. (2005). Key concepts in genetic epidemiology. *Lancet*, 366(9489), 941–951. doi:10.1016/S0140-6736(05)67322-9 4
- Campbell, P. (2008). Escape from the impact factor. *Ethics in Science and Environmental Politics*, 8(1), 5–7. doi:10.3354/esep00078 258
- Cannata, N., Merelli, E. and Altman, R. B. (2005). Time to organize the bioinformatics resourceome. *PLoS Computational Biology*, 1(7), e76. doi:10.1371/journal.pcbi.0010076 58
- Carlson, C. S., Eberle, M. A., Kruglyak, L. *et al.* (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990), 446–452. doi:10.1038/nature02623 8
- Cassman, M. (2005). Barriers to progress in systems biology. *Nature*, 438(7071), 1079. doi:10.1038/4381079a 76
- Chadwick, D. and Inman, G. (2009). Attribute Aggregation in Federated Identity Management. *Computer*, 42(5), 33–40. doi:10.1109/MC.2009.143 214

- Chadwick, D. W., Zhao, G., Otenko, S. *et al.* (2008). PERMIS: a modular authorization infrastructure. *Concurrency and Computation: Practice and Experience*, 20(11), 1341–1357. Online ISSN: 1532-0634. doi:10.1002/cpe.1313 214
- Chain, P. S. G., Grafham, D. V., Fulton, R. S. *et al.* (2009). Genome project standards in a new era of sequencing. *Science*, 326(5950), 236–237. doi:10.1126/science.1180614 55
- Chakravarti, A. (1999). Population genetics—making sense out of sequence. *Nature Genetics*, 21(1 Suppl), 56–60. doi:10.1038/4482 7
- Chandras, C., Weaver, T., Zouberakis, M. *et al.* (2009). Models for financial sustainability of biological databases and resources. *Database*, 2009. Published online October 23. doi:10.1093/database/bap017 42
- Chen, W., Liang, L. and Abecasis, G. (2008). GWAS GUI: Graphical browser for the results of whole-genome association studies with high-dimensional phenotypes. *Bioinformatics*. doi:10.1093/bioinformatics/btn600 193
- Cheung, K.-H., Kashyap, V., Luciano, J. S. *et al.* (2008). Semantic mashup of biomedical data. *Journal of Biomedical Informatics*, 41(5), 683–6. doi:10.1016/j.jbi.2008.08.003 264
- Church, D. M., Lappalainen, I., Sneddon, T. P. *et al.* (2010). Public data archives for genomic structural variation. *Nature Genetics*, 42(10), 813–814. doi:10.1038/ng1010-813 28
- Church, G. M. (2005). The Personal Genome Project. *Molecular Systems Biology*, 1, 2005.0030. doi:10.1038/msb4100040 13
- Cochrane, G., Akhtar, R., Aldebert, P. *et al.* (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, 36(Database issue), D5–D12. doi:10.1093/nar/gkm1018 20
- Cockerill, M. (2004). Delayed impact: ISI's citation tracking choices are keeping scientists in the dark. *BMC Bioinformatics*, 5(1), 93. doi:10.1186/1471-2105-5-93 219
- Collins, F. S., Guyer, M. S. and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science*, 278(5343), 1580–1. doi:10.1126/science.278.5343.1580 7
- Conrad, D., Pinto, D., Redon, R. *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289), 704–712. doi:10.1038/nature08516 12
- Costello, M. J. (2009). Motivating Online Publication of Data. *Bioscience*, 59(5), 418–427. doi:10.1525/bio.2009.59.5.9 19, 220

- Cox, T. M. (1999). Mendel and his legacy. *QJM: An International Journal of Medicine*, 92(4), 183–186. 4
- Cupples, L. A., Arruda, H. T., Benjamin, E. J. *et al.* (2007). The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics*, 8(Suppl 1), S1. doi:10.1186/1471-2350-8-S1-S154
- Dalgleish, R., Flicek, P., Cunningham, F. *et al.* (2010). Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Medicine*, 2(4), 24. doi:10.1186/gm145 45
- Daub, J., Gardner, P. P., Tate, J. *et al.* (2008). The RNA WikiProject: community annotation of RNA families. *RNA*, 14(12), 2462–4. doi:10.1261/rna.1200508 57
- De Roure, D. and Goble, C. (2009). Software Design for Empowering Scientists. *IEEE Software*, 26(1), 88–95. doi:10.1109/MS.2009.22 84
- Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107–113. doi:10.1145/1327452.1327492 53
- Deatrich, D., Liu, S., Payne, C. *et al.* (2008). Managing Petabyte-Scale Storage for the ATLAS Tier-1 Centre at TRIUMF. *22nd Annual International Symposium on High Performance Computing Systems and Applications*, pages 167–171. doi:10.1109/HPCS.2008.27 53
- Dellavalle, R. P., Hester, E. J., Heilig, L. F. *et al.* (2003). Information science. Going, going, gone: lost Internet references. *Science*, 302(5646), 787–8. doi:10.1126/science.1088234 221
- den Dunnen, J. T. and Antonarakis, S. E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human Mutation*, 15(1), 7–12. doi:10.1002/(SICI)1098-1004(200001)15:1;1-AID-HUMU4;3.0.CO;2-N 45
- den Dunnen, J. T., Sijmons, R. H., Andersen, P. S. *et al.* (2009). Sharing data between LSDBs and central repositories. *Human Mutation*, 30(4), 493–495. doi:10.1002/humu.20977 44
- Dermitzakis, E. T. and Clark, A. G. (2009). Life after GWA studies. *Science*, 326(5950), 239–240. doi:10.1126/science.1182009 10
- Doctorow, C. (2008). Welcome to the petacentre. *Nature*, 455(7209), 16–21. doi:10.1038/455016a 52
- Dowell, R. D., Jokerst, R. M., Day, A. *et al.* (2001). The Distributed Annotation System. *BMC Bioinformatics*, 2, 7. doi:10.1186/1471-2105-2-7 70

- Dudley, J. T. and Butte, A. J. (2010). In silico research in the era of cloud computing. *Nature Biotechnology*, 28(11), 1181–1185. doi:10.1038/nbt1110-1181 53, 88
- Dwinell, M., Worthey, E., Shimoyama, M. *et al.* (2008). The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Research*, pages D744–D749. doi:10.1093/nar/gkn842 32
- Editors (2006). The Impact Factor Game. *PLoS Medicine*, 3(6), e291. doi:10.1371/journal.pmed.0030291 219
- Editors (2008a). Positively disruptive. *Nature Genetics*, 40(2), 119. doi:10.1038/ng0208-119 61
- Editors (2008b). Prepare for the deluge. *Nature Biotechnology*, 26(10), 1099. doi:10.1038/nbt1008-1099 52
- Editors (2008c). Standardizing data. *Nature Cell Biology*, 10(10), 1123–1124. doi:10.1038/ncb1008-1123 54
- Eilbeck, K., Lewis, S. E., Mungall, C. J. *et al.* (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), R44. doi:10.1186/gb-2005-6-5-r44 72
- Elliott, P., Peakman, T. C. and Biobank, U. (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology*, 37(2), 234–44. doi:10.1093/ije/dym276 9
- Elnitski, L. L., Shah, P., Moreland, R. T. *et al.* (2007). The ENCODEdb portal: simplified access to ENCODE Consortium data. *Genome Research*, 17(6), 954–959. doi:10.1101/gr.5582207 59
- Enserink, M. (2009). Scientific publishing. Are you ready to become a number? *Science*, 323(5922), 1662–1664. doi:10.1126/science.323.5922.1662 225
- Feenstra, I., Fang, J., Koolen, D. A. *et al.* (2006). European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. *European Journal of Medical Genetics*, 49(4), 279–91. doi:10.1016/j.ejmg.2005.10.131 27
- Feigenbaum, L., Herman, I., Hongsermeier, T. *et al.* (2007). The Semantic Web in action. *Scientific American*, 297(6), 90–97. doi:10.1038/scientificamerican1207-90 264
- Field, D., Sansone, S.-A., Collis, A. *et al.* (2009). 'Omics data sharing. *Science*, 326(5950), 234–236. doi:10.1126/science.1180598 71
- Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine. Available online at <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>. 262

- Finn, R. D., Stalker, J. W., Jackson, D. K. *et al.* (2007). ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, 23(12), 1568–1570. doi:10.1093/bioinformatics/btl650 173
- Firth, H. V., Richards, S. M., Bevan, A. P. *et al.* (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4), 524–533. doi:10.1016/j.ajhg.2009.03.010 27
- Fisher, P., Hedeler, C., Wolstencroft, K. *et al.* (2007). A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Research*, 35(16), 5625–5633. doi:10.1093/nar/gkm623 84
- Flicek, P., Aken, B. L., Ballester, B. *et al.* (2010). Ensembl's 10th year. *Nucleic Acids Research*, 38(Database issue), D557–D562. doi:10.1093/nar/gkp972 16
- Fokkema, I. F. A. C., den Dunnen, J. T. and Taschner, P. E. M. (2005). LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. *Human Mutation*, 26(2), 63–68. doi:10.1002/humu.20201 46
- Fortier, I., Burton, P. R., Robson, P. J. *et al.* (2010). Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *International Journal of Epidemiology*, 39(5). doi:10.1093/ije/dyq139 144
- Foster, I. (2005). Service-oriented science. *Science*, 308(5723), 814–817. doi:10.1126/science.1110411 69
- Foster, I., Kesselman, C. and Tuecke, S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Applications*, 15(3), 200–222. doi:10.1177/109434200101500302 69
- Foster, M. W. and Sharp, R. R. (2007). Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Reviews Genetics*, 8(8), 633–9. doi:10.1038/nrg2124 227
- Fredman, D., Munns, G., Rios, D. *et al.* (2004a). HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Research*, 32(Database issue), D516–D519. doi:10.1093/nar/gkh111 24
- Fredman, D., Siegfried, M., Yuan, Y. P. *et al.* (2002). HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research*, 30(1), 387–391. doi:10.1093/nar/30.1.387 24
- Fredman, D., White, S. J., Potter, S. *et al.* (2004b). Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, 36(8), 861–866. doi:10.1038/ng1401 11

- Frishauf, P. (2009). Reputation systems: a new vision for publishing and peer review. *J Participat Med*, 1(1), e13a. Published online at <http://www.jopm.org/opinion/commentary/2009/10/21/reputation-systems-a-new-vision-for-publishing-and-peer-review>. Archived by WebCite® at <http://www.webcitation.org/5t59wGUzq>. 203
- Galperin, M. Y. and Cochrane, G. R. (2009). Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Research*, 37(Database issue), D1–D4. doi:10.1093/nar/gkn942 16
- Gardner, M. J., Hall, N., Fung, E. *et al.* (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906), 498–511. doi:10.1038/nature01097 31
- Garfield, E. (1955). Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111. doi:10.1126/science.122.3159.108 219
- Garfield, E. (1969). British quest for uniqueness versus American egocentrism. *Nature*, 223, 763. doi:10.1038/223763b0 223
- Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association journal*, 161(8), 979–80. 219
- Garten, Y. and Altman, R. B. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*, 10(Suppl 2), S6. doi:10.1186/1471-2105-10-S2-S6 43
- Garwood, K., McLaughlin, T., Garwood, C. *et al.* (2004a). PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, 5(1), 68. doi:10.1186/1471-2164-5-68 286
- Garwood, K., Taylor, C., Runte, K. *et al.* (2004b). Pedro: a configurable data entry tool for XML. *Bioinformatics*, 20(15), 2463–2465. doi:10.1093/bioinformatics/bth251 286
- Ge, H., Walhout, A. J. M. and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics*, 19(10), 551–560. doi:10.1016/j.tig.2003.08.009 18
- Giardine, B., Riemer, C., Hefferon, T. *et al.* (2007a). PhenCode: connecting ENCODE data with mutations and phenotype. *Human Mutation*, 28(6), 554–562. doi:10.1002/humu.20484 44
- Giardine, B., van Baal, S., Kaimakis, P. *et al.* (2007b). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Human Mutation*, 28(2), 206. doi:10.1002/humu.9479 43

- Gilad, Y., Pritchard, J. K. and Thornton, K. (2009). Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics*, 25(10), 463–471. doi:10.1016/j.tig.2009.09.003 13
- Gilbert, N. (2008). Researchers criticize genetic data restrictions. *Nature News*. Published online September 4. doi:10.1038/news.2008.1083 60
- Goble, C. and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5), 687–693. doi:10.1016/j.jbi.2008.01.008 18, 68, 73
- Goble, C., Stevens, R., Hull, D. *et al.* (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9(6), 506–517. doi:10.1093/bib/bbn034 85
- Goble, C. A., Bhagat, J., Aleksejevs, S. *et al.* (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 38(suppl 2), W677–W682. doi:10.1093/nar/gkq429 84
- Goldstein, D. B. (2009). Common genetic variation and human traits. *The New England Journal of Medicine*, 360(17), 1696–8. doi:10.1056/NEJMp0806284 10
- Good, B. and Wilkinson, M. (2006). The Life Sciences Semantic Web is Full of Creeps! *Briefings in Bioinformatics*, 7(3), 275–286. doi:10.1093/bib/bbl025 79
- Gretarsdottir, S., Sveinbjörnsdottir, S., Jonsson, H. H. *et al.* (2002). Localization of a susceptibility gene for common forms of stroke to 5q12. *American Journal of Human Genetics*, 70(3), 593–603. doi:10.1086/339252 6
- Gretarsdottir, S., Thorleifsson, G., Manolescu, A. *et al.* (2008). Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Annals of Neurology*, 64(4), 402–9. doi:10.1002/ana.21480 10
- Grubb, S. C., Maddatu, T. P., Bult, C. J. *et al.* (2009). Mouse phenome database. *Nucleic Acids Research*, 37(Database issue), D720–D730. doi:10.1093/nar/gkn778 32
- Gudbjartsson, D. F., Arnar, D. O., Helgadottir, A. *et al.* (2007). Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*, 448(7151), 353–357. doi:10.1038/nature06007 10
- Gulcher, J. R., Jónsson, P., Kong, A. *et al.* (1997). Mapping of a familial essential tremor gene, FET1, to chromosome 3q13. *Nature Genetics*, 17(1), 84–87. doi:10.1038/ng0997-84 6
- Guldener, U., Munsterkotter, M., Kastenmuller, G. *et al.* (2005). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*, 33(Database Issue), D364–D368. doi:10.1093/nar/gki053 30

- Guttmacher, A. E. and Collins, F. S. (2002). Genomic medicine—a primer. *The New England Journal of Medicine*, 347(19), 1512–20. doi:10.1056/NEJMra012240 1, 6
- Guttmacher, A. E. and Collins, F. S. (2003). Welcome to the genomic era. *The New England Journal of Medicine*, 349(10), 996–998. doi:10.1056/NEJMe038132 1
- Guttmacher, A. E. and Collins, F. S. (2005). Realizing the promise of genomics in biomedical research. *Journal of the American Medical Association*, 294(11), 1399–1402. doi:10.1001/jama.294.11.1399 1
- Guttmacher, A. E., Nabel, E. G. and Collins, F. S. (2009). Why data-sharing policies matter. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40), 16894. doi:10.1073/pnas.0910378106 320
- Harding, P., Johansson, L. and Klingenstein, N. (2008). Dynamic Security Assertion Markup Language: Simplifying Single Sign-On. *IEEE Security & Privacy*, 6(2), 83–85. doi:10.1109/MSP.2008.31 208, 228
- Hardison, R. C., Chui, D. H., Riemer, C. R. *et al.* (1998). Access to a syllabus of human hemoglobin variants (1996) via the World Wide Web. *Hemoglobin*, 22(2), 113–127. 43
- Harris, T. W., Antoshechkin, I., Bieri, T. *et al.* (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, 38(Database issue), D463–467. doi:10.1093/nar/gkp952 29
- Heidorn, B. (2008). Shedding light on the dark data in the long tail of science. *Library trends*, 57(2), 280–299. doi:10.1353/lib.0.0036 19
- Hertz-Fowler, C., Peacock, C., Wood, V. *et al.* (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research*, 32(Database Issue), D339–D343. doi:10.1093/nar/gkh007 31
- Hey, T. and Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, 308(5723), 817–821. doi:10.1126/science.1110410 68
- Higgs, D. R., Vickers, M. A., Wilkie, A. O. *et al.* (1989). A review of the molecular genetics of the human alpha-globin gene cluster. *Blood*, 73(5), 1081–1104. 4
- Hindorff, L. A., Sethupathy, P., Junkins, H. A. *et al.* (2009a). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9362–9367. doi:10.1073/pnas.0903103106 9, 39
- Hindorff, L. A., Sethupathy, P., Junkins, H. A. *et al.* (2009b). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9362–9367. doi:10.1073/pnas.0903103106 196

- Hirschhorn, J. N. (2009). Genomewide association studies—illuminating biologic pathways. *The New England Journal of Medicine*, 360(17), 1699–701. doi:10.1056/NEJMp0808934 10
- Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2), 95–108. doi:10.1038/nrg1521 5, 7
- Hirschman, J., Berardini, T. Z., Drabkin, H. J. *et al.* (2010). A MOD(ern) perspective on literature curation. *Molecular Genetics and Genomics*, 283(5), 415–425. doi:10.1007/s00438-010-0525-8 29
- Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9), 1047–1051. doi:10.1038/ng.f.217 57
- Holt, R., Subramanian, G., Halpern, A. *et al.* (2002). The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. *Science*, 298(5591), 129–149. doi:10.1126/science.1076181 31
- Homer, N., Szelinger, S., Redman, M. *et al.* (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8), e1000167. doi:10.1371/journal.pgen.1000167 60, 184, 185, 242, 250
- Hong, E., Balakrishnan, R., Dong, Q. *et al.* (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Research*, 36(Database issue), D577–581. doi:10.1093/nar/gkm909 30
- Hood, L., Heath, J. R., Phelps, M. E. *et al.* (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696), 640–643. doi:10.1126/science.1104635 18
- Horaitis, O. and Cotton, R. G. H. (2004). The challenge of documenting mutation across the genome: the Human Genome Variation Society approach. *Human Mutation*, 23(5), 447–452. doi:10.1002/humu.20038 44
- Horaitis, O., Talbot, C. C. J., Phommavanh, M. *et al.* (2007). A database of locus-specific databases. *Nature Genetics*, 39(4), 425. doi:10.1038/ng0407-425 43
- Howe, D., Costanzo, M., Fey, P. *et al.* (2008). The future of biocuration. *Nature*, 455(7209), 47–50. doi:10.1038/455047a 55, 57, 63
- Hubbard, T. J. P., Aken, B. L., Ayling, S. *et al.* (2009). Ensembl 2009. *Nucleic Acids Research*, 37(Database issue), D690–D697. doi:10.1093/nar/gkn828 35
- Hulbert, E. M., Smink, L. J., Adlem, E. C. *et al.* (2007). T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Research*, 35(Database issue), D742–D746. doi:10.1093/nar/gkl933 46

- Hull, D., Pettifer, S. R. and Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Computational Biology*, 4(10), e1000204. doi:10.1371/journal.pcbi.1000204 15
- Hull, D., Wolstencroft, K., Stevens, R. *et al.* (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue), W729–W732. doi:10.1093/nar/gkl320 84
- Hunter, D. J., Kraft, P., Jacobs, K. B. *et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7), 870–874. doi:10.1038/ng2075 109
- Iafrate, A. J., Feuk, L., Rivera, M. N. *et al.* (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9), 949–951. doi:10.1038/ng1416 11, 26
- Ioannidis, J. P. A., Allison, D. B., Ball, C. A. *et al.* (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2), 149–155. doi:10.1038/ng.295 56
- Jacobs, K. B., Yeager, M., Wacholder, S. *et al.* (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41(11), 1253–1257. doi:10.1038/ng.455 60
- Johnson, A. and O'donnell, C. (2009). An Open Access Database of Genome-wide Association Results. *BMC Medical Genetics*, 10(1), 6. doi:10.1186/1471-2350-10-6 186
- Jones, A., Lister, A., Hermida, L. *et al.* (2009). Modeling and Managing Experimental Data Using FuGE. *OMICS A Journal of Integrative Biology*, 13(3), 239–51. doi:10.1089/omi.2008.0080 126, 146
- Jones, A. R., Miller, M., Aebersold, R. *et al.* (2007). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotechnology*, 25(10), 1127–1133. doi:10.1038/nbt1347 71
- Jones, A. R. and Paton, N. W. (2005). An analysis of extensible modelling for functional genomics data. *BMC Bioinformatics*, 6, 235. doi:10.1186/1471-2105-6-235 123
- Kalow, W. (2002). Pharmacogenetics and personalised medicine. *Fundamental clinical pharmacology*, 16(5), 337–342. doi:10.1046/j.1472-8206.2002.00109.x 42
- Kasprzyk, A., Keefe, D., Smedley, D. *et al.* (2004). EnsMart: A Generic System for Fast and Flexible Access to Biological Data Access to Biological Data. *Genome Research*, 14(1), 160–169. doi:10.1101/gr.1645104 157
- Kaye, J., Heeney, C., Hawkins, N. *et al.* (2009). Data sharing in genomics — re-shaping scientific practice. *Nature Reviews Genetics*, 10(5), 331–335. doi:10.1038/nrg2573 48

- Kerem, B., Rommens, J. M., Buchanan, J. A. *et al.* (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922), 1073–1080. doi:10.1126/science.2570460 5
- King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2), 173–199. doi:10.1177/0049124107306660 258
- Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560), 1662–1664. doi:10.1126/science.1069492 18
- Klein, R. J., Zeiss, C., Chew, E. Y. *et al.* (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720), 385–9. doi:10.1126/science.1109557 8, 9
- Koike, A., Nishida, N., Inoue, I. *et al.* (2009). Genome-wide association database developed in the Japanese Integrated Database Project. *Journal of Human Genetics*, 54(9), 543–546. doi:10.1038/jhg.2009.68 50
- Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nature Genetics*, 27(3), 234–236. doi:10.1038/85776 3
- Kyrpides, N. C. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nature Biotechnology*, 27(7), 627–632. doi:10.1038/nbt.1552 57
- Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288), 488–489. doi:10.1038/464488a 259
- Langella, S., Hastings, S., Oster, S. *et al.* (2008). Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment. *Journal of the American Medical Informatics Association*, 15(3), 363–373. doi:10.1197/jamia.M2662 229
- Langmead, B., Schatz, M., Lin, J. *et al.* (2009). Searching for SNPs with cloud computing. *Genome Biology*, 10(11), R134. doi:10.1186/gb-2009-10-11-r134 53
- Laporte, R., Marler, E., Akazawa, S. *et al.* (1995). The death of biomedical journals. *British Medical Journal*, 310(6991), 1387–1390. 62
- Laursen, L. (2009). Fake Facebook pages spin web of deceit. *Nature*, 458(7242), 1089. doi:10.1038/news.2009.398 205
- Lawrence, P. A. (2003). The politics of publication. *Nature*, 422(6929), 259–61. doi:10.1038/422259a 219
- Leary, R., Kinde, I., Diehl, F. *et al.* (2010). Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. *Science Translational Medicine*, 2(20), 20ra14. doi:10.1126/scitranslmed.3000702 13

- Lehmann, H. and Kynoch, P. A. M. (1976). *Human haemoglobin variants and their characteristics*. North-Holland Publishing Company, Amsterdam. 43
- Lenffer, J., Nicholas, F. W., Castle, K. *et al.* (2006). OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Research*, 34(Database issue), D599–D601. doi:10.1093/nar/gkj152 41
- Letovsky, S. I., Cottingham, R. W., Porter, C. J. *et al.* (1998). GDB: the Human Genome Database. *Nucleic Acids Research*, 26(1), 94–99. 23
- Li, H., Handsaker, B., Wysoker, A. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–9. doi:10.1093/bioinformatics/btp352 189
- Li, W. H. and Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics*, 129(2), 513–523. 3
- Lister, R., Pelizzola, M., Dowen, R. H. *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), 315–322. doi:10.1038/nature08514 13
- Little, J., Higgins, J. P. T., Ioannidis, J. P. A. *et al.* (2009). STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. *PLoS Medicine*, 6(2), e22. doi:10.1371/journal.pmed.1000022 37
- Litton, J.-E., Muilu, J., Bjorklund, A. *et al.* (2003). Data modeling and data communication in GenomEUtwin. *Twin Research*, 6(5), 383–390. doi:10.1375/136905203770326385 69
- Louie, B., Mork, P., Martin-Sanchez, F. *et al.* (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics*, 40(1), 5–16. doi:10.1016/j.jbi.2006.02.007 2
- Lowrance, W. W. and Collins, F. S. (2007). Identifiability in genomic research. *Science*, 317(5838), 600–602. doi:10.1126/science.1147699 61
- Lunshof, J. E., Chadwick, R., Vorhaus, D. B. *et al.* (2008). From genetic privacy to open consent. *Nature Reviews Genetics*, 9(5), 406–411. doi:10.1038/nrg2360 61
- Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S. *et al.* (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2), 219–232. doi:10.1016/0092-8674(91)90613-4 11
- Macmullen, W. J. (2007). Facets and measures of gene ontology annotation quality in model organism databases. *Proceedings of the American Society for Information Science and Technology*, 43(1), 1–7. doi:10.1002/meet.14504301260 55

- Maglott, D., Ostell, J., Pruitt, K. D. *et al.* (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(Database issue), D26–D31. doi:10.1093/nar/gkl993 57
- Maier, D., Wymore, F., Sherlock, G. *et al.* (2008). The XBabelPhish MAGE-ML and XML Translator. *BMC Bioinformatics*, 9(1). doi:10.1186/1471-2105-9-28 75
- Mailman, M. D., Feolo, M., Jin, Y. *et al.* (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10), 1181–1186. doi:10.1038/ng1007-1181 48
- Maler, E. (2009). The design of everyday identity. *Online Information Review*, 33(3), 443–457. doi:10.1108/14684520910969899 210, 244
- Maler, E. and Reed, D. (2008). The Venn of Identity: Options and Issues in Federated Identity Management. *IEEE Security & Privacy*, 6(2), 16–23. doi:10.1109/MSP.2008.50 216
- Mallon, A.-M., Blake, A. and Hancock, J. M. (2008). EuroPhenome and EMPReSS: online mouse phenotyping resource. *Nucleic Acids Research*, 36(Database issue), D715–D718. doi:10.1093/nar/gkm728 36
- Malone, J., Holloway, E., Adamusiak, T. *et al.* (2010). Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8), 1112–1118. doi:10.1093/bioinformatics/btq099 192, 198
- Manion, F. J., Robbins, R. J., Weems, W. A. *et al.* (2009). Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. *BMC Medical Informatics and Decision Making*, 9, 31. doi:10.1186/1472-6947-9-31 229, 261
- Manolio, T. A. (2008). Biorepositories—at the bleeding edge. *International Journal of Epidemiology*, 37(2), 231–3. doi:10.1093/ije/dym282 9
- Manolio, T. A., Collins, F. S., Cox, N. J. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753. doi:10.1038/nature08494 10
- Manolio, T. A., Rodriguez, L. L., Brooks, L. *et al.* (2007). New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genetics*, 39(9), 1045–1051. doi:10.1038/ng2127 48
- Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133–141. doi:10.1016/j.tig.2007.12.007 12
- Marshall, C. R., Noor, A., Vincent, J. B. *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*, 82(2), 477–488. doi:10.1016/j.ajhg.2007.12.009 27

- Martin, S., Hohman, M. M. and Liefeld, T. (2005). The impact of Life Science Identifier on informatics data. *Drug Discovery Today*, 10(22), 1566–1572. doi:10.1016/S1359-6446(05)03651-2 170
- McCarroll, S. A. (2008). Extending genome-wide association studies to copy-number variation. *Human Molecular Genetics*, 17(R2), R135–R142. doi:10.1093/hmg/ddn282 12
- McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, 39(7 Suppl), S37–S42. doi:10.1038/ng2080 27
- McKusick, V. A. (1966). *Mendelian Inheritance in Man, A Catalog of Autosomal Dominant, Autosomal Recessive, and X-linked Phenotypes*. Johns Hopkins University Press, Baltimore, MD, 1st edition. 40
- McPherson, J. D. (2009). Next-generation gap. *Nature Methods*, 6(11s), S2–S5. doi:10.1038/nmeth.f.268 52
- McWilliam, H., Valentin, F., Goujon, M. *et al.* (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Research*, 37(Web Server issue), W6–W10. doi:10.1093/nar/gkp302 86
- Mellor, S., Scott, K., Uhl, A. *et al.* (2002). Model-Driven Architecture. *Advances in Object-Oriented Information Systems*, pages 233–239. doi:10.1007/3-540-46105-1_3 70
- Mons, B., Ashburner, M., Chichester, C. *et al.* (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biology*, 9, R89. doi:10.1186/gb-2008-9-5-r89 57
- Mons, B. and Velterop, J. (2009). Nano-Publication in the e-science era. *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*. Available online at http://www.nbic.nl/uploads/media/Nano-Publication_BarendMons-JanVelterop.pdf [Accessed 2010-06-10]. Archived by WebCite® at <http://www.webcitation.org/5qSZBHvRi>. 258
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 114. doi:10.1109/N-SSC.2006.4785860 53
- Muilu, J., Peltonen, L. and Litton, J.-E. (2007). The federated database—a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe. *European Journal of Human Genetics*, 15(7), 718–723. doi:10.1038/sj.ejhg.5201850 69
- Muller, H.-M., Kenny, E. E. and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11), e309. doi:10.1371/journal.pbio.0020309 43

- Mungall, C. J. and Emmert, D. B. (2007). A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13), i337–i346. doi:10.1093/bioinformatics/btm189 143, 281
- Murphy, S. N., Weber, G., Mendis, M. *et al.* (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2), 124–30. doi:10.1136/jamia.2009.000893 83
- Nelson, B. (2009). Data sharing: Empty archives. *Nature*, 461(7261), 160–163. doi:10.1038/461160a 19
- Nenadic, A., Zhang, N., Yao, L. *et al.* (2007). Levels of Authentication Assurance: an Investigation. *Third International Symposium on Information Assurance and Security*, pages 155–160. doi:10.1109/IAS.2007.88 215
- Neylon, C. (2009). Head in the clouds: Re-imagining the experimental laboratory record for the web-based networked world. *Automated Experimentation*, 1(1), 3. doi:10.1186/1759-4499-1-3 62
- Neylon, C. and Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biology*, 7(11), e1000242. doi:10.1371/journal.pbio.1000242 219
- O’Brien, S. (2009). Stewardship of Human Biospecimens, DNA, Genotype, and Clinical Data in the GWAS Era. *Annual Review of Genomics and Human Genetics*, 10(1), 193–209. doi:10.1146/annurev-genom-082908-150133 61
- Oinn, T., Addis, M., Ferris, J. *et al.* (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17), 3045–3054. doi:10.1093/bioinformatics/bth361 84
- Oldenburg, H. (1665). Epistle Dedicatory. *Philosophical Transactions*, 1(1-22). doi:10.1098/rstl.1665.0001 15
- Olson, M., Hood, L., Cantor, C. *et al.* (1989). A common language for physical mapping of the human genome. *Science*, 245(4925), 1434–1435. doi:10.1126/science.2781285 23
- Oppliger, R. (2004). Microsoft .NET Passport and identity management. *Information Security Technical Report*, 9(1), 26–34. doi:10.1016/S1363-4127(04)00013-5 210
- Osborne, J. D., Lin, S. and Kibbe, W. A. (2007). Other riffs on cooperation are already showing how well a wiki could work. *Nature*, 446(7138), 856. doi:10.1038/446856a 57
- Paskin, N. (2000). E-Citations: actionable identifiers and scholarly referencing. *Learned Publishing*, 13(3), 159–166. doi:10.1087/09531510050145308 170
- Paskin, N. (2005). Digital Object Identifiers for scientific data. *Data Science Journal*, 4, 12–20. doi:10.2481/dsj.4.12 221

- Pautasso, C., Zimmermann, O. and Leymann, F. (2008). RESTful web services vs. "big" web services: making the right architectural decision. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 805–814. doi:10.1145/1367497.1367606 262
- Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association*, 299(11), 1335–1344. doi:10.1001/jama.299.11.1335 37
- Peterson, J. and Campbell, J. (2010). Marker papers and data citation. *Nature Genetics*, 42(11), 919–919. doi:10.1038/ng1110-919 221
- Phillips, J., Chilukuri, R., Frago, G. *et al.* (2006). The caCORE Software Development Kit: streamlining construction of interoperable biomedical information services. *BMC Medical Informatics and Decision Making*, 6, 2. doi:10.1186/1472-6947-6-2 82
- Plesance, E. D., Cheetham, R. K., Stephens, P. J. *et al.* (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278), 191–196. doi:10.1038/nature08658 13
- Plesance, E. D., Stephens, P. J., O'Meara, S. *et al.* (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278), 184–190. doi:10.1038/nature08629 13
- Post, L. J. G., Roos, M., Marshall, M. S. *et al.* (2007). A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. *Bioinformatics*, 23(22), 3080–7. doi:10.1093/bioinformatics/btm461 265
- Prainsack, B., Reardon, J., Hindmarsh, R. *et al.* (2008). Personal genomes: Misdirected precaution. *Nature*, 456(7218), 34–35. doi:10.1038/456034a 61
- Qian, Y., Tchuvatkina, O., Spidlen, J. *et al.* (2009). FuGEFlow: data model and markup language for flow cytometry. *BMC Bioinformatics*, 10, 184. doi:10.1186/1471-2105-10-184 146
- Qiu, J. (2008). Scientific publishing: Identity crisis. *Nature News*, 451(7180), 766. doi:10.1038/451766a 223
- Quackenbush, J. (2004). Data standards for 'omic' science. *Nature Biotechnology*, 22(5), 613–4. doi:10.1038/nbt0504-613 76
- Quackenbush, J. (2006). Standardizing the standards. *Molecular Systems Biology*, 2, 2006.0010. doi:10.1038/msb4100052 81
- Quackenbush, J., Stoeckert, C., Ball, C. *et al.* (2006). Top-down standards will not serve systems biology. *Nature*, 440(7080), 24. doi:10.1038/440024a 76

- Rajeevan, H., Cheung, K.-H., Gadagkar, R. *et al.* (2005). ALFRED: An Allele Frequency Database for Microevolutionary Studies. *Evolutionary Bioinformatics Online*, 1, 1–10. 25
- Rayner, T. F., Rocca-Serra, P., Spellman, P. T. *et al.* (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7, 489. doi:10.1186/1471-2105-7-489 75
- Recordon, D. and Reed, D. (2006). OpenID 2.0: a platform for user-centric identity management. *DIM '06: Proceedings of the second ACM workshop on Digital identity management*, pages 11–16. doi:10.1145/1179529.1179532 211, 311
- Redon, R., Ishikawa, S., Fitch, K. *et al.* (2006). Global variation in copy number in the human genome. *Nature*, 444(7118), 444–54. doi:10.1038/nature05329 4
- Reich, D. E. and Goldstein, D. B. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14), 8119–8123. doi:10.1073/pnas.95.14.8119 3
- Renear, A. H. and Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325(5942), 828–832. doi:10.1126/science.1157784 63
- Rhead, B., Karolchik, D., Kuhn, R. M. *et al.* (2010). The UCSC Genome Browser database: update 2010. *Nucleic Acids Research*, 38(Database issue), D613–D619. doi:10.1093/nar/gkp939 16
- Riikonen, P. and Vihinen, M. (1999). MUTbase: maintenance and analysis of distributed mutation databases. *Bioinformatics*, 15(10), 852–859. 46
- Riordan, J. R., Rommens, J. M., Kerem, B. *et al.* (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922), 1066–1073. doi:10.1126/science.2475911 5
- Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788), 847–856. doi:10.1038/35015718 7
- Robinson, P., Köhler, S., Bauer, S. *et al.* (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *American Journal of Human Genetics*. doi:10.1016/j.ajhg.2008.09.017 249
- Rommens, J. M., Iannuzzi, M. C., Kerem, B. *et al.* (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922), 1059–1065. doi:10.1126/science.2772657 5
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. doi:10.1037/0033-2909.86.3.638 252

- Ruttenberg, A., Rees, J., Samwald, M. *et al.* (2009). Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in Bioinformatics*, 10(2), 193–204. doi:10.1093/bib/bbp004 264
- Sagotsky, J. A., Zhang, L., Wang, Z. *et al.* (2008). Life Sciences and the web: a new era for collaboration. *Molecular Systems Biology*, 4, 201. doi:10.1038/msb.2008.39 78
- Saltz, J., Oster, S., Hastings, S. *et al.* (2006). caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics*, 22(15), 1910–1916. doi:10.1093/bioinformatics/btl272 82
- Salzberg, S. L. (2007). Genome re-annotation: a wiki solution? *Genome Biology*, 8(1), 102. doi:10.1186/gb-2007-8-1-102 57
- Sander, C. (2000). Genomic medicine and the future of health care. *Science*, 287(5460), 1977–1978. doi:10.1126/science.287.5460.1977 1
- Sanguhl, K., Berlin, D. S., Altman, R. B. *et al.* (2008). PharmGKB: understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4), 539–551. doi:10.1080/03602530802413338 42
- Sankararaman, S., Obozinski, G., Jordan, M. I. *et al.* (2009). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9), 965–967. doi:10.1038/ng.436 60
- Sansom, C. (2010). Up in a cloud? *Nature Biotechnology*, 28(1), 13–15. doi:10.1038/nbt0110-13 53
- Sansone, S.-A., Rocca-Serra, P., Brandizi, M. *et al.* (2008). The First RSBI (ISA-TAB) Workshop: "Can a Simple Format Work for Complex Studies?". *OMICS A Journal of Integrative Biology*, 12(2), 143–149. doi:10.1089/omi.2008.0019 71, 75
- Sansone, S.-A., Rocca-Serra, P., Tong, W. *et al.* (2006). A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. *OMICS A Journal of Integrative Biology*, 10(2), 164–71. doi:10.1089/omi.2006.10.164 75
- Schatz, M. (2009). CloudBurst: Highly Sensitive Read Mapping with MapReduce. *Bioinformatics*, 25(11), 1363–1369. doi:10.1093/bioinformatics/btp236 53
- Schekman, R. (2009). PNAS takes action regarding breach of NIH embargo policy on a PNAS paper. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40), 16893. doi:10.1073/pnas.0910317106 320
- Scherer, S. W., Lee, C., Birney, E. *et al.* (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, 39, S7–S15. doi:10.1038/ng2093 23, 26, 120

- Schwenk, K., Padilla, D., Bakken, G. *et al.* (2009). Grand challenges in organismal biology. *Integrative and Comparative Biology*, 49(1), 7–14. doi:10.1093/icb/icp034 18
- Sebat, J., Lakshmi, B., Troge, J. *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683), 525–528. doi:10.1126/science.1098918 11, 26
- Seringhaus, M. and Gerstein, M. (2006). The Death of the Scientific Paper. *The Scientist*, 20(9), 25. Available online at <http://papers.gersteinlab.org/e-print/paperdeath/preprint.pdf>. 62
- Seringhaus, M. R. and Gerstein, M. B. (2007). Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, 8, 17. doi:10.1186/1471-2105-8-17 58, 62
- Shadbolt, N., Hall, W. and Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3), 96–101. doi:10.1109/MIS.2006.62 79
- Shaikh, T. H., Gai, X., Perin, J. C. *et al.* (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Research*, 19(9), 1682–90. doi:10.1101/gr.083501.108 114
- Sharp, A. J., Locke, D. P., McGrath, S. D. *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, 77(1), 78–88. doi:10.1086/431652 26
- Sherry, S. T., Ward, M. H., Kholodov, M. *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308–311. doi:10.1093/nar/29.1.308 24
- Shields, P. G. (2000). Publication bias is a scientific problem with adverse ethical outcomes: the case for a section for null results. *Cancer epidemiology, biomarkers prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 9(8), 771–772. 40
- Shimoyama, M., Hayman, G. T., Laulederkind, S. J. F. *et al.* (2009). The rat genome database curators: who, what, where, why. *PLoS Computational Biology*, 5(11), e1000582. doi:10.1371/journal.pcbi.1000582 55
- Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2), 85–94. doi:10.1087/2009202 63
- Sinnott, R. O., Doherty, T., Gray, N. *et al.* (2009). Semantic security: specification and enforcement of semantic policies for security-driven collaborations. *Studies in Health Technology and Informatics*, 147, 201–11. doi:10.3233/978-1-60750-027-8-201 214, 236
- Sistrom, C. L. and Garvan, C. W. (2004). Proportions, odds, and risk. *Radiology*, 230(1), 12–9. doi:10.1148/radiol.2301031028 104

- Smalheiser, N. R. and Torvik, V. I. (2009). Author Name Disambiguation. In Cronin, B., editor, *Annual Review of Information Science and Technology*, volume 43, pages 287–313. Information Today, Inc., Medford, New Jersey, USA. Available online at http://arrowsmith.psych.uic.edu/arrowsmith_uic/tutorial/ARIST_preprint.pdf. Archived by WebCite® at <http://www.webcitation.org/5qhu5g2Y3>. 223
- Smedley, D., Haider, S., Ballester, B. *et al.* (2009). BioMart – biological queries made easy. *BMC Genomics*, 10(22). doi:10.1186/1471-2164-10-22 34, 157, 173
- Smedley, D., Schofield, P., Chen, C. K. *et al.* (2010). Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *Database*, 2010, baq014. doi:10.1093/database/baq014 58
- Smedley, D., Swertz, M. A., Wolstencroft, K. *et al.* (2008). Solutions for data integration in functional genomics: a critical assessment and case study. *Briefings in Bioinformatics*, 9(6), 532–544. doi:10.1093/bib/bbn040 77
- Smith, B., Ashburner, M., Rosse, C. *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11), 1251–1255. doi:10.1038/nbt1346 72
- Smith, C. L., Goldsmith, C.-A. W. and Eppig, J. T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1), R7. doi:10.1186/gb-2004-6-1-r7 36
- Smith, V. (2009). Data publication: towards a database of everything. *BMC Research Notes*, 2(1), 113. doi:10.1186/1756-0500-2-113 16, 19
- Spellman, P. T., Miller, M., Stewart, J. *et al.* (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9), research0046.1–0046.9. doi:10.1186/gb-2002-3-9-research0046 70
- Stajich, J. E., Block, D., Boulez, K. *et al.* (2002). The Bioperl Toolkit: Perl modules for the life sciences. *Genome Research*, 12(10), 1611–8. doi:10.1101/gr.361602 153
- Stein, L. (2002). Creating a bioinformatics nation. *Nature*, 417(6885), 119–120. doi:10.1038/417119a 17
- Stein, L. D. (1997). How Perl Saved the Human Genome Project. *The Perl Journal*, 1(2). Published online at <http://www.drdobbs.com/architect/184410424>. Archived by WebCite® at <http://www.webcitation.org/5nvh2aRtg>. 76
- Stein, L. D. (2003). Integrating biological databases. *Nature Reviews Genetics*, 4(5), 337–345. doi:10.1038/nrg1065 66

- Stein, L. D. (2008). Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics*, 9(9), 678–688. doi:10.1038/nrg2414 68, 83
- Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology*, 11(5), 207. doi:10.1186/gb-2010-11-5-207 53
- Stein, L. D., Mungall, C., Shu, S. *et al.* (2002). The Generic Genome Browser: a building block for a model organism system database. *Genome Research*, 12(10), 1599–1610. doi:10.1101/gr.403602 34
- Stein, R., Dittmers, K., Fahl, K. *et al.* (2004). Arctic (palaeo) river discharge and environmental change: evidence from the Holocene Kara Sea sedimentary record. *Quaternary Science Reviews*, 23(11-13), 1485–1511. doi:10.1016/j.quascirev.2003.12.004 222
- Stenson, P., Mort, M., Ball, E. *et al.* (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine*, 1(1), 13. doi:10.1186/gm13 41
- Stevens, R., Goble, C. A. and Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1(4), 398–414. doi:10.1093/bib/1.4.398 36
- Stevens, R. D., Robinson, A. J. and Goble, C. A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19(Suppl 1), i302–304. doi:10.1093/bioinformatics/btg1041 84
- Stockinger, H., Attwood, T., Chohan, S. *et al.* (2008). Experience using web services for biological sequence analysis. *Briefings in Bioinformatics*, 9(6), 493–505. doi:10.1093/bib/bbn029 262
- Stranger, B. E., Forrest, M. S., Dunning, M. *et al.* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813), 848–853. doi:10.1126/science.1136678 127
- Sugawara, H., Ogasawara, O., Okubo, K. *et al.* (2008). DDBJ with new system and face. *Nucleic Acids Research*, 36(Database issue), D22–24. doi:10.1093/nar/gkm889 20
- Swertz, M., van der Velde, K. J., Tesson, B. *et al.* (2010). XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biology*, 11(3), R27. doi:10.1186/gb-2010-11-3-r27 77, 146
- Swertz, M. A., Brock, E. O. D., van Hijum, S. A. F. T. *et al.* (2004). Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases. *Bioinformatics*, 20(13), 2075–2083. doi:10.1093/bioinformatics/bth206 77

- Swertz, M. A. and Jansen, R. C. (2007). Beyond standardization: dynamic software infrastructures for systems biology. *Nature Reviews Genetics*, 8(3), 235–243. doi:10.1038/nrg2048 77, 257
- Szalay, A. and Gray, J. (2001). The world-wide telescope. *Science*, 293(5537), 2037–2040. doi:10.1126/science.293.5537.2037 65
- Szalay, A. S., Kunszt, P. Z., Thakar, A. *et al.* (2000). Designing and mining multi-terabyte astronomy archives: the Sloan Digital Sky Survey. *ACM SIGMOD Record*, 29(2), 451–462. doi:10.1145/335191.335439 54
- Tabor, H. K., Risch, N. J. and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, 3(5), 391–397. doi:10.1038/nrg796 5
- Tan, W., Foster, I. and Madduri, R. (2008). Combining the Power of Taverna and caGrid: Scientific Workflows that Enable Web-Scale Collaboration. *IEEE Internet Computing*, 12(6), 61–68. doi:10.1109/MIC.2008.120 86
- Taylor, C. F., Field, D., Sansone, S.-A. *et al.* (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8), 889–896. doi:10.1038/nbt.1411 56, 75
- Taylor, C. F., Paton, N. W., Garwood, K. L. *et al.* (2003). A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, 21(3), 247–254. doi:10.1038/nbt0303-247 286
- The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073. doi:10.1038/nature09534 11
- The BioMoby Consortium (2008). Interoperability with Moby 1.0—It’s better than sharing your toothbrush! *Briefings in Bioinformatics*, 9(3), 220–231. doi:10.1093/bib/bbn003 84
- The BioSapiens Network of Excellence (2005). BioSapiens: a European network for integrated genome annotation. *European Journal of Human Genetics*, 13(9), 994–997. doi:10.1038/sj.ejhg.5201470 87
- The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696), 636–640. doi:10.1126/science.1105136 59
- The FlyBase Consortium (1994). FlyBase — the *Drosophila* database. *Nucleic Acids Research*, 22(17), 3456–3458. doi:10.1093/nar/22.17.3456 30
- The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. *Nucleic Acids Research*, 36(Database issue), D440–D444. doi:10.1093/nar/gkm883 36

- The International Cancer Genome Consortium (2010). International network of cancer genome projects. *Nature*, 464(7291), 993–998. doi:10.1038/nature08987 13
- The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063), 1299–1320. doi:10.1038/nature04226 7, 220
- The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. doi:10.1038/35057062 1
- The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928–933. doi:10.1038/35057149 7
- The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, 38(Database issue), D142–D148. doi:10.1093/nar/gkp846 64
- The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. doi:10.1038/nature05911 48
- The Wellcome Trust Case Control Consortium (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289), 713–720. doi:10.1038/nature08979 12
- Thorisson, G., Lancaster, O., Free, R. *et al.* (2008). HGVbaseG2P: a central genetic association database. *Nucleic Acids Research*, 37(Database issue), D797–D802. doi:10.1093/nar/gkn748 51
- Thorisson, G. A. (2009a). Accreditation and attribution in data sharing. *Nature Biotechnology*, 27(11), 984–985. doi:10.1038/nbt1109-984b 222
- Thorisson, G. A. (2009b). Atom web feeds, the AtomPub protocol and G2P databases. Published online at <http://www.gen2phen.org/post/atom-web-feeds-atompub-protocol-and-g2p-databases>. Archived by WebCite® at <http://www.webcitation.org/5ndKYWUua>. 173
- Thorisson, G. A. (2009c). Web service architecture: simple REST vs complex SOAP. Published online at <http://www.gen2phen.org/post/web-service-architecture-simple-rest-vs-complex-soap>. Archived by WebCite® at <http://www.webcitation.org/5ndKPyUzz>. 194
- Thorisson, G. A., Muilu, J. and Brookes, A. J. (2009). Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nature Reviews Genetics*, 10(1), 9–18. doi:10.1038/nrg2483 65
- Thorisson, G. A., Smith, A. V., Krishnan, L. *et al.* (2005). The International HapMap Project Web site. *Genome Research*, 15(11), 1592–1593. doi:10.1101/gr.4413105 26

- Thornton, J. and the BioSapiens Network (2009). Annotations for all by all – the BioSapiens network. *Genome Biology*, 10(2), 401. doi:10.1186/gb-2009-10-2-401 87
- Torvik, V. I. and Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3), 1–29. doi:10.1145/1552303.1552304 223
- Tweedie, S., Ashburner, M., Falls, K. *et al.* (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, 37(Database issue), D555–D559. doi:10.1093/nar/gkn788 30
- Ubois, J. (2003). Online reputation systems. *Release 1.0 Esther Dyson's Monthly Report*, 21(9), 1–33. Available online at <http://cdn.oreilly.com/radar/r1/10-03.pdf>. Archived by WebCite® at <http://www.webcitation.org/5lKQ0QD3W>. Accessed January 11, 2010. 203
- Valentin, F., Squizzato, S., Goujon, M. *et al.* (2010). Fast and efficient searching of biological data resources—using EB-eye. *Briefings in Bioinformatics*, 11(4), 375–384. doi:10.1093/bib/bbp065 66
- van Driel, M. A., Bruggeman, J., Vriend, G. *et al.* (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5), 535–542. doi:10.1038/sj.ejhg.5201585 41
- Venter, J. C., Adams, M. D., Myers, E. W. *et al.* (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351. doi:10.1126/science.1058040 1
- von Elm, E., Moher, D., Little, J. *et al.* (2009). Reporting genetic association studies: the STREGA statement. *Lancet*, 374(9684), 98–100. doi:10.1016/S0140-6736(09)61265-4 37
- Wain, L. V., Armour, J. A. L. and Tobin, M. D. (2009). Genomic copy number variation, human health, and disease. *Lancet*, 374(9686), 340–350. doi:10.1016/S0140-6736(09)60249-X 12
- Waldrop, M. (2008). Wikiomics. *Nature*, 455(7209), 22–25. doi:10.1038/455022a 57
- Walker, F. O. (2007). Huntington's disease. *Lancet*, 369(9557), 218–228. doi:10.1016/S0140-6736(07)60111-1 4
- Wang, D. G., Fan, J.-B., Siao, C.-J. *et al.* (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, 280(5366), 1077–1082. doi:10.1126/science.280.5366.1077 7
- Wang, J., Song, L., Grover, D. *et al.* (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation*, 27(4), 323–329. doi:10.1002/humu.20307 26

- Wang, X., Gorlitsky, R. and Almeida, J. S. (2005). From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nature Biotechnology*, 23(9), 1099–1103. doi:10.1038/nbt1139 81
- Warner, S. (2010). Author Identifiers in Scholarly Repositories. *arXiv*. Available online at <http://arxiv.org/abs/1003.1345v1>. 240
- Washington, N. L., Haendel, M. A., Mungall, C. J. *et al.* (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, 7(11), e1000247. doi:10.1371/journal.pbio.1000247 249
- Watt, J., Sinnott, R., Jiang, J. *et al.* (2009). Tool Support for Security-Oriented Virtual Research Collaborations. In *2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pages 419–424. doi:10.1109/ISPA.2009.49 214
- Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44(3), 388–396. 5
- Weitzner, D. J. (2006). In Search of Manageable Identity Systems. *IEEE Internet Computing*, 10(6), 84–86. doi:10.1109/MIC.2006.127 210
- Weitzner, D. J. (2007). Whose Name Is It, Anyway? Decentralized Identity Systems on the Web. *IEEE Internet Computing*, 11(4), 72–76. doi:10.1109/MIC.2007.95 210
- Wheeler, D. L., Barrett, T., Benson, D. A. *et al.* (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Database issue), D13–D21. doi:10.1093/nar/gkm1000 16, 25
- Wilkinson, M. D. and Links, M. (2002). BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics*, 3(4), 331–41. doi:10.1093/bib/3.4.331 84
- Wolfson, M., Wallace, S., Masca, N. *et al.* (2010). DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual-level data without sharing the data. *International Journal of Epidemiology*, Advance access, published online July 14. doi:10.1093/ije/dyq111 251
- Wolinsky, H. (2008). What's in a name? *EMBO Reports*, 9(12), 1171–1174. doi:10.1038/embor.2008.217 259
- Yeager, M., Orr, N., Hayes, R. B. *et al.* (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5), 645–649. doi:10.1038/ng2022 109, 110
- Yee, K.-P., Swearingen, K., Li, K. *et al.* (2003). Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 5, pages 401–408. doi:10.1145/642611.642681 170

- Yu, W., Clyne, M., Dolan, S. M. *et al.* (2008a). GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9, 205. doi:10.1186/1471-2105-9-205 38
- Yu, W., Clyne, M., Khoury, M. *et al.* (2010). Phenopedia and Genopedia: Disease-centered and Gene-centered Views of the Evolving Knowledge of Human Genetic Associations. *Bioinformatics*, 26(1), 145–146. doi:10.1093/bioinformatics/btp618 38
- Yu, W., Gwinn, M., Clyne, M. *et al.* (2008b). A navigator for human genome epidemiology. *Nature Genetics*, 40(2), 124–125. doi:10.1038/ng0208-124 37
- Yu, W., Wulf, A., Yesupriya, A. *et al.* (2008c). HuGE Watch: tracking trends and patterns of published studies of genetic association and human genome epidemiology in near-real time. *European Journal of Human Genetics*, 16(9), 1155–1158. doi:10.1038/ejhg.2008.95 4
- Zerhouni, E. A. and Nabel, E. G. (2008). Protecting aggregate genomic data. *Science*, 322(5898), 44. doi:10.1126/science.1165490 60
- Zhang, J., Feuk, L., Duggan, G. E. *et al.* (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic and Genome Research*, 115(3-4), 205–214. doi:10.1159/000095916 26
- Zhang, Y., De, S., Garner, J. R. *et al.* (2010). Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics*, 3(1). doi:10.1186/1755-8794-3-1 196