From data to function: Functional modeling of poultry genomics data[1]

F. M. McCarthy*†[2] and E. Lyons†‡§

*Department of Veterinary Science and Microbiology, †BIO5 Institute, and ‡Department of Plant Sciences, University of Arizona, Tucson 85721; and §iPlant Collaborative, Tucson, AZ 85721

ABSTRACT One of the challenges of functional genomics is to create a better understanding of the biological system being studied so that the data produced are leveraged to provide gains for agriculture, human health, and the environment. Functional modeling enables researchers to make sense of these data as it reframes a long list of genes or gene products (mRNA, ncRNA, and proteins) by grouping based upon function, be it individual molecular functions or interactions between these molecules or broader biological processes, including metabolic and signaling pathways. However, poultry researchers have been hampered by a lack of functional annotation data, tools, and training to use these data and tools. Moreover, this lack is becoming more critical as new sequencing technologies enable us to generate data not only for an increasingly diverse range of species but also individual genomes and populations of individuals. We discuss the impact of these new sequencing technologies on poultry research, with a specific focus on what functional modeling resources are available for poultry researchers. We also describe key strategies for researchers who wish to functionally model their own data, providing background information about functional modeling approaches, the data and tools to support these approaches, and the strengths and limitations of each. Specifically, we describe methods for functional analysis using Gene Ontology (GO) functional summaries, functional enrichment analysis, and pathways and network modeling. As annotation efforts begin to provide the fundamental data that underpin poultry functional modeling (such as improved gene identification, standardized gene nomenclature, temporal and spatial expression data and gene product function), tool developers are incorporating these data into new and existing tools that are used for functional modeling, and cyberinfrastructure is being developed to provide the necessary extendibility and scalability for storing and analyzing these data. This process will support the efforts of poultry researchers to make sense of their functional genomics data sets, and we provide here a starting point for researchers who wish to take advantage of these tools.

Key words: gene expression, microarray, data analysis, sequence analysis

2013 Poultry Science 92:2519–2529
http://dx.doi.org/10.3382/ps.2012-02808

INTRODUCTION

The advent of "omics" technologies enabled researchers to undertake genome-wide surveys of gene expression (e.g., using microarrays, proteomics, or Serial Analysis of Gene Expression). However, these approaches also result in long lists of differentially expressed genes that do not per se provide useful information about the biological system being studied. Instead, researchers must rely on biological modeling to understand how these gene expression lists provide insights into their biological systems (Cordero et al., 2007; McCarthy et al., 2007a). Because biological modeling relies on annotation and analysis tools, many model organisms developed Model Organism Databases that provide centralized reference gene sets with standardized gene nomenclature and functional annotation to support comparative and functional modeling on a genome-wide scale (Baxevanis, 2011). This in turn provides the core data used by bioinformatics tool developers who develop tools for functional modeling of gene expression data sets.

The development of next generation sequencing techniques for transcription profiling democratized functional genomics studies by enabling researchers to study an even broader range of species instead of focusing on species that have well-defined microarray platforms available. However, this same technique puts further pressure on the development of annotations and functional modeling tools that can support a much larger range of species and predict functions for novel genes identified by this same technique. As a result, while the

Received September 27, 2012.
Accepted October 16, 2012.
[1]Presented as part of the Experimental Design for Poultry Production and Genomics Research Symposium at the Poultry Science Association's annual meeting in Athens, Georgia, July 12, 2012.
[2]Corresponding author: fionamcc@email.arizona.edu

gap between data and knowledge is closing in several well-studied species, the need for fundamental annotation to support functional modeling in a broad range of species is critical.

This change in the way gene expression studies are done is borne out by a cursory examination of the type of gene expression data submitted to gene expression repositories. Currently (August, 2012) the National Center for Biotechnology Information Gene Expression Omnibus Database (Barrett and Edgar, 2006) and its European partner the ArrayExpress Archive (Brazma et al., 2006) together contain 5,427 avian gene expression data sets. These data sets are predominantly based upon microarray platforms, mostly from chicken but also from turkey and zebra finch. However, an inspection of these records reveals that next generation sequencing data sets are available for avian species such as Northern bobwhite quail, house finch, duck, and rock doves. Moreover, the Genome 10K Project is already sequencing more than 50 additional bird genomes (Genome 10K Community of Scientists, 2009), an effort that will provide reference genomes for future bird gene expression studies. The recent announcement that this initiative completed sequencing of the white goose genome (http://www.genomics.cn/en/news/show_news?nid=98853) means that there is a representative genome available for most major poultry genomes: chicken, quail, turkey, duck, and goose. Although the chicken genome remains the best studied and annotated bird genome, information from each bird genome has the power to inform each of the other genomes through comparative genomic analyses (Dalloul et al., 2010; Warren et al., 2010). If researchers are to leverage these sequence data into information concerning key production traits for poultry, it is necessary that they are able to translate their functional genomic data sets into information about function so that they can produce gains for agriculture and human nutrition.

The following sections describe key aspects of annotation and functional modeling with a particular focus on resources that support functional modeling of poultry data sets. Each section will provide background information about the type of data and tools as well as strategies for incorporating these data into a functional model. However, it is important to stress that there is no one method for functional modeling but rather this must be informed by the biological system itself and the type of biological questions that the system lends itself to studying. Moreover, no amount of functional modeling can overcome deficiencies in experimental design. Other manuscripts published as part of the 2012 Poultry Science Association Symposium Experimental Design for Poultry Production and Genomics Research addressed aspects of experimental design and analysis and should also be considered, as appropriate. Instead, this manuscript should be viewed as a starting primer for modeling of functional genomics data sets, with further iterations and insights driven by existing research knowledge.

Functional Modeling

The key concept underlying functional modeling is that biologists are better able to conceptualize and understand the dynamics of complex biological systems if they can move from long, differentially expressed gene lists to larger concepts such as biological processes (e.g., development) or pathways (e.g., specific metabolic and signaling events). This approach enables the researchers to group their 3,000 differentially expressed transcripts from an array (or more, from RNASeq) into 10 to 30 functional categories. Reducing the number of elements that are relevant to common molecular functions, physiological processes, and pathways resolves the data into biological components that the researcher is familiar with and also avoids confusion caused by nomenclature of genes and their gene products. For example, whereas some researchers may know the key genes involved in specific processes, it is unlikely that they are familiar with several thousand gene names that are produced as part of a differentially expressed data set. Similarly, this abstraction helps navigate across different species where gene nomenclature is not consistent. This problem of understanding a large number of genes and their gene products is further complicated in poultry (as it is with all agricultural species) because there is no standard way of naming genes or their subsequent gene products, and names may vary between database, publication, and research groups. More recently, the establishment of an international Chicken Gene Nomenclature Committee has begun the work of assigning standardized gene nomenclature (Burt et al., 2009); however, this work has not yet affected legacy gene array annotation files, nor is it extended to include other poultry and avian species.

If researchers are to move from molecular and "omic" studies to understanding the phenotypes of their systems and how they is affected by changes to the environment, researchers must be able to readily translate gene sets into physiological processes. Functional modeling of large data sets enables data to be restructured and reexamined in biological terms that make sense to the researcher. Functional modeling of larger, "omics" data sets is underpinned by fundamental annotation data and requires bioinformatics tools that automate the analysis using these data and visualize the results in a way that makes sense to the biologist. We stress that these data and tools cannot replace the biological knowledge of the researcher—they can only assist her or him to view it in a way that enables a deeper understanding of the system being studied. In addition, before describing these general approaches to functional modeling, it is first important to consider the underlying data used in functional modeling and the strengths

and limitations for modeling poultry data sets using existing annotations.

Data to Support Functional Modeling

One of the current limitations facing all but a handful of model organisms is a lack of sufficient underlying annotation data to support functional modeling. This is a 2-fold problem. First, there are a large number of gene products (including those that are clearly identified) that are not well characterized, and we do not know anything about their function. These gene products also include novel (newly identified) and computationally predicted genes (often lineage specific). Information about their function requires a mixture of bioinformatic analysis and development of high-throughput functional assays are required to keep pace with the discovery of novel genes (Roberts, 2004). Second, there is a further subset of gene products where elements of their function have been studied, but these data are not readily accessible. These genes require annotation efforts to extract the available information that is known about these genes and their gene products. In addition to these 2 limitations, we must also consider that advances in functional genomics technologies (such as the development of RNASeq expression analysis and increased sensitivity of proteomic detection) require new analysis tools and the requisite computational resources and knowledge to use these tools. As a preface to discussing functional modeling approaches, we will briefly discuss the opportunities and limitations of current efforts to provide data and resources to support functional modeling in poultry.

Cyberinfrastructure to Support High-Throughput Analysis: iAnimal. iAnimal (http://genepro.cshl.edu/ianimal/) is a concept project based on the cyberinfrastructure and web portal of the iPlant Collaborative. Briefly, the iPlant Discovery Environment (Goff et al., 2011) is a web-based portal that enables researchers to manage their data, share it among collaborators, analyze it using different computing resources (as appropriate), and integrate and share new analysis tools. Although most of the data management and analytical applications are agnostic with regard to the organismal source of the data, each research community has specific needs. The iAnimal portal is designed to provide the animal sciences with animal-specific resources based on iPlant cyberinfrastructure. RNASeq analysis is an excellent example of why cyberinfrastructure, such as iAnimal, is needed by research community. A typical RNASeq workflow requires moving large sets of raw sequencing reads, cleaning them to remove poor-quality reads, digital normalization to remove redundant data, possibly converting sequence formats (e.g., sanger fastq encoding vs. Illumina fastq encoding), aligning the reads to a reference genome, viewing the alignments, and identifying differential expression between different experiences. Such a workflow is usually run from the command-line step by step, and may require substantial computing resources to process. Having a prebuilt workflow that is assessable from the web, through which it is easy to also upload data for processing, means that more time can be spent analyzing the results and inferring biological meaning than trying to figure out how to properly format a command line argument and locating the necessary computing resources. National shared cyberinfrastructure means that computing specialists maintain both the hardware and software, and knowledge specialists help support and train the research community. In addition, such cyberinfrastructure may facilitate and foster collaborations by providing the means for people to easily share data, algorithm, knowledge, expertise, and experience.

An example of how the cyberinfrastructure of iPlant/iAnimal is fundamentally changing the scientific workflow is their data store. This system is based on iRODS software (http://www.irods.org) and creates a single data-storage resource accessible by any platform with Internet access. One of its main features from a user's standpoint is high-performance parallel file transfer that permits reliable, resilient, and high-throughput transfer of very large data files. Our tests (Table 1; http://tinyurl.com/8u85xnb) show that moving data between academic institutions connected through backbone of Internet2 may be faster than the traditional method of mailing hard drives for large sets of data.

Table 1. iPlant Data Store transfer speeds between the University of California–Berkeley and University of Arizona (UA)[1]

Source	Copy method	Time to transfer 1 G of data (s)
CD	cp	320
Berkeley server	scp	150
External hard drive	cp	36
USB2.0 flash	cp	30
iPlant Data Store (iDS)	Berkeley iput to iDS/iget to UA	18
Second internal hard drive	cp	15

[1]These tests were performed during regular academic working hours on December 8, 2011. All data were transferred to the internal hard drive (7.2 kB) of a MacPro located at the University of Arizona. Where possible, 100 GB of data were copied between devices. The copy methods are Unix/Linux commands for "copy" (cp) and "secure copy" (scp).

Currently the Discovery Environment contains the Ontologizer Gene Ontology (**GO**) term enrichment analysis tool (Grossmann et al., 2007), which launches from the Ontologizer website and allows users to add their own GO annotations. Other applications of the Discovery Environment include tools to support analysis of next generation sequencing alignment and analysis, phylogenetics, QTL, and genome-wide association studies, sequence alignments, and motif identification.

Comparative Genomics in Avians. As researchers move to larger and more complex data sets, it is necessary that they are provided with tools to assist them with making sense of these data so that they are able to provide gains for agriculture. While next generation sequencing platforms are driving cyberinfrastructure requirements for genome assembly and gene identification, the need for functional annotation to support modeling of gene expression and phenotypic data sets is also increasing. Not only are we now able to work on a much more diverse range of species (all of which require functional annotation), but we are also rapidly identifying novel gene elements (that have no functional information) and we are working with more complex data sets (e.g., integrating gene expression, metabolomics, SNP analysis, and phenotypic data sets).

One way to rapidly provide more functional data is leveraging what is already known from other organisms. The key to mapping genetic information between genomes is identifying orthologous sets of genes. One platform that has tools for doing those analyses quickly and easily is CoGe (http://genomevolution.org). CoGe is a web-based platform focused on providing a rich set of tools for managing, analyzing, and comparing genomic data across all domains of life (Lyons and Freeling, 2008). The avian genomics community and CoGe development group are working together to provide a comprehensive set of avian, reptile, and related vertebrate genomes for use in comparative genomics. This includes managing multiple versions of various genomes including alternative assemblies and various sequence maskings, prebuilt data sets of orthologous genes, which include links for downloading sequences and analyzing neighboring genomic regions for conserved noncoding sequence and ultraconserved elements, links for rapidly building phylogenetic trees, and other tools for rapidly finding and extracting data of interest. Whereas data and tools are essential for the discovery process, so are training and collaborative tools. The avian genomics and CoGe groups are developing a set of tutorials focused specifically on avian comparative genomics, as well as leveraging iPlant's (Goff et al., 2011) online community forums to create a space where avian researchers can get help with their various analyses by posting questions that are answered by other community members. In addition, CoGe is developing support for storing and visualizing expression data in a comparative genomics context, and will work closely with the avian community members to ensure that its tools are compatible with and best serve their research needs.

Annotation and Bio-Ontologies. Genomic annotation involves both the identification and demarcation of functional elements within the genome (structural annotation) and associating functional descriptions with these elements (functional annotation). The traditional paradigm for doing genomic annotation is based on genomic sequencing done by centralized genome sequencing centers that use their established genome annotation pipelines. Typically, structural annotation is done as part of the final assembly stages, whereas functional annotation is not always included as part of this same process, nor is there a single, standard procedure to add functional information. If, as part of structural annotation, genes are related to homologous or orthologous genes in related species, then initial gene nomenclature and functional annotation may also be assigned on this basis. Another common approach for assigning gene function is to analyze genes or their gene products for conserved functional motifs and domains, which also provides a good "first pass" basis for functional annotation. However, this traditional model of genome annotation is changing as new sequencing technologies move the role of genomic annotation into the domain of smaller research groups, which may not have the capacity or expertise to provide comprehensive and up-to-date genomic annotation. Democratization of genome sequencing, along with reduced research funding, is also seeing a shift away from central databases to distributed systems that can continue to update and refine genome annotation after the initial sequencing effort is complete.

It was a combination of the lack of any standardized method for functional annotation and the subsequent inability to share functional information between existing model organism databases that led to the development of the GO for functional annotation (Lewis, 2005). Ontologies are structured, controlled vocabularies that not only define the terms but also the relationships between these terms. For example, the GO defines gene product function in terms of Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). A single term (e.g., "GO:0006917 induction of apoptosis") has a definition ("a process that directly activates any of the steps required for cell death by apoptosis") but is also related to other functional terms; for example, "GO:0042267 natural killer cell mediated cytotoxicity" is a more specific type of "GO:0006917 induction of apoptosis." Ontologies are used in biology to ensure that annotation is done consistently between different groups, thereby promoting data sharing. Additionally, the defined structure of the ontology enables rapid computational analysis of ontology data, systematically traversing the structure using the relationships between terms. The ability of the GO to describe function in a way that is computationally tractable resulted in a rapid uptake of the GO by the bioinformatics community as they developed tools to do functional modeling. The development of tools that use the GO data and its usefulness for modeling large

functional genomics data sets in turn drove efforts to provide the underlying GO annotation data to support an increasing number of species. Rhee et al. (2008) provided an excellent review of the use of the GO, highlighting key points of which a researcher using the GO should be aware.

However, whereas the GO is considered the premier bio-ontology, it is not the only bio-ontology. The National Center of Biomedical Bio-ontologies reports 324 bio-ontologies in its BioPortal interface (http://bioportal.bioontology.org/) and the Open Biological and Biomedical Ontologies (http://www.obofoundry.org/) lists 112 ontologies. Several among these are important to genomic annotation and functional modeling. The Sequence Ontology is designed to describe functional and physical features of genomic sequence (e.g., genes, CpG islands, QTL) and to facilitate comparative analyses of genomes (e.g., comparing repeat elements; Mungall et al., 2011). Other ontologies describe specific functional aspects, such as the Pathway Ontology and the Molecular Interaction ontology. Because many genes can only be functionally characterized based on expression patterns, and the number of novel genes is increasing with the application of next-generation sequencing technologies, ontologies that describe cell and tissue expression (e.g., Cell Ontology and the BRENDA Tissue Ontology) are also increasingly useful for functional genomics analysis. There are several different anatomy ontologies, and development of a chicken anatomy ontology is underway by researchers at the AgBase databases and the Roslin Institute. The development of multiple ontologies to support annotation and analysis of phenotypic data (Mungall et al., 2010) is also underway, and these projects are likely to enable large-scale and comparative analysis of phenotypic traits across species in the near future.

However, despite the availability or recent development of bio-ontologies, not all of these ontologies are as frequently used as the GO. Whereas the GO was one of the first bio-ontologies to be developed, there has also been a continual effort to annotate data to the GO, ensuring that data exists for analysis tools. Moreover, there is a 2-fold and complementary approach toward annotation to the GO. In the first approach, manual biocuration of the published papers provides detailed, species-specific functional data. Although this is necessarily time-consuming and costly, it provides a core of "gold standard" annotations that is used to test computational annotation tools and to develop analysis tools that use the GO data. These same data can also be transferred to other species where there is known functional orthology. The second approach is to develop computational pipelines for providing species-independent "first-pass" functional annotation data for a large number of gene products. This approach rapidly provides breadth of annotation (i.e., most gene products have at least some GO annotation), but this annotation lacks species-specific, detailed functional information. Despite these limitations of computationally derived GO, the ability to rapidly add at least a first-pass functional annotation for new species is one of the reasons that the GO is widely used for analysis of an increasing number of species. As data acquisition increases due to next generation sequencing and the use of other "omics" approaches, there is a critical need to develop similar high-throughput annotation pipelines for other data types and other bio-ontologies.

Currently chicken is the only poultry species with a manual GO annotation effort (McCarthy et al., 2007b); computationally derived GO annotations are provided by both the European Bioinformatics Institute Gene Ontology Annotation (EBI GOA) Project (Dimmer et al., 2012) and by curators at AgBase (McCarthy et al., 2011; Table 2). To date (September 2012) there are 327,528 GO annotations for 67,735 chicken gene products and 62.6% of these annotations are computationally derived. In comparison, the next most commonly studied poultry species, turkey, has 99,429 GO annotations for 12,970 gene products and 99.9% of these are computationally derived, with the small remainder being opportunistic annotation of turkey gene products identified while annotating related gene products from other species. This lack of functional data for most poultry species inhibits modeling of functional genomics data sets, and the lack of dedicated literature annotation means that many of the annotations produced will lack detail. However, recent work at AgBase to expand functional annotation to other agriculturally

Table 2. Summary of Gene Ontology (GO) annotation for chicken gene products[1]

Source	No. of GO annotations	No. of gene products
Manual GO annotation, GOA	3,169	563
Manual GO annotation, AgBase	121,161	40,272
Computational GO annotation, GOA	79,033	16,691
Computational GO annotation, AgBase	204,969	39,291
Total[2]	327,528	67,735

[1]The GO annotations for chicken were provided by AgBase and the EBI GOA Project. The GO annotation numbers are current as at September 1, 2012.

[2]Note that totals are not additive as gene products are annotated using both manual and computational approaches.

important species will provide some targeted manual biocuration for turkey gene products over the next few years.

Whereas sequence analysis provides one rapid and broad-based source of functional annotations, another approach is to transfer annotations between functional orthologs (Gaudet et al., 2011). Using this approach, it is possible to transfer the more detailed, manually derived GO annotations to other poultry species where it is possible to identify a clear 1:1 ortholog with genes in other bird species. This same approach is used to transfer information about pathways (Moriya et al., 2007), protein interactions (Yu et al., 2004; Huang et al., 2007), and gene nomenclature (Burt et al., 2009). However this method presupposes 2 things: first, that functional orthologs are identified, and second, that there is a core set of annotations in the reference species that can be transferred to the second species. A limitation to this approach is that it is not always easy to identify true functional orthologs in newly assembled genomes; typically, reciprocal sequence matches (using BLAST or BLAT) are used, which does not take into account synteny and may not always identify true orthologs from paralogs. Methods to more accurately and easily identify orthologs in newly sequenced genomes would contribute to both functional and structural annotation. In addition, developing a core set of high-quality annotations in a reference species (such as chicken), would enable these data to be transferred to other poultry species where a functional ortholog is identified. However, in practice, functional information about an ortholog is not limited to a single species; chicken genes that do not have any functional information may be well studied in turkey or pigeon, and identifying the best sources of annotation to support poultry functional modeling may mean focusing on other bird species.

Annotation is a continual process. Just as reannotating a genome will result in new gene sets and new exon-intron boundary annotations, functional annotation is also continually improving, updating, and renewing. New functional information is added, rules for assigning computational annotations are continually reviewed, and obsolete data are removed (Rhee et al., 2008). For example, reannotation of the FHCRC Chicken 13K cDNA v.2.0 microarray (GPL 1836) with updated GO annotations changed the GO functions that were identified as being differentially expressed during *Salmonella enterica* infection of chickens (van den Berg et al., 2010). Thus, it is important to know the source of the annotation data used in modeling (where it was obtained from and when), to ensure that the latest and most up-to-date data are used.

Functional Modeling Strategies and Analysis Tools

Typical approaches for functional modeling of large data sets include examining GO summaries to ascertain the overall function, GO Enrichment Analysis to determine which functions are statistically enriched in a data set, pathways analysis, and network analysis. These approaches are complementary, visualizing the same data in different ways that are only partly overlapping, and we will discuss each approach below.

GO Summary Using Slim Sets. The initial output of a statistical analysis to identify differentially expressed genes is often a daunting list of genes or gene products. For arrays, this may be a list of several thousand transcripts, whereas for proteomics and RNASeq data sets this number is often orders of magnitude larger. A useful first step is to summarize this list by grouping the gene products into functional categories. As of August 2012 there are 38,120 GO terms (ontology version 1.3493), so clearly using the GO can result in a large number of functional categories. Instead, GO Slim sets are truncated forms of the GO—terms that are manually picked to represent broader biological categories—that may be used for summarizing function (Rhee et al., 2008). There are currently 6 different GO Slim sets (Table 3) that have been developed by different groups, and 3 of these are specifically designed for bacteria, plants, and yeast. A fourth Slim set may be useful in some specialized instances but has not been updated since 2002 and is not recommended for general use. Among the remaining 2 GO Slim sets that are useful for summarizing poultry data, there are significant differences in the number of GO terms used and therefore the results will be remarkably different. The Generic GO Slim set (developed by the GO Consortium) has a total of 127 terms, whereas the PIR GO Slim set (developed by Darren Natale from the Protein Information Resource) contains 464 GO terms. It is strongly recommended that researchers investigate the use of both GO Slim sets to see which terms may best suit their experimental conditions (e.g., metabolism or immune function or development terms).

AgBase provides the GOSlimViewer tool (McCarthy et al., 2007b), an online tool that enables researchers to upload a truncated GO annotation set for their gene list and summarize it using a selected GO Slim set. The results and presented as a tab-separated table that may be charted in any format using standard Excel functions. Summarized functions are presented as Molecular Function, Biological Process, and Cellular Component, and these may be charted independently or combined. AgBase also provides a file that outlines the GO SlimViewer accession details for each of the GO Slim terms. This file can be used to identify the gene products summarized to each of the GO summary terms. This enables the researcher not only to summarize function but to retrieve the genes categorized into any one of these functions.

More recently a tool for summarizing GO annotations for specific gene sets is REVIGO (Supek et al., 2011). This tool is markedly different from GO Slim summaries in that it uses a clustering algorithm to develop a representative subset of GO terms and may result in more informative functional terms. In addi-

Table 3. Properties of Gene Ontology (GO) Slim sets[1]

GO Slim set	Developed by	Date updated	No. GO terms
Generic	GO Consortium	September 2012	148 total terms 70 BP terms 43 MF terms 35 CC terms
GOA Whole Proteome	N. Mulder, M. Pruess (EBI GOA)	November 2002	62 total terms 23 BP terms 27 MF terms 12 CC terms
PIR	Darren Natale (PIR)	September 2012	464 total terms 204 BP terms 75 MF terms 185 CC terms
Plant Slim	TAIR	August 2012	100 total terms 46 BP terms 27 MF terms 27 CC terms
Yeast Slim	SGD	September 2012	167 total terms 100 BP terms 43 MF terms 24 CC terms
TIGR Prokaryote	Michelle Gwinn-Giglio (UMD)	August 2009	202 total terms 202 BP terms

[1]The GO Slim sets are used to summarize GO function to broader terms in the ontologies. Several Slim sets are available, and they are shown here along with the number of GO terms each contains (BP = biological process, MF = molecular function, CC = cellular component). Developers include biocurators from the GO Consortium, EBI GOA, the Protein Information Resource (PIR), The *Arabidopsis* Information Resource (TAIR), Saccharomyces Genome Database (SGD), and University of Maryland (UMD).

tion, the resulting GO term sets are visualized in multiple ways, including downloadable tables and different graphic displays.

GO Enrichment Analysis. The GO enrichment analysis uses statistical analyses to identify GO terms that are enriched in a functional data set relative to a background set. For example, researchers may want to identify functions overrepresented in a treatment compared with a control set, or in sick compared with healthy data sets. This differs from GO Summary using Slim set because a term may represent a large proportion of the summarized function, but if it is also commonly found in the background set, then it will not be considered enriched. This type of analysis accounts for the fact that not all GO terms are equally represented; for example, GO:0008152 metabolic process and GO:0005488 binding are more commonly found because they are general terms well represented by computational mapping.

There are 3 main approaches to doing functional enrichment (Huang et al., 2009a). The first is singular enrichment analysis (**SEA**), which was initially established by older enrichment analysis tools and is proven to be effective. In SEA enrichment, a differentially expressed set of genes is compared with a background. Examples of enrichment tools that use SEA are GOStat (Beissbarth and Speed, 2004), Onto-tools (Khatri et al., 2007), and DAVID (Huang et al., 2009b). In contrast, gene set enrichment analysis—sometimes referred to as also Parametric Analysis of Gene Set Enrichment or PAGE—examines all genes in a set (e.g., all genes in an array) and their expression values and calculates enrichment against a randomly shuffled background. DAVID and AgriGO (Du et al., 2010) enable the user to do gene set enrichment analysis. This latter approach avoids arbitrary expression cut-off values and allows for minimally changing genes but has the disadvantage of sometimes producing very long GO term lists. One method for dealing with long lists of enriched functions is to cluster these functions based upon semantic similarity or, for ontologies terms, based on their relationships. Although not many tools currently support poultry analyses, REVIGO (Supek et al., 2011) allows users to input a simple list of GO terms and enrichment values from other tools.

A seemingly bewildering array of GO Enrichment Analysis tools is available, and the GO Consortium website provides a large but not comprehensive list of these tools. However, only a fraction of these enrichment tools are able to analyze poultry data, and the shorter list of relevant enrichment tools is available as part of the AgBase Educational Resources section (Table 4); every effort is made to keep this list up to date. The more commonly used GO Enrichment tools among this list include Database for Annotation, Visualization, and Integrated Discovery (**DAVID**; Huang et al., 2009b), AgriGO (Du et al., 2010), GOStat (Beissbarth and Speed, 2004), and Onto-Tools package (Khatri et al., 2007). The list of tools that can be used to analyze poultry functional genomics data sets can be expanded by considering enrichment tools that allow users to upload their own GO data [e.g., GOStat, Funcassociate 2.0 (Berriz et al., 2009), Onto-Tools]. These allow researchers to analyze nontraditional species but also enable researchers to add their own GO annotations and include these data in their analysis.

Although the ability to analyze data from the species of interest is a primary consideration for selection

Table 4. Gene Ontology (GO) enrichment analysis tools that support poultry data analysis[1]

Enrichment tool	Platform	Species	Upload your own GO
AgriGO	Online	Chicken	Yes
BiNGO	Cytoscape plugin	Wide range	Yes
CLASSIFI	Online	Wide range	No
DAVID	Online	Wide range of species	No
FuncAssociate	Online	Chicken	Yes
GENECODIS	Online	Chicken	No
GeneMerge	Online	Chicken	No
GFINDer: Genome Function	Online	Many Affy arrays	No
GOEAST	Online	Includes Affy, Agilent, Illumina arrays	No
GOstat	Online	All UniProt species	Yes
GraphWeb	Online	Chicken	No
Onto-Compare	Online (login)	Wide range	Yes
Onto-Express	Online (login)	Wide range	Yes
Ontologizer	Webstart/desktop	Limited	Yes

[1]Whereas many enrichment tools support analysis of chicken data, other tools allow researchers to upload their own data sets for analysis.

of a GO Enrichment Analysis tool, many other pragmatic factors contribute to the selection of the analysis tool. For example, many of the tools mentioned here are available as web-based interfaces. However, the Onto-Tools package has an online login and uses a Java plugin that may not be popular for those who want a simpler interface. The tools use different methods for mapping accessions to GO annotations and are on different update cycles, which affects how much functional annotation is included in the analysis and may influence the choice of tool. Critically, tools should have a prefiltering step to remove any GO annotation with a qualifier (e.g., "NOT" annotations, for review, see Rhee et al., 2008); however, this is not always easy for a user to determine. Another feature that is popular is the option to analyze multiple data types; for example, DAVID combines GO, pathways, and interaction analyses.

Another consideration is that all of the enrichment tools developed to date are based upon microarray data sets. Current algorithms do not handle biological replicates very well; that is, correcting for multiple testing of differential expression in RNASeq (e.g., using the R package DESeq) creates very small differentially expressed data sets. Development and application of programs that account for RNASeq biases is critical. Likewise, there are inherent biases in tools for modeling RNASeq expression data that are not considered in existing tools based upon microarray platforms (e.g., GO terms that have a higher than average number of shorter transcripts are more likely to be falsely overrepresented). To our knowledge, there is only one enrichment tool that accounts for these biases (Young et al., 2010), although this tool is not currently available as a web-based interface.

Pathways and Network Analysis. Whereas GO analyses focus on individual and larger biological processes, pathways and network analyses describe different aspects of functional annotation. It is worthwhile to clarify that the GO Biological Process terms include some pathways but not all; in addition, it includes other processes that are not pathways (e.g., development, immune function, and so on). Therefore, although a GO enrichment analysis may include some pathways, this may not be a comprehensive list.

The main pathways databases are Reactome (Croft et al., 2011) and KEGG (Tanabe and Kanehisa, 2012), and most commercial pathways analysis tools use data from both of these databases. Notably, almost all pathways data for poultry are based upon identifying orthologs from other species involved in pathways and almost all of these data are transferred automatically without manual review. The exception is Reactome, which has a manual biocuration effort for chicken pathways (Gillespie et al., 2011). This means that pathways data will be as good as our ability to clearly identify orthologs between poultry and mammalian species. Further, these data will not account for species-specific variation to generic pathways. The Reactome database has an online tool that allows researchers to analyze pathways enrichment from gene expression data and there are commercially available tools that also use pathways data from public databases and do pathways enrichment analyses. Several of the newer commercial pathways analysis packages combine GO and pathways analysis. Another option is the freely available Pathways Express software from the Onto-Tools (Khatri et al., 2007) suite of programs.

Whereas all pathways are networks, not all networks are pathways. Network analysis looks at the interactions between elements in a gene set using molecular interaction; whereas pathways can be considered as a series of interactions, they are also directional and produce a clear outcome or products. Networks analysis (at least for protein-protein interactions) is not directional and can point to key gene products in the biological system and predict the effects of perturbation of this system. Network analysis relies on molecular interaction data and currently for poultry most of these data are transferred from other species based upon orthology or homology rather than direct experimental evidence. Interaction data may be obtained from a large number of molecular interaction databases, but more

recently the International Molecular Exchange (IMEx) consortium is working to produce standard file formats for data sharing across these databases and developing standard annotation procedures (Orchard et al., 2012).

The most commonly used, freely available network analysis software is Cytoscape, (http://www.cytoscape.org/), which also visualizes pathway data. Cytoscape is used to find active subnetworks/pathway within an expression data set, and this software has many additional plugins (or "apps") available to do network visualization, GO analysis clustering, and so forth. Commercial software for functional analysis of gene expression data will also often include network analysis.

A special case instance of network analysis that relates to many poultry gene expression studies is analysis of host-pathogen interactions using network analysis. This is critical for many disease-related studies in poultry but is hindered by lack of data to support these analyses. Although specific molecular interaction data are sparse for poultry species, even fewer data exist for host-pathogen molecular interactions and most of these data focuses on human-related pathogens. The Host Pathogen Interaction Database specifically integrates experimental interaction data from several public databases into a single, nonredundant web-accessible resource and does contain the available chicken-pathogen interaction data (Kumar and Nanduri, 2010).

Concluding Thoughts on Functional Modeling

The final part in the process of functional modeling is bringing together the information obtained from the different modeling strategies to create a coherent whole. Because this process relies on a detailed understanding of both the biological system being studied and the experimental design, it cannot be programmed into any functional analysis tool but rather relies on the researcher's biological knowledge. However, some considerations may be helpful. First, functional modeling is both complementary and iterative. This means that information gained using any one modeling approach can be used to make informed choices about other modeling approaches or to better focus the initial modeling. For example, if a GO summary analysis indicates a preponderance of transcripts summarized to a general term such as metabolism (GO:0008152 metabolic process), the researcher may consider doing a pathways analysis identify the specific metabolic pathways represented in this subset of transcripts. Second, researchers should note that not all analysis tools will accept/identify all of the data in their data set. Typically a certain proportion of gene products may not included in the analysis because their accessions are not recognized or they have no functional annotation. For example, EST sequences represented on arrays may not be included in network analysis because they cannot be linked to a gene or because that gene has no interaction data. It is important that this data loss is minimized and accounted for where possible when developing an overall model. Third, researchers need to consider not only aspects of their functional modeling that are already known about their system but also new information gained during the analysis process. Finally, we reiterate that functional modeling must be driven by the biology of the system being studied, rather than by the results of any bioinformatics analysis.

As poultry researchers increasingly adopt new sequencing technologies, we will have more opportunities to refine genome annotations for these species. However, with the move from traditional functional genomics platforms (such as arrays) to newer gene expression technologies based on transcriptome sequencing and proteomics, there is an increasing need to apply these data not only to answer fundamental questions about poultry production but also for reusing these same data to capture more information to support functional modeling (such as structural annotation and tissue expression data). Analysis of gene expression using RNASeq, because it relies on alignment to an existing genome, requires that we know the gene within the genome and can correctly assemble and align expressed transcripts. Moreover, like all functional genomics technologies, RNASeq comes with its own inherent biases, limitations, and capacity that need to be considered when doing bioinformatic tool development to support functional modeling. In addition to providing more annotation to support functional modeling tool and resources, there is also the need to integrate multiple types of data from multiple sources: data from genome sequencing, orthologies, expression, pathways, and interactions all need to be combined in a way that enables researchers insights into the biology of their system of interest. With improved data integration comes the ability develop new tools that more easily integrate genotype, phenotype, statistical prediction, and functional modeling.

ACKNOWLEDGMENTS

Annotations and resources provided by the AgBase databases is supported by Agriculture and Food Research Initiative Competitive Grant no. 2011-67015-30332 from the USDA National Institute of Food and Agriculture. Eric Lyons and the CoGe browser is supported by the iPlant Collaborative, a National Science Foundation Plant Cyberinfrastructure Program (#DBI-0735191) award, and the Betty and Gordon Moore Foundation. We especially acknowledge the members of the iPlant Education, Outreach and Training Group at the DNA Learning Center of Cold Spring Harbor Laboratory (Cold Spring Harbor, NY) who developed the iAnimal portal described in this manuscript. iAnimal is currently available for evaluation and comment by the animal research community at http://genepro.cshl.edu/ianimal/.

REFERENCES

Barrett, T., and R. Edgar. 2006. Gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 411:352–369.

Baxevanis, A. D. 2011. The importance of biological databases in biological discovery. Curr. Protoc. Bioinformatics Chapter 1: Unit 1 1.

Beissbarth, T., and T. P. Speed. 2004. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20:1464–1465.

Berriz, G. F., J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth. 2009. Next generation software for functional trend analysis. Bioinformatics 25:3043–3044.

Brazma, A., M. Kapushesky, H. Parkinson, U. Sarkans, and M. Shojatalab. 2006. Data storage and analysis in ArrayExpress. Methods Enzymol. 411:370–386.

Burt, D. W., W. Carre, M. Fell, A. S. Law, P. B. Antin, D. R. Maglott, J. A. Weber, C. J. Schmidt, S. C. Burgess, and F. M. McCarthy. 2009. The Chicken Gene Nomenclature Committee report. BMC Genomics 10(Suppl. 2):S5.

Cordero, F., M. Botta, and R. A. Calogero. 2007. Microarray data analysis and mining approaches. Brief Funct. Genomic Proteomic 6:265–281.

Croft, D., G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. 2011. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res. 39(Database issue):D691–D697.

Dalloul, R. A., J. A. Long, A. V. Zimin, L. Aslam, K. Beal, A. Blomberg Le, P. Bouffard, D. W. Burt, O. Crasta, R. P. Crooijmans, K. Cooper, R. A. Coulombe, S. De, M. E. Delany, J. B. Dodgson, J. J. Dong, C. Evans, K. M. Frederickson, P. Flicek, L. Florea, O. Folkerts, M. A. Groenen, T. T. Harkins, J. Herrero, S. Hoffmann, H. J. Megens, A. Jiang, P. de Jong, P. Kaiser, H. Kim, K. W. Kim, S. Kim, D. Langenberger, M. K. Lee, T. Lee, S. Mane, G. Marcais, M. Marz, A. P. McElroy, T. Modise, M. Nefedov, C. Notredame, I. R. Paton, W. S. Payne, G. Pertea, D. Prickett, D. Puiu, D. Qioa, E. Raineri, M. Ruffier, S. L. Salzberg, M. C. Schatz, C. Scheuring, C. J. Schmidt, S. Schroeder, S. M. Searle, E. J. Smith, J. Smith, T. S. Sonstegard, P. F. Stadler, H. Tafer, Z. J. Tu, C. P. Van Tassell, A. J. Vilella, K. P. Williams, J. A. Yorke, L. Zhang, H. B. Zhang, X. Zhang, Y. Zhang, and K. M. Reed. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. PLoS Biol. 8:e1000475.

Dimmer, E. C., R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler. 2012. The UniProt-GO Annotation database in 2011. Nucleic Acids Res. 40(Database issue):D565–D570.

Du, Z., X. Zhou, Y. Ling, Z. Zhang, and Z. Su. 2010. agriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res. 38(Web Server issue):W64–W70.

Gaudet, P., M. S. Livstone, S. E. Lewis, and P. D. Thomas. 2011. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief. Bioinform. 12:449–462.

Genome 10K Community of Scientists. 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100:659–674.

Gillespie, M., V. Shamovsky, and P. D'Eustachio. 2011. Human and chicken TLR pathways: Manual curation and computer-based orthology analysis. Mamm. Genome 22:130–138.

Goff, S. A., M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S. J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzione. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Front. Plant Sci. 2:34.

Grossmann, S., S. Bauer, P. N. Robinson, and M. Vingron. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics 23:3024–3031.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4:44–57.

Huang, T. W., C. Y. Lin, and C. Y. Kao. 2007. Reconstruction of human protein interolog network using evolutionary conserved network. BMC Bioinformatics 8:152.

Khatri, P., C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A. L. Tarca, and S. Draghici. 2007. Onto-Tools: New additions and improvements in 2006. Nucleic Acids Res. 35(Web Server issue):W206–W211.

Kumar, R., and B. Nanduri. 2010. HPIDB—A unified resource for host-pathogen interactions. BMC Bioinformatics 11(Suppl. 6):S16.

Lewis, S. E. 2005. Gene Ontology: Looking backwards and forwards. Genome Biol. 6:103.

Lyons, E., and M. Freeling. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53:661–673.

McCarthy, F. M., S. M. Bridges, and S. C. Burgess. 2007a. GOing from functional genomics to biological significance. Cytogenet. Genome Res. 117:278–287.

McCarthy, F. M., S. M. Bridges, N. Wang, G. B. Magee, W. P. Williams, D. S. Luthe, and S. C. Burgess. 2007b. AgBase: A unified resource for functional analysis in agriculture. Nucleic Acids Res. 35(Database issue):D599–D603.

McCarthy, F. M., C. R. Gresham, T. J. Buza, P. Chouvarine, L. R. Pillai, R. Kumar, S. Ozkan, H. Wang, P. Manda, T. Arick, S. M. Bridges, and S. C. Burgess. 2011. AgBase: Supporting functional modeling in agricultural organisms. Nucleic Acids Res. 39(Database issue):D497–D506.

Moriya, Y., M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(Web Server issue):W182–W185.

Mungall, C. J., C. Batchelor, and K. Eilbeck. 2011. Evolution of the Sequence Ontology terms and relationships. J. Biomed. Inform. 44:87–93.

Mungall, C. J., G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner. 2010. Integrating phenotype ontologies across multiple species. Genome Biol. 11:R2.

Orchard, S., S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, G. Cesareni, A. Chatraryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stumpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob. 2012. Protein interaction data curation: The International Molecular Exchange (IMEx) consortium. Nat. Methods 9:345–350.

Rhee, S. Y., V. Wood, K. Dolinski, and S. Draghici. 2008. Use and misuse of the gene ontology annotations. Nat. Rev. Genet. 9:509–515.

Roberts, R. J. 2004. Identifying protein function—A call for community action. PLoS Biol. 2:E42.

Supek, F., M. Bosnjak, N. Skunca, and T. Smuc. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE 6:e21800.

Tanabe, M., and M. Kanehisa. 2012. Using the KEGG database resource. Curr. Protoc. Bioinformatics Chapter 1: Unit 1 12.

van den Berg, B. H., F. M. McCarthy, S. J. Lamont, and S. C. Burgess. 2010. Re-annotation is an essential step in systems biology modeling of functional genomics data. PLoS ONE 5:e10642.

Warren, W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Kunstner, S. Searle, S. White, A. J. Vilella, S. Fairley, A. Heger, L. Kong, C. P. Ponting, E. D. Jarvis, C. V. Mello, P. Minx, P. Lovell, T. A. Velho, M. Ferris, C. N. Balakrishnan, S. Sinha, C. Blatti, S. E. London, Y. Li, Y. C. Lin, J. George, J. Sweedler, B. Southey, P. Gunaratne, M. Watson, K. Nam, N. Backstrom, L. Smeds, B. Nabholz, Y. Itoh, O. Whitney, A. R. Pfenning, J. Howard, M. Volker, B. M. Skinner, D. K. Griffin, L. Ye, W. M. McLaren, P. Flicek, V. Quesada, G. Velasco, C. Lopez-Otin, X. S. Puente, T. Olender, D. Lancet, A. F. Smit, R. Hubley, M. K. Konkel, J. A. Walker, M. A. Batzer, W. Gu, D. D. Pollock, L. Chen, Z. Cheng, E. E. Eichler, J. Stapley, J. Slate, R. Ekblom, T. Birkhead, T. Burke, D. Burt, C. Scharff, I. Adam, H. Richard, M. Sultan, A. Soldatov, H. Lehrach, S. V. Edwards, S. P. Yang, X. Li, T. Graves, L. Fulton, J. Nelson, A. Chinwalla, S. Hou, E. R. Mardis, and R. K. Wilson. 2010. The genome of a songbird. Nature 464:757–762.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack. 2010. Gene ontology analysis for RNA-seq: Accounting for selection bias. Genome Biol. 11:R14.

Yu, H., N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. 2004. Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. Genome Res. 14:1107–1118.