Mamm Genome (2012) 23:550–558
DOI 10.1007/s00335-012-9408-0

BioGPS and GXD: mouse gene expression data—the benefits and challenges of data integration

Martin Ringwald · Chunlei Wu · Andrew I. Su

Received: 9 April 2012 / Accepted: 21 June 2012 / Published online: 31 July 2012

Abstract Mouse gene expression data are complex and voluminous. To maximize the utility of these data, they must be made readily accessible through databases, and those resources need to place the expression data in the larger biological context. Here we describe two community resources that approach these problems in different but complementary ways: BioGPS and the Mouse Gene Expression Database (GXD). BioGPS connects its large and homogeneous microarray gene expression reference data sets via plugins with a heterogeneous collection of external gene centric resources, thus casting a wide but loose net. GXD acquires different types of expression data from many sources and integrates these data tightly with other types of data in the Mouse Genome Informatics (MGI) resource, with a strong emphasis on consistency checks and manual curation. We describe and contrast the "loose" and "tight" data integration strategies employed by BioGPS and GXD, respectively, and discuss the challenges and benefits of data integration. BioGPS is freely available at http://biogps.org. GXD is freely available through the MGI web site (www.informatics.jax.org) or directly at www.informatics.jax.org/expression.shtml.

M. Ringwald (✉)
The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA
e-mail: ringwald@informatics.jax.org

C. Wu · A. I. Su (✉)
The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA
e-mail: asu@scripps.edu

Introduction

Microarray and more recently RNA-Seq technology have revolutionized the field of gene expression analysis and enabled researchers to systematically interrogate gene expression levels on a genome scale. Many groups have used these high-throughput technologies to interrogate gene expression across large-scale reference data sets. For example, several groups have independently profiled expression patterns in diverse anatomic tissues in human, mouse, rat, and pig (Hornshoj et al. 2007; Son et al. 2005; Su et al. 2002; Walker et al. 2004). These reference compendia have proven to be very useful in biomedical research to infer biological roles for otherwise uncharacterized genes.

However, microarray and RNA-Seq experiments reveal only part of the expression profile for a given gene. Other assay types provide complementary insights into expression patterns at the RNA and protein level. For example, Northern and Western blot analyses reveal the number and sizes of transcripts and proteins, and RNA in situ hybridization and immunohistochemistry provide detailed spatial expression information with a potential resolution at the cellular level. Furthermore, expression information from mouse mutants can reveal important information about molecular mechanisms of differentiation and disease. Therefore, it is necessary to integrate data from different types of expression assays, to capture expression information from wild-type and mutant mice, and to combine these data with pertinent genetic and phenotypic information.

Reference gene expression data are also useful in the context of other online biomedical resources. There are hundreds, if not thousands, of other gene-centric sites for mouse as well as orthologs in other organisms that present

diverse data on, for example, genomic variation, protein interactions, literature mining, genetic models, and many other aspects of gene function. Integrating expression data with this broader landscape of genomic resources is also critical and potentially quite powerful.

In this article we explore the themes described above and discuss the challenges and benefits of data integration by focusing on two complementary database resources: BioGPS and the Gene Expression Database (GXD).

BioGPS

BioGPS (and its precursors) were initially created as a mechanism to disseminate one reference gene expression data set online. Several dozen diverse mouse tissues were profiled to characterize the normal transcriptome in a data set commonly referred to as the "Gene Atlas." This data set was based on high-throughput gene expression arrays starting with the commercial mouse U74A and human U95A Affymetrix arrays (Su et al. 2002), followed by custom whole-genome human (GNF1H) and mouse (GNF1M) arrays based on Celera predictions (Su et al. 2004), and finally updated using the commercial MOE430 Affymetrix array (Lattin et al. 2008).

Reference gene expression data sets like the Gene Atlas have served as invaluable resources for the analysis of genome-scale profiling experiments. To ensure as wide and broad utilization of these data as possible, the native Affymetrix probe set identifiers were mapped to commonly used gene identifiers and gene annotations from sources including NCBI Gene, Ensembl, Gene Ontology, and UniProt. This basic mapping enables a simple gene-centric online query interface to query and access these data.

Researchers who generate genome-scale data are often faced with a diverse list of candidate genes and the goal of translating that list into a testable hypothesis. Not surprisingly then, users of these initial systems often sought to go beyond the few sets of resources that were directly integrated within the system. While the Gene Atlas reference gene expression data serves as one useful resource for inferring gene function, there are hundreds, if not thousands, of other online gene-centric resources that provide valuable information for researchers on their gene or genes of interest. Data integration among all these online resources is a key challenge in bioinformatics and biomedical research. After devoting significant effort to integrating data from a few key resources directly into the site, it quickly became apparent that this method would not scale with the size of the small developer team.

As one approach to addressing the challenge of data integration, we developed a gene portal called BioGPS (http://biogps.org) (Wu et al. 2009). BioGPS leverages the principle of crowdsourcing as a mechanism for identifying and aggregating useful gene-centric resources. The architecture behind BioGPS is conceptually quite simple. The vast majority of online gene-centric databases use one of a few gene identifiers as their "primary key." For example, NCBI Gene, Ensembl, RefSeq, and UniProt identifiers are among the most commonly used for online biological databases. When a user chooses to view the web page for a specific gene or protein of interest, these identifiers are often passed between the browser and the server as a parameter in the website's URL.

BioGPS generalizes this pattern by defining a "URL Template" for every gene-centric resource available in the BioGPS plugin library, where the specific gene identifier is replaced by a variable for the identifier type. When a BioGPS user views a certain gene, the appropriate gene identifier is retrieved and the URL template is rendered to a working URL. As an example, the website for the International Gene Trap Consortium (IGTC) is at http://www.genetrap.org/. The IGTC page for the mouse gene *Cdk2* (NCBI Gene ID: 12566) is located at http://www.genetrap.org/cgi-bin/annotation.py?entrez=12566. Therefore, the generic URL template for the IGTC BioGPS plugin is expressed as http://www.genetrap.org/cgi-bin/annotation.py?entrez={{EntrezGene}}. Using this simple system for wrapping existing online gene-centric databases, the BioGPS plugin library currently contains over 250 publicly shared plugins. This library spans a diverse collection of protein, genetics, literature, pathway, and expression resources, and in aggregate provides an expansive index of gene-specific web pages.

By wrapping these gene-centric resources within plugins using this simple URL template system, BioGPS offers two distinct benefits to users.

First, BioGPS embraces the concept of crowdsourcing by encouraging direct contributions from the community of users. Most directly, community members can *explicitly* contribute by registering new BioGPS plugins. Anyone who has a BioGPS account has permission to create a new plugin, and over 75 people have contributed one or more plugins in BioGPS. Resources that are accessible through BioGPS plugins span a wide range of categories, including model organism databases, pathway resources, reagent providers, literature mining, genetics resources, and expression databases. BioGPS effectively crowdsources the creation and maintenance of its gene-centric plugin library.

BioGPS also accepts *implicit* contributions from the community by aggregating usage data across all users. For example, BioGPS can easily rank all plugins according to their popularity among fellow users, thereby providing a community view of plugin utility. When a user searches the BioGPS plugin library for a specific keyword, BioGPS sorts the matching plugins according to popularity (and, by

extension, utility). For example, when searching for "splicing" plugins, BioGPS identifies the Alternative Splicing Gallery (Leipzig et al. 2004) as the most relevant resource by community consensus, differentiating it from the other splicing plugins with less community adoption.

Second, the BioGPS interface is flexible enough to tailor the BioGPS gene report to each individual user. Since geneticists have different use cases than protein biochemists and systems biologists, it is natural that each user community would be interested in a different collection of BioGPS plugins. BioGPS enables users to individually define which data sources should be aggregated into their personalized gene report "layout" (Fig. 1). Each layout is defined by a collection and positioning of BioGPS plugins.

While BioGPS offers a high degree of user customizability, its data *aggregation* capabilities should be distinguished from true data *integration*. The simple and lightweight plugin interface allows for easy plugin registration and plugin rendering, but the content of each plugin window is essentially treated as a black box. Researchers can look up the information displayed in each window but they cannot perform queries across the different resources. To truly enable data integration, plugin content should be parsed (or presented by the plugin content provider) with semantically encoded data. This additional step of data structure would allow users to combine pieces of data from multiple plugin sources into a single analysis or visualization.

Richer data integration is one of the emphases for BioGPS moving forward. As an integration platform in its current state, BioGPS encompasses almost the entire landscape of gene-centric online databases. Its main disadvantage, however, is that the connections that can be drawn between those resources are relatively weak. We refer to this approach as "loose data integration." In the next section we explore a second case study focusing on "tight data integration," which has complementary advantages and disadvantages.

The mouse Gene Expression Database (GXD)

Both BioGPS and GXD are built on a foundation of providing gene expression data to the community, but beyond that basic shared goal these two projects diverge. BioGPS provides expression data for only a handful of data sets on a single technology (Affymetrix microarrays), focusing on linking to a heterogeneous collection of external gene-centric resources. In contrast, GXD stores different types of expression data from many different sources and focuses on a deep integration with genetic, functional, and phenotypic information for the laboratory mouse as an integral part of the Mouse Genome Informatics (MGI) resource (Eppig et al. 2012; Finger et al. 2011; Smith et al. 2007b).

Database scope and current data content

GXD currently includes data from RNA in situ hybridization, in situ knock-in reporter, immunohistochemistry, Northern blot, Western blot, and RT-PCR experiments.

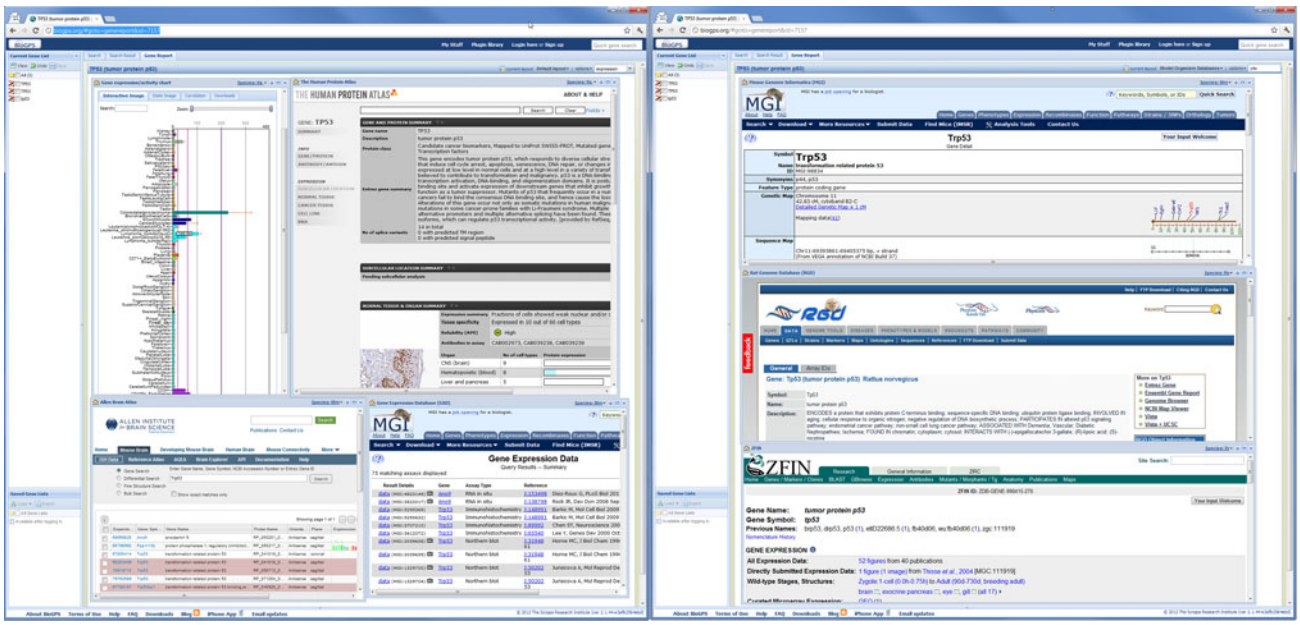

Fig. 1 Example BioGPS layouts: *left focused* on expression resources, *right focused* on model organism databases

Focusing on endogenous gene expression, the database covers all developmental stages and expression data from wild-type and mutant mice. Data are acquired from the literature, from electronic data submissions, and through collaboration with projects that generate expression data at a large scale. All these data are carefully reviewed by scientific curators and annotated using standard nomenclature and ontologies to ensure proper data integration and data maintenance. Wherever possible, database records are associated with images showing the primary expression results. New data are being added to GXD on a daily basis. At this point, GXD has indexed over 19,000 publications with respect to their expression data content. The database holds over 223,300 images and more than 1.1 million annotated expression results from over 57,000 assays for more than 13,000 genes. This includes expression data from over 1,600 mouse mutants. As part of the larger Mouse Genome Informatics (MGI) resource, GXD combines its expression data with other genetic, functional, phenotypic, and disease-oriented data. Due to the deep data integration, users can search for expression data and images in many different ways using a variety of biologically and biomedically relevant parameters.

Data curation, integration, and maintenance

GXD integrates expression data at many points. A typical database record and some of the salient integration points are illustrated in Fig. 2. In general, data integration involves much more than merely collecting data from different sources. It requires the identification of common objects and the proper assembly of these objects into a larger framework in which the links between objects are correct and properly maintained.

Gene objects are a good example in this regard. Different publications can refer to the same gene using different names, or they can use the same name to refer to different genes. Therefore, an essential curatorial task is to determine which gene was studied in a given expression assay. One way to identify the gene is to analyze the probe information. Thus, the curation and maintenance of correct probe-to-gene associations is another crucial data annotation and integration task. This applies to data curation from the literature as well as to data obtained from large-scale projects. For example, efforts which performed RNA in situ hybridization screens for thousands of genes, such as EurExpress (Diez-Roux et al., 2011), GenePaint (Visel et al. 2004), GUDMAP (McMahon et al. 2008), and BGEM (Magdaleno et al. 2006), developed probes for each gene at the start of their projects. Several years had passed until these large-scale efforts were completed and their data became available to GXD. Because the information about genes and gene models had evolved during that time, GXD reanalyzed all probe-to-gene associations for these data sets to ensure that they are correct and up to date. Furthermore, correct probe-to-gene associations will be maintained as we move from one genome build to the next. In this way, the expression information will remain associated with the correct genes and properly connected with all the other rapidly accumulating data about these genes.

Data integration at the probe level and determination of whether different assays from different sources used the same nucleotide or antibody probe are also important. For expression data from mutant mice, it is essential to assign the data to the correct mutant allele. This integrates all expression data for a given mutant (and distinguishes them from wild-type data). One of the most important and challenging curation tasks is the standardized recording of expression patterns. Expression patterns are annotated using extensive anatomical ontologies that list the anatomical entities for each developmental stage in a hierarchical way (Bard et al. 1998; Hayamizu et al. 2005). This allows a standardized description of expression patterns and the integration of expression results from assays with differing spatial resolution.

Even integration at the image level is useful, as illustrated by the following examples:

(1) As originally proposed for the collaboration between GXD and the Edinburgh Mouse Atlas Project (Ringwald et al. 1994), GXD makes all its data and annotations available to EMAGE (Richardson et al. 2010) so that pertinent in situ and immunohistochemistry data can be mapped into the 3D atlas. Images in GXD have pointers to the corresponding spatially mapped images in EMAGE. (2) Images in GXD link to corresponding entries in EurExpress and GenePaint where users can take advantage of the high-resolution images and zoom capabilities provided by these sites. Several research groups have analyzed high-resolution images from GenePaint and published papers based on these studies. GXD integrates these data with the data provided by GenePaint.

The integration points described above are also required for combining expression data with other types of data in MGI and for maintaining additional links to external resources. Gene objects are major hubs in this regard, providing access to information such as chromosomal location, Gene Ontology (GO) annotations, protein structure data, links to orthologous genes from other species, and, by transitivity, to data for these genes in external resources. Allele objects and standardized anatomical entities integrate expression data with genetic, phenotypic, and disease information. Cross references between anatomical ontologies from different species have been and are being established (Haendel et al. 2009; Hayamizu et al. 2012) to enable the comparative analysis of data that

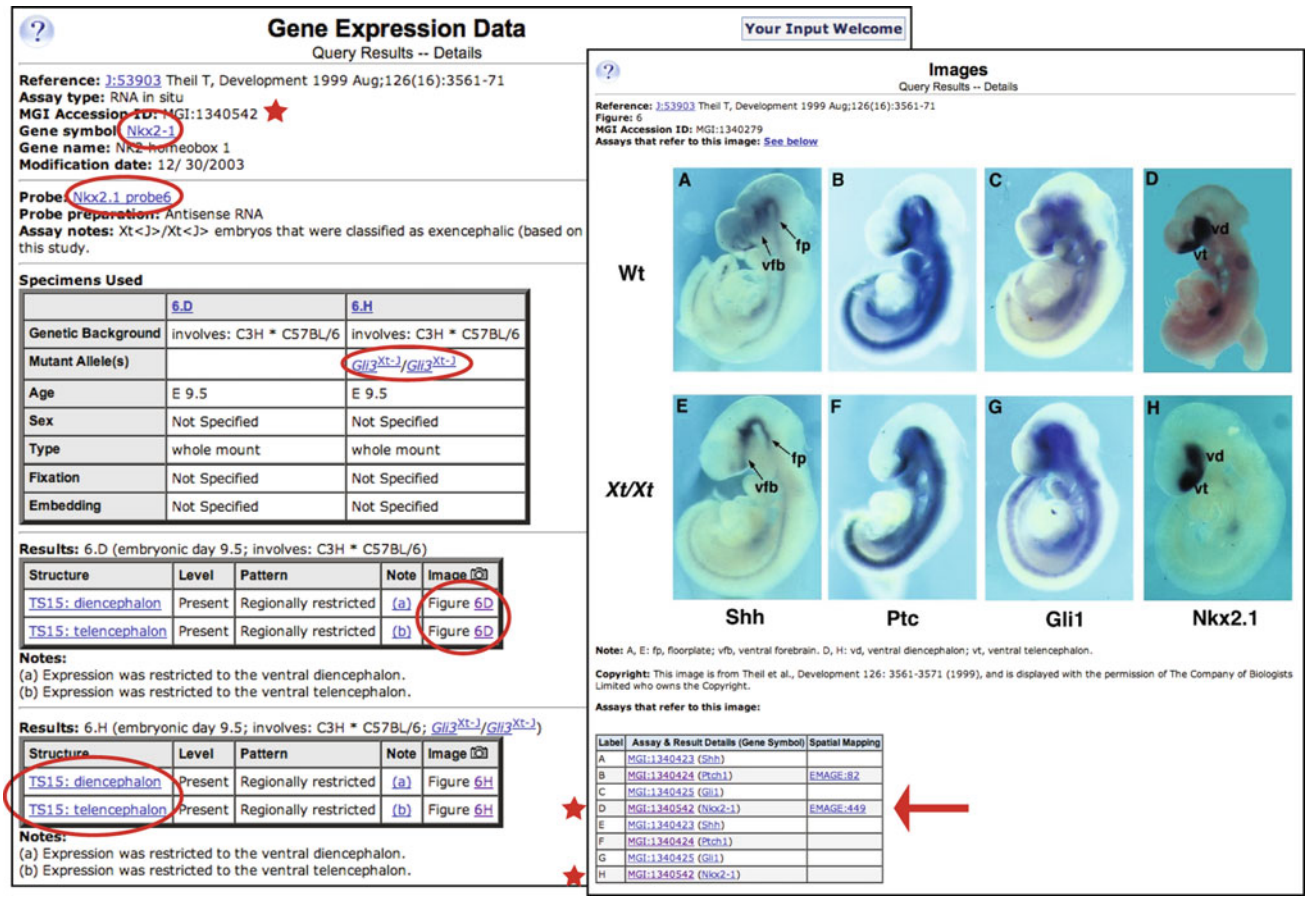

Fig. 2 GXD assay record (*left*) and associated image page (*right*). A RNA in situ hybridization experiment for *Nkx2-1* is shown. Two specimens were analyzed, one from a wild-type mouse and one from a *Gli3* mutant mouse. Results are recorded for each specimen and links to the corresponding image data are provided. *Red circles* and *stars* indicate some of the salient points of data integration: genes, probes, assays, mutant alleles, anatomical structures, and images. To preserve context, the image page (*right*) displays the whole figure of the publication. The table at the *bottom* provides links to the assay records for each individual image pane. As indicated by *stars*, *panes D* and *H* are annotated in the assay record shown on the *left*. For images that have been spatially mapped into the Edinburgh Mouse Atlas, the table provides a link to the corresponding entry in EMAGE (*arrow*)

pertain to anatomy such as gene expression, phenotypic, and pathological information.

Search capabilities

Mammalian organisms and their associated research data are very complex. Therefore, one main objective of data integration is to enable powerful search capabilities. Figure 3 illustrates one of GXD's query forms and some of the search capabilities made possible through the data curation and integration work. Users can search for expression data for specific genes or for sets of genes such as genes located in a specified genomic interval or genes whose products perform a given molecular function. They can search for genes that have been detected (or not detected) at particular developmental stages and/or in specific anatomical structures. They can filter for data from specific types of assays. They can search for expression data from mice that have been mutated in a particular gene. Furthermore, they can combine any of the search parameters mentioned above and thus easily perform queries that are quite complex but highly relevant for biomedical research. For example, a search for all genes, or for all transcription factors, located within a specific chromosomal region and found to be expressed in a given tissue at a particular developmental stage could be very helpful in identifying disease candidate genes that have been mapped to a particular genetic interval. Based on the integrated data representation in GXD and MGI, query capabilities will continue to be expanded and enhanced to support contemporary research. For example, MGI's "Cre Portal" encodes reporter patterns using the same anatomical terms by which expression data are being recorded, and the phenotypic information in MGI will be integrated via anatomical ontology terms as well. This will permit the development of tools that allow researchers to correlate spatiotemporal allele, expression,

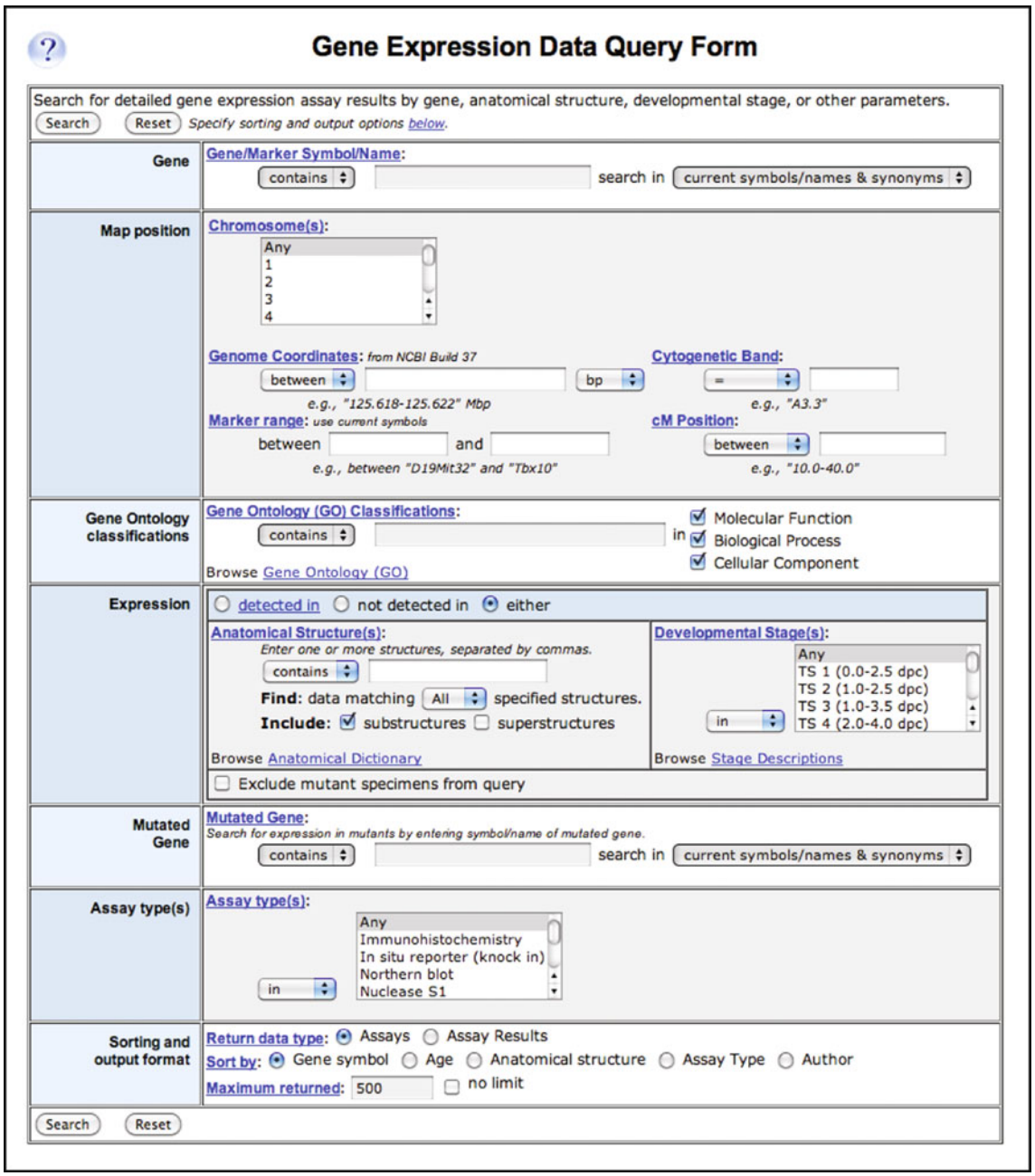

Fig. 3 The Gene Expression Data Query form illustrates some of GXD's search utilities. Researchers can use many different search parameters and combinations to perform concise queries for expression data

and phenotype patterns. Such tools will be essential for the analysis of conditional mutants.

The GXD BioMart is another tool to query GXD data. It provides rapid access to gene expression results. Search returns can be customized and the results can be downloaded for further analysis. Importantly, the GXD BioMart can be interconnected with other BioMarts to enable queries across the combined resources. For example, in the context of the International Knockout Mouse Consortium (IKMC) Web Portal (Oakley et al. 2011; Ringwald et al. 2011), expression data can be combined with data from the IKMC to search for embryonic stem (ES) cells in which genes have been targeted that are expressed at a particular stage and/or in a specific tissue (http://www.knockoutmouse.org/martsearch). BioMart technology provides an easy way to combine data from different resources via shared database objects. In the example above, data from the IKMC and GXD BioMart are combined via gene objects (using MGI gene IDs). It is important to note that the federated query works only because the targeted ES cell data and expression data are annotated to the same gene objects, i.e., because these integration points have already been established. Furthermore,

the correct pairing of ES cells and expression data depends entirely on the ES cell-to-gene and gene-to-expression data associations provided by the IKMC and GXD BioMarts. This is true for the BioMart approach to data integration in general: it requires existing integration hubs, and the quality of search returns is entirely dependent on correct data annotations within the resources it combines.

In this context, it is worth noting that the core data of BioGPS, the microarray reference data sets, could also be exposed via a BioMart which then can be interconnected with the other BioMarts, including the ones provided by GXD, EMAGE (Stevenson et al. 2011), and EurExpress (Oakley et al. 2011).

Discussion

Mouse expression data provide important insights into the molecular mechanisms of differentiation, health, and human disease. The importance of integrating these expression data and placing them in a larger biological context cannot be overstated. The rate of biomedical research is increasing at an explosive pace, and consequently the proportion of all biological knowledge that is known by any single individual is shrinking. Increasingly, our individual ability to interpret and analyze data will need to be supplemented by informatics tools to aggregate and integrate data from many different sources.

BioGPS and GXD pursue different but complementary approaches to data integration. They represent examples at each end of the data aggregation–integration spectrum. Whereas the BioGPS plugin approach casts a very wide net, starting out with aggregated data views and proceeding to data aggregation (referred to above as 'loose" integration), GXD pursues true ("tight") data integration within a smaller data domain.

Both approaches have advantages and disadvantages. GXD's approach requires strong efforts in data curation, quality controls, and data maintenance for each data set, and significant software development work for the incorporation and representation of new types of data. The payoffs are access to data from the published literature that would not be readily available without GXD's curation effort, full integration of heterogeneous data from many sources, and computability, enabling complex search capabilities and unified views of data and search results. In contrast, the BioGPS plugin model employs a very simple and easy framework for adding new resources, employing a crowdsourcing model for extending BioGPS that does not require any centralized developer effort. However, while this approach is strong in terms of scalability and adaptability, it achieves only "loose" data integration with limited query and integration capabilities.

Of course, it would be ideal if one could achieve a high level of data integration (like GXD) with ease and breadth (like BioGPS). However, that is simply not possible. There is no free lunch. Complex data require concise and complex query capabilities. As pointed out above, approaches such as BioGPS and BioMart rely on existing integration points and on the proper maintenance of the resources with which they are interconnected. The question then is: what could make the task of data integration easier?

Some of the answers to this question are surprisingly simple. Authors are most knowledgeable about their experiments and thus in the best position to report their data in a reasonably complete, accurate, and more standardized way. While less and less primary data can be published in research papers, they can be submitted to pertinent public databases, possibly in conjunction with journal publications (submitters would receive accession numbers that can be cited in publications). Such mechanisms are already in place. For example, array expression data can be submitted to NCBI-GEO (Barrett et al. 2011) or ArrayExpress (Parkinson et al. 2011) using standardized data formats such as MIAME and MAGE (Brazma et al. 2001; Rayner et al. 2006). Likewise, GXD is accepting electronic submissions for the types of expression data it collects, see http://www.informatics.jax.org/mgihome/GXD/GEN/gxd_submission_guidelines.shtml. Ideally, the meta-data in these submissions would also be annotated in a standardized way through the use of established controlled vocabularies and ontologies (Smith et al. 2007a). Indeed, the use of biomedical ontologies for structuring biological knowledge and data is expanding in popularity and adoption (Musen et al. 2012). While electronic submissions would still require some review by database curators and potentially some mapping to data objects and ontology terms, they would significantly speed up the process of data acquisition and integration.

An even better approach would be to generate data in a systematic and standardized way. The BioGPS "Gene Atlas" expression data set is a good example in this regard. Other examples include the EurExpress, GenePaint, Allen Brain Atlas (Lein et al. 2007), and ENCODE (ENCODE Project Consortium 2011) projects, as well as the EUMODIC (Gates et al. 2011) and KOMP2 projects that employ standard operating procedures for the generation and analysis of mouse phenotyping data.

These projects illustrate the importance and economy of large, homogeneous baseline data sets. However, it will not be feasible to generate similar data sets for every mouse mutant, let alone for every conditional mutant. Much valuable data will continue to be generated by conventional laboratories that study specific biological systems in depth. Thus, it will remain important to integrate expression data (or other types of data) from many different sources.

Electronic data submissions in conjunction with publications, as described above, would strongly facilitate the acquisition and integration of data from conventional laboratories.

However, even under the best of circumstances, there are limits to what information can and should be integrated. Furthermore, biomedical research and its experimental methods are evolving rapidly. Therefore, the quick and adaptive BioGPS approach will remain very helpful. It will benefit and can become more sophisticated as more data become available in a highly integrated format.

Finally, it is worth considering that the future solution for data integration among biological resources may involve broader initiatives to facilitate data integration online. According to the W3C, the primary organization for defining web standards: "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries." This mission of the Semantic Web echoes the goals of data integration in biology. Although there are many technical and sociological challenges to adoption within the biological community, several initiatives [e.g., RDFScape (Splendiani 2008) and Bio2rdf (Belleau et al. 2008)] have begun to explore this approach to data integration.

Since there is no clear consensus within the biological community (and perhaps since the diversity of needs among community members precludes consensus), we believe that the bioinformatics community will and should continue to explore a wide variety of data integration approaches.

Acknowledgments The authors thank Drs. Joel Richardson, Constance Smith, and Benjamin Good for their helpful comments and discussions on the manuscript. The authors also thank all the members of the GXD and BioGPS teams for their dedicated work, as well as the members of other MGI projects for their contributions to GXD and to the larger MGI Resource. The authors acknowledge support from the National Institute of General Medical Sciences (GM083924 to AIS) and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD062499 to MR).

References

Bard JB, Kaufman MH, Dubreuil C, Brune RM, Burger A, Baldock RA, Davidson DR (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. Mech Dev 74:111–120

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall AK, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 39:D1005–D1010

Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 41(5):706–716

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365–371

Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S et al (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol 9(1):e1000582

ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9(4):e1001046

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, the Mouse Genome Database Group (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res 40(1):D881–D886

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2011) The mouse Gene Expression Database (GXD): 2011 update. Nucleic Acids Res 39(1):D835–D841

Gates H, Mallon AM, Brown SD, EUMODIC Consortium (2011) High-throughput mouse phenotyping. Methods 53(4):394–404

Haendel MA, Gkoutos GV, Lewis SE, Mungall CJ (2009) Uberon: towards a comprehensive multi-species anatomy ontology. Presented at the International Conference on Biomedical Ontology (ICBO), 26 July 2009. Available at Nature Precedings, http://precedings.nature.com/documents/3592/version/1

Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M (2005) The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biol 6:R29

Hayamizu TF, de Coronado S, Fragoso G, Sioutos N, Kadin JA, Ringwald M (2012) The mouse-human anatomy ontology mapping project. Database (Oxford) 2012:bar066

Hornshoj H, Conley LN, Hedegaard J, Sorensen P, Panitz F, Bendixen C (2007) Microarray expression profiles of 20,000 genes across 23 healthy porcine tissues. PLoS One 11:e1203

Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Ck Glass, Hume DA, Kellie S, Sweet MJ (2008) Expression analysis of G protein-coupled receptors in mouse macrophages. Immunome Res 4:5

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ et al (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445:168–176

Leipzig J, Pevzner P, Heber S (2004) The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. Nucleic Acids Res 32:3977–3983

Magdaleno S, Jensen P, Brumwell CL, Seal A, Lehman K, Asbury A, Cheung T, Cornelius T, Batten DM, Eden C, Norland SM, Rice DS, Dosooye N, Shakya S, Mehta P, Curran T (2006) BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. PLoS Biol 4:e86

McMahon AP, Aronow BJ, Davidson DR, Davies JA, Gaido KW, Grimmond S, Lessard JL, Little MH, Potter SS, Wilder EL, Zhang P, GUDMAP Project (2008) GUDMAP: the genitourinary developmental molecular anatomy project. J Am Soc Nephrol 19:667–671

Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B, NCBO team (2012) The National Center for Biomedical Ontology. J Am Med Inform Assoc 19:190–195

Oakley DJ, Iyer V, Skarnes WC, Smedley D (2011) BioMart as an integration solution for the International Knockout Mouse Consortium. Database (Oxford) 2011:bar028

Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E,

Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 39:D1002–D1004

Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Irizarry RA, Liu J, Maier DS, Miller M, Petersen K, Quackenbush J, Sherlock G, Christian J, Stoeckert CJ, White J, Whetzel PL, Wymore F, Parkinson H, Sarkans U, Catherine A, Ball CA, Brazma A (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7:489

Richardson L, Venkataraman S, Stevenson P, Yang Y, Burton N, Rao J, Fisher M, Baldock RA, Davidson DR, Christiansen JH (2010) EMAGE mouse embryo spatial gene expression database: 2010 update. Nucleic Acids Res 38:D703–D709

Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, Richardson JE, Nadeau JH, Davidson D (1994) A database for mouse development. Science 265:2033–2034

Ringwald M, Iyer V, Mason JC, Stone KR, Tadepally HD, Kadin JA, Bult CJ, Eppig JT, Oakley DJ, Briois S, Stupka E, Maselli V, Smedley D, Liu S, Hansen J, Baldock R, Hicks GG, Skarnes WC (2011) The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. Nucleic Acids Res 39(1):D849–D855

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007a) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255

Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2007b) The mouse Gene Expression Database (GXD): 2007 update. Nucleic Acids Res 35:D618–D623

Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N, Khan J (2005) Database of mRNA gene expression of profiles of multiple human organs. Genome Res 15(3):443–450

Splendiani A (2008) RDFScape: Semantic Web meets systems biology. BMC Bioinformatics 9(Suppl 4):S6

Stevenson P, Richardson L, Venkataraman S, Yang Y, Baldock R (2011) The BioMart interface to the eMouseAtlas gene expression database EMAGE. Database (Oxford) 2011:bar029

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465–4470

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067

Visel A, Thaller C, Eichele G (2004) GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acids Res 32:D552–D556

Walker JR, Su AI, Self DW, Hogenesch JB, Lapp H, Maier R, Hoyer D, Bilbe G (2004) Applications of a rat multiple tissue gene expression data set. Genome Res 14(4):742–749

Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, Su AI (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10:R130