CASIMIR: Coordination and Sustainability of International Mouse Informatics Resources

John M. Hancock, Paul N. Schofield, Christina Chandras, Michael Zouberakis, Vassilis Aidinis, Damian Smedley, Nadia Rosenthal, Klaus Schughart, The CASIMIR Consortium

***Abstract*—In recent years the European Commission has funded an increasing number of functional genomics projects aimed at using the mouse as a model of human disease. Many of these projects are producing large data volumes. A recently funded programme, CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources) aims to make recommendations on the most efficient way to integrate these datasets. In Summer 2007 CASIMIR carried out a questionnaire survey of relevant EC-funded projects to determine their current use of data integration technologies and standards. This report describes the consortium's aims, its achievements so far, the results of the survey and initial conclusions deriving from it.**

I. CASIMIR

The need for integration of data sets is well established in the computer science, bioinformatics and high throughput biology communities. However it is less well-established amongst bench biologists whose primary interest is hypothesis-driven experimental science and do not have experience of propagating large data sets to the wider community with a view to integrated analysis.

Manuscript received June 26, 2008. (Write the date on which you submitted your paper for review.) This work was supported in part by the Sixth Framework Programme CASIMIR under Grant FP6-037811 (European Union).

J. M. Hancock is with the Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire OX11 0RD, U.K. (phone: +44 1235 841014; fax: +44 1235 841210; e-mail: j.hancock@har.mrc.ac.uk).

P. N. Schofield is with the Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY U.K. (e-mail: PS@mole.bio.cam.ac.uk).

C. Chandras is with the B.S.R.C. Alexander Fleming, Vari, Greece (e-mail: chandras@fleming.gr).

M. Zouberakis is with the B.S.R.C. Alexander Fleming, Vari, Greece (e-mail: zouberakis@fleming.gr).

V. Aidinis is with the B.S.R.C. Alexander Fleming, Vari, Greece (e-mail: v.aidinis@fleming.gr).

D. Smedley is with the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K. (e-mail: damian@ebi.ac.uk).

N. Rosenthal is with the Mouse Biology Unit, European Molecular Biology Laboratory (EMBL), 00016 Monterotondo, Italy (e-mail: rosenthal@embl-monterotondo.it)

K. Schughart is with Experimental Mouse Genetics, Helmholtz Centre for Infection Research, Germany (e-mail: klaus.schughart@helmholtz-hzi.de).

CASIMIR Partners: University of Cambridge, Cambridge, UK; MRC Harwell, Oxfordshire, UK; MRC, Edinburgh, UK; EBI, Hinxton, UK; EMBL, Monterotondo, Italy; BSRC Fleming, Vari, Greece; GSF National Research Center for Environment and Health, Neuherberg, Germany; Helmholtz-Zentrum fuer infektionsforschung GmbH, Braunschweig, Germany; CNR-Consiglio Nazionale delle Richerche-Instituto di Biologia Cellulare, Monterotondo, Italy; Geneservice Limited, Cambridge, UK

Over the last few years, the European Commission has supported an increasing number of functional genomics projects focusing on the use of the laboratory mouse as a model of human disease. The mouse has numerous advantages as a disease model including mammalian physiology and anatomy, short generation time and a well-developed genetic toolkit allowing, amongst other manipulations, knocking out and knocking in of genes, production of tissue specific knockouts, and production of point mutations [1].

Mouse projects funded by the European Commission encompass methods for mouse phenotyping (EUMORPHIA: http://www.eumorphia.org/), archiving and distributing mutant mouse lines (EMMA: http://www.emmanet.org), large scale phenotyping of mouse lines (EUMODIC: http://www.eumodic.org/), systematic generation of knockouts of a significant proportion of all mouse genes (EUCOMM: http://www.eucomm.org), mapping of gene expression domains in mouse embryos (EurExpress: http://www.eurexpress.org), development of mouse models to investigate human immunological disease (MUGEN: http://www.mugen-noe.org), a database of images of mouse pathology (PATHBASE: http://www.pathbase.net) and numerous others (see http://www.prime-eu.org/euromouseiiprojects.htm for a fuller listing of current and recent projects). The diversity of these projects is so great that the Commission has also funded Coordination Actions both to provide an overview of the activities of the various projects and to provide means for wider dissemination of the results. These have included PRIME (Priorities For Mouse Functional Genomics Research Across Europe: http://www.prime-eu.org/), a priority-setting organisation, and CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources: http://www.casimir.org.uk) which is aimed at recommending standards to allow data sharing and integration between the different projects.

CASIMIR is an important initiative because bioinformatics is often not given enough thought when projects of this kind are planned. As a consequence, there is a risk that data will not be stored or preserved in a form amenable to future use or integration into other data sets. This would result in a massive waste of resources.

Like other EU-funded initiatives, the work of CASIMIR is organised into a number of "Work Packages". The non-administrative work packages of CASIMIR deal with the

following areas, all of which are critical for the optimal integration of biological databases: data representation (in particular the use of ontologies to represent complex biological information), technical issues concerning database compatibility and interoperability, data acquisition, curation and ownership, the integration of biological collections and material resources into the data network, and user interactions.

As a result of discussion at the CASIMIR meeting in Corfu on October 3-6, 2007, the Data Representation Work Package has a particular interest in the representation of phenotype information using ontologies and how this might be linked to descriptions of human disease. The phenotype of an organism is "the observable properties of an organism that are produced by the interaction of the genotype and the environment". Phenotypic attributes take a wide variety of forms, ranging from simple measures such as body weight or life span through quantitative measurements such as blood glucose concentration to more subjective observations such as aggressiveness or nervousness. Mouse bioinformaticians, in common with bioinformaticians working on other organisms which can act as models of human diseases, face two major challenges. The first of these is how to represent the diverse features that fall under the general heading of phenotype using a single semantic formalism, and the second is designing a system that will allow the phenotypes of mouse lines to be related to diseases seen in humans.

There are currently two ontology schemes in use to describe mouse phenotype data: the Mammalian Phenotype ontology (MP) [2] and the combinatorial schema that uses the Quality Ontology (PATO) as its core component [3], [4]. Both of these approaches have their advantages and disadvantages and it seems likely that a single unifying framework will need to be developed which maps MP terms to PATO-style descriptions so that the two schemes become interchangeable (for example see http://www.mugen-noe.org/database where some such mappings are implemented). CASIMIR's particular interest is in the development of means to map mouse phenotypes to human diseases. Disease is a complex concept, and most diseases have a variety of features which can be considered to be phenotypic attributes comparable to those measured in mice. However disease may be diagnosed on the basis of the presence of only a subset of these phenotypes, although some may be essential. In order to effectively map mice showing a particular constellation of phenotypes to one or more human diseases it will therefore be necessary to produce descriptions of diseases in terms of their component phenotypes. This will require engagement of ontologists with clinicians with an interest in these issues. CASIMIR aims to stimulate this area of research by holding a meeting at the Nobel Forum in Stockholm in December 2008. This area will be addressed in more detail in the accompanying paper by Schofield *et al.* [5].

The development of an appropriate ontological framework will contribute greatly to improving the semantic interoperability between mouse databases. A related problem is to develop frameworks that improve syntactic interoperability using freely available tools that are relatively easy to install. To date the consortium has investigated the implementation of a set of approaches which, in combination, allow integration of a group of databases ranging from large core databases to relatively experimental ones. The solutions investigated are Web Services [6], BioMart [7], MOLGENIS [8] and TAVERNA [9]. BioMart allows joint querying of a set of databases by generating a denormalised schema for each database which can then be queried by the Mart software. BioMart also has a built in facility to generate Web Services for any given Mart. In principle any set of relational databases can be queried in this way, although this is less tractable for large and/or complex databases. Disadvantages of the approach include the need to re-generate the BioMart table(s) at intervals, meaning that information is not up-to-date for rapidly changing databases, and the lack of semantic mapping between fields. MOLGENIS creates software wrappers around existing databases enabling automated data access from R, Java and Web Services. Once data sources have been made interoperable some sort of client is then required to make the integrated querying possible. We have used TAVERNA, the MyGrid [10] workflow management system, to integrate databases through a mixture of Web Services, BioMart and MOLGENIS technologies to illustrate the potential for the use of these relatively straightforward technologies in integrating mouse databases.

A related task has been carried out by the "User Interaction" work package, which has developed use cases for the types of complex queries that biologists might wish to make using sets of databases. The aim of these use cases is to inform the design of interfaces, queries and analysis tools from the perspective of the end user. The group decided upon apparently relatively straightforward queries: "What is the function of genes X, Y & Z?", "Which information is available in various databases on these genes?" and "Does a group of selected genes exhibit common functional features?" In a first step, the use cases allow identification of proper gene IDs in various databases, using either the correct names or synonyms. A particular attribute of the current use case is that lists of genes can be generated, expanded and combined in a shopping cart fashion. The functional information on a given gene or a list of genes may then be retrieved from a number of databases. In addition, queries can be made that compile common features of groups of genes from various databases, e.g. expression patterns, and human or mouse diseases or phenotypes associated with particular genes.

Given the wide variety of databases available [11], a critical infrastructural issue for the more widespread use of web services in the integration of biological data is the

availability of information on the contents of databases and the services they provide. The work package on The Integration of Biological Collections And Material Resources Into The Data Network is therefore leading the development of a "database of databases" which is intended to provide this information across the domain of mouse functional genomics. A preliminary version of this database is currently available at http://bioit.fleming.gr/mrb. As part of the process of developing this database, CASIMIR has also discussed the development of a Minimum Information criterion for describing databases and benchmarking criteria for identifying areas of relative strength in a given database.

The final area of interest for CASIMIR is ensuring that the maximum amount of data of the best possible quality is placed in public databases. Data submission faces a number of barriers that limit the submission of data, such as perceptions concerning the consequences of database submission on intellectual property rights and patentability. The aims of the Work Package on Data Acquisition, Curation And Ownership are to:

- Examine current practice in existing databases regarding data quality assurance, traceability, provenance and reach a community consensus of best practice.
- Assess the range of curatorial practices and annotation strategies and their costs and compare the advantages and disadvantages of human expert curation and annotation with respect to the aims of the database
- Gather information concerning Intellectual Property Rights (IPR) concerns from the participants and other stakeholders, compare practices between different funding agencies, companies and institutions and conduct round table talks specifically aimed at bringing together representatives of all these groups to discuss a common approach to IPR constraints on data submission.
- Make recommendations as to how the community might be persuaded to contribute at least publicly funded data to public databases.
- Investigate the potential for public/private domains in large databases as a potential source of funding.

II. The CASIMIR Questionnaire

As a first step towards developing its recommendations, CASIMIR carried out a survey in summer 2007 to ascertain the sorts of database activities carried out by currently-active EC-funded mouse functional genomics projects and whether they are currently making use of community standards such as ontologies and minimum information standards for reporting experimental data. In the following sections we summarise the results of the survey and discuss their consequences in the context of integration of these large projects into the wider data network.

A. Methodology

The questions included in the questionnaire are shown in Table I. The questionnaire was circulated to a panel of recipients. As well as EC-funded projects, these included bioinformatics representatives of projects funded by Europe-wide institutions (the European Commission and European Molecular Biology Laboratory) as well as contacts in other databases, many in the USA, which act as a control group and give the results a broader perspective. Results were gathered using a custom web form accessible via the CASIMIR web site (http://www.casimir.org.uk). The list of EC-funded projects targeted and a detailed description of results can be found on the CASIMIR web site at http://www.casimir.org.uk/qresults.

TABLE 1.
QUESTIONNAIRE QUESTIONS

Question No.	Question
1	Are you using a relational database, object database or flat files?
2	If relational, what is your chosen RDBMS (Relational Database Management System)?
3	Is your database providing external links to other on-line resources; possibly via URL/HTTP (if yes please name them)?
4	Supported/Installed Web Services (if yes please name them)?
5	Please list the sorts of data entities you store (e.g. protein sequence data, mouse strain information etc...)
6	Can you provide a brief 'explanatory' description/schema of your data/data structure?
7	Are you willing to provide a entity relationship diagram and would you be willing to provide it under an open source license?
8	Are you currently using or do you intend to use any ontologies or controlled vocabularies to describe your data?
9	Do you plan to expand your use of ontologies in future?
10	Do you use OBO ontologies?
11	Do you perceive the need for additional ontologies to serve your domain of knowledge?
12	Do you make use of Minimum Information standards (such as MIAME for microarray experiments) to describe any data? If so, which ones? If you do not make use of these standards, are you likely to do so in future?
13	Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR?
14	What do you perceive as the main limiting factor in data representation/interoperability etc. in European bioinformatics databases?

B. Overview of Responses

28 responses were received, of which 11 were from the 13 targeted EC-funded projects (85% response rate). In the analysis the responses from the EC-funded projects were combined with responses from databases funded by the other pan-European funding agency, the EMBL, to give a broad picture of the state of European-funded databases. Detailed results are available from the CASIMIR web site.

The results suggest that in general European projects are

well-placed to respond to the challenges of integration but that some issues need to be addressed. Relatively few projects are relying on flat-file formats for storing data - most are using relational or object technology (Questions 1&2). In this they are consistent with practice on the non-European-funded projects that responded to the questionnaire. We asked if databases were willing to make their relational schemas publicly available (Question 7). Most were willing to do so but some were not. The main argument from those databases not willing to make their schemas public was that they did not wish to do so before publishing a journal article on their database, after which most were willing to publish their schemas. We therefore conclude that most databases operate in a spirit of openness. Most databases in the survey provide external links to data in other databases (Question 3), linking them into the wider data network at the level of the user of the web interface.

The range of data being stored in the databases we involved in the questionnaire, addressed by Question 5, is wide and covers most of the areas that are important in modern biology. Question 5 returned a wide variety of terms which indicate the wide spread of data types, from genomic and proteomic (DNA sequence, Protein sequence, Gene name, Gene structure, Protein feature, Gene/protein function, Transcript sequence, Gene regulation, other genome features); gene expression data (from gene expression arrays and in situ hybridization); systems biology information at the level of pathways, DNA-protein interaction and systems models; cell lines and chemical interventions applied to them; information on individual mice and mouse lines and strains, including breeding history, genetic manipulations applied to them genotype, phenotype and pathology data and information concerning the welfare regulatory regime under which they were kept; more complex data types such as images and their metadata and full descriptions and comparisons of ontologies; and information on researchers, publications and user requests.

An increasingly important route for making data accessible to external “power” users is the implementation of web services. Less than half of the EC-funded databases we involved currently had web services available (Question 4) although the proportion (44%) was higher than for the non-European Commission or EMBL-funded databases (25%). A significant proportion declared an intention to implement web services (31% for EC+EMBL-funded databases, 25% for the others) but a large group also declared no intention to do so (25% of EC+EMBL-funded projects and 50% of others). This may reflect an opinion that web services are of no obvious value to the users of a given database. This might change over time as more and more useful implementations making use of web services are demonstrated, for example the demonstration projects being developed by CASIMIR.

An essential element for developing the potential for applications that mine data across multiple databases is consistent nomenclature. In the biological sciences the development of domain-specific ontologies, particularly the Gene Ontology [12] has played an important role in widening the acceptance and use of consistent nomenclature in biological databases. Consistent nomenclature across databases demands use of the same core set of broadly accepted ontologies by all databases. The OBO foundry family of ontologies, which developed from the original GO concept, is intended to act as a set of consistent, broadly orthogonal ontologies for the biological sciences [13]. We therefore asked about the use of ontologies in our database set and whether they favoured OBO foundry ontologies. A majority of databases currently use ontologies to represent their data but a significant minority do not. Some (exclusively in this sample amongst the non-EC-funded databases) use in-house controlled vocabularies (CVs) rather than ontologies. When asked if they intended to expand their use of ontologies, the majority said yes but a few again said no indicating that there is a core of resistance to the use of ontologies. This may be because they are not seen to be necessary, or because some developers find them difficult to implement. In Question 10 we asked if databases made use of OBO ontologies. A slim majority did so, but a proportion did not and either developed their own or used some nomenclatures not part of the OBO “family”, such as NCBI Taxonomy. At least one respondee was unaware whether the ontologies they used were OBO ontologies. It would seem that a valuable way forward in this area would be the development of a forum involving OBO and other ontology providers that could work towards a self-consistent set of usable ontologies. Increased involvement with the user community (defined here as the database managers and programmers who might be expected to implement ontologies) may also be worthwhile.

In Question 11 we asked whether additional ontologies were needed to improve databases’ data representation. Some of the areas mentioned by responders are already the subject of ontology development - specifically phenotype, general anatomy and gene products (although the exact meaning of the latter response is unclear). The responses may reflect ignorance of what is available or dissatisfaction over lack of clarity or over-complexity in these areas.

The last specific area investigated by the questionnaire was the use of Minimum Information (MI) standards. MI standards define the information that needs to be collected to adequately describe specific types of high throughput, functional genomics experiment. The original example was MIAME for microarray-based gene expression experiments [14], but numerous standards are now under development by various communities, many under the auspices of the MIBBI (Minimum Information for Biological and Biomedical Investigations; mibbi.sourceforge.net/ [15]) consortium. Relatively few of the responding databases currently implemented MI standards - in nearly all cases this was MIAME although one implements MISFISHIE (Minimum

Information Specification For In Situ Hybridization and Immunohistochemistry Experiments) [16]. It is likely that the uptake of MI standards protocols will increase as they become available for more areas. As with all such computational tools, it will be important that these are easy to use as well as powerful.

Finally we asked two open questions, the aim being to elicit opinions on the most important areas in which development was needed to further database interoperability. Many of the areas mentioned in these responses also emerge in the discussion above. However a theme that clearly emerges is the need for overarching advisory bodies that can help individual database managers and programmers design their databases optimally for data integration, recommend on standards, and so on. Some technical needs were also raised, specifically a resource providing mappings between equivalent IDs that would enable mapping of data from different databases. Another technical suggestion was the establishment of a "database of databases" that could be automatically queried to provide information on issues such as accessibility of web services or usage of ontologies in a specific database. This has been acted on through Work Package 7 of CASIMIR.

III. Conclusions

An increasing number of large projects, generating high volumes of functional genomics data, are being established in Europe to exploit the mouse as a model of human disease [1]. It is crucial that the best use is made of these large data sets. To do this, it is essential that any large project of this kind establishes a database which can be integrated into the wider mouse data network. Since its initiation in February 2007 CASIMIR has played a significant role in catalysing the integration of mouse Functional Genomics and related data across Europe and worldwide. Many of its meetings and workshops have included participants from outside the EU, including the USA, Canada, Japan and Australia. As a Coordination Action, CASIMIR's main role is to promote interaction and develop policy. Many of the directions the project, and particularly the data representation work package, has taken have been informed by the questionnaire carried out during Summer 2007 and described in this paper. The questionnaire was designed to investigate the current state of the art in European-funded projects, to identify strengths and weaknesses, and to drive further discussions under the auspices of CASIMIR, leading to a set of recommendations on how to facilitate the data integration process. Any such process should be compatible with developments world-wide, where in the US (through projects such as caBIG [17]), Japan (through a new initiative to integrate all RIKEN's biological databases [18]) and Australia (http://www.ncris.dest.gov.au/capabilities/integrated_biological_systems.htm) major data integration initiatives are being established.

It is clear from the questionnaire results that considerable progress is being made towards better integration of mouse data but that there are some areas where more work needs to be done, notably in the further development of some standard tools such as ontologies, minimum information check-lists and a database registry, but also in demonstrating the utility and ease-of-use of currently available tools for database integration. Aims for the second half of the project include publishing the results of these initial discussions and producing recommendations to the European Commission on how large-scale European projects should develop data storage solutions and the importance of bioinformatics in such projects.

In recent years the "bottom-up" approach to developing standards through community consensus has proved to be the most effective way of establishing usable data standards and resources, such as ontologies tailor-made to the needs of that community. Global adoption will only happen if standards are easy to apply and meet the current and projected requirements of the community. Projects such as CASIMIR and the Gene Ontology can act as forums for the generation of community consensus and represent an important social integration of the resources and expertise within the biological community. It is hopefully through initiatives like this we can move to a seamless data network in the life sciences with all the power that will bring.

Acknowledgment

The authors thank CASIMIR (funded by the European Commission under contract number LSHG-CT-2006-037811) for financial support.

References

[1] N. Rosenthal and S. Brown, "The mouse ascending: perspectives for human-disease models," *Nat Cell Biol*, vol. 9, pp. 993-999, 2007.

[2] C. L. Smith, C. A. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information," *Genome Biol*, vol. 6, p. R7, 2005.

[3] T. Beck, A.-M. Mallon, H. Morgan, A. Blake, and J. M. Hancock, "Using ontologies to annotate large-scale mouse phenotype data," *BMC Bioinformatics*, vol. Accepted for Publication, 2008.

[4] G. V. Gkoutos, E. C. J. Green, A.-M. Mallon, J. M. Hancock, and D. Davidson, "Using ontologies to describe mouse phenotypes," *Genome Biol*, vol. 6, p. R8, 2005.

[5] P. N. Schofield, G. V. Gkoutos, J. Sundberg, J. M. Hancock, The CASIMIR Consortium "One Medicine: Integrating mouse and human disease phenotypes", *8th IEEE International Conference on Bioinformatics and Bioengineering*, to be published.

[6] The World Wide Web Consortium, "Web Service Activity," http://www.w3.org/2002/ws, 2002.

[7] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney, "EnsMart: A Generic System for Fast and Flexible Access to Biological Data," Genome Res, vol. 14, pp. 160-169, 2004.

[8] M. A. Swertz, E. O. De Brock, S. A. Van Hijum, A. De Jong, G. Buist, R. J. Baerends, J. Kok, O. P. Kuipers, and R. C. Jansen, "Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases," Bioinformatics, vol. 20, pp. 2075-2083, 2004.

[9] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," Nucleic Acids Res, vol. 34, pp. W729-W732, 2006.

[10] R. D. Stevens, A. J. Robinson, and C. A. Goble, "myGrid: personalised bioinformatics on the information grid," *Bioinformatics*, vol. 19, pp. i302-i304, 2003.

[11] J. M. Hancock and A.-M. Mallon, "Phenobabelomics--mouse phenotype data resources," *Brief Funct Genomic Proteomic*, vol. 6, pp. 292-301, 2007.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nat.Genet.*, vol. 25, pp. 25-29, 2000.

[13] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, OBI_Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, pp. 1251-1255, 2007.

[14] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME) - towards standards for microarray data," *Nat Genet*, vol. 29(4), pp. 365-371, 2001.

[15] C. F. Taylor, "Standards for reporting bioscience data: a forward look," *Drug Discov Today*, vol. 12, pp. 527-533, 2007.

[16] E. W. Deutsch, C. A. Ball, G. S. Bova, A. Brazma, R. E. Bumgarner, D. Campbell, H. C. Causton, J. Christiansen, D. Davidson, L. J. Eichner, Y. A. Goo, S. Grimmond, T. Henrich, M. H. Johnson, M. Korb, J. C. Mills, A. Oudes, H. E. Parkinson, L. E. Pascal, J. Quackenbush, M. Ramialison, M. Ringwald, S. A. Sansone, G. Sherlock, C. J. J. Stoeckert, J. Swedlow, R. C. Taylor, L. Walashek, Y. Zhou, A. Y. Liu, and L. D. True, "Development of the Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)," *OMICS*, vol. 10, pp. 205-208, 2006.

[17] S. Oster, S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. Covitz, K. Shanbhag, I. Foster, and J. Saltz, "caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research," in *J Am Med Inform Assoc*, 2007.

[18] T. Toyoda and A. Wada, "Omic space: coordinate-based integration and analysis of genomic phenomic interactions," *Bioinformatics*, vol. 20, pp. 1759-1765, 2004.