BMC Genomics

Database

AgBase: a functional genomics resource for agriculture

Fiona M McCarthy*†[1,7], Nan Wang†[2,7], G Bryce Magee[2,7], Bindu Nanduri[1,7], Mark L Lawrence[1], Evelyn B Camon[3], Daniel G Barrell[3], David P Hill[4], Mary E Dolan[4], W Paul Williams[5], Dawn S Luthe[6,7], Susan M Bridges†[2,7] and Shane C Burgess†[1,7]

Address: [1]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, P.O. Box 1600, Mississippi State, MS 39762, USA, [2]Department of Computer Science and Engineering, Bagley College of Engineering, P.O. Box 9637, Mississippi State University, MS 39762, USA, [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [4]Mouse Genome Informatics, The Jackson Laboratory 600 Main Street, Bar Harbor, ME 04609, USA, [5]USDA ARS Corn Host Plant Resistance Research Unit, Box 5367, Mississippi State University, MS 39762, USA, [6]Department of Biochemistry and Molecular Biology, P.O. Box 9650, Mississippi State University, MS 39762, USA and [7]Institute for Digital Biology, Mississippi State University, MS 39762, USA

Email: Fiona M McCarthy* - fmccarthy@cvm.msstate.edu; Nan Wang - nw43@msstate.edu; G Bryce Magee - gbm1@msstate.edu; Bindu Nanduri - bnanduri@cvm.msstate.edu; Mark L Lawrence - lawrence@cvm.msstate.edu; Evelyn B Camon - camon@ebi.ac.uk; Daniel G Barrell - dbarrell@ebi.ac.uk; David P Hill - dph@informatics.jax.org; Mary E Dolan - Mary_Dolan@umit.maine.edu; W Paul Williams - wpwilliams@msa-msstate.ars.usda.gov; Dawn S Luthe - dsluthe@ra.msstate.edu; Susan M Bridges - bridges@cse.msstate.edu; Shane C Burgess - burgess@cvm.msstate.edu

* Corresponding author †Equal contributors

Published: 08 September 2006

BMC Genomics 2006, **7**:229 doi:10.1186/1471-2164-7-229

Received: 18 April 2006
Accepted: 08 September 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/229

Abstract

Background: Many agricultural species and their pathogens have sequenced genomes and more are in progress. Agricultural species provide food, fiber, xenotransplant tissues, biopharmaceuticals and biomedical models. Moreover, many agricultural microorganisms are human zoonoses. However, systems biology from functional genomics data is hindered in agricultural species because agricultural genome sequences have relatively poor structural and functional annotation and agricultural research communities are smaller with limited funding compared to many model organism communities.

Description: To facilitate systems biology in these traditionally agricultural species we have established "AgBase", a curated, web-accessible, public resource http://www.agbase.msstate.edu for structural and functional annotation of agricultural genomes. The AgBase database includes a suite of computational tools to use GO annotations. We use standardized nomenclature following the Human Genome Organization Gene Nomenclature guidelines and are currently functionally annotating chicken, cow and sheep gene products using the Gene Ontology (GO). The computational tools we have developed accept and batch process data derived from different public databases (with different accession codes), return all existing GO annotations, provide a list of products without GO annotation, identify potential orthologs, model functional genomics data using GO and assist proteomics analysis of ESTs and EST assemblies. Our journal database helps prevent redundant manual GO curation. We encourage and publicly acknowledge GO annotations from researchers and provide a service for researchers interested in GO and analysis of functional genomics data.

Conclusion: The AgBase database is the first database dedicated to functional genomics and systems biology analysis for agriculturally important species and their pathogens. We use experimental data to improve structural annotation of genomes and to functionally characterize gene products. AgBase is also directly relevant for researchers in fields as diverse as agricultural production, cancer biology, biopharmaceuticals, human health and evolutionary biology. Moreover, the experimental methods and bioinformatics tools we provide are widely applicable to many other species including model organisms.

Background

The genomes of agriculturally important organisms are sequenced [1-3] or being sequenced [4,5] not only due to their economic importance but also because many are biomedical models [6-10] or zoonotic pathogens and bioterrorism agents [11]. However, after genome sequencing it is critical to identify and demarcate the functional elements in the genome (structural annotation) and to link these genomic elements to biological function (functional annotation). Current genome assemblies have several thousands of gaps, causing bad gene model predictions due to missing exons and splice sites. Statistics for the chicken and cow genomes compared with the human, mouse and rat genomes (Table 1) reveal fundamental problems in genomic structural and functional annotation in livestock genomes. Livestock genomes will always have low build numbers compared with model organisms such as human and mouse and yet they have comparable numbers of genes (UniGene). A relatively large proportion of these genes in these species are electronically predicted. Another problem is that, compared to human and mouse, the chicken and cow have 10-fold fewer ESTs to aid in structural annotation and functional analysis. These statistics, combined with smaller funding bases and resources for manual genome structural annotation, suggest that the human and mouse paradigm for genome structural annotation is unlikely to be successful for agricultural species [12].

For functional annotation, the GO is the *de facto* standard and its use for modeling microarray and other functional genomics data is growing exponentially [13]. However, this growth in the use of GO is not seen in agricultural species (Figure 1) because of poor GO annotation (Table 1). Gramene [14] and TIGR [15] provide annotations for grasses and microbes, respectively, but most GO annotations for other agriculturally important species are provided by the European Bioinformatics Institute Gene Ontology Annotation project (EBI-GOA) [16]. EBI-GOA annotates proteins in the UniProt Knowledgebase (UniProtKB) only. However, agricultural species have an order of magnitude fewer entries in UniProtKB than human and mouse. Many proteins in agricultural species are still only electronically predicted and reside in the UniProt Archive (UniParc) database, which EBI-GOA does not annotate. Moreover, most GO annotations that do exist for agricultural proteins are "inferred from electronic annotation" (IEA). IEA is usually only applied to broad GO terms and results in very general superficial GO functional information. More detailed functional annotations require expert human curation of experimental evidence, typically from peer-reviewed literature. The rat genome (22) is an interesting example of what can be achieved in GO annotation

Table 1: Comparison of human, mouse, rat, chicken and bovine genome statistics.

Species	Genome Build	Genes (UniGene)	ESTs
H. sapiens	36.1	86 803	7 741 746
M. musculus	36.1	67 096	4 719 380
R. norvegicus	3.4	51 564	871 147
G. gallus	1.1	30 470	588 288
B. taurus	2.1	41 986	1 039 059
Species	**No. Proteins (NRPD)**	**No. Proteins (UniProtKB)**	**% 'Predicted' Proteins**
H. sapiens	299 863	73 932	5.9
M. musculus	184 110	61 537	15.5
R. norvegicus	52 857	14 885	29.9
G. gallus	29 763	7 291	47.9
B. taurus	52 425	10 059	57.1
Species	**All GO Associations**	**Non-IEA Associations**	**% IEA Associations**
H. sapiens	266 785	50 612	81
M. musculus	327 082	517 032	60.2
R. norvegicus	73 783	10 285	86.1
G. gallus	29 963	3 058	89.8
B. taurus	39 832	9 484	76.2

Current annotation statistics for selected genomes (15/03/06) are shown. The build number was obtained from NCBI, the estimated number of gene products is based on UniGene numbers [38] and EST numbers are obtained from ESTdb. The number of proteins in the UniProtKB database is under represented for agricultural species. To estimate the proportion of predicted genes in the genome, the number of gene predictions is expressed as a percentage of the total of number of genes both predicted and from UniGene. GO statistics are obtained from GO association files using *GOProfiler* (available from AgBase).

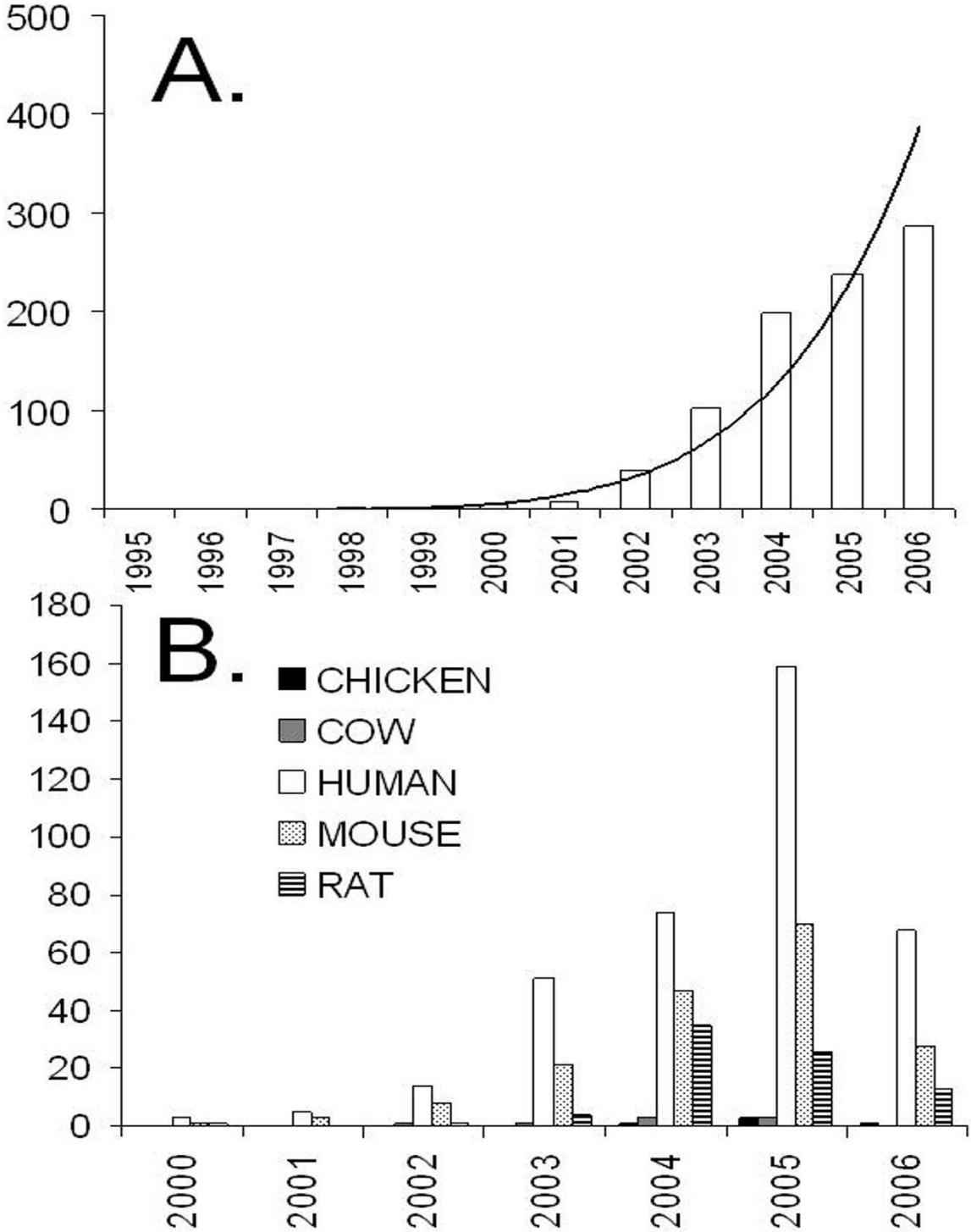


Figure 1

Papers referencing GO by species. The number of papers referencing GO, as determined from PubMed (06/09/06). GO annotation has become the accepted standard for functional annotation [13] and its use is growing exponentially (A). Despite this, GO annotation has been minimally used in chicken and cow (B), in part this is because of smaller numbers of livestock researchers, but also using GO annotation in livestock first requires researchers to functionally annotate their own data.

through a community's concerted effort. The rat genome was published only 8 months prior to the chicken genome. Like the chicken, but unlike human and mouse, rat has relatively few proteins in the UniProtKB and a high proportion of "predicted" proteins in UniParc. However, there are twice as many GO annotations for rat than currently exist for chicken and fewer of these annotations are IEA. Consequently, there is a concomitant growth in rat publications using GO to model microarray and other functional genomics data (Figure 1).

The current state of agricultural genome annotation hinders its utility for systems biology modeling of microarray and other functional genomics datasets. To fully utilize agricultural genome sequence data requires further, computationally accessible, structural and functional annotation. Here we describe "AgBase", a unified resource dedicated to enabling genome-wide structural and functional annotation and modeling of microarray and other functional genomics data in agricultural species. AgBase integrates structural and functional annotations and provides tools in an easy-to-use pipeline, allowing agricultural and biomedical researchers to rapidly and effectively model and derive biological significance from microarray and other functional genomics datasets.

Construction and content

The AgBase server is a dual Xeon 3.0 processor with a 800 Mhz FSB, 4 GB of Ram and five 146 GB hard drives in a RAID-5 configuration. The operating system is Windows 2000 Server. AgBase has a dedicated tape backup system with a total storage capacity of 3.2 TB native and 6.4 TB compressed. The backup software is Veritas Netback. A full backup is done each weekend, an archive backup once a month, and incremental backups nightly.

AgBase is implemented using the mySQL 4.1 database management system, NCBI Blast, and scripts written in Perl CGI. The schema is a protein centric design that is an adaptation of the Chado schema with extensions to accommodate storage of expressed peptide sequence tags (ePSTs). The entity relationship (ER) model for primary objects in the database for each protein is given as supplementary data [see Additional file 1]. A separate schema is implemented for ePST data. Data that is generated in-house includes AgBase GO annotations, the AgBase gene association files and ePSTs. External data that is integrated into the database includes the Gene Ontology, the UniProt database, EBI-GOA and the NCBI Entrez Taxonomy.

The GO annotations are generated by manual curation of the literature and by sequence similarity (GO evidence code ISS) using the GOanna tool followed by manual inspection of the alignments that are produced. AgBase biocurators are trained in a GO curation course that is held periodically. All literature-based AgBase GO annotations are quality checked to GO Consortium standards. The ePSTs are generated using a proteogenomic mapping pipeline implemented in Perl. The pipeline integrates information from experimental proteomics experiments and annotated genomes. Results are visualized using the Apollo genome browser to allow curation by scientists. Each ePST is quality checked by AgBase Biocurators. The generation of ePSTs is discussed in the experimental structural annotation section below.

Users can access protein information by protein name, gene name, GO term, taxon, a variety of accession numbers, or via BLAST searches. The AgBase tools also access the AgBase database. AgBase is updated from external sources every three months and locally generated data is loaded as it is generated. Gene association files of gene products annotated by AgBase are accessible in a tab-delimited format to facilitate data exchange.

We have purposely followed the paradigm of multi-species databases suggested by Stein [17] and the Reactome database [18] and are currently focused on plants and animals whose genomes are, or will be, sequenced and microbial pathogens and parasites that have significant economic impact on agricultural production and zoonotic disease. AgBase has four main aims (discussed in detail below): (1) to provide experimentally derived structural annotations of agricultural genomes; (2) to provide highly curated, GO functional annotations; (3) to promote the use of standardized nomenclature in agricultural species; (4) to develop computational pipelines for processing and using structural and functional annotations.

Utility

The AgBase database is intended as a resource to assist functional genomics in agricultural species and the tools provided support analysis of large scale datasets. To this end, we provide both experimentally derived structural annotation and functional data in a unified resource. While agriculturally important organisms may have other resources that provide structural annotation or GO annotations, AgBase is unique because (1) the structural data provided is experimentally derived; (2) the structural and functional data is provided from a unified resource; and (3) tools for analysis of this data are freely available via AgBase. The AgBase interface allows users to search for information in several ways. The Text Search performs an exact substring search on the selected database. To facilitate data sharing, searching based on commonly used accession numbers and identifiers is supported in addition to BLAST searches. Multiple query searches are also available.

Discussion

Experimental structural annotation

The use of experimental data for genome annotation is critical for conclusive identification of the functional sequences within genomes, accurate description of intron/exon structures and determination of the potential products from each gene in different tissues and cellular states [19]. Through AgBase we make available improved structural annotation of agriculturally important genomes from experimental confirmation of electronically predicted proteins/open reading frames, especially via proteogenomic mapping [19-23].

Proteogenomic mapping generates expressed peptide sequence tags (ePSTs) [23]. These ePSTs are derived by identifying novel protein fragments through proteomics, aligning these to the genome sequence and extending to the nearest 3' stop codon. We have used the proteogenomic mapping pipeline to generate ePSTs for a prokaryote (*Pasteurella multocida*) and a eukaryote (chicken). *P. multocida*, or chicken "fowl cholera", is a bovine respiratory disease pathogen and human zoonosis. Although the *P. multocida* genome was sequenced in 2001 [24] and is considered well annotated, our proteogenomic pipeline identified 202 ePSTs that had identifiable methionine start codons [see Additional file 2]. One of these is a 130 amino acid ePST that was identified by six different peptides and is located in a 704 bp intergenic region between accA and guaA in the Pm70 genome [see Additional file 3]. The ePST has 60% identity and 74% similarity at the protein level with the 114 amino acid hypothetical protein HD_1218 (Genbank accession AAP96060) from *Haemophilus ducreyi* (a major cause of human genital ulcer disease [chancroid] in humans). A database of ePSTs identified from chicken and *P. multocida* is publicly accessible via the proteogenomics link on the AgBase homepage. The ePST database is fully searchable either by text or Blast searching. Text-searchable fields include taxonID, genome build, chromosome or chromosomal location. Public submissions to the ePST database are cited by submitter name.

Generating ePSTs is time consuming and labor intensive. To facilitate structural annotation we have developed a proteogenomic mapping pipeline for generation of ePSTs (available from AgBase by request). The pipeline for prokaryotes currently includes a visualization component (we currently use Apollo [25]) that allows the researcher to view the ePSTs in context in the genome. In eukaryote genomes it is possible that the extension is carried beyond a splice signal producing an ePST that includes intronic DNA. We are currently in the process of extending the pipeline to detect splice signals and to show alignments with ESTs in the visualizations to address this shortcoming.

To ensure that structural data is based on high quality proteomics identifications, we have developed a method for assigning probabilities to mass spectral identifications during proteogenomic mapping [26]. Assigning probabilities to mass spectral identifications is important because one issue associated with tandem mass spectral searching against databases is false positive and false negative peptide identifications. Moreover, all of our proteomics data is submitted to the PRIDE database [27]. Mass spectrometry data submitted to PRIDE is further curated for inclusion in UniProtKB, where it is available for uploading into genome browsers, for example Ensembl [28]. To add value to the structural annotations provided by AgBase and enhance biological modeling, we have also developed methods and tools for assigning GO annotations (see below).

Functional annotation

Many gene products from agriculturally important organisms have no GO annotation. Practically, this means that experimentalists working with these species must provide their own GO annotations if they wish to use GO to model their microarray and other functional genomics data. While those best qualified to functionally annotate a gene product may be those who work directly with it [29], few experimentalists can devote the time and resources needed to learn the intricacies of GO biocuration. To facilitate functional modeling in agricultural organisms, we are actively GO annotating chicken, cow, sheep and catfish gene products.

While EBI-GOA uses an electronic mapping strategy to rapidly provide GO annotations for a large number of gene products, these are IEA mappings that rely on curated information from SwissProt, InterPro and the Enzyme Commission (EC) databases [16]. Many agricultural gene products are 'predicted' products based on gene prediction algorithms (Table 1) and do not exist in these curated databases. However, GO annotation can be assigned based on human interpretation of sequence and/or structural similarities (ISS) with well-studied and already GO-annotated gene products. By definition, such gene products can only be annotated to ISS or IEA since they have no experimental functional data as yet.

Our GO annotation strategy first provides *breadth* by focusing on the large proportion of gene products that currently exist in the UniParc database and have no GO annotation. Since predicted proteins represent approximately half of the gene products from newly sequenced genomes (Table 1.), being able to provide GO annotations for these gene products complements the GO annotations provided by EBI-GOA and dramatically improves our ability to model functional genomics data. We are doing a "first-pass" ISS annotation of chicken, cow and

sheep gene products that currently have no GO annotation (using manual inspection of BLAST alignments and, where possible, established orthology). In addition, we have developed DDF-MudPIT [30], a high-throughput, proteomics-based method that simultaneously confirms expression and experimentally determines the cellular component of gene products [30]. We next provide finer and more precise functional annotations (i.e. improved GO *depth*) by curating literature. All of our GO annotations are prioritized based on our experimental needs. One example is our recent proteomics model of B-cell development in the chicken bursa of Fabricius [23]. Initially we were hampered because few chicken proteins had any GO functional annotation. We annotated 142 chicken proteins, including curation of 24 PubMed articles. These GO annotations were used to refine cell differentiation, proliferation and cell death modeling in the developing bursa.

To date (02/15/2006) we have provided GO annotations for chicken, cow, sheep and channel catfish (Table 2). For evidence codes other than IEA (which we do not do), in our first nine months MSU-AgBase made a comparable number of GO associations to chicken as EBI-GOA (Figure 2). Our biocurators collaborate with those at EBI-GOA to provide a single, publicly accessible GO annotation file for both chicken and cow. AgBase-derived GO annotations to gene products that exist in UniProtKB are incorporated with EBI-GOA annotations and added to UniProtKB. For completeness, in addition to our own GO annotations, we include EBI-GOA derived GO annotations (including IEA) in AgBase. The source of these annotations is shown in the protein detail page (Figure 3). Currently AgBase is the only source of GO annotations for UniParc agricultural proteins.

We actively educate, encourage and seek out researchers in the scientific communities to contribute their own GO annotations. We help these researchers properly format their annotations and they are acknowledged for their annotations on the protein detail page for the gene product in AgBase. As public annotations are submitted to AgBase, research-directed GO annotations from the research community will be acknowledged on the protein detail page. We will also supply maize gene product annotations to Gramene [31] and MaizeGDB [32]. To avoid duplication of effort in literature curation, we developed a journal database (JDB) based on PubMed identity number (PMID). JDB tracks all PubMed articles used as a source for manual GO annotations. The JDB will aid collaborative GO annotation as it can be used for quality control for GO annotations among interested groups. Non-biocurators may access the JDB as a guest user.

Nomenclature

While the GO does not specifically deal with gene nomenclature, unified and unique nomenclature for orthologous genes is essential. Where possible, chicken genes will be assigned nomenclature based on orthologous human nomenclature [33]. We use human orthologs to provide standardized gene symbols to chicken and cow gene products during the process of making GO associations. A chicken gene nomenclature committee is at a formative stage; as yet, no corresponding committee exists for cow.

Tool development

We have developed freely available computational tools to help researchers use the GO to derive biological significance from their microarray and other functional genomics data. These tools are designed as part of an integrated pipeline to batch process input. In order to improve interoperability with different types of data, the tools accept several input formats. Our GO annotation suite of tools is available online via the Tools link at the AgBase homepage. These tools can also be used for non-agricultural organisms, including newly sequenced species and those without complete genome sequence available. The steps to analyze a microarray or other functional genomics dataset are:

1. Enter the list of accession numbers into *GORetriever* to return all existing GO annotations available for that dataset (Figure 4). *GORetriever* also provides a list of proteins without GO annotation. The researcher then enters this second list into the *GOanna* tool.

Table 2: AgBase GO annotations by species and evidence code.

Species	GO Associations	No. Proteins Annotated	UniParc Proteins Annotated (ISS)	No. Papers Curated
Chicken	1 007	142	80	24
Cow	4 411	382	4	13
Sheep	316	61	0	0
Channel Catfish	19	3	0	2

We aim to increase the coverage of GO annotations in agriculturally important species and we are currently GO annotating chicken, cow, sheep and channel catfish. To improve GO coverage, we determine which proteins currently have no GO annotations and use *GOanna* to do a 'first-pass' annotation based on sequence homology (ISS). We have also provided GO annotations via literature curation for chicken, cow and channel catfish. We do not currently provide IEA annotations.

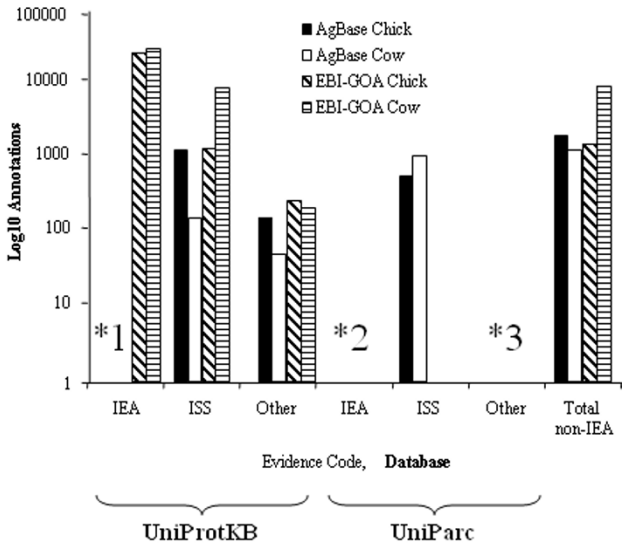


Figure 2
A comparison of chicken and cow GO annotations from AgBase and EBI-GOA. We are currently focused on providing GO annotations for chicken and cow gene products and we collaborate with EBI-GOA to provide a combined GO gene association file for each of these species. The number of GO annotations for chicken and cow is represented here based on GO evidence code; details about the GO evidence codes can also be found on the GO Consortium homepage [37]. (1) Unlike EBI-GOA, AgBase does not currently annotate to IEA. (2) In newly sequenced genomes, such as cow and chicken, a large proportion of gene products are not represented in the UniProtKB database (Table 1) and are not annotated by EBI-GOA. To complement the EBI-GOA annotation effort and provide breadth of coverage, we identify the expression of these 'predicted' gene products *in vivo* and, where possible, provide GO annotations. (3) By definition, there is no published literature for these 'predicted' proteins and they can only be GO annotated using either IEA or ISS.

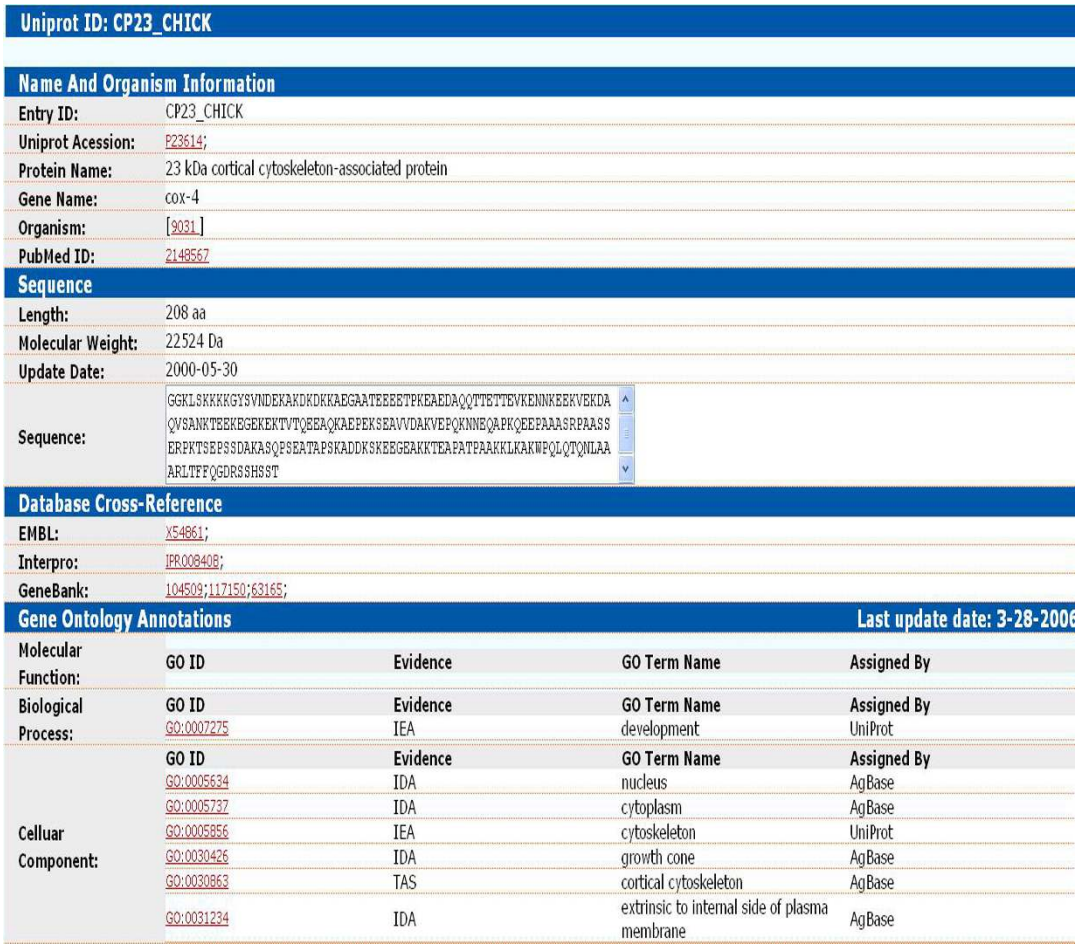
2. *GOanna* accepts either a list of IDs or a user defined file of sequences in FASTA format and does a Blast search against databases containing only annotated proteins. The user can choose the number of Blast hits to retrieve per query and set the "Evalue" threshold (NCBI thresholds are the default). The *GOanna* output file contains hyperlinks that direct the user to the original Blast alignment (Figure 5) so that the user can make their own value judgments for their ISS GO annotations.

3. After ISS annotation, the user may choose to annotate the data further by curating published literature. We provide advice on GO annotation and are developing a mechanism for researchers to be publicly acknowledged for GO annotations they submit to AgBase.

4. After annotation, the researcher can then use *GOSlimViewer* to summarize GO data for each of the three ontologies in chart form (Figure 6). *GOSlimViewer* accepts a text-based file created from the above pipeline as input and, using a user-specified GO Slim, returns a text simple text file. This file can be opened and charted in Excel to obtain publication quality figures.

We are committed to developing tools and pipelines to maximize the payoff gained from expensive high-throughput microarray and other functional genomics experiments. We have designed tools that may be applied across a diverse range of species, including microbes, parasites, viruses, plants and animals. For example, the same tools used to model B-cell development in the chicken allowed us to formulate experimental models for disease resistance in maize. We identified 1,522 unique proteins from the developing maize rachis (cob) using a combination of MudPIT and 2-D electrophoresis. In addition, rachis proteins from *Aspergillus flavus* resistant and susceptible lines were compared by differential gel electrophoresis: seventy-three proteins that were more abundant in resistant lines (1.5-fold or greater). Using the tools described above we divided these over-expressed proteins into four categories: abiotic stress proteins; antioxidant enzymes; enzymes in the phenylpropanoid pathway leading to flavonoid and lignin biosynthesis; and proteins with various other metabolic functions. Analyses of these data will help us formulate testable hypotheses regarding the role of the maize rachis in resistance to *A. flavus* infection and aflatoxin accumulation.

Finally, we also develop tools for agriculturally important species that do not yet (or may never have) their genomes sequenced. Researchers working with such species often rely on ESTs and EST assemblies for functional analysis. However, most ESTs (and microarrays derived from these) are not associated with GO annotation. *GOanna* accepts FASTA files and can be used associate GO function with ESTs. Another tool to enable EST modeling is the *ProtIDer* tool (freely available by request to AgBase). *ProtIDer* is a homology-based search program that provides an automated pipeline for the proteomic analysis of ESTs and EST assemblies from TIGR [34-36]. The *ProtIDer* tool compares EST assemblies or singleton ESTs to the UniProtKB and uses high-matching proteins to correct sequencing errors and to annotate the sequence. We have tested *ProtIDer* using data obtained from channel catfish, an organism which currently has only 1,108 protein records in the NRPD, but has 45,622 ESTs available from the dbEST (03/20/06). Tandem mass spectra obtained from channel catfish ovary cells was used to search against three databases: all catfish entries in the NRPD (cfNRPD); a database of highly homologous proteins (hpDB) that come from NRPD as a result of TBLASTN-searching the NRPD with all



Uniprot ID: CP23_CHICK

Name And Organism Information

Entry ID:	CP23_CHICK
Uniprot Acession:	P23614;
Protein Name:	23 kDa cortical cytoskeleton-associated protein
Gene Name:	cox-4
Organism:	[9031]
PubMed ID:	2148567

Sequence

Length:	208 aa
Molecular Weight:	22524 Da
Update Date:	2000-05-30
Sequence:	GGKLSKKKKGYSVNDEKAKDKDKKAEGAATEEEETPKEAEDAQQTTETTEVKENNKEEKVEKDA QVSANKTEEKEGEKEKTVTQEEAQKAEPEKSEAVVDAKVEPQKNNEQAPKQEEPAAASRPAASS ERPKTSEPSSDAKASQPSEATAPSKADDKSKEEGEAKKTEAPATPAAKKLKAKWPQLQTQNLAA ARLTFFQGDRSSHSST

Database Cross-Reference

EMBL:	X54861;
Interpro:	IPR008408;
GeneBank:	104509;117150;63165;

Gene Ontology Annotations Last update date: 3-28-2006

	GO ID	Evidence	GO Term Name	Assigned By
Molecular Function:				
Biological Process:	GO ID	Evidence	GO Term Name	Assigned By
	GO:0007275	IEA	development	UniProt
Celluar Component:	GO ID	Evidence	GO Term Name	Assigned By
	GO:0005634	IDA	nucleus	AgBase
	GO:0005737	IDA	cytoplasm	AgBase
	GO:0005856	IEA	cytoskeleton	UniProt
	GO:0030426	IDA	growth cone	AgBase
	GO:0030863	TAS	cortical cytoskeleton	AgBase
	GO:0031234	IDA	extrinsic to internal side of plasma membrane	AgBase

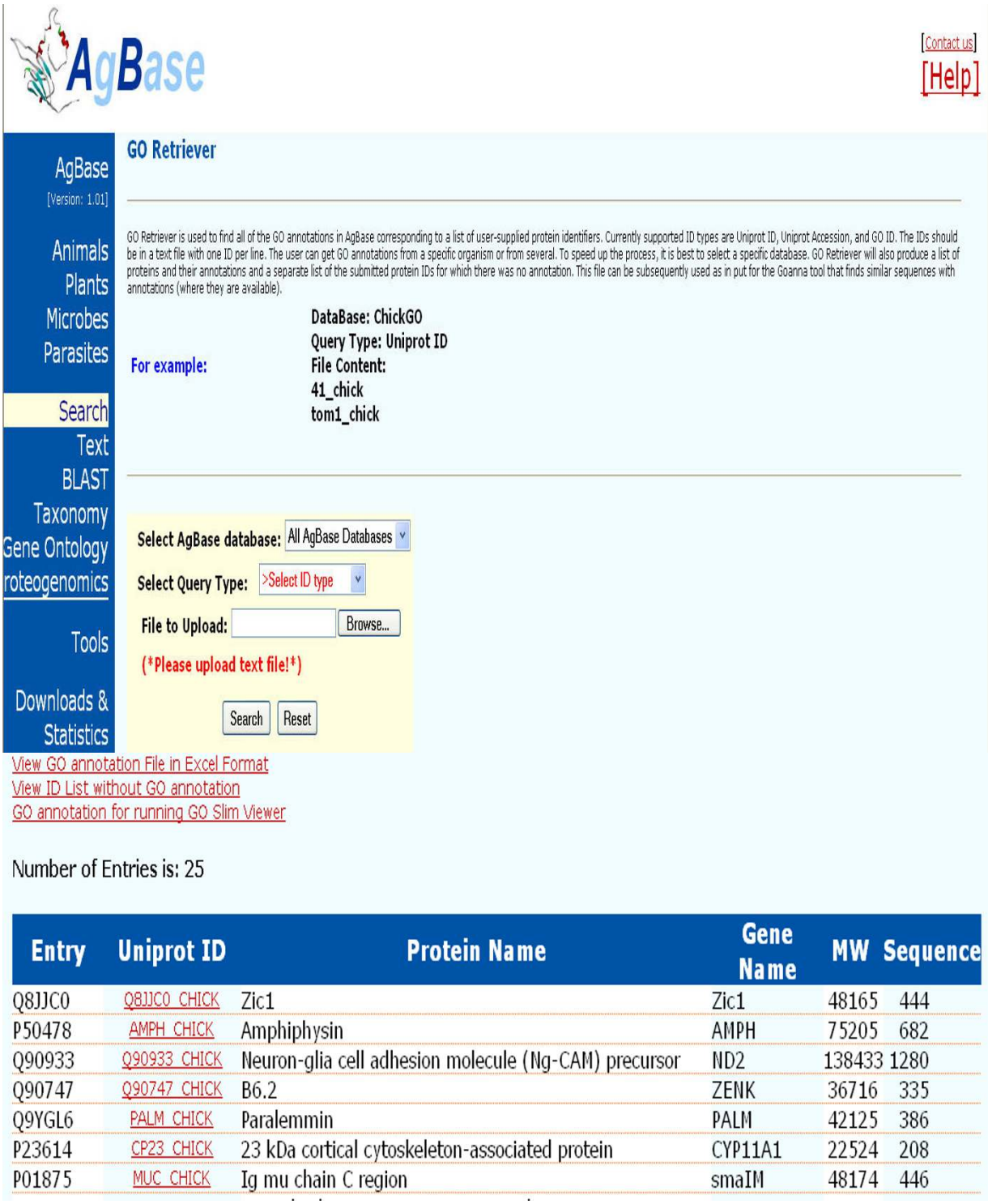
Figure 3

The AgBase protein detail page. The AgBase protein detail page shows proteins and their GO annotation. The GO annotation terms are interactive links and the source of the GO annotation is acknowledged. Protein sequence is displayed in a text accessible window and where possible, links to other databases are cross-referenced.

catfish ESTs generated using *ProtIDer*; and the ESTs themselves translated in all 6 frames (cfESTDB). We identified 1001 proteins and ESTs [see Additional file 4]: 10 from cfNRPD (4 were ribosomal proteins); 48 from the hpDB (only 5 of which were ribosomal) and 962 from cfESTDB. These approaches provide complementary annotation information. Not all of the cfNRPD entries are yet represented in the EST databases; the hpDB allowed us to identify highly conserved proteins and searching the cfESTDB directly indicates ESTs that may be translated. When we used ISS to the hpDB to make GO associations to catfish ESTs we found that the GO terms were distributed over the cellular component, although the biological process had a larger proportion of gene products annotated to response to stimulus and cell communication. From this initial data we will focus on modeling cell communication pathways in developing channel catfish ovary cells.

Future developments

We are building upon the tools and resources already available at AgBase. The proteogenomics pipeline is being extended to allow more informative visualization of ePSTs in context within the genome and alignment with ESTs and orthologous sequences from other organisms. We will continue to generate ePSTs for newly sequenced agricultural genomes and will also continue to add GO annotations for agriculturally important organisms. We are working to improve the representation of agricultural gene products in the UniProtKB and a tangible example of this is the recent addition of experimentally confirmed



Entry	Uniprot ID	Protein Name	Gene Name	MW	Sequence
Q8JJC0	Q8JJC0_CHICK	Zic1	Zic1	48165	444
P50478	AMPH_CHICK	Amphiphysin	AMPH	75205	682
Q90933	Q90933_CHICK	Neuron-glia cell adhesion molecule (Ng-CAM) precursor	ND2	138433	1280
Q90747	Q90747_CHICK	B6.2	ZENK	36716	335
Q9YGL6	PALM_CHICK	Paralemmin	PALM	42125	386
P23614	CP23_CHICK	23 kDa cortical cytoskeleton-associated protein	CYP11A1	22524	208
P01875	MUC_CHICK	Ig mu chain C region	smaIM	48174	446

Figure 4
GORetriever. *GORetriever* takes a list of accession numbers or IDs and fetches the existing GO annotation for these products. A list of IDs for which there is currently no GO annotation is also returned and may be used as input for *GOanna* (Figure 5). An example of a chicken protein and its corresponding matches is shown.

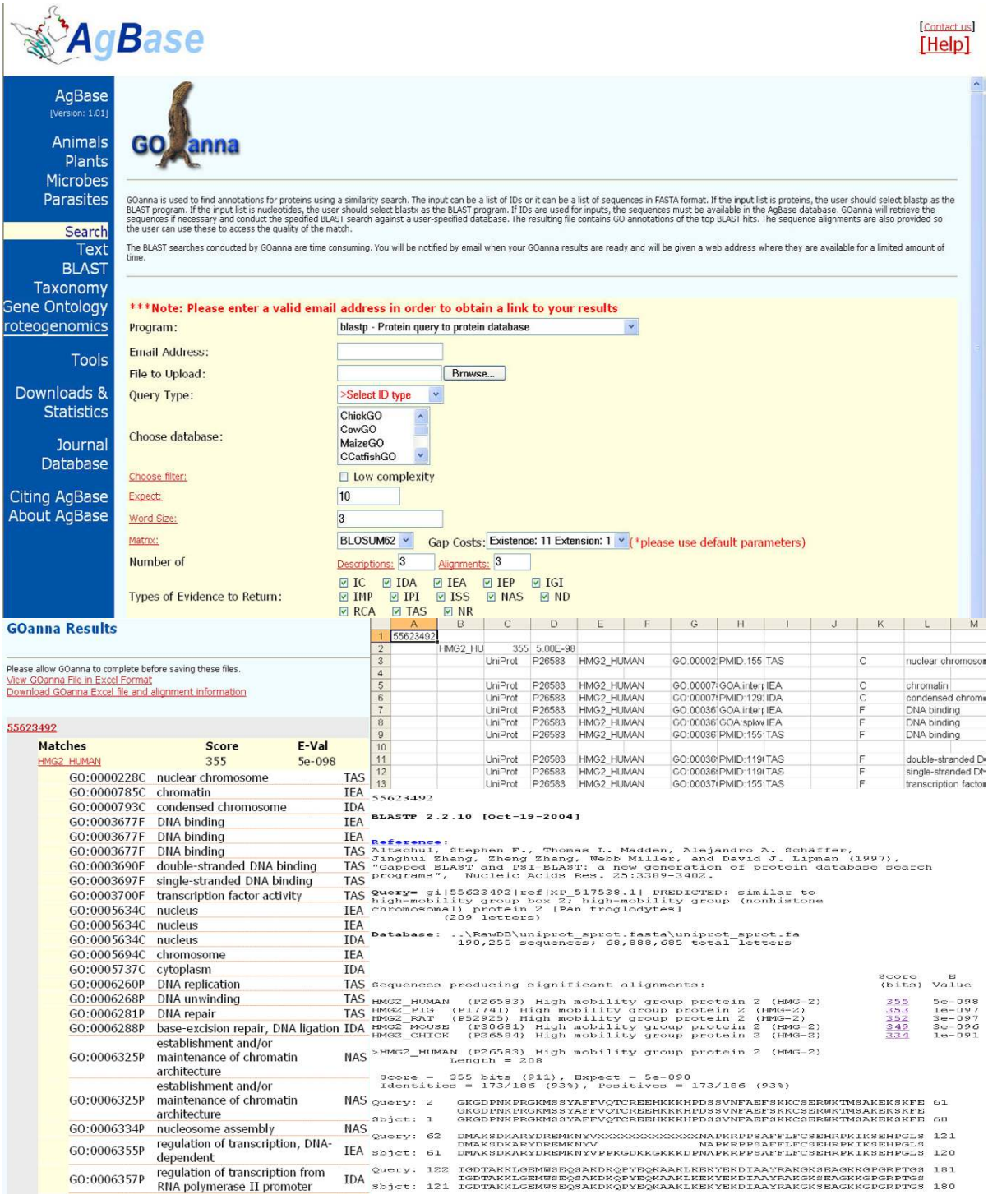


Figure 5

GOanna. *GOanna* allows a user to make GO annotations based on sequence similarity. The user inputs a file of IDs or sequences and the tool does a Blast search against a user-specified database of GO annotated gene products using user-defined parameters. The output is shown both at the web interface and as a downloadable file that contains hyperlinks to the BlastP alignments.

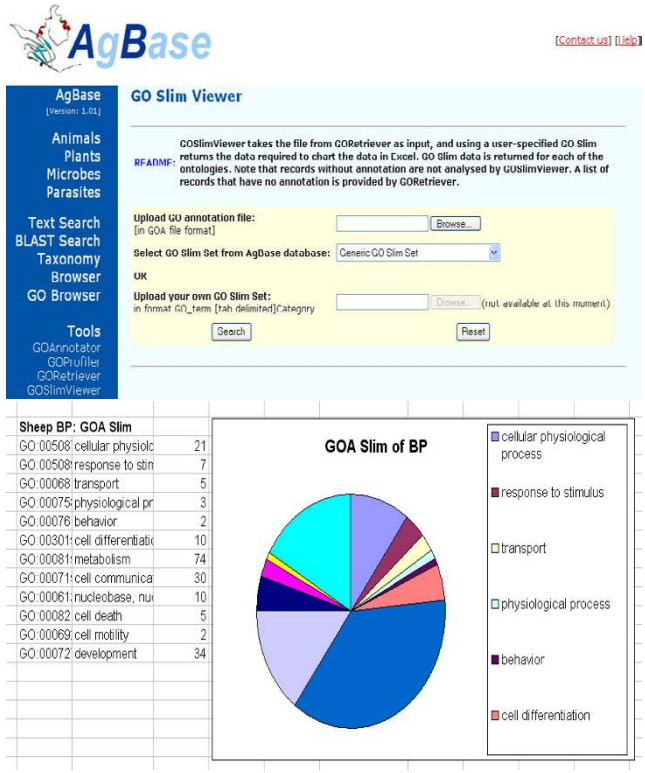


Figure 6
GOSlimViewer. *GOSlimViewer* takes a list of list of GO numbers generated from the GO Retriever program (A) and using a user-defined slim, creates an Excel compatible file that can be used for visualization of the results (B).

chicken 'predicted' gene proteins [23] added to the UniProtKB database.

Conclusion

We have improved the structural annotation of agriculturally important genomes by experimentally confirming 8,704 predicted proteins in chicken and cow (PRIDE submissions numbers pending) and 723 ePSTs from chicken and *P. multocida*. In our first nine months (04/22/05–03/20/06) we have provided 5,762 new GO annotations to 759 proteins from five different species. While most of our GO annotations are ISS (97%), we have also manually curated 42 PubMed references. We have developed a suite of tools to associate GO annotation with experimental data and to provide higher-order summaries of the data, and a tool to aid EST analysis. Users external to MSU account for more than one third of the hits recorded at the AgBase website.

Availability and requirements

Access to the AgBase databases is via http://www.agbase.msstate.edu/ and access to data is unrestricted. The tools we have developed are either freely available online at AgBase or by contacting us via the link provided at the AgBase website. The help pages provide information about how to use these tools or technical support can be obtained directly by contacting us. AgBase is an on-going project and interaction with the user community is vital for its success. We encourage the submission of data, correction of errors, and suggestions for making AgBase of greater use, including ideas for new computational tools. Our biocurators make every effort to maintain data integrity by linking data with researchers, references and methods.

Authors' contributions

FMM developed the GO associations, assisted with the conception and design of the AgBase databases and website, analyzed and interpreted channel catfish data and drafted the manuscript. NW developed the GO and ePSTs databases, developed and implemented the AgBase website, assisted with the conception of the AgBase databases and contributed to tool development. GBM developed the JDB and contributed to tool development. BN designed, analyzed and interpreted the *P. multocida* data, assisted with concepts in tool development and critically reviewed the manuscript. MLL designed, analyzed and interpreted the *P. multocida* data and critically reviewed the manu-

script. EBC provided assistance and quality assessment of gene associations and assisted with making AgBase data publicly available. DGB provided tools and strategies for making the initial gene associations and provided quality assurance. DPH provided GO training, mentoring and quality assurance of AgBase GO annotations. MED developed computer strategies for mapping GO terms to chicken gene products and developed the initial ChickGO database. WPW and DSL designed the maize experiments, and analyzed and interpreted the maize data. DSL also contributed to development of the manuscript. SMB developed the conception and design of the AgBase databases and website, assisted with concepts in tool development, analyzed and interpreted channel catfish and maize data and critically reviewed the manuscript. SCB developed the conception and design of the AgBase databases and website, developed concepts for the bioinformatics tools, analyzed and interpreted channel catfish and *P. heamolytica* data and critically reviewed the manuscript.

Additional material

Additional File 1

Entity relationship (ER) model of the AgBase database*. The AgBase schema design is protein centric. In addition to the taxonomy identifier, sequence, and GO annotations, mappings to number of identifiers are maintained for each protein.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-229-S1.pdf]

Additional File 2

ePSTs identified from* P. multocida *using the proteogenomic pipeline developed at AgBase*.* P. multocida *Pm70 ePSTs identified by proteogenomic mapping. For all 202 ePSTS the original MS peptide and its length, the extended ePST and its length as well as the genome start and end positions of the ePST are shown.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-229-S2.pdf]

Additional File 3

Location of ePST47 in* P. multocida *Pm70 genome*. Region of PM70 genome coding for accA and guaA genes and the 5' end of PMO291 (A). The position in genome is on the left hand side; DNA sequence is shown above; amino acid sequence below; direction of reading frames shown by arrows. The ePST47 is located in the intergenic region between accA and guaA. The figure is taken directly from NCBI nucleotide database graph view with the accession number NC_002663. An expanded view of intergenic region between accA and guaA including ePST47, which runs in the opposite direction to accA, guaA and PMO291 (arrowed) is also shown (B). The overlapping peptides used to identify ePST47 are indicated using bold type (PRIDE accession number pending). This region of the PM70 genome also has 60% identity (* indicates identical nucleotides) and 74% similarity at the protein level with the* Haemophilus ducreyi *hypothetical protein HD_1218 (Genbank accession AAP96060).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-229-S3.pdf]

Additional File 4

Channel catfish proteins and ESTs identified using the* ProtIDer *tool*. The* ProtIDer *tool is designed for use with species that have large numbers of ESTs but few proteins in the NRPD.* ProtIDer *matches ESTs and EST assemblies to highly homologous proteins from other species by TBLASTN-searching the NRPD with all catfish ESTs. For example, the channel catfish genome sequence is not available and there are only 1,108 NRPD entries for this species. When we analyzed channel catfish ovary tissue we were only able to identify 10 proteins from NRPD. Using* ProtIDer *to create a database of highly homologous proteins resulted in a five-fold increase in the number of proteins identified in this experiment. The proteins identified from the channel catfish NRPD database (cfNRPD), highly homologous database (hpDB) and EST databases are shown here. The* ProtIDer *tool is available from AgBase upon request.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-229-S4.pdf]

Acknowledgements

We would like to thank MGI and EBI-GOA for their continued help and support with the Gene Ontology aspects of this manuscript. Financial support for our projects has come from the USDA NRI, MSU Office of Research (MAFES contribution number J-10924), MSU Bagley College of Engineering, MSU College of College of Veterinary Medicine and the MSU Life Science and Biotechnology institute. The authors thank T. Pechan, D. Kunec, B. van den Berg, A.M. Cooksey, A. Shack, C. Doffitt, and E. Dimmer for help with preparing the manuscript.

References

1. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, Dodgson JB, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, Magrini V, McPherson JD, Miner TL, Minx P, Nash WE, Nhan MN, Nelson JO, Oddy LG, Pohl CS, Randall-Maher J, Smith SM, Wallis JW, Yang SP, Romanov MN, Rondelli CM, Paton B, Smith J, Morrice D, Daniels L, Tempest HG, Robertson L, Masabanda JS, Griffin DK, Vignal A, Fillon V, Jacobbson L, Kerje S, Andersson L, Crooijmans RP, Aerts J, van der Poel JJ, Ellegren H, Caldwell RB, Hubbard SJ, Grafham DV, Kierzek AM, McLaren SR, Overton IM, Arakawa H, Beattie KJ, Bezzubov Y, Boardman PE, Bonfield JK, Croning MD, Davies RM, Francis MD, Humphray SJ, Scott CE, Taylor RG, Tickle C, Brown WR, Rogers J, Buerstedde JM, Wilson SA, Stubbs L, Ovcharenko I, Gordon L, Lucas S, Miller MM, Inoko H, Shiina T, Kaufman J, Salomonsen J, Skjoedt K, Wong GK, Wang J, Liu B, Yu J, Yang H, Nefedov M, Koriabine M, Dejong PJ, Goodstadt L, Webber C, Dickens NJ, Letunic I, Suyama M, Torrents D, von Mering C, Zdobnov EM, Makova K, Nekrutenko A, Elnitski L, Eswara P, King DC, Yang S, Tyekucheva S, Radakrishnan A, Harris RS, Chiaromonte F, Taylor J, He J, Rijnkels M, Griffiths-Jones S, Ureta-Vidal A, Hoffman MM, Severin J, Searle SM, Law AS, Speed D, Waddington D, Cheng Z, Tuzun E, Eichler E, Bao Z, Flicek P, Shteynberg DD, Brent MR, Bye JM, Huckle EJ, Chatterji S, Dewey C, Pachter L, Kouranov A, Mourelatos Z, Hatzigeorgiou AG, Paterson AH, Ivarie R, Brandstrom M, Axelsson E, Backstrom N, Berlin S, Webster MT, Pourquie O, Reymond A, Ucla C, Antonarakis SE, Long M, Emerson JJ, Betran E, Dupanloup I, Kaessmann H, Hinrichs AS, Bejerano G, Furey TS, Harte RA, Raney B, Siepel A, Kent WJ, Haussler D, Eyras E, Castelo R, Abril JF, Castellano S, Camara F, Parra G, Guigo R, Bourque G, Tesler G, Pevzner PA, Smit A, Fulton LA, Mardis ER, Wilson RK: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432(7018):**695-716.
2. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436(7052):**793-800.
3. Sonstegard TS, van Tassell CP: **Bovine genomics update: making a cow jump over the moon.** *Genet Res* 2004, **84(1):**3-9.

4. Barbazuk WB, Bedell JA, Rabinowicz PD: **Reduced representation sequencing: a success in maize and a promise for other plant genomes.** *Bioessays* 2005, **27(8):**839-848.
5. Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B, Chumley F, Dvorak J, Iwanaga M, Keller B, Li W, McCombie WR, Ogihara Y, Quetier F, Sasaki T: **A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium.** *Genetics* 2004, **168(2):**1087-1096.
6. Anthony RV, Scheaffer AN, Wright CD, Regnault TR: **Ruminant models of prenatal growth restriction.** *Reprod Suppl* 2003, **61:**183-194.
7. Harris A: **Towards an ovine model of cystic fibrosis.** *Hum Mol Genet* 1997, **6(13):**2191-2194.
8. McMillen IC, Adam CL, Muhlhausler BS: **Early origins of obesity: programming the appetite regulatory system.** *J Physiol* 2005, **565(Pt 1):**9-17.
9. Prather RS, Hawley RJ, Carter DB, Lai L, Greenstein JL: **Transgenic swine for biomedicine and agriculture.** *Theriogenology* 2003, **59(1):**115-123.
10. Steffen DJ, Elliott GS, Leipold HW, Smith JE: **Congenital dyserythropoiesis and progressive alopecia in Polled Hereford calves: hematologic, biochemical, bone marrow cytologic, electrophoretic, and flow cytometric findings.** *J Vet Diagn Invest* 1992, **4(1):**31-37.
11. Kahn LH: **Confronting zoonoses, linking human and veterinary medicine.** *Emerg Infect Dis [serial on the Internet]* 2006-0956.htm. Available from http://www.cdc.gov/ncidod/EID/vol12no04/05-0956.htm
12. Eyras E, Reymond A, Castelo R, Bye JM, Camara F, Flicek P, Huckle EJ, Parra G, Shteynberg DD, Wyss C, Rogers J, Antonarakis SE, Birney E, Guigo R, Brent MR: **Gene finding in the chicken genome.** *BMC Bioinformatics* 2005, **6(1):**131.
13. Lewis SE: **Gene Ontology: looking backwards and forwards.** *Genome Biol* 2005, **6(1):**103.
14. Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR: **Gramene, a tool for grass genomics.** *Plant Physiol* 2002, **130(4):**1606-1613.
15. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res* 2003, **31(1):**371-373.
16. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32(Database issue):**D262-6.
17. Stein L: **What's Next for Bioinformatics?** *The Scientist* 2005, **19(10):**31.
18. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Res* 2005, **33(Database issue):**D428-32.
19. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.** *Genome Biol* 2005, **6(1):**R9.
20. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation.** *Proteomics* 2004, **4(1):**59-77.
21. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, Kodira CD, Major J, Wang S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM: **The complete genome and proteome of Mycoplasma mobile.** *Genome Res* 2004, **14(8):**1447-1461.
22. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method.** *J Mol Biol* 2004, **338(5):**1027-1036.
23. McCarthy FM, Cooksey AC, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a Whole Organ using Proteomics: the Avian Bursa of Fabricius.** *Proteomics* 2006, **6:**2759-2771.
24. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V: **Complete genomic sequence of Pasteurella multocida, Pm70.** *Proc Natl Acad Sci U S A* 2001, **98(6):**3460-3465.
25. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol* 2002, **3(12):**RESEARCH0082.
26. Kunec D, Nanduri B, Hanson LA, Burgess SC: **Experimental Annotation of the Herpesvirus Genome: May 28 - June 1; Seattle, WA.** ; 2006.
27. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: the proteomics identifications database.** *Proteomics* 2005, **5(13):**3537-3545.
28. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T: **Ensembl 2004.** *Nucleic Acids Res* 2004, **32(Database issue):**D468-70.
29. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33(17):**5691-5702.
30. McCarthy FM, Burgess SC, van den Berg BH, Koter MD, Pharr GT: **Differential detergent fractionation for non-electrophoretic eukaryote cell proteomics.** *J Proteome Res* 2005, **4(2):**316-324.
31. Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L: **Gramene: a resource for comparative grass genomics.** *Nucleic Acids Res* 2002, **30(1):**103-105.
32. Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V: **MaizeGDB, the community database for maize genetics and genomics.** *Nucleic Acids Res* 2004, **32(Database issue):**D393-7.
33. Crittenden L, Bitgood J, Burt D: **Genetic nomenclature guide. Chick.** *Trends Genet* 1995:33-34.
34. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Res* 2005, **33(Database issue):**D71-4.
35. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucleic Acids Res* 2001, **29(1):**159-164.
36. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: **The TIGR gene indices: reconstruction and representation of expressed gene sequences.** *Nucleic Acids Res* 2000, **28(1):**141-145.
37. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32(Database issue):**D258-61.
38. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31(1):**28-33.