1-1-2009

Improving structural and functional annotation of the chicken genome

Teresia Buza

IMPROVING STRUCTURAL AND FUNCTIONAL

ANNOTATION OF THE CHICKEN GENOME

By

Teresia Buza

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Veterinary Medical Science
in the College of Veterinary Medicine

Mississippi State, Mississippi

December 2009

IMPROVING STRUCTURAL AND FUNCTIONAL

ANNOTATION OF THE CHICKEN GENOME

By

Teresia Buza

Approved:

Shane C. Burgess Associate Dean and Professor of CVM Basic Sciences Director of Life Sciences and Biotechnology Institute, Co-director of Institute for Digital Biology (Major Professor and Director of Dissertation)	Fiona M. McCarthy Assistant Professor of CVM Basic Sciences (Co-major Professor)
Larry A. Hanson Professor of CVM Basic Sciences (Graduate Coordinator of CVM Basic Sciences)	Susan M Bridges Professor of Computer Science and Engineering Co-director of Institute for Digital Biology (Committee Member)
Daniel G. Peterson Associate Professor of Plant and Soil Sciences Associate Director of Life Sciences and Biotechnology Institute (Committee Member)	Bindu Nanduri Assistant Professor of CVM Basic Sciences (Committee Member)

Kent H. Hoblet
Dean and Professor of the
College of Veterinary Medicine

Name: Teresia Buza

Date of Degree: December 11, 2009

Institution: Mississippi State University

Major Field: Veterinary Medical Science

Major Professor: Dr. Shane C. Burgess

Title of the Study: IMPROVING STRUCTURAL AND FUNCTIONAL ANNOTATION OF THE CHICKEN GENOME

Pages in Study: 69

Candidate for Degree of Doctor of Philosophy

Chicken is an important non-mammalian vertebrate model organism for biomedical research, especially for vaccine production and the study of embryology and development. Chicken is also an important agricultural species and major food source for high-quality protein worldwide. In addition, chicken is an important model organism for comparative and evolution genomics. Exploitation of this genome as a biomedical model is hindered by its incomplete structural and functional annotation. This incomplete annotation makes it difficult for researchers to model their functional genomics datasets. Improving structural and functional annotation of the chicken genome will allow researchers to derive biological meaning from their functional genomics datasets.

The objectives of this study were to identify proteins expressed in multiple chicken tissues, to functionally annotate experimentally confirmed proteins expressed in

different chicken tissues, to quantify and assess the Gene Ontology (GO) annotation quality, and to facilitate functional annotation of microarray data.

The results of this research have proven to be fundamental resource for improving the structural and functional annotation of chicken genome. Specifically, we have improved the structural annotation of the chicken genome by adding support to predicted proteins. In addition, we have improved the functional annotation of the chicken genome by assigning useful biological information to proteomics datasets and the whole genome chicken array. The Gene Ontology Annotation Quality (GAQ) and Array GO Mapper (AGOM) tools developed in this study will sustainably continue to facilitate functional modeling of chicken arrays and high-throughput experimental datasets from microarray and proteomics studies. The ultimate positive impact of these results is to facilitate the field of biomedical research with useful information for comparative biology, better understanding of chicken biological systems, diseases, drug discovery and eventually development of therapies.

Keywords: Genome annotation, Gene Ontology, proteomics, GO annotation quality, microarray

DEDICATION

I dedicate this dissertation to my parents, the late Mr. Stephen Mrema & Mrs. Felistas Mrema, my husband Joram and my wonderful children Joe, Janeth and Steve

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Shane Burgess, my advisor and mentor, for accepting me into his program which made me build my capacity in the exciting world of genomics and proteomics. His compassion, support and professional guidance have elevated my desire to prosper in the world of cutting-edge research.

I am also grateful to my co-adviser, Dr. Fiona McCarthy for introducing me to the Gene Ontology, a de facto standard for biological functional modeling. Most importantly, her consistently recommendations and suggestions have been very useful for the completion of research work reported in this dissertation.

I am also thankful to my dissertation committee members, Drs. Susan Bridges, Daniel Peterson and Bindu Nanduri. I have heartfelt gratitude for their invaluable suggestions and hard work in reviewing this dissertation. The bioinformatics and bio-computing courses offered by Dr. Susan Bridges and the genome and genomics course offered by Dr. Daniel Peterson have been crucial throughout my entire doctoral study. These courses gave me an insight towards my future career.

Special thanks go to my colleagues in Dr. Burgess' lab (especially Divyaswetha Peddinti, Shyamesh Kumar and Lakshmi Pillai) for providing support and peaceful

working environment while sharing the same office room. They have been my friends indeed.

My sincere gratitude is also extended to the Department of Basic Sciences, College of Veterinary Medicine and the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service for providing the physical and financial resources for this research.

My heartfelt thanks go to my loving mother, the late Mrs. Felistas Mrema who passed away on August 16, 2009 when I was preparing this dissertation. Her everlasting prayers, sacrifice and compassion pushed me to work hard in order to achieve my academic goals. Also my sincere gratitude goes to my brothers Ewald, Simon, Onesfor and Fred, my sisters Betty, Regina, Brigiter, Frida and Ann-Joyce. Their encouragement and dedicated support in many ways have been essential for my entire academic life and career voyage.

My warm thanks go to my husband Joram for his encouragement and assistance. My endless gratitude goes to my wonderful children, Joe, Janeth and Steve, for their love, patience, understanding and encouragement. Their cheerful support during tough and difficult times and their ideas on how to move on, bring me much joy and pride. They have been my happiness and my friends indeed. My gratitude to them cannot be expressed in words.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Chicken (*Gallus gallus*) is an important non-mammalian vertebrate model organism for biomedical research, especially for vaccine production and the study of embryology and development. Chicken is also an important agricultural species and major food source for high-quality protein worldwide. Chicken was the first avian and the first agricultural animal to have its genome sequenced in 2004. After the chicken genome sequence was released scientists started to interpret the raw sequence data into useful biological information, a process known as genome annotation. This process involves comprehensive genome structural annotation mostly performed using automatic tools (*ab-initio method*) to identify structural elements such as open reading frames (ORFs) and their localization, structural description of genes, location of regulatory motifs, protein coding regions, characterization of putative protein products and other features in the primary genomic sequence.

The next step after genome structural annotation is functional annotation - a process where both biological experiments and *in silico* analysis are used to attach biological information to the identified genomic elements. As a biomedical model species, detailed annotation of the chicken genome sequence greatly facilitates

comparative genome studies to accelerate the process of finding the causes of human diseases, drug discovery and therapies development.

The chicken genome sequence provides opportunities for combining new technologies for functional profiling of genome scale experiments. Proteomics and microarray studies (aka: transcriptomics) are among the new technologies currently used to realize biological meaning from the chicken genome sequence. However, exploitation of this genome as a biomedical model organism is hindered by its poor structural and functional annotation because researchers find it difficult to model their proteomics and transcriptomics datasets to biological systems. About 42% of the chicken proteins that have been predicted by *ab initio* methods have not been confirmed experimentally and therefore, there is no functional information that is associated with these proteins. On the other hand, the chicken genome array, which enables researchers to simultaneously monitor genome-wide expression profiles, is associated with little structural and/or less detailed functional annotation information.

Our central objective is to improve the structural and functional annotation of the chicken genome. In this objective we want to confirm the expression of chicken predicted proteins *in vivo*. In addition we would like to provide researchers with tools and biological functional information for modeling their proteomics (proteins) and transcriptomics (microarray) datasets. To achieve the central objective we implemented the following specific objectives: (1) to identify chicken predicted proteins expressed in multiple tissues; (2) to use Gene Ontology (GO) standards to functionally annotate experimentally confirmed proteins that were expressed in multiple chicken tissues; (3) to develop a tool that will help assess and track

improvement of the functional annotation quality in chicken and other eukaryotes; and (4) to facilitate functional annotation of chicken microarray data by developing GO mapping tool(s).

CHAPTER 2

REVIEW OF PERTINENT LITERATURE

Importance of chicken

Chicken (*Gallus gallus*) is an important agricultural species and major food source for high-quality protein worldwide. In the United States alone, more than 9 billion chicken are produced for meat yearly with a value exceeding $20 billion [1]. Chicken was the first farm animal and non-mammalian vertebrate to have its genome completely sequenced in 2004 [2]. The completion of this genome has raised the status of chicken as an important animal model for biomedical research [3] especially in the fields of evolution [4,5], immunology [6], oncology [7-9], virology [10-13], embryology and development [14,15], as well as comparative genomics[16-19].

Genome structural annotation

The chicken genome contains 1.2 billion base pairs of DNA divided into 40 chromosomes of different lengths, designated as large macro-chromosomes (Chr. 1–5), intermediate chromosomes (Chr. 6–10) and micro-chromosomes (Chr. 11–38) as well as sex chromosomes Z and W [20,21]. Unlike mammals, the chicken males are homogametic (Z/Z), while the females are heterogametic (Z/W). After genome sequencing and assembly processes are completed, researchers start to convert the

sequence into a meaningful information that relates to the biology of the organism [22]. The first phase in the genome annotation is known as *structural annotation*. This process uses gene prediction tools to identify the open reading frames (ORFs) and their localization [23,24], gene structure [25,26], protein coding genes [27,28] and regulatory motifs [29].

The sequencing of the chicken genome, in combination with advances in computing technology, has resulted in rapid advances in discovery of genes and other functional elements. The structural annotation statistics of the current assembly of the chicken genome (http://www.ncbi.nlm.nih.gov/, Build 2.1, 03/14/2009) estimates 19,936 genes that encode nearly 34,209 proteins. Complete structural annotation is an essential tool as chicken researchers investigate the biology of this potent biomedical model organism. Improved genome annotation has been realized in other organisms through a combination of comparative and *ab initio* gene prediction algorithms [27]. While structural annotation identifies the functional elements, it should be distinguished from the identifying functions of the elements (refer section 2.3).

Genome structural annotation by proteomics approach

Incomplete structural annotation of the chicken genome poses a challenge to researchers who want to derive value from their experimental datasets. Currently, 42% of chicken gene products are based on computational predictions which lack any experimental information (http://www.ncbi.nlm.nih.gov/; 03/14/2009). These gene products need to be experimentally confirmed through a series of functional genomics experiments such as proteomics. Mass spectrometry is a technology in the field of

proteomics that produces tandem mass spectra (MS/MS) to enable scientists to identify and quantify the entire complement of proteins (proteome) in a complex biological sample [27,30,31]. Traditionally, proteomics relies on matching peptide sequences from a protein database with experimental MS/MS spectra to identify proteins. Given MS/MS spectra, programs such as SEQUEST can identify the peptide which produced it by comparing the experimental MS/MS spectra (found in sample) against theoretical spectra (*in-silico* generated) and return best matches in form of amino acid sequences [23,32,33].

High-throughput expression proteomics for rapid experimental structural annotation have been demonstrated in a study which involved multiple chicken tissues [34]. A limitation of proteomics is that it can only detect peptides of proteins present in database. If the protein is not in the searched database, it will never be identified, despite its presence in the sample. This is a significant problem with the newly sequenced or poorly annotated genomes which have just a fraction of known or predicted proteins in the public databases. However, biological modeling of high-throughput datasets requires that we know all the components in the complex biological system of an organism. This can be partially achieved through comparative genome analysis between poor and better annotated genomes.

At the genomic level, tandem mass spectrometry (MS/MS) allows researchers to experimentally validate computationally predicted open reading frames in a high-throughput manner [31,35,36] as well as making novel gene predictions [23,37]. This procedure is known as *proteogenomics* [31,38]. Proteogenomics matches

experimental MS/MS spectra against genomic sequences [30,35,39] and can even involve multiple genome alignments to make gene predictions [23,37]. Searching or aligning a whole eukaryotic genome such as that of chicken (1.2 Gb) for novel gene predictions require faster tools such as BLAT [40]. BLAT is a BLAST-Like Alignment Tool which has been observed to be more accurate and 500 times faster than popular existing tools for mRNA/DNA alignments and 50 times faster for protein alignments [40]. BLAT uses an index of all non-overlapping *K-mers* in the genome and this can fit inside the RAM of normal computers.

Genome functional annotation

The function of genomic elements is determined through a process known as *functional annotation*. In this process the gene or gene products are linked with functional information using Gene Ontology (GO) standards [41]. Gene ontology contains standardized vocabularies of terms that are organized into three categories representing molecular functions, biological processes, and cellular component [41,42] Basically, molecular function terms describe the biochemical activity performed by a gene product (e.g. kinase activity) whereas biological process terms describe the ordered assembly of more than one molecular function (e.g. limb development) and cellular component terms describe the cellular location (e.g. nucleus). It is good to note that GO annotations are always based on the characteristics of gene products, even though it may be the gene that is cited in the annotation [43].

Various groups such as AgBase [44] and UniProtKB [45] continuously annotate chicken gene products with GO. This helps researchers to access already existing functional information. Other groups such as NetAffx [46] annotates the gene products linked to probesets in Affymetrix chicken arrays. The current statistics of chicken GO annotation (http://www.geneontology.org/GO.current.annotations.shtml; 03/14/2009) shows that there are 64,093 GO annotations associated with 16,353 proteins. The fraction of proteins associated with GO is nearly 48% (chicken build 2.1) and over 98% of these annotations are inferred from electronic annotation (IEA). The IEA annotations are obtained using InterPro tool and InterProScan software [47] which searches the protein sequences to identify signatures from the InterPro member databases i.e. Pfam [48], PROSITE [49], PRINTS [50], ProDom [51], SMART [52], TIGRFAMs [53], PIRSF [54,55], SUPERFAMILY [56], Gene3D [57], and PANTHER [54]. By using InterProScan software, the AgBase [44] biocurators have been able to provide a *breadth* of GO annotation coverage for a poorly annotated chicken genome. It should be noted the most of IEA annotations represent general functions of a gene product in contrast with the direct experimental-based annotation of functional literature which provides detailed, organism specific functional annotation [45,58]. So far, less than 1% of chicken GO annotations are based on direct experimental evidence.

Functional annotation of gene products by orthology method

In comparative genomics the transfer of functional annotations from one species to the other is one of the main applications of comparative genomics

[34,59,60]. The key concern is that the annotators need to know where the functions are transferred from. Orthology is currently the most logical way of assigning functions to gene products when there is no direct experimental evidence available [34]. The term orthology describes the evolutionary relationship between homologous genes in different species that have been derived from a single gene in the last common ancestor [61], and since orthologous pairs have minimum level of evolutionary separation between them, they are more likely to retain a common function [62,63]. Orthology is different from the paralogy - the latter describes the relationship between two genes that arose through duplication within the same species and may not have the same function [61]. After speciation, if an ortholog undergoes duplication in one species, the resulting orthologs are referred to as inparalogs [63,64], indicating paralogs that arose through a gene duplication event after speciation. Paralogs are also commonly referred to as outparalogs especially in cases where the inparalog term is used. [63,65]. Unlike outparalogs, inparalogs can form a group of genes that together are orthologous to a gene in another species.

There are various tools for ortholog prediction [66-68], ortholog databases and search tools [60,65,67]. Software such as Biomart [69] can be used to search orthologs for a given set of gene products and also retrieve the GO annotations for the orthologs. Biocurators at AgBase [44] continuously use orthology to annotate chicken predicted proteins. These predicted proteins are normally not assigned any GO during the assembly process because they are not experimentally confirmed but have only been predicted by *ab initio* methods. However, transferring of functional information

should be done with caution because, for most species most of their GO annotations are from electronic prediction. The most reliable GO annotations to be transferred are the ones associated with experimental evidence codes, a method always adopted in a previous study [34] by AgBase biocurators [44].

The quality of GO annotation

Gene Ontology (GO) vocabularies [41,42] have been widely used in various species to facilitate proteome [70,71] and microarray [72-75] data interpretation. The statistics of GO annotation as submitted by various GO consortium members show great variation in terms of the amount of information represented by specific projects (http://www.geneontology.org/GO.current.annotations.shtml). As reported in a previous study, looking only at the volume of annotations does not directly help researchers to correlate the amount and quality of GO annotations especially between different sets of gene products in different species [76]. One common characteristic feature of all annotations is the large fraction of electronic annotations. However, proportionally, some species such as human and mouse have more experimental and manually checked computational annotations than, for example, chicken. Experimental GO annotations which are obtained from literature by skilled biologists generates high-quality reliable information that is more accurate, reliable and detailed than electronic annotation [45,58]. Nevertheless, reading literature is very time consuming and more labor-intensive. Computational GO annotation continues to be the most rigorous method for annotation of the high-throughput data generated from microarray and proteomics studies.

Efforts to maintain the quality of available annotations is extremely important. Some quality measures that have been used in GO annotation are mainly based on maintaining the consistency and accuracy of annotation [45,58,77]. These measures are employed to minimize variability in annotation between curators, make sure all the necessary fields are complete in an annotation, check for data integrity and updating annotations based on a combination of evidence codes. Frequently, evidence codes linked with annotations has been used to measure the quality of annotation. For example, some groups replaces annotations with IEA evidence code with non-IEA based annotation if the term is in the same ontology [77].

Integration of features of GO annotation such as the number of annotations (breadth), the level of annotation detail (depth) and the evidence for the annotation (quality) has been recommended as a more precise way of assessing the quality of GO annotation [76], and at the same time monitoring the quality of GO over time. In this method the evidence codes are ranked based on whether they represent direct experimental evidence or indirect evidence. For example, direct experimental evidence codes such as IDA, IMP, IGI, IPI and EXP shown on Table 2.1 can be given higher ranks than the computational evidence codes because the functional information associated with these codes has been proven by specific direct experiments. Consideration of the depth of GO annotation will give a direct guide to researchers because GO is organized as a hierarchy of terms in a Directed Acyclic Graph (DAG) [78]. In this structure more general term such as 'growth' lead to more specific terms such as 'organ growth', 'heart growth', and 'cardiac muscle tissue

growth', allowing gene products to be annotated to any level of specificity as the biological understanding allows.

Functional annotation of chicken array

Microarray technologies such as cDNA [79,80] and oligonucleotide probe [81] arrays have emerged as important tools in functional genomics for global analysis of gene expression and biological systems in chicken. A number of microarray screening platforms have been developed to study differential gene expression occurring in chicken as a response to different challenges and stimuli [6,82]. In the chicken research community, microarrays are used for a wide range of applications including not only gene expression analysis [83,84] but also exon expression analysis [85-87], novel transcript discovery [88], genotyping [89,90], resequencing [91,92] and in identification of transcription factors along with their respective binding sites [93].

The common problem in microarray data analysis is biological interpretation of the results. The Gene Ontology (GO) [41,42,58] has been the *de facto* functional annotation method for array modeling [73-75,94]. In GO, the proteins are the one annotated to either molecular function, biological process or cellular component but the annotations are assigned to the respective gene that codes the protein being annotated. Most microarrays generated by the chicken research community are deposited in the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) [95]. These microarray data can easily be browsed, queried and retrieved for further studies [96]. Most of the information

represented on the chicken microarray platform at GEO is insufficient for biological interpretation of any results obtained from microarray studies. One of the reasons for this is that chicken microarrays are mostly developed from cDNA and ESTs. These cDNA and ESTs have not been structurally linked to protein accessions that can be annotated to GO. Of all the chicken arrays, the Affymetrix GenChip chicken genome array has been annotated to GO [46].

The Affymetrix GeneChip chicken genome array has been extensively used in different studies such as gene expression profiling in chicken and avian viruses [84,97-99]. The current array (NetAffx build 29) contains coverage of 37,703 probesets for spotting 32,774 transcripts corresponding to nearly 28,000 chicken genes. In addition, it contains 689 probesets for detecting 684 transcripts from 17 avian viruses. NetAffx [46] links probesets on Affymetrix GenChip microarrays to GO and has developed GO mining tool to give a picture of GO graph relationships [94]. However, these annotations are far from complete because they lack important features such as references used to make functional assertions (See: http://www.affymetrix.com/support/support_result.affx). The annotations of this array, if improved, can facilitate annotation of other arrays and even experimental microarray datasets because of its comprehensive coverage of transcripts, genes and GO information. In addition, this array is linked to cross reference (over six different types of gene identifiers and protein accessions) that can be used to facilitate mapping to similar accessions from other chicken arrays and experimental datasets. Identifiers that are represented and may be used for mapping are Probe set ID linked to GenBank

mRNA, Gene Symbol, UniGene ID, Entrez Gene ID, Ensembl gene ID, SwissProt accession, RefSeq Protein ID, RefSeq Transcript ID and InterPro, all in separate columns. Improving the amount and quality of GO annotation linked to gene products represented on the Affymetrix GenChip chicken genome array may form a more comprehensive database for chicken microarray structural and functional annotation.

References

1. USDA: U.S. Department of Agriculture National Agricultural Statistics Service. 2008. Poultry slaughter: 2007 annual summary. In.; 2008.

2. Chicken-Genome: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004, 432(7018):695-716.

3. Stern CD: The chick; a great model system becomes even greater. *Dev Cell* 2005, 8(1):9-17.

4. Temperley ND, Berlin S, Paton IR, Griffin DK, Burt DW: Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. *BMC Genomics* 2008, 9:62.

5. Nikolaidis N, Makalowska I, Chalkia D, Makalowski W, Klein J, Nei M: Origin and evolution of the chicken leukocyte receptor complex. *Proc Natl Acad Sci U S A* 2005, 102(11):4057-4062.

6. Smith J, Speed D, Hocking PM, Talbot RT, Degen WG, Schijns VE, Glass EJ, Burt DW: Development of a chicken 5 K microarray targeted towards immune function. *BMC Genomics* 2006, 7:49.

7. Buza JJ, Burgess SC: Modeling the proteome of a Marek's disease transformed cell line: a natural animal model for CD30 overexpressing lymphomas. *Proteomics* 2007, 7(8):1316-1326.

8. Shack LA, Buza JJ, Burgess SC: The neoplastically transformed (CD30hi) Marek's disease lymphoma cell phenotype most closely resembles T-regulatory cells. *Cancer Immunol Immunother* 2008, 57(8):1253-1262.

9. Thanthrige-Don N, Abdul-Careem MF, Shack LA, Burgess SC, Sharif S: Analyses of the spleen proteome of chickens infected with Marek's disease virus. *Virology* 2009, 390(2):356-367.

10. Borges MB, Caride E, Jabor AV, Malachias JM, Freire MS, Homma A, Galler R: Study of the genetic stability of measles virus CAM-70 vaccine strain after serial

passages in chicken embryo fibroblasts primary cultures. *Virus Genes* 2008, 36(1):35-44.

11. Kaffashi A, Shrestha S, Browning GF: Evaluation of chicken anaemia virus mutants as potential vaccine strains in 1-day-old chickens. *Avian Pathol* 2008, 37(1):109-114.

12. Natesan S, Kataria JM, Dhama K, Rahul S, Bhardwaj N: Biological and molecular characterization of chicken anaemia virus isolates of Indian origin. *Virus Res* 2006, 118(1-2):78-86.

13. Zsak L, Strother KO, Day JM: Development of a polymerase chain reaction procedure for detection of chicken and turkey parvoviruses. *Avian Dis* 2009, 53(1):83-88.

14. Feng Y, Zhang S, Peng X, Yuan J, Yang Y, Zhan H, Gong Y: Expression analysis of genes putatively involved in chicken gonadal development. *Acta Biol Hung* 2007, 58(2):163-172.

15. Porter TE, Ghavam S, Muchow M, Bossis I, Ellestad L: Cloning of partial cDNAs for the chicken glucocorticoid and mineralocorticoid receptors and characterization of mRNA levels in the anterior pituitary gland during chick embryonic development. *Domest Anim Endocrinol* 2007, 33(2):226-239.

16. Castelo R, Reymond A, Wyss C, Camara F, Parra G, Antonarakis SE, Guigo R, Eyras E: Comparative gene finding in chicken indicates that we are closing in on the set of multi-exonic widely expressed human genes. *Nucleic Acids Res* 2005, 33(6):1935-1939.

17. Nie W, O'Brien PC, Ng BL, Fu B, Volobouev V, Carter NP, Ferguson-Smith MA, Yang F: Avian comparative genomics: reciprocal chromosome painting between domestic chicken (Gallus gallus) and the stone curlew (Burhinus oedicnemus, Charadriiformes)--an atypical species with low diploid number. *Chromosome Res* 2009, 17(1):99-113.

18. Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, Joens LA, Konkel ME, Angly F, Dinsdale EA, Edwards RA *et al*: Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS One* 2008, 3(8):e2945.

19. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RP, Groenen MA, Deryusheva S, Gaginskaya E, Carre W *et al*: Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics* 2008, 9:168.

20. Burt DW: Origin and evolution of avian microchromosomes. *Cytogenet Genome Res* 2002, 96(1-4):97-112.

21. ICGSC: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004, 432(7018):695-716.

22. Stein L: Genome annotation: from sequence to biology. *Nat Rev Genet* 2001, 2(7):493-503.

23. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ: Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* 2006, 7(4):R35.

24. Hirosawa M, Isono K, Hayes W, Borodovsky M: Gene identification and classification in the Synechocystis genomic sequence by recursive gene mark analysis. *DNA Seq* 1997, 8(1-2):17-29.

25. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997, 268(1):78-94.

26. van Baren MJ, Koebbe BC, Brent MR: Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences. *Curr Protoc Bioinformatics* 2007, Chapter 4:Unit 4 8.

27. Li J, Riehle MM, Zhang Y, Xu J, Oduol F, Gomez SM, Eiglmeier K, Ueberheide BM, Shabanowitz J, Hunt DF *et al*: Anopheles gambiae genome reannotation through synthesis of ab initio and comparative gene prediction algorithms. *Genome Biol* 2006, 7(3):R24.

28. Liu J, Gough J, Rost B: Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet* 2006, 2(4):e29.

29. Jolly ER, Chin CS, Herskowitz I, Li H: Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis. *BMC Bioinformatics* 2005, 6:275.

30. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O: Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res* 2009, 19(1):128-135.

31. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD *et al*: Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* 2007, 17(9):1362-1377.

32. Edwards NJ: Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol Syst Biol* 2007, 3:102.

33. Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ: Exploiting proteomic data for genome annotation and gene model validation in Aspergillus niger. *BMC Genomics* 2009, 10:61.

34. Buza TJ, McCarthy FM, Burgess SC: Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC Genomics* 2007, 8:425.

35. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J *et al*: Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res* 2008, 18(7):1133-1142.

36. Jaffe JD, Berg HC, Church GM: Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004, 4(1):59-77.

37. Xia J, Radford C, Guo X, Magor KE: Immune gene discovery by expressed sequence tag analysis of spleen in the duck (Anas platyrhynchos). *Dev Comp Immunol* 2007, 31(3):272-285.

38. Sigdel TK, Sarwal MM: The proteogenomic path towards biomarker discovery. *Pediatr Transplant* 2008.

39. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD: Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct Genomic Proteomic* 2008, 7(1):50-62.

40. Kent WJ: BLAT--the BLAST-like alignment tool. *Genome Res* 2002, 12(4):656-664.

41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.

42. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, 32(Database issue):D258-261.

43. Lomax J: Get ready to GO! A biologist's guide to the Gene Ontology. *Brief Bioinform* 2005, 6(3):298-304.

44. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC: AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res* 2007, 35(Database issue):D599-603.

45. Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R: An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 2005, 6 Suppl 1:S17.

46. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA: NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 2003, 31(1):82-86.

47. Mulder N, Apweiler R: InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 2007, 396:59-70.

48. Sonnhammer EL, Eddy SR, Durbin R: Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997, 28(3):405-420.

49. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: The PROSITE database. *Nucleic Acids Res* 2006, 34(Database issue):D227-230.

50. Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN: The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res* 1998, 26(1):304-308.

51. Corpet F, Servant F, Gouzy J, Kahn D: ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 2000, 28(1):267-269.

52. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* 2000, 28(1):231-234.

53. Haft DH, Selengut JD, White O: The TIGRFAMs database of protein families. *Nucleic Acids Res* 2003, 31(1):371-373.

54. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ *et al*: The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005, 33(Database issue):D284-288.

55. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P *et al*: PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res* 2004, 32(Database issue):D112-114.

56. Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan N: SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res* 2002, 30(1):289-293.

57. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C: Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res* 2008, 36(Database issue):D414-418.

58. 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):D262-266.

59. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL: InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 2008, 36(Database issue):D263-266.

60. Eyre TA, Wright MW, Lush MJ, Bruford EA: HCOP: a searchable database of human orthology predictions. *Brief Bioinform* 2007, 8(1):2-5.

61. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19(2):99-113.

62. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.

63. Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314(5):1041-1052.

64. Sonnhammer EL, Koonin EV: Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 2002, 18(12):619-620.

65. O'Brien KP, Remm M, Sonnhammer EL: Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 2005, 33(Database issue):D476-480.

66. van der Heijden RT, Snel B, van Noort V, Huynen MA: Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 2007, 8:83.

67. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006, 34(Database issue):D363-368.

68. Wright MW, Eyre TA, Lush MJ, Povey S, Bruford EA: HCOP: the HGNC comparison of orthology predictions search tool. *Mamm Genome* 2005, 16(11):827-828.

69. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005, 21(16):3439-3440.

70. Carvalho PC, Fischer JS, Chen EI, Domont GB, Carvalho MG, Degrave WM, Yates JR, 3rd, Barbosa VC: GO Explorer: A gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Sci* 2009, 7:6.

71. Ngoka LC: Sample prep for proteomics of breast cancer: proteomics and gene ontology reveal dramatic differences in protein solubilization preferences of radioimmunoprecipitation assay and urea lysis buffers. *Proteome Sci* 2008, 6:30.

72. Alvesalo J, Greco D, Leinonen M, Raitila T, Vuorela P, Auvinen P: Microarray analysis of a Chlamydia pneumoniae-infected human epithelial cell line by use of gene ontology hierarchy. *J Infect Dis* 2008, 197(1):156-162.

73. Ochs MF, Peterson AJ, Kossenkov A, Bidaut G: Incorporation of gene ontology annotations to enhance microarray data analysis. *Methods Mol Biol* 2007, 377:243-254.

74. Osborne JD, Zhu LJ, Lin SM, Kibbe WA: Interpreting microarray results with gene ontology and MeSH. *Methods Mol Biol* 2007, 377:223-242.

75. Papachristoudis G, Diplaris S, Mitkas PA: SoFoCles: Feature filtering for microarray classification based on Gene Ontology. *J Biomed Inform* 2009.

76. Buza TJ, McCarthy FM, Wang N, Bridges SM, Burgess SC: Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res* 2008, 36(2):e12.

77. Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G *et al*: Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol* 2004, 135(2):745-755.

78. del Pozo A, Pazos F, Valencia A: Defining functional distances over gene ontology. *BMC Bioinformatics* 2008, 9:50.

79. Afrakhte M, Schultheiss TM: Construction and analysis of a subtracted library and microarray of cDNAs expressed specifically in chicken heart progenitor cells. *Dev Dyn* 2004, 230(2):290-298.

80. Burnside J, Neiman P, Tang J, Basom R, Talbot R, Aronszajn M, Burt D, Delrow J: Development of a cDNA array for chicken gene expression analysis. *BMC Genomics* 2005, 6(1):13.

81. Li X, Chiang HI, Zhu J, Dowd SE, Zhou H: Characterization of a newly developed chicken 44K Agilent microarray. *BMC Genomics* 2008, 9:60.

82. Sarson AJ, Read LR, Haghighi HR, Lambourne MD, Brisbin JT, Zhou H, Sharif S: Construction of a microarray specific to the chicken immune system: profiling gene

expression in B cells after lipopolysaccharide stimulation. *Can J Vet Res* 2007, 71(2):108-118.

83. Heidari M, Huebner M, Kireev D, Silva RF: Transcriptional profiling of Marek's disease virus genes during cytolytic and latent infection. *Virus Genes* 2008, 36(2):383-392.

84. Masker K, Golden A, Gaffney CJ, Mazack V, Schwindinger WF, Zhang W, Wang LH, Carey DJ, Sudol M: Transcriptional profile of Rous Sarcoma Virus transformed chicken embryo fibroblasts reveals new signaling targets of viral-src. *Virology* 2007, 364(1):10-20.

85. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE: Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol* 2007, 8(4):R64.

86. Grigoryev DN, Ma SF, Shimoda LA, Johns RA, Lee B, Garcia JG: Exon-based mapping of microarray probes: recovering differential gene expression signal in underpowered hypoxia experiment. *Mol Cell Probes* 2007, 21(2):134-139.

87. Xing Y, Kapur K, Wong WH: Probe selection and expression index computation of Affymetrix Exon Arrays. *PLoS ONE* 2006, 1:e88.

88. Cao W, Epstein C, Liu H, DeLoughery C, Ge N, Lin J, Diao R, Cao H, Long F, Zhang X *et al*: Comparing gene discovery from Affymetrix GeneChip microarrays and Clontech PCR-select cDNA subtraction: a case study. *BMC Genomics* 2004, 5(1):26.

89. Butcher LM, Meaburn E, Liu L, Fernandes C, Hill L, Al-Chalabi A, Plomin R, Schalkwyk L, Craig IW: Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav Genet* 2004, 34(5):549-555.

90. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I, Schalkwyk LC: Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* 2005, 6(1):52.

91. Corless CE, Kaczmarski E, Borrow R, Guiver M: Molecular characterization of Neisseria meningitidis isolates using a resequencing DNA microarray. *J Mol Diagn* 2008, 10(3):265-271.

92. Lebet T, Chiles R, Hsu AP, Mansfield ES, Warrington JA, Puck JM: Mutations causing severe combined immunodeficiency: detection with a custom resequencing microarray. *Genet Med* 2008, 10(8):575-585.

93. Chung HR, Kostka D, Vingron M: A physical model for tiling array analysis. *Bioinformatics* 2007, 23(13):i80-86.

94. Cheng J, Sun S, Tracy A, Hubbell E, Morris J, Valmeekam V, Kimbrough A, Cline MS, Liu G, Shigeta R *et al*: NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics* 2004, 20(9):1462-1463.

95. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA *et al*: NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009, 37(Database issue):D885-890.

96. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y: GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics* 2008, 24(23):2798-2800.

97. Lee SI, Lee WK, Shin JH, Han BK, Moon S, Cho S, Park T, Kim H, Han JY: Sexually dimorphic gene expression in the chick brain before gonadal differentiation. *Poult Sci* 2009, 88(5):1003-1015.

98. Schippert R, Schaeffel F, Feldkaemper MP: Microarray analysis of retinal gene expression in chicks during imposed myopic defocus. *Mol Vis* 2008, 14:1589-1599.

99. Li H, Gilbert ER, Zhang Y, Crasta O, Emmerson D, Webb KE, Jr., Wong EA: Expression profiling of the solute carrier gene family in chicken intestine from the late embryonic to early post-hatch stages. *Anim Genet* 2008, 39(4):407-424.

Table 2.1 The current evidence codes approved by the Gene Ontology consortium

Types of evidence codes	Description
Experimental Evidence Codes	
1. EXP	Inferred from Experiment
2. IDA	Inferred from Direct Assay
3. IPI	Inferred from Physical Interaction
4. IMP	Inferred from Mutant Phenotype
5. IGI	Inferred from Genetic Interaction
6. IEP	Inferred from Expression Pattern
Computational Analysis Evidence Codes	
7. ISS	Inferred from Sequence or Structural Similarity
8. ISO	Inferred from Sequence Orthology
9. ISA	Inferred from Sequence Alignment
10. ISM	Inferred from Sequence Model
11. IGC	Inferred from Genomic Context
12. RCA	inferred from Reviewed Computational Analysis
Author Statement Evidence Codes	
13. TAS	Traceable Author Statement
14. NAS	Non-traceable Author Statement
Curator Statement Evidence Codes	
15. IC	Inferred by Curator
16. ND	No biological Data available
Automatically-assigned Evidence Codes	
17. IEA	Inferred from Electronic Annotation
Obsolete Evidence Codes	
18. NR	Not Recorded

Source: http://www.geneontology.org/GO.evidence.shtml

CHAPTER 3

EXPERIMENTAL CONFIRMATION AND FUNCTIONAL ANNOTATION OF PREDICTED PROTEINS IN THE CHICKEN GENOME[1]

[1] Reprint from T.J. Buza, F.M. McCarthy, and S.C. Burgess. 2007. Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC Genomics* **8:** 42. This article is available from: http://www.biomedcentral.com/1471-2164/8/425

BMC Genomics

BioMed Central

Research article

Open Access

Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome

Teresia J Buza[†1,2], Fiona M McCarthy*[†1,2] and Shane C Burgess[1,2,3,4]

Address: [1]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA, [2]Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA, [3]Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS 39762, USA and [4]Mississippi Agricultural and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA

Email: Teresia J Buza - tbuza@cvm.msstate.edu; Fiona M McCarthy* - fmccarthy@cvm.msstate.edu; Shane C Burgess - burgess@cvm.msstate.edu

* Corresponding author †Equal contributors

Published: 19 November 2007

BMC Genomics 2007, **8**:425 doi:10.1186/1471-2164-8-425

Received: 7 August 2007
Accepted: 19 November 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/425

Abstract

Background: The chicken genome was sequenced because of its phylogenetic position as a non-mammalian vertebrate, its use as a biomedical model especially to study embryology and development, its role as a source of human disease organisms and its importance as the major source of animal derived food protein. However, genomic sequence data is, in itself, of limited value; generally it is not equivalent to understanding biological function. The benefit of having a genome sequence is that it provides a basis for functional genomics. However, the sequence data currently available is poorly structurally and functionally annotated and many genes do not have standard nomenclature assigned.

Results: We analysed eight chicken tissues and improved the chicken genome structural annotation by providing experimental support for the *in vivo* expression of 7,809 computationally predicted proteins, including 30 chicken proteins that were only electronically predicted or hypothetical translations in human. To improve functional annotation (based on Gene Ontology), we mapped these identified proteins to their human and mouse orthologs and used this orthology to transfer Gene Ontology (GO) functional annotations to the chicken proteins. The 8,213 orthology-based GO annotations that we produced represent an 8% increase in currently available chicken GO annotations. Orthologous chicken products were also assigned standardized nomenclature based on current chicken nomenclature guidelines.

Conclusion: We demonstrate the utility of high-throughput expression proteomics for rapid experimental structural annotation of a newly sequenced eukaryote genome. These experimentally-supported predicted proteins were further annotated by assigning the proteins with standardized nomenclature and functional annotation. This method is widely applicable to a diverse range of species. Moreover, information from one genome can be used to improve the annotation of other genomes and inform gene prediction algorithms.

Background

After genome sequencing, genome annotation is critical to denote and demarcate the functional elements in the genome (structural annotation) and to link these genomic elements to biological function (functional annotation). Structural annotation of newly sequenced genomes begins during the final stages of genome assembly with electronic prediction of open reading frames (ORFs) [1-3]. Sequencing consortiums typically release these predicted genes and their translated products into public databases, where they account for the majority of data for the newly sequenced species [4,5] and are critical for high-throughput wet lab functional genomics (microarray and proteomics) experiments [4,6]. The NCBI Non-Redundant Protein Database (NRPD) and the UniProt Archive (UniParc) do not directly provide functional annotation for these predicted ORFs. The highly curated UniProt Knowledgebase (UniProtKB) database [7] displays functional annotation from the European Bioinformatics Institute Gene Ontology Annotation (EBI-GOA) Project [8], but does not include predicted gene products until there is experimental evidence for their *in vivo* expression. Thus, despite being critical for functional genomics experiments, most data from a newly sequenced genome does not have even preliminary functional annotation. This problem is exacerbated as other public resources such as Ensembl [9]. Entrez Gene [10] and Affymetrix Netaffx [11] use data from UniProtKB or the EBI-GOA Project as their functional annotation source.

GO has become the de facto standard for functional annotation [12]. Annotations are attributed to sources (e.g. a PubMed ID) and to the type of evidence used to make the association (indicated by evidence codes; Table 1). Many of the evidence codes describe direct species-specific experimental evidence such as "inferred from direct assay" (IDA), "physical interaction" (IPI), "mutant phenotype" (IMP) or "genetic interaction" (IGI). Other evidence codes refer to indirect lines of evidence such as functional motifs and structural or sequence similarity. However, by definition, there can be no direct experimental evidence available for determining the function of predicted gene products. Instead, adding GO annotations based upon indirect evidence such as "inferred from electronic annotation" (IEA) or "inferred from structural/sequence similarity" (ISS) provide the first significant and valuable increases in the breadth of annotations for functional modelling.

Although most GO annotations for newly sequenced species are the IEA-based annotations provided by the EBI-GOA Project [8], these IEA annotations do not initially include the gene products predicted during sequence assembly. Moreover, while IEA annotations are based on functional motifs and sequences, the most rigorous way of assigning function when there is no direct experimental evidence available, is based on strict orthology. Orthology is one of the central concepts of comparative genome analysis. By definition orthologs are genes or proteins in two or more species that share significant similarity, and are thought to have diverged from a common ancestral gene that existed in their last common ancestor [13-17]. Since orthologous pairs have minimum level of evolutionary separation between them, they are more likely to retain a common function. Determination of orthology relations assists knowledge transfer between species and can be used to improve both structural and functional annotation in organisms that have less annotation.

A number of ortholog prediction methods and search tools are available [9,18-20]. However, the number of proteins from one species that is considered to be part of the same orthologous group varies from one method to another due to different algorithms employed and species included in the methods [14]. For example, Homologene [21] does orthology analyses by comparing protein sequences using the BLASTP tool and then matching the sequences using phylogenetic trees built from sequence similarity and synteny, where possible. Ensembl [9] first uses BLASTP and the Smith-Waterman algorithm to identify putative orthologs by reciprocal BLAST analysis and synteny evidence. Inparanoid [17] is based on pairwise similarity scores and it detects best-best hits between sequences from two different species to form the main orthologous group to which other sequences (in-paralogs) are added only if they are closely related. Treefam (Tree families) [18] uses phylogeny based on Ensembl datasets and clusters genes (and corresponding gene products) from multiple organisms into groups that are all descended from a single ancestor gene. In order to obtain good coverage and reliable predicted orthologs, various methods should be integrated [13].

Comparative genome analysis also requires standardized nomenclature. By identifying orthologs of experimentally supported proteins, standardized nomenclature can be added. Committees for standardized nomenclature exist for human and mouse gene and gene products [22] and chicken researchers have followed suit [23] and will use human nomenclature for orthologous chicken genes.

In this work we analysed nine chicken tissues using a three-stage combined high throughput proteomics and computational biology approach to derive "expressed protein sequence tags" (ePSTs) to improve structural annotation by experimentally supporting the *in vivo* expression of computationally predicted chicken proteins [24]. We then used orthology to add standardized gene nomenclature and GO annotations (by transferring func-

Table 1: Gene Ontology evidence codes

Code	Description	Example
Direct experimental evidence codes		
IDA	Inferred from Direct Assay	enzyme assays
		in vitro reconstitution
		immunofluorescence
		cell fractionation
		physical interaction/binding assay
IGI	Inferred from Genetic Interaction	"traditional" genetic interactions such as suppressors, synthetic lethals, etc.
		functional complementation
		rescue experiments
		inference about one gene drawn from the phenotype of a mutation in a different gene
IMP	Inferred from Mutant Phenotype	any gene mutation/knockout
		overexpression/ectopic expression of wild-type or mutant genes
		anti-sense experiments
		RNAi experiments
		specific protein inhibitors
		polymorphism or allelic variation
IPI	Inferred from Physical Interaction	2-hybrid interactions
		co-purification
		co-immunoprecipitation
		ion/protein binding experiments
IEP	Inferred from Expression Pattern	transcript levels (e.g. Northerns, microarray data)
		protein levels (e.g. Western blots)
Indirect evidence codes		
NAS	Non-traceable Author Statement	Database entries that don't cite a paper
TAS	Traceable Author Statement	original experiments are traceable through that article
IC	Inferred by Curator	inferred by a curator from other GO annotations
IGC	Inferred from Genomic Context	operon structure
		syntenic regions
		pathway analysis
		genome-scale analysis of processes
NR	Not Recorded	used for annotations done before curators began tracking evidence types, not used for new annotations
ND	No biological Data available	"unknown" molecular function, biological process, cellular component
IEA	Inferred from Electronic Annotation	"hits" in sequence similarity searches, if they have not been reviewed by curators; transferred from database records, if not reviewed by curators
ISS	Inferred from Sequence or Structural Similarity	sequence similarity (homologue of/most closely related to)
		recognized domains
		structural similarity
		Southern blotting
		protein features, predicted or observed (e.g. hydrophobicity, sequence composition)
RCA	Inferred from Reviewed Computational Analysis	predictions based on large-scale experiments (e.g. genome-wide two-hybrid)
		predictions based on integration of large-scale datasets of several types
		text-based computation (e.g. text mining)

tional annotations based on direct experimental evidence for corresponding human and mouse orthologs).

Results

Identification of predicted proteins

In total, we identified 7,809 proteins from the analyzed tissues (see additional file 1), corresponding to 51% of the chicken predicted proteins in NCBI (01/08/2007). In doing so, we also obtained data about the tissue expression patterns of these proteins (Figure 1A). By setting $P \leq 0.05$ as a threshold for peptide identification we were able to identify 48,583 peptides that had scores above the threshold in the real database and 438 in the reversed database, giving a peptide false discovery rate (FDR) of

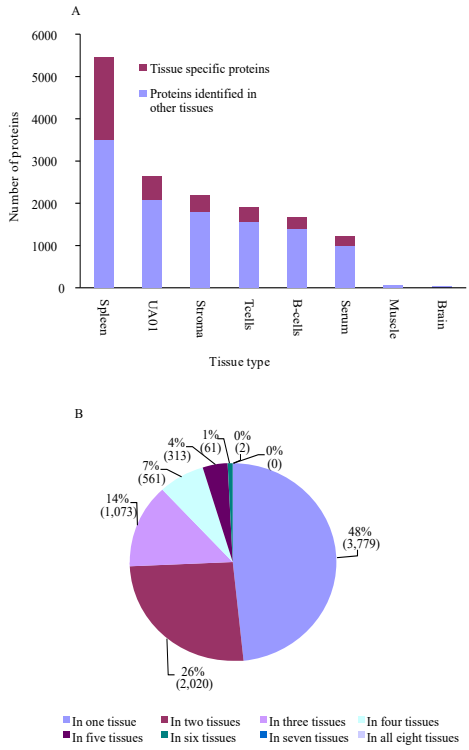

Figure 1
Chicken *predicted* proteins identified from different tissues. Proteomic based analysis was used to demonstrate the *in vivo* expression of electronically predicted chicken proteins. (A) The number of predicted chicken proteins identified from each tissue, with the proportion of proteins that were identified in more than one tissue indicated. (B) The majority of proteins were identified in more than one tissue.

0.9% on the real database. The protein FDR was 1%, equivalent to 78 proteins from this dataset. This FDR is better than recently reported rates [25] and although 4,567 (58%) of the protein identifications in this study were based on single-peptide matches, the low FDR provides a high degree of confidence in these identifications. In other studies, nearly 98% of proteins identified by a single peptide match have been predicted to be correctly identified [26]. Moreover, 44% of the single-peptide matches were identified independently in more than one tissue, providing further evidence for their *in vivo* expression. Interestingly, we identified 30 proteins that were only electronically predicted or hypothetical translations in human.

Not surprisingly, more predicted proteins were identified by mass spectrometry when Differential Detergent Fractionation (DDF) was used as the method for protein isolation, as previously reported [27]. This means that muscle and brain tissues, two tissues which would normally be expected to have the highest number of identified proteins, had the fewest predicted proteins (61 and 36, respectively). We found that 52% of the identified proteins were expressed in more than one tissue (Figure 1B), and their independent identification in multiple tissues lends validity to their *in vivo* expression in chicken. The protein identification and mass spectrometry data has been submitted to the PRoteomic IDEntifications database (PRIDE; [28]), accession numbers 1621–1626, 1654 & 1655.

ID mapping

One of the most time consuming tasks in high-throughput experiments is navigating among different database identifiers. To assist researchers with their data analysis and facilitate data sharing we mapped all identified proteins to UniParc, IPI (International Protein Index), Entrez Gene and Ensembl identifiers (see additional file 2). Only 80% of the identified proteins were mapped to Ensembl IDs. This may be because Ensembl has a different gene prediction method [9] to that of NCBI and not all of the NCBI predicted proteins are represented in Ensembl.

Ortholog identification

We identified human or mouse orthologs for 77% (6,008) of the identified chicken predicted proteins (Figure 2A) and 86% of these orthologs are predicted by more than one ortholog prediction method (Figure 2B). Since each of these tools use different methods for ortholog prediction, orthologs predicted by more than one method are more likely to be accurately predicted.

Standardized nomenclature

The use of standardized nomenclature facilitates comparative biology and aids modelling of functional genomics data. We assigned 5,064 (65%) chicken predicted proteins with HGNC (Human Genome Organization (HUGO) Gene Nomenclature Committee) approved gene symbols and names based on their human or mouse orthologs (see additional file 3). Although it has been agreed to base chicken gene nomenclature on human nomenclature guidelines [23] it is only relatively recently that there has been a concerted effort to provide standardized nomenclature for chicken genes, and the majority of chicken gene products are not named according to standardized nomenclature guidelines. We have assigned standardized nomenclature to chicken genes on a large scale as part of a high-throughput experimental annotation effort.

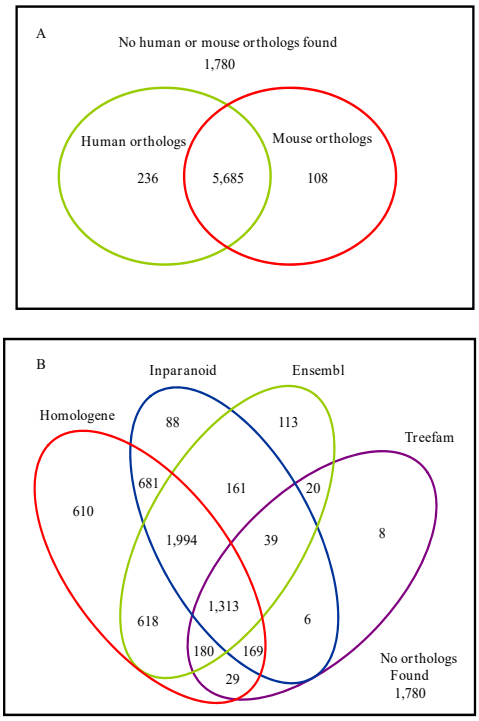

Figure 2
Chicken – human/mouse orthologs. **(A)** The number of identified *predicted* proteins that had either human or mouse 1:1 orthologs. **(B)** Distribution of orthologs identified by different orthology prediction methods. The 4 most commonly used ortholog prediction tools are Homologene, Ensembl, InParanoid and Treefam. Human/mouse orthologs were identified for 77% of the identified chicken proteins (see additional file 3).

Functional Annotation

To functionally annotate the *predicted* proteins we mapped them to the GO annotations for human and mouse orthologs that are based on direct experimental evidence codes (Table 1). We GO annotated 1,651 (21%) chicken *predicted* proteins with 8,213 associations. These GO annotations are summarized based on cellular component (Figure 3), molecular function (Figure 4) and biological process (Figure 5). These GO annotations represent an increase of 8% over the current chicken GO annotations (EBI-GOA, 04/25/2007) and a doubling of chicken non-IEA annotations. These GO annotations are publicly available via the AgBase database [5] and will enter the pipeline to be submitted to the EBI-GOA Project.

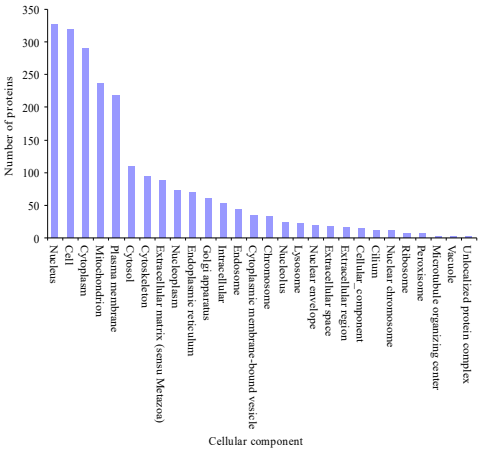

Figure 3
Overview of cellular component transferred to orthologous chicken predicted proteins. The GO annotations are summarized to broad terms of cellular component. These GO annotations are publicly available via the AgBase database [4].

Discussion

Here we demonstrate a combined approach to provide experimental-based structural annotations and functional annotations based on orthology. The workflow we have

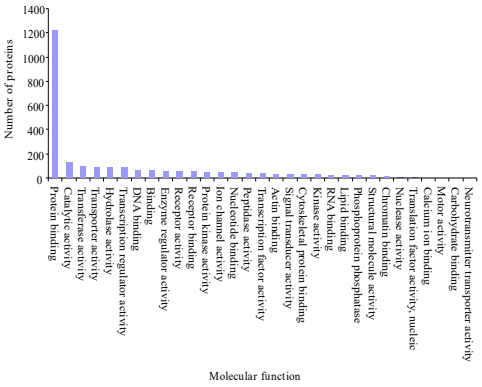

Figure 4
Overview of molecular function transferred to orthologous chicken predicted proteins. The GO annotations are summarized to broad terms of molecular function. These GO annotations are publicly available via the AgBase database [4].

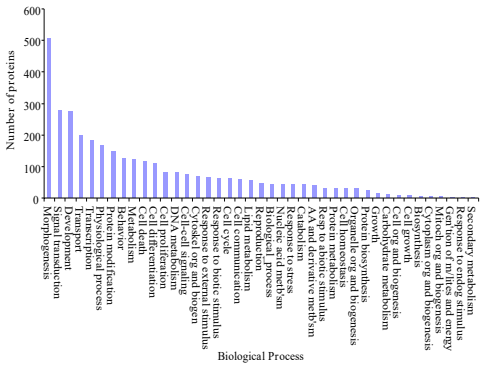

Figure 5
Overview of biological processes transferred to orthologous chicken predicted proteins. The GO annotations are summarized to broad terms of biological processes. These GO annotations are publicly available via the AgBase database [4].

developed relies on using proteomics to survey a range of tissues from the species of interest. Newer structural annotation pipelines include the use of ESTs and mRNA in their computational models. We are proposing an analogous method that would include experimental support at the protein level while providing information that can be used to improve structural annotation in the species being studied, provide information to improve annotation in other species and be used to improve open reading frame prediction algorithms. In addition, providing information about tissue specificity and preliminary functional information based on sequence analysis will facilitate analysis of future functional genomics studies.

The chicken genome was sequenced because of its importance as a non-mammalian vertebrate model, its use as a biomedical model to study embryology and [29,30] development and its agricultural importance. A major step that follows after genome sequencing is structural and functional annotation (denoting and demarcating the functional elements in the genome and link these genomic elements to biological function, respectively). When we began the work described in this manuscript only 53% of chicken proteins were known to be expressed *in vivo*, with the remainder being electronically predicted using in silico methods. Moreover, only 52% of chicken gene products had any GO annotations and, although genes predicted during genome assembly may be the bulk of the data for a newly sequenced species, these predicted gene products are not automatically assigned any GO annotation.

The parameters we have used in this study provide strong support for protein expression *in vivo*. In particular, the parameter DeltaCn is a measure of specificity of the match within the database used and a DeltaCn value 0.1 ensures that a peptide is distinctly different from other peptides within the same database. However, a single peptide match to a predicted protein does not necessarily provide evidence that the annotation for the entire open reading frame is accurate; this can only be confirmed by accumulating more mass spectra data and accounting for the detectable peptides within the genome [31]. While some of the predicted proteins we identified were identified on the basis of a single peptide, 44% of these proteins were expressed in more than one tissue, providing additional evidence for their *in vivo* expression. In a typical proteomics experiment 20–67%-of proteins are identified by a single peptide match [26,32,33]. Calculation of false discovery rate has been used to validate peptide or proteins identifications [32,34-37], including proteins identified by a single peptide match. In one study, 90% of the proteins identified by a single peptide were validated by immunoassay detection [33].

By analysis of multiple tissues we maximize the number of predicted proteins identified and provide tissue expression data for these identified proteins. Also, identifying predicted proteins in more than one experiment (52% of the chicken proteins identified were detected in more than one tissue) provides additional confidence that the predicted protein is expressed *in vivo*. In addition, 30 proteins were only electronically predicted or hypothetical translations in human. Identifying these proteins in chicken is additional information to support, not only the expression of these proteins in chicken but also in human based on orthology.

The least number of proteins were identified from the muscle and brain tissues. However, this does not necessarily reflect the biological complexity of these tissues but is more likely a reflection of the different protein extraction method used for these two tissues and amount of sample analyzed.

In addition to providing experimental support for the *in vivo* expression of chicken predicted proteins, we used strict 1:1 orthology with human and mouse genes to provide the identified proteins with standardized gene nomenclature based on established nomenclature guidelines and functional annotations based on the best available data. Since by definition predicted proteins have no direct experimental evidence, assignation of GO annotation for these proteins can be done using either IEA or ISS. While IEA is provided for a large range of organisms by the EBI-GOA Project, this annotation effort does not include predicted proteins and IEA annotations tend to be broad

descriptions of function (e.g. "protein binding"). The most rigorous way to assign function in the absence of direct experimental evidence is by strict orthology.

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Orthologs are, by definition, more likely to share functional similarity [38] and orthology can be used to reliably infer function to their co-orthologs. We determined chicken orthologous genes that pair with human and mouse genes. Since there is no a 'gold standard' method for orthologs identification [14], we integrated different published orthology identification methods that could possibly increase the breadth of orthologs identified. We were able to identify human or mouse orthologs for 77% of the identified chicken proteins. This figure, however, is better than the number that could have been obtained when using only one method (see additional file 3). For example from the total number of identified chicken predicted proteins (7,809), only 71%, 57%, 57% and 23% could have been identified by Homologene, Inparanoid, Ensembl and Treefam, respectively. Each of these methods use different procedures and orthologs identified by more than one method have been reported to be more consistent and reliable [14].

In addition to the experimentally supported predicted proteins that have human or mouse orthologs, there are a further 1,780 predicted proteins that we identified in this study. We are in the process of providing GO functional annotation for these proteins based on sequence similarity to other GO annotated gene products and functional motifs and domains and this information will be also be made publicly available.

Standardized nomenclature is becoming increasingly important with the large amounts of data released by sequencing projects, gene expression microarrays and proteomics. This information will facilitate comparative and functional genomics studies in both avians and mammals. Moreover, assigning functional annotation based on orthology is more robust than using sequence similarity alone [14]. This is because the higher level of functional conservation between orthologous proteins makes orthology highly relevant for protein function prediction. Thus our 8% increase in chicken GO annotated proteins is a significant improvement.

Conclusion

We demonstrate the value of proteomics to experimentally support the in-vivo expression of electronically predicted proteins of a newly sequenced genome. We assigned standardized nomenclature and GO functional annotations for these newly confirmed proteins. The approach we have developed facilitates comparative and functional genomics studies and may be applied to improve the annotations of a diverse range of newly sequenced genomes.

Methods

Tissues and protein extraction

Proteins were isolated from several different tissues in a series of experiments. Bursal B cells and stromal cells were isolated from bursas collected from five 21-day-old Ross 508 mixed sex chickens, muscle from the Pectoralis Major muscle of six 42 day old female chickens, brain from six 42 day old female chickens, spleen from eighteen 7- and 8-day-old advanced intercross Fayoumi and Leghorn mixed sex chickens, T cells from peripheral blood mononuclear cells (PBMC) obtained from adult Ross 508 mixed sex chickens, serum from 20-day-old Ross 508 male chickens. The disease virus-transformed cell line, MDCC-UA01 (obtained from Dr M. Parcells, University of Delaware) was grown as described [39]. Proteins were isolated using Differential Detergent Fractionation (DDF) [27] for each of the tissues except muscle and brain. For the muscle and brain samples, the samples were immediately frozen at -80°C. The samples were then allowed to warm to -21°C and solubilized in lysis buffer (7 M urea, 2 M thiourea, 4% CHAPSO, 8 mM PMSF) with repetitive pulsed sonication on ice. Note that the DDF method has been shown to yield more proteins than a single step lysis of tissues (as used for muscle and brain) [27].

Proteomics

All solubilized proteins were identified by 2-dimensional liquid chromatography tandem mass spectrometry (2-DLCMS/MS) exactly as previously described [24,27]. Briefly, protein mixtures are trypsin digested and the peptides desalted prior to strong cation exchange followed by reverse phase liquid chromatography coupled directly in line with ESI ion trap MS. A flow rate of 3 μL/min was used for both SCX and RP columns. A salt gradient was applied in steps of 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 57, 64, 71, 79, 90, 110, 300, and 700 mM ammonium acetate in 5% ACN, 0.1% formic acid and the resultant peptides loaded directly into the sample loop of a 0.18 × 100 mm BioBasic C18 reverse phase liquid chromatography column of a Proteome X workstation (ThermoElectron). The reverse phase gradient used 0.1% formic acid in ACN and increased the ACN concentration in a linear gradient from 5% to 30% in 30 min and then 30% to 65% in 9 min followed by 95% for 5 min and 5% for 15 min.

A database containing only chicken proteins that have been electronically predicted was prepared by parsing the chicken RefSeq entries (chicken gene build 2.1, 01/08/2007) for records with an XP prefix (14,676 proteins). The XP prefix is used to indicate proteins that have been predicted using the GNOMON pipeline. Redundancies were minimized by using the RefSeq dataset rather than the

dataset from the Non-redundant Protein Database. The RefSeq database contained 19,500 chicken proteins but only the 14,676 GNOMON predicted proteins were used in this study. Trypsin digestion was applied in silico to the predicted protein database including mass changes due to cysteine-carboxyamidomethylation and methionine oxidation.

The MS2 spectra were then used to search the non-redundant predicted protein database using Cluster 3.2 (Bioworks Browser 3.2, Thermo Electron, San Jose, CA). The peptide (MS precursor ion) mass tolerance was set to 1.4 and the groups scan to 1.0. Peptide molecular range was set to 600–3500. Only peptides ≥ 6 amino acids in length that had cross correlation (Xcorr) scores of 1.5, 2.0 and 2.5 (for +1, +2, and +3 charge state, respectively) and DeltaCn of > 0.1 [25,40,41] were considered matches. To quantify the peptide false discovery rate (FDR), we used the reverse database function in Bioworks 3.2 to search all MS2 spectra against a reversed version of our predicted proteins database using the same search criteria described above. Prior to calculating the FDR, we calculated the probability of each peptide match from both real and reversed database based on the product of XCorr and DeltaCn and set a cut-off of $P \leq 0.05$ for individual peptide identifications. With this probability as the cut-off, we calculated the FDR using the expected proportion E(V) of incorrect identifications from correct identifications (R) [36]: FDR = E(V)/R. Proteins were identified based on the peptides that pass the above criteria.

ID Mapping

Proteins identified by SEQUEST search algorithm have a Genbank identifier (gi) and RefSeq identifiers. In order to facilitate data sharing with public databases and ortholog determination we mapped the identified proteins to corresponding identifiers from UniProt Archive (UniParc), the International Protein Index (IPI), Entrez Gene and Ensembl protein identifiers using either different online tools for ID mapping [42-45] or an in-house Perl script (MapProtID.pl) to match different ID datasets. In cases where the program could not find an identifier, we used gi or RefSeq numbers to manually search co-identifiers in the UniParc [46], IPI [47], Entrez [48] or Ensembl [49] databases.

Ortholog Prediction

Chicken-human orthologs were downloaded from the HGNC (Human Genome Organization (HUGO) Gene Nomenclature Committee) Comparison of Orthology Predictions (HCOP) site [50] using the HCOP search tool [20,51]. HCOP integrates and displays the orthology assertions made by different ortholog prediction methods such as Ensembl [9], Homologene [21,52], Inparanoid [17], MGI (Mouse Genome Informatics) [53] and Treefam [18]. In cases where we could not identify chicken-human orthologs we manually checked Homologene [52], Inparanoid [54] or Ensembl [49,55] in order to obtain the most recent data. Chicken-mouse orthologs were downloaded only from Homologene, Inparanoid and Ensembl because HCOP does not predict chicken-mouse orthologs

Standardized Nomenclature

Standardized gene nomenclature is vital for effective scientific communication [22] and chicken researchers have agreed to use human nomenclature for orthologous chicken genes [23]. In this study we assigned chicken standardized nomenclature based on HGNC approved gene symbols and names that were associated with the human or mouse orthologs. We manually check the existence of each symbol and name in the HGNC nomenclature database before transferring it to chicken. In cases where the human or mouse gene symbol or name was not found or withdrawn from HGNC, no symbol or name was assigned to the chicken co-ortholog. To distinguish chicken from human genes the symbol assigned to chicken gene products are all in lowercases except for the first letter, as is the convention for mouse.

Functional Annotation

Since orthologs are presumed to have the same function, useful functional information can be extracted from other species when annotating orthologous gene products with unknown functions. To provide GO annotation for the identified chicken predicted proteins, we downloaded the human and mouse GO annotations from either the European Bioinformatics Institute GO annotation project (EBI-GOA: 03/12/2007) or searched Ensembl [49] using Biomart [43,55]. We assigned the chicken predicted proteins the GO annotations of human and mouse orthologs that are only based on direct experimental evidence codes (Table 1) and each chicken GO annotation was assigned an ISS GO evidence code, as per usual GO annotation procedure.

Public Availability of Data

Experimentally supported predicted proteins will be shared with the NCBI database, standardized nomenclature made available to both the NCBI and UniProt databases and GO annotations made available publicly via AgBase, the EBI-GOA Project and the GO Consortium. Assigned GO annotations are publicly available via the AgBase database [5] and will be submitted to the EBI-GOA Project. A summary of these GO annotations was obtained by mapping the associated GO terms to the Generic GOSlim Sets [56] using GOSlimViewer [4,5].

Authors' contributions

TJB contributed in the data generation, analysis of results and writing the draft of the manuscript. Both FMM and SCB contributed in the formulation, design of the study and manuscript preparation. All authors read and approved the final manuscript.

Additional material

Additional file 1

***Proteins identified by DDF-MudPIT and their distribution by tissue type**. Column 1 shows the RefSeq numbers of the identified chicken predicted proteins, column 2 indicates the corresponding predicted protein names (assigned by NCBI). Columns 3–8 shows the different types of tissue/cells used in this study and + and - indicate the presence or absence of the proteins in the specified tissue/cell, respectively.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-425-S1.xls]

Additional file 2

***Database identifiers for the predicted proteins**. RefSeq and gi identifiers (columns 1 & 2) are cross-referenced with other database identifiers for each of the identified chicken proteins.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-425-S2.xls]

Additional file 3

***Chicken-human/mouse orthologs predicted by different tools**. Using either the human or mouse orthologs shown in column 3, a standardized gene symbol and name (column 4 & 5) was assigned to 5,064 (65%) of the predicted proteins identified in this study. Columns 6–10 list the orthology prediction tools that were used to predict the human or mouse orthologs.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-425-S3.xls]

Acknowledgements

We wish to acknowledge Alan Shack, Joram Buza, Amanda Cooksey and Bart van den Berg assistance with collecting the tissue samples and protein extraction & Tibor Pechan for MS/MS analysis. The authors acknowledge financial support from Mississippi Agricultural and Forestry Experiment Station, Mississippi State University.

References

1. Alexandersson M, Cawley S, Pachter L: **SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model.** *Genome Res* 2003, **13(3):**496-502.
2. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14(5):**942-950.
3. Wu JQ, Shteynberg D, Arumugam M, Gibbs RA, Brent MR: **Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing.** *Genome Res* 2004, **14(4):**665-671.
4. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC: **AgBase: a unified resource for functional analysis in agriculture.** *Nucleic acids research* 2007, **35(Database issue):**D599-603.
5. McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC: **AgBase: a functional genomics resource for agriculture.** *BMC genomics* 2006, **7:**229.
6. Azuaje F, Al-Shahrour F, Dopazo J: **Ontology-driven approaches to analyzing data in functional genomics.** *Methods Mol Biol* 2006, **316:**67-86.
7. **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35(Database issue):**D193-7.
8. 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.
9. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, et : **Ensembl 2007.** *Nucleic acids research* 2007, **35(Database issue):**D610-7.
10. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35(Database issue):**D26-31.
11. Cheng J, Sun S, Tracy A, Hubbell E, Morris J, Valmeekam V, Kimbrough A, Cline MS, Liu G, Shigeta R, Kulp D, Siani-Rose MA: **NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis.** *Bioinformatics* 2004, **20(9):**1462-1463.
12. Lewis S, Ashburner M, Reese MG: **Annotating eukaryote genomes.** *Curr Opin Struct Biol* 2000, **10(3):**349-354.
13. Chen F, Mackey AJ, Stoeckert CJ Jr., Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic acids research* 2006, **34(Database issue):**D363-8.
14. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: **Benchmarking ortholog identification methods using functional genomics data.** *Genome biology* 2006, **7(4):**R31.
15. Li L, Stoeckert CJ Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9):**2178-2189.
16. O'Brien KP, Westerlund I, Sonnhammer EL: **OrthoDisease: a database of human disease orthologs.** *Human mutation* 2004, **24(2):**112-119.
17. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *Journal of molecular biology* 2001, **314(5):**1041-1052.
18. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R: **TreeFam: a curated database of phylogenetic trees of animal gene families.** *Nucleic acids research* 2006, **34(Database issue):**D572-80.
19. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic acids research* 2005, **33(Database issue):**D476-80.
20. Wright MW, Eyre TA, Lush MJ, Povey S, Bruford EA: **HCOP: the HGNC comparison of orthology predictions search tool.** *Mamm Genome* 2005, **16(11):**827-828.
21. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35(Database issue):**D5-12.
22. Wright MW, Bruford EA: **Human and orthologous gene nomenclature.** *Gene* 2006, **369:**1-6.
23. Crittenden LB, Bitgood JJ, Burt DW, Ponce de Leon FA, Tixier-Boichard M: **Nomenclature for naming loci, alleles, linkage groups, and chromosomes to be used in poultry genome publications and databases.** *The Second International Workshop on Poultry Genome Mapping in Prague* 1994.
24. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius.** *Proteomics* 2006, **6(9):**2759-2771.
25. Balgley BM, Laudeman T, Yang L, Song T, Lee CS: **Comparative Evaluation of Tandem MS Search Algorithms Using a Target-Decoy Search Strategy.** *Mol Cell Proteomics* 2007, **6(9):**1599-1608.
26. Higdon R, Kolker E: **A predictive model for identifying proteins by a single peptide match.** *Bioinformatics* 2007, **23(3):**277-280.

27. 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.
28. 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.
29. Burt DW: **Chicken genome: Current status and future opportunities .** In *Genomes* Edited by: Sussman HE, Smit MA. Cold Harbor Laboratory Press ; 2006:221-236.
30. McPherson JD, Dodgson J, R. K, Pourquié O: **Proposal to sequence the genome of chicken. World Wide Web (http://www.nih.gov/science/models/gallus/ChickenGenomeWhitePaper.pdf).** 2003 .
31. Sanders WS, Bridges SM, McCarthy FM, Nanduri B, Burgess SC: **Prediction of peptides observable by mass spectrometry applied at the experimental set level,.** *BMC Bioinformatics,* 2007, **8(Suppl 7)(S23):**.
32. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD, Pevzner PA: **Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation.** *Genome Res* 2007, **17(9):**1362-1377.
33. Lowenthal MS, Mehta AI, Frogale K, Bandle RW, Araujo RP, Hood BL, Veenstra TD, Conrads TP, Goldsmith P, Fishman D, Petricoin EF 3rd, Liotta LA: **Analysis of albumin-associated peptides and proteins from ovarian cancer patients.** *Clinical chemistry* 2005, **51(10):**1933-1945.
34. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.** *Nature methods* 2007, **4(3):**207-214.
35. Nesvizhskii AI, Aebersold R: **Interpretation of shotgun proteomic data: the protein inference problem.** *Mol Cell Proteomics* 2005, **4(10):**1419-1440.
36. Nesvizhskii AI, Vitek O, Aebersold R: **Analysis and validation of proteomic data generated by tandem mass spectrometry.** *Nature methods* 2007, **4(10):**787-797.
37. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM: **Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study.** *Nature biotechnology* 2006, **24(3):**333-338.
38. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19(2):**99-113.
39. Dienglewicz RL, Parcells MS: **Establishment of a lymphoblastoid cell line using a mutant MDV containing a green fluorescent protein expression cassette.** *Acta Virol* 1999, **43(2-3):**106-112.
40. Eng JK, McCormack AL, Yates JR, III: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *J Am Soc Mass Spectrom* 1994, **5:**976-989.
41. Liu T, Qian WJ, Gritsenko MA, Xiao W, Moldawer LL, Kaushal A, Monroe ME, Varnum SM, Moore RJ, Purvine SO, Maier RV, Davis RW, Tompkins RG, Camp DG 2nd, Smith RD: **High dynamic range characterization of the trauma patient plasma proteome.** *Mol Cell Proteomics* 2006, **5(10):**1899-1913.
42. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R: **IDconverter and IDClight: conversion and annotation of gene and protein IDs.** *BMC bioinformatics* 2007, **8:**9.
43. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics (Oxford, England)* 2005, **21(16):**3439-3440.
44. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic acids research* 2006, **34(Database issue):**D187-91.
45. **Batch Retrieval:PIR - Protein Information Resource** [http://pir.georgetown.edu/pirwww/search/idmapping.shtml]
46. **UniProt Archive Database** [http://www.pir.uniprot.org/database/archive.shtml]
47. **International Protein Index database** [http://www.ebi.ac.uk/IPI/IPIhelp.html]
48. **Entrez cross-database search** [http://www.ncbi.nlm.nih.gov/sites/entrez]
49. **Ensembl Genome Browser** [http://www.ensembl.org/Gallus_gallus/index.html]
50. **HGNC Comparison of Orthology Predictions search tool** [http://www.genenames.org/cgi-bin/hcop.pl]
51. Eyre TA, Wright MW, Lush MJ, Bruford EA: **HCOP: a searchable database of human orthology predictions.** *Briefings in bioinformatics* 2007, **8(1):**2-5.
52. **Homologene: A homology resource** [http://www.ncbi.nlm.nih.gov/HomoloGene/]
53. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, et : **The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.** *Nucleic acids research* 2005, **33(Database issue):**D471-5.
54. **Inparanoid: Eukaryotic Ortholog Groups** [http://inparanoid.sbc.su.se]
55. **BioMart: Data mining tool** [http://www.ensembl.org/biomart/martview]
56. **Generic GOSlim set** [http://www.geneontology.org/GO_slims/goslim_generic.obo]

CHAPTER 4

GENE ONTOLOGY ANNOTATION QUALITY ANALYSIS IN MODEL EUKARYOTES[1]

[1] Reprint from T. J. Buza, F..M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess. 2008. Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res* 36: e12. This article is available from: http://nar.oxfordjournals.org/cgi/content/full/36/2/e12.a

Published online 10 January 2008

Nucleic Acids Research, 2008, Vol. 36, No. 2 ***e12***
doi:10.1093/nar/gkm1167

Gene Ontology annotation quality analysis in model eukaryotes

Teresia J. Buza[1,3], Fiona M. McCarthy[1,3,*], Nan Wang[2,3], Susan M. Bridges[2,3,4] and Shane C. Burgess[1,3,4,5]

[1]Department of Basic Sciences, [2]Department of Computer Science and Engineering, [3]Institute of Digital Biology, [4]Mississippi Agricultural and Forestry Experiment Station and [5]Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi 39762, USA

Received September 4, 2007; Revised November 30, 2007; Accepted December 18, 2007

ABSTRACT

Functional analysis using the Gene Ontology (GO) is crucial for array analysis, but it is often difficult for researchers to assess the amount and quality of GO annotations associated with different sets of gene products. In many cases the source of the GO annotations and the date the GO annotations were last updated is not apparent, further complicating a researchers' ability to assess the quality of the GO data provided. Moreover, GO biocurators need to ensure that the GO quality is maintained and optimal for the functional processes that are most relevant for their research community. We report the GO Annotation Quality (*GAQ*) score, a quantitative measure of GO quality that includes breadth of GO annotation, the level of detail of annotation and the type of evidence used to make the annotation. As a case study, we apply the *GAQ* scoring method to a set of diverse eukaryotes and demonstrate how the *GAQ* score can be used to track changes in GO annotations over time and to assess the quality of GO annotations available for specific biological processes. The *GAQ* score also allows researchers to quantitatively assess the functional data available for their experimental systems (arrays or databases).

INTRODUCTION

Elucidation of the complete human genome sequence (1,2) was a watershed event for both biology and computer science. As more genome sequence projects have been initiated, the amount of biological data and number of databases have proliferated (3,4). Methods for high-throughput, genome-wide analysis of biological systems have been developed and applied to an increasing number of organisms. Foremost among these techniques are functional genomics using microarrays and proteomics. The current challenge for functional genomics experiments is to translate large lists of genes or gene products into biologically relevant models. The Gene Ontology (GO) (5,6) was developed in part to answer this problem and has since become the *de facto* method for functional annotation of gene products (7).

GO annotations are provided by literature curation or by computational analysis that must be continually updated by human biocurators. For example, the European Bioinformatics Institute GO Annotation (EBI-GOA) Project (8) currently provides annotations for over 122 199 different species; GO annotations for all but 33 of these organisms have been generated by mapping functional motifs and domains to GO terms ['inferred by electronic annotation' (IEA) annotations] (9). These IEA annotations account for more than 90% of GO annotations and the basis for these annotations is continually reviewed so that all IEA annotations are updated on a weekly basis. Moreover, IEA annotations are generalized to apply to a diverse range of species and usually only represent very broad functions such as 'protein binding' and 'enzyme binding'. In effect, this means that as functional genomics data is modeled using GO annotation, there are no curated GO annotations for many gene products and a large proportion of the remaining data describes only very broad biological concepts.

One axiom of GO is that the amount of functional information for any gene product varies from species to species, depending on the literature and databases available for different species. To assist researchers and biocurators with assessing the overall species-specific GO annotation quality of a particular dataset we developed the GO Annotation Quality (*GAQ*) score. The *GAQ* score is a quantitative measure of the GO annotation of a set of

*To whom correspondence should be addressed. Tel: +1 662 325 5859; Fax: +1 662 325 1031; Email: fmccarthy@cvm.msstate.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

gene products (e.g. all annotated proteins in a species) based on the number of GO annotations available, the level of detail of the annotation and the types of evidence used to make these GO annotations. We demonstrate the utility of the *GAQ* score by comparing the current state of GO annotation in nine taxonomically diverse eukaryotes, by quantifying the improvement in GO annotation for two biomedical model species (chicken and mouse) relative to the time a dedicated GO annotation effort commenced for each species, and by demonstrating how the *GAQ* score can be used by biocurators to better direct GO annotation efforts and facilitate comparative functional annotation.

MATERIALS AND METHODS

The *GAQ* score

The overall GO annotation quality of a set of gene products is related to the coverage of gene products with GO annotation (breadth), the level of detail of GO annotation (depth), the types of evidence used to make these GO annotations (GO evidence code) and the completeness of the annotations based on how much of the current literature containing relevant information has been annotated.

We used quantitative information from breadth, depth and GO evidence code to derive a quantitative measure of GO annotation quality which we call the *GAQ* score. We define the *GAQ* score for an annotation (a) as the product of its depth in the ontology (*Dd*) and the evidence code rank (*ECR*) of the annotation:

$$GAQ(a) = ECR_a \quad Dd_a$$

The *GAQ* score for a set of gene products (S) with a total of A GO annotations is defined as:

$$GAQ(S) = \sum_{a=1}^{A} (ECR_a \quad Dd_a)$$

The 'breadth' in this study is defined as 'the number of annotations assigned to each of the gene products in the dataset.' Note that, in some cases, it may be more informative to compute a separate GAQ score for each of the three GO ontologies and to consider the 'breadth of annotation' for each ontology. When considering the annotation, breadth of a specific gene product should be evaluated separately for each ontology.

GO annotation 'depth' is quantified by the depth of each GO annotation term within the ontology structure. The gene ontologies are structured as directed acyclic graphs (DAGs) where each 'leaf' term represents the most detailed level of information in relation to the parent level. Therefore, DAG depth from the root to an annotation term a (child node) is an indicator of the level of functional detail captured in the annotation. It has recently been argued that DAG structural levels are not good indicators of specificity for GO terms when grouping terms for functional analysis and that information theory can be used to partition GO terms into groups with similar specificity as measured by information content (10). However, this approach results in different groupings of terms for different species and would make cross-species comparisons very difficult. We have chosen to use DAG depth because we feel it gives the best overall view of the level of annotation detail, it is easily understood and because it facilitates comparison of annotation levels among different species. Since the GO ontologies are DAGs and not trees, there may be several paths from a child term to the root node. We define the GO DAG depth (*Dd*) of an annotation term as the length of the longest path from the term to its top-level parent in the ontology (either 'molecular function', 'biological process' or 'cellular compartment'). We use the longest path rather than the shortest because the 'true path rule' used by the Gene Ontology (http://www.geneontology.org/GO.annotation.shtml#general) implies annotation to all parents on any path to the root. Note that different GO annotations will have different path lengths (which represent granularity) and that such annotations depends on the type of experiment performed, the amount of literature available for the gene product in question and the species being annotated. Therefore, a less granular GO term does not equate to a lesser annotation. We also define the *Dd* for an entire ontology as the sum of the *Dd* for each term in the ontology. Likewise, the average *Dd* for ontology is the *Dd* of all the terms divided by the number of terms in the ontology.

Each GO annotation indicates the type of evidence used to make that annotation and we initially assigned each GO term an evidence code rank (*ECR*) on a scale of 1 to 5 based on whether the evidence was direct or indirect (Table 1). However, like the GO itself, evidence code usage is evolving and we expect that *ECR*s will change over time. To test how any change in the *ECR* will affect the *GAQ* score we also used two other ranking systems to calculate *GAQ* (Supplementary Data). The average *ECR* for a species is a reflection of how much of the GO annotation is based on direct experimental evidence.

The breadth of annotations for a set of gene products (for example all annotated gene products for a species) can be measured in two ways. First, the total *GAQ* score for the set is an indication of both the number of products annotated and the quality of the annotation. In order to evaluate the breadth of annotation for each annotated gene product, we also define the *meanGAQ* score for a set of gene products as the *GAQ* score for the set divided by the total number of gene products (n) annotated:

$$meanGAQ(S) = GAQ/n$$

The *meanGAQ* for a species is defined as the *meanGAQ* for all annotated gene products for that species.

Two in-house Perl scripts (DAGdepth.pl and GAQ.pl) have been implemented to determine the *Dd* of a given GO term and the *GAQ* score for a set of gene products.

GO annotation statistics for model eukaryotes

We obtained GO annotation statistics for nine species that have a dedicated GO annotation effort (Table 2). The number of GO annotations for each species, number of gene products that have annotations and percentage

Table 1. GO evidence codes and their corresponding rank used for this study.

Code	Code definition	Evidence code rank
IDA	Inferred from Direct Assay	5
IGI	Inferred from Genetic Interaction	5
IMP	Inferred from Mutant Phenotype	5
IPI	Inferred from Physical Interaction	5
IC	Inferred by Curator	4
TAS	Traceable Author Statement	4
IEP	Inferred from Expression Pattern	3
RCA	Inferred from Reviewed Computational Analysis	3
IGC	Inferred from Genomic Context	3
ISS	Inferred from Sequence or Structural Similarity	2
IEA	Inferred from Electronic Annotation	2
NAS	Non-traceable Author Statement	2
NR	Not Recorded	1
ND	No Biological data available	0

Direct experimental evidence codes (IDA, IMP, IGI and IPI) are ranked higher than indirect evidence codes. The IC and TAS evidence codes are based on expert judgment (of either the GO annotator or the researcher, respectively). The IEP, IGC and RCA codes refer to functions inferred from expression pattern, genomic context and reviewed computation analysis, respectively, and rank lower than direct functional evidence. The ISS evidence code is used for annotations made based on structural or sequence similarities. In contrast, the IEA evidence code is used for annotations that depend on automated transfer of annotations. Since some IEA annotations assigned by some groups may be of the same quality as ISS annotations assigned by other groups we assigned the same rank to both codes. NAS refers to uncited statements in reviewed articles and this data is not readily traced or the author may be referring to experiments done in a different species. The NR evidence code is a historical artifact of the GO and is used for older GO annotations made before the evidence code ontology was developed; since the evidence source is unrecorded, it must be presumed to be of lesser rank. ND is assigned where there are no biological data available. Other ranking systems used in this study are outlined in Supplementary data 1.

Table 2. GO annotation statistics.

Species	Number of GO annotations	Number of annotated gene products	Number of annotations per gene product	% IEA	Lc
Bt	85 316	22 812	4	96	193
Ce	72 558	12 171	6	90	723
Dm	83 615	11 363	7	65	3546
Dr	102 202	31 106	3	98	527
Gg	56 745	16 230	3	96	123
Hs	167 889	34 118	5	69	13 361
Mm	179 696	34 886	5	59	7834
Rn	113 012	27 954	4	88	2933
Sc	64 770	5536	12	54	6123

Current GO statistics (as at 05/05/2007) for *B. taurus* (Bt), *C.elegans* (Ce), *D. melanogaster* (Dm), *D. renio* (Dr), *G. gallus* (Gg), *H. sapiens* (Hs), *M. musculus* (Mm), *R. norvegicus* (Rn) and *S. cerevisiae* (Sc). The number of GO annotations, annotations per gene products and percentage non-IEA annotations are obtained from EBI-GOA. Literature curated (Lc) figures are obtained by parsing the total number of PubMed records in the GO association files.

of GO annotations that are IEA were all obtained from EBI-GOA statistics (http://www.ebi.ac.uk/GOA/proteomes.html; 05/05/2007). A quantitative measure of the literature curated to the GO (*Lc*) for each species was

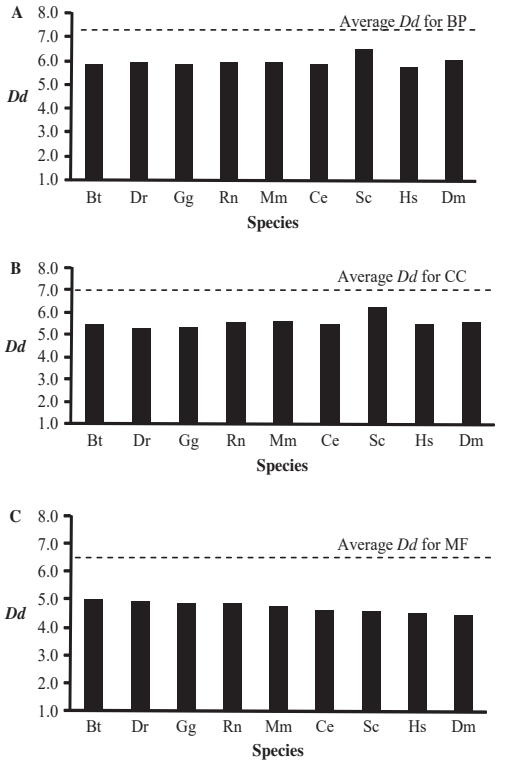

Figure 1. The DAG depth (*Dd*) for each Gene Ontology. The overall average *Dd* (dashed line) was determined for all GO terms in each ontology (as at 05/052007). GO term *Dd*s were compared to mean *Dd* of each species for (**A**) Biological Process (BP), (**B**) Cellular Component (CC) and (**C**) Molecular Function (MF). The species represented are *B. taurus* (Bt), *D. renio* (Dr), *G. gallus* (Gg), *R. norvegicus* (Rn), *M. musculus* (Mm), *C. elegans* (Ce), *S. cerevisiae* (Sc), *H. sapiens* (Hs) and *D. melanogaster* (Dm).

obtained by downloading the EBI-GOA gene association file and counting the number of different literature entries for each of the species. However, none of these statistics allow a quantitative comparison of 'how well' a species is GO annotated. To capture this information, we computed the average *Dd* for each species for each ontology (Figure 1), the mean *ECR* for all annotations for each species (Figure 2) and the *meanGAQ* for the set of all annotated gene products for each of the species (Figure 3).

To compare the overall *GAQ* scores between species, we constructed *GAQ* matrices by pair-wise comparison of mean *GAQ* scores for all species (Table 3). Each entry in the table is the ratio of the *GAQ* scores of the species listed with each column divided by that of the species listed with each row.

Measuring *GAQ* over time

It may be useful to know the *GAQ* score for a species of interest or even to compare *GAQ* scores between two species. Obviously, care must be taken when

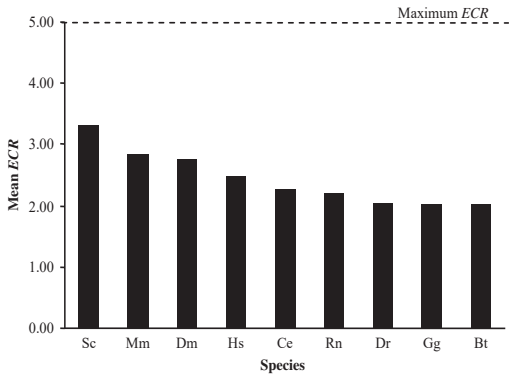

Figure 2. The evidence code rank (*ECR*) for each species. GO evidence codes were ranked based on how closely they describe direct experimental evidence (Table 1) and current GO annotations were evaluated based upon these rankings. The maximum *ECR*, based on direct experimental evidence, is five. The species represented are *S. cerevisiae* (Sc), *M. musculus* (Mm), *D. melanogaster* (Dm), *H. sapiens* (Hs), *R. norvegicus* (Rn), *C. elegans* (Ce), *B. taurus* (Bt), *G. gallus* (Gg) and *D. renio* (Dr). The founder species (Sc, Mm, Dm), with a longer history of GO annotation, have the highest average *ECR*s. Other evidence code rankings were also used (Supplementary Data).

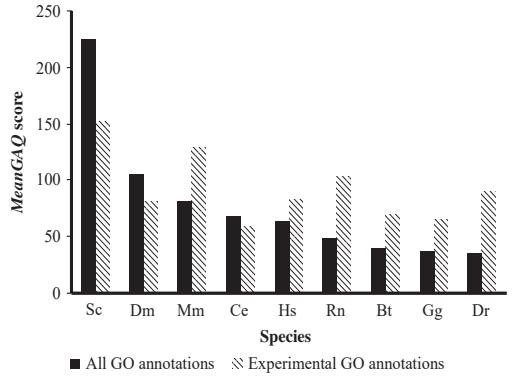

Figure 3. Mean GO Annotation Quality (*GAQ*) scores for each species. To quantify GO annotation quality, we combined annotations (number of annotations per gene product), 'depth' (*Dd*) and evidence quality (*ECR*) to create the GO Annotation Quality (*GAQ*) score. The average *GAQ* score for *S. cerevisiae* (Sc), *D. melanogaster* (Dm), *M. musculus* (Mm), *H. sapiens* (Hs), *C. elegans* (Ce), *R. norvegicus* (Rn), *B. taurus* (Bt), *G. gallus* (Gg) and *D. renio* (Dr) (as at 05/05/2007) is shown. GO annotation founder species have higher overall *meanGAQ* scores than species with more recent GO annotation efforts. Higher scores are found in *Sc*, *Mm*, *Rn* and *Dr*, when computing *meanGAQ* scores from annotations made using only direct experimental evidence codes.

comparing functional annotations between species, however, because each species has its own set of literature that contains data that can be annotated directly for that species. The *GAQ* score is also useful for tracking how GO annotations may be improving with time (especially relative to changes in the ontology) for a given species of interest. Improving species-specific *GAQ* scores indicate improving functional annotation, which can be used with more confidence by researchers to model their genes or gene products to derive biological value. We used *GAQ* scores to measure the change in *GAQ* in chicken (which has only recently been actively GO annotated) and mouse (one of the GO founder species) for the first 5 years of each species' respective GO annotation (Figure 4). Since the date of each GO annotation is recorded, we obtained annotations for each time period by parsing the chicken and mouse gene association files. The IEA annotations were excluded from this study because all IEA annotations are updated on a monthly basis and the date of these annotations changes to reflect this updating.

Assessing *GAQ* scores for different areas of the GO

Since each species has its own body of functional information that can be annotated to the GO, and because some species are specifically used as model organisms for particular physiologic processes, we hypothesize that some sub-areas of the GO have more comprehensive annotation than others and that annotation cannot proceed uniformly across the entire GO. To test our hypothesis, we calculated the *meanGAQ* (excluding IEA annotations) for sub-areas of the chicken and mouse GO Biological Process Ontology (Table 4). We first summarized the annotations to Generic GOSlim terms using the GoSlimViewer tool at AgBase (11). Generic GOSlim terms are a subset of the GO ontologies and provide a summary level view of annotation in different major categories.

Assessing *GAQ* using available functional literature

The amount of functional literature available for curation to the GO varies for each species and estimating the amount of literature available for a species is difficult. We estimated the total PubMed entries available for a species by using that species' scientific name, common name or taxonomy identifier. To estimate the amount of functional literature that could contain GO annotation data we used both Gene Reference Into Function (GeneRIF) (12) entries and GOPubMed (13). To determine the amount of literature curated to the GO (*Lc*) in each species we counted the number of unique PubMed identifiers recorded in the species' gene association file (Table 2). The proportion of literature that contains functional data suitable for GO annotation varied significantly by species but in every case the percentage of available literature that has already been annotated using the GO is a small fraction of the functional literature available (Table 5).

RESULTS

GO annotation statistics of the study species

While it might be expected that organisms with the longest history of active GO annotation would have the most comprehensive GO annotations, the number of GO annotations does not accurately reflect the overall GO annotation quality (*GAQ*) for a species. This is because so many GO annotations are based on nondirect

Table 3. The *GAQ* matrix obtained from pairwise comparison of *meanGAQ* scores for each species.

Species	meanGAQ	Sc	Dm	Mm	Ce	Hs	Rn	Bt	Gg	Dr
	meanGAQ(1)	**225**	**105**	**81**	**68**	**64**	**49**	**41**	**37**	**36**
Sc	**225**	**1.0**	0.5	0.4	0.3	0.3	0.2	0.2	0.2	0.2
Dm	**105**	2.1	**1.0**	0.8	0.6	0.6	0.5	0.4	0.4	0.3
Mm	**81**	2.8	1.3	**1.0**	0.8	0.8	0.6	0.5	0.5	0.4
Ce	**68**	3.3	1.5	1.2	**1.0**	0.9	0.7	0.6	0.5	0.5
Hs	**64**	3.5	1.6	1.3	1.1	**1.0**	0.8	0.6	0.6	0.6
Rn	**49**	4.6	2.1	1.7	1.4	1.3	**1.0**	0.8	0.8	0.7
Bt	**41**	5.5	2.6	2.0	1.7	1.6	1.2	**1.0**	0.9	0.9
Gg	**37**	6.1	2.8	2.2	1.8	1.7	1.3	1.1	**1.0**	1.0
Dr	**36**	6.3	2.9	2.3	1.9	1.8	1.4	1.1	1.0	**1.0**
	meanGAQ(2)	**152**	**81**	**128**	**59**	**83**	**103**	**70**	**65**	**90**
Sc	**152**	**1.0**	0.5	0.8	0.4	0.5	0.7	0.5	0.4	0.6
Dm	**81**	1.9	**1.0**	1.6	0.7	1.0	1.3	0.9	0.8	1.1
Mm	**128**	1.2	0.6	**1.0**	0.5	0.6	0.8	0.5	0.5	0.7
Ce	**59**	2.6	1.4	2.2	**1.0**	1.4	1.7	1.2	1.1	1.5
Hs	**83**	1.8	1.0	1.5	0.7	**1.0**	1.2	0.8	0.8	1.1
Rn	**103**	1.5	0.8	1.2	0.6	0.8	**1.0**	0.7	0.6	0.9
Bt	**70**	2.2	1.2	1.8	0.8	1.2	1.5	**1.0**	0.9	1.3
Gg	**65**	2.3	1.2	2.0	0.9	1.3	1.6	1.1	**1.0**	1.4
Dr	**90**	1.7	0.9	1.4	0.7	0.9	1.1	0.8	0.7	**1.0**

Species represented are *S. cerevisiae* (Sc), *D. melanogaster* (2 Dm), *M. musculus* (Mm), *H. sapiens* (Hs), *C. elegans* (Ce), *R. norvegicus* (Rn), *B. taurus* (Bt), *G. gallus* (Gg) and *D. renio* (Dr). The *meanGAQ* scores are based on number of gene products associated with the GO terms. *meanGAQ(1)* is based on all species' GO annotations, *meanGAQ(2)* is based on annotations made using only direct experimental evidence codes and in each case the *meanGAQ* is shown in bold at the top of each matrix. Where a species is compared to itself, the value will necessarily be one and these values are also marked in bold. A value >1 indicates that the species has higher *meanGAQ* score than the one it is compared against. For example, on average the *meanGAQ* score for the mouse gene products are two folds higher than that of chicken. Yeast consistently has the highest rates of *meanGAQ* scores when compared to each of the other organisms.

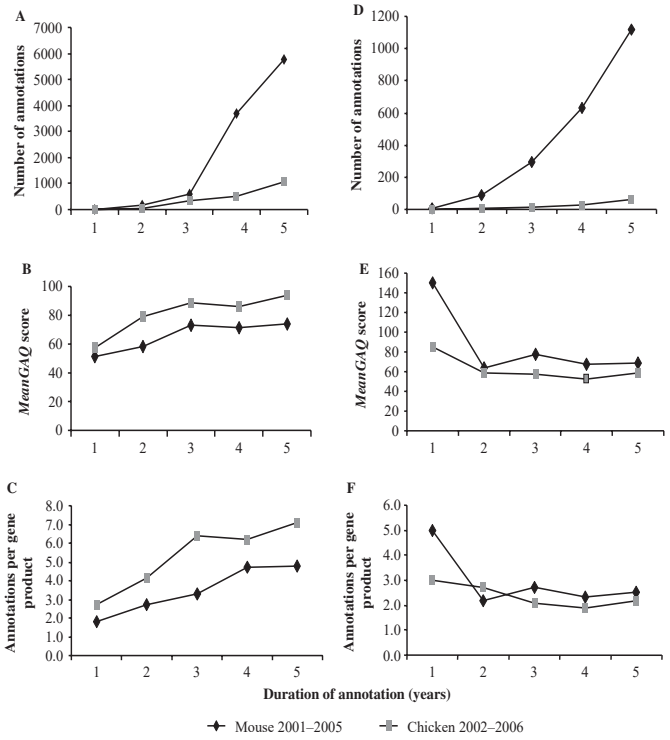

Figure 4. Change in GO annotations and GAQ score over time. Chicken and mouse were chosen as two species with a dedicated GO annotation effort that started at different times. Number of annotations, meanGAQ scores and annotations per gene product derived from all non-IEA annotations (**A**, **B** & **C**) and from annotations made using only direct evidence codes (**D**, **E** & **F**) are shown.

Table 4. The 20 top-ranked chicken biological processes and the mouse *GAQ* score for these processes.

Biological process	Chicken		Mouse	
	meanGAQ	Rank	*meanGAQ*	Rank
Ion transport	46	1	44	9
DNA metabolic process	36	2	51	4
Response to biotic stimulus	31	3	60	2
Cell death	26	4	47	7
Anatomical structure morphogenesis	25	5	65	1
Multicellular organismal development	24	6	48	6
Lipid metabolic process	23	7	41	11
Nucleic acid metabolic process	22	8	41	11
Amino acid and derivative metabolic process	22	8	44	9
Cell cycle	22	8	41	11
Signal transduction	22	8	44	9
Transcription	21	9	54	3
Protein modification process	21	9	44	9
Cytoskeleton organization and biogenesis	19	10	45	8
Embryonic development	19	10	33	18
Response to stress	19	10	25	25
Metabolic process	18	11	29	22
Translation	18	11	32	19
Cell differentiation	18	11	40	12
Catabolic process	17	12	30	21

meanGAQ scores were calculated for sub-areas of the Biological Process ontology in both chicken and mouse (excluding IEA annotations). The 20 top-ranked chicken biological processes (as summarized by the Generic GOSlim using the GoSlimViewer) are shown along with the calculated *GAQ* score for the chicken gene products currently described by these processes. The corresponding mouse *meanGAQ* score for the same sub-area and its ranking is also shown.

Table 5. Assessment of literature for GO annotation.

Species	PubMed (*L*)	% Functional literature (*Lf*)		% *Lc*
		GeneRIF	GOPubMed	
Bt	301 568	0.49	4.01	0.06
Ce	15 920	7.73	104.22	4.54
Dm	61 488	7.81	27.63	5.77
Dr	9058	15.51	157.01	5.82
Gg	143 170	0.71	9.58	0.09
Hs	10 018 771	1.10	0.10	0.13
Mm	902 076	5.73	1.82	0.87
Rn	2 125 874	1.01	0.72	0.14
Sc	83 543	4.00	22.60	7.33

For consistency we searched in NCBI the total number of PubMed available for a species (*L*) by using the species' scientific name, common name and/or taxonomy identifier. Species represented are *B. taurus* (Bt), *C.elegans* (Ce), *D. melanogaster* (Dm), *D. renio* (Dr), *G. gallus* (Gg), *H. sapiens* (Hs), *M. musculus* (Mm), *R. norvegicus* (Rn) and *S. cerevisiae* (Sc). The amount of functional literature (*Lf*) is from the geneRIF database and GOPubMed. GeneRIFs are often extracted directly from the document that is identified by the PubMed ID while GoPubMed is a knowledge-based search engine for biomedical texts. The amount of curated literature (*Lc*) is computed as the number of Pubmed IDs recorded in GO annotation (EBI-GOA; 5 May 2007). The percentage of *Lf* and *Lc* is computed based on *L* available for a species.

experimental evidence (e.g. ISS and IEA). For example, zebrafish has more annotations than two of the 'founder' species (fruitfly, yeast), but a much smaller percentage of these annotations are based on direct experimental evidence (Table 2). Moreover, each species has its own body of direct experimental evidence that can be used for functional annotation and each group annotating to the GO have prioritized their annotation efforts based on their resources and the needs of the scientific community that they serve.

The *GAQ* score

The overall average *Dd* of Biological Process is 7.1, Cellular Component is 6.9 and Molecular Function is 6.1 (dashed line in Figure 1). In general, we found that there is very little variation for *Dd* between the species, although *Saccharomyces cerevisiae* (Sc) has a higher average *Dd* for both Biological Process and Cellular Component ontologies when compared to the other species. Also, the mean *ECR* for each species is higher in yeast, mouse and fruitfly, the founder species of GO annotation (Figure 2). This is expected because these species have the earliest dedicated, literature biocuration effort.

The *meanGAQ* score was calculated from all GO annotations and compared to that obtained from annotations that are only based on direct experimental evidence codes (Figure 3). Intuitively, *GAQ* scores should reflect the amount of dedicated GO annotation effort in each species. Yeast, fruitfly and mouse have the highest overall *meanGAQ* scores. This is expected because these three species (the GO founder species) have the longest effort of GO annotation. However, cow is an interesting exception to this trend as the effort to annotate bovine gene products is relatively new, yet it has slightly higher *GAQ* scores than chicken. We expect that this is because, as a mammalian species, cow benefits more from the transfer of GO annotations from other species such as mouse and human.

To compare the magnitude of *meanGAQ* scores between different species we used a *GAQ* matrix (Table 3). A score of 1 means that the two species compared in the pair-wise comparison have equal *GAQ* scores. A score >1 means that the species listed in column has better quality annotation than the one it is compared against in the corresponding row. Yeast consistently has the highest *meanGAQ* when compared to each of the other organisms. Although by no means completely GO annotated, yeast may be considered as the current 'gold standard' species for *GAQ*.

Measuring *GAQ* over time

Since the structure of the GO DAG, the available functional literature and the investment and effort in GO annotation change over time, it is desirable to be able to compare GO annotation progress over time. We compared the progression of annotation and *GAQ* scores in chicken and mouse (Figure 4; Supplementary Data). As we expected, based on the investment in GO annotation for these species, the number of annotations for both species increased over time (Figure 4A and D), with mouse annotations showing a rapid increase after the

third year of annotation. Interestingly, although mouse has more annotations, chicken has higher overall *meanGAQ* scores (Figure 4B). But mouse has a higher *meanGAQ* score when using only annotations based on direct experimental evidence codes (Figure 4E) are used in the calculation. The *meanGAQ* score is directly proportional to the numbers of annotations per gene product (Figure 4C and F) rather than overall numbers of GO annotations.

Assessing *GAQ* scores for different areas of the GO

By using the *meanGAQ* score to evaluate specific regions of the Biological Process ontology, we found that some regions of the GO have more comprehensive annotation than others (Table 4). This also applies when either comparing GO annotation within a species (chicken). In general, chicken *meanGAQ* scores for the 20 highest-ranked regions of the Biological Process ontology are lower when compared to those of mouse. The exception is ion transport.

Assessing *GAQ* using available functional literature

By estimating the amount of literature available for annotation to the GO, we were able to assess what proportion of functional literature has been curated. Since it is difficult to assess how much functional literature is available, we used two different methods to estimate the amount of functional literature (*Lf*) that is available (Table 5). Some 'model species' (e.g. mouse and rat) have a low *Lf* while *Caenorhabditis elegans* and *D. renio* have a high *Lf*. However, while the *Lf* differs from one species to another, in all cases the percentage of literature curated (*Lc*) is very small. This is partially due to the amount of time and resources it takes to do literature curation but also because the amount of literature available is increasing dramatically.

DISCUSSION

Oftentimes it is difficult for researchers to assess the quality of functional annotation associated with their gene expression arrays or proteomics databases and it is often not easy to determine when they were last updated. Ideally, an overall assessment of the current GO annotation status for a genome would include the average number of GO annotations per gene. However, for many species the number of genes is not known or the number of reported genes differs significantly depending on the source used. This problem is compounded when comparing different species because it is even more difficult to find comparable information for a diverse range of species. Moreover, the number of GO annotations does not provide information about the quality of the available GO annotations. We developed the *GAQ* score as a quantitative measure of GO quality.

The *GAQ* score is derived from the number of GO annotations (breadth), DAG depth (*Dd*) and GO Evidence Code Rankings (*ECR*). In this instance, when we are discussing the 'breadth of annotation' we are referring to the total number of annotations assigned to each of the gene products in the dataset of interest. However, the overall *GAQ* score for a dataset provides little information about GO annotation for individual genes. For example, when GO annotations for mouse or chicken POLA1 and BASP1 are combined, there are 37 GO annotations for the mouse proteins and 34 GO annotations for the chicken proteins (Table 6). While this is a comparable number of GO annotations, the BASP1 mouse protein has annotations for each of the three ontologies while chicken BASP1 has no molecular function and the majority of GO annotations are to cellular component. The mouse BASP1 protein has fewer GO annotations but greater GO annotation breadth.

Table 6. Example of breadth of GO annotations for mouse and chicken.

Gene product	Total annotations	Number of annotations		
		MF	BP	CC
Mouse POLA1	33	14	12	7
Chicken POLA1	27	9	11	7
Mouse BASP1	4	1	1	2
Chicken BASP1	7	0	1	6
Mouse Total	37	15	13	9
Chicken Total	34	9	12	13

Using the number of GO annotations as a measure of annotation breadth shows the overall GO annotation breadth of a dataset but does not reflect the annotation breadth of individual gene products. In this example mouse and chicken GO annotations are obtained from EBI-GOA (6 November 2007) for polymerase (DNA directed), alpha 1 (POLA1) and brain abundant, membrane attached signal protein 1 (BASP1) for each GO ontology. The three GO are molecular function (MF), biological process (BP) and cellular component (CC). Although the overall number of GO annotations is comparable for both species, the chicken BASP1 GO annotations are predominately CC annotations. When examined individually, the mouse BASP1 has better GO annotation breadth as there are annotations to all three ontologies. The UniProtKB accession numbers for the proteins are: chicken POLA1–Q59J86; mouse POLA1–P33609; chicken BASP1–P23614; and mouse BSAP1–Q91XV3.

The GO DAGs are designed so that the more detailed terms are deeper in the structure. As expected, none of the species in this study reach the average *Dd* for any of the three ontologies. Even comprehensively GO-annotated orthologs from different species have different *Dd*, reflecting the type of experiments performed in each species, the amount of species-specific literature available for that gene and inter-species variation in gene function. However, while a less granular GO term does not equate to a lesser annotation, it does mean less detailed functional information. The only way to assess the maximum granularity possible for a species is to have completed literature annotation for each of the gene products of interest; this is not possible nor is it currently possible to accurately and quantitatively assess the amount of granularity currently available in comparison to the functional detail available in current literature. Despite these practical limitations, our method still provides a quantitative measure of GO annotation that enables researchers to assess the *GAQ* of a specific dataset at a given time.

It is unlikely that any one species will have direct experimental evidence to be annotated to the most detailed

(or deepest) GO terms across the enormous range of the GO. Detailed GO annotation relies on continued funding of new and existing annotation efforts, including support for developing the GO, maintaining existing data and database resources and updating existing GO annotations. Literature curation to the GO across a wider range of different species will provide more detailed and species-specific information in addition to informing functional annotation in closely related species.

Our *ECR* also reflected the importance of species-specific GO annotation. However, GO evidence code usage changes over time and the IEA and ISS evidence codes are particularly broad. To assess how the *ECR* may skew results we did additional analyses using different ranking systems (Supplementary Data) but the *meanGAQ* showed little change. We hypothesized that annotations based on direct experimental support will provide the 'best-case scenario' for assessing the GAQ and this is supported by our results (Figure 3). The use of GO evidence codes is evolving and that ranking GO evidence codes should be done knowledgably and to best suit the needs of specific datasets, questions and requirements.

To test the *GAQ* score we measured the GO annotation effort over a period of time and we also assessed GO quality for different sub-areas of the GO for both chicken and mouse. We chose chicken and mouse because they represent two species that we expected to have very different bodies of literature (based on the fact that the mouse is a purely model organism while the chicken is an agricultural species as well as a biomedical model). Moreover, the mouse and chicken GO annotation efforts started at different times and their annotation efforts employed different strategies for annotating literature; moreover, as a GO founding species, mouse annotators were heavily involved in the development of the GO during this period. By tracking *GAQ* score over time, we observed that for the first 5 years of GO annotation effort mouse had more annotations than chicken, but chicken had a higher average *GAQ* score. The mouse annotation effort focuses on biocurating the latest available literature while the biocurators for chicken gene products annotate all the literature for specific gene products, so that initially the average number of annotations per gene product is higher in chicken than that of mouse (eight compared to five). However, when only annotations based on direct experimental evidence are considered, mouse has a higher *meanGAQ* score, reflecting the early emphasis on literature biocuration in this species. A high GAQ score does not necessarily mean the most direct experimental knowledge has been captured for a species; it is more a general annotation coverage. Nevertheless, the improvement of the chicken *GAQ* over time demonstrates the effectiveness of a gene product-directed literature curation effort for newly sequenced species.

By using the *GAQ* score to quantitatively assess GO annotation for different sub-areas of the GO we show that GO annotation does not progress evenly across the ontology. This is in part due to differences in experimental literature available for each species and in part due to the focus of the GO annotation efforts. Analysis of sub-areas is useful as many research projects are directed at specific functional processes. By determining the quality of functional annotation available for different species, researchers may choose to target their research for experimental models that have the best-curated functional data for the processes they are studying.

The ability to assess what functional literature is available for a particular species is very difficult and it was this lack of accessibility for functional data that could be compared across species that initially drove the development of the GO (5). PubMed contains most of the published papers but one of the problems we faced is how to accurately assess the amount of literature (*L*) and functional literature (*Lf*) available for a specific species. We used GeneRIF (12) and GOPubMed (13) to estimate *Lf*. The GeneRIF database contains statements about the function of a gene and each geneRIF entry links to the PubMed ID and the gene name. While anyone may add GeneRIFs, National Library of Medicine (NLM) curators also add GeneRIFs and it may be this effort that skews GeneRIFs numbers to favor human, mouse and rat publications while other species are under-represented. GOPubMed is a sophisticated tool that combines PubMed searching with controlled vocabulary terms and does not have the same species as GeneRIFs. However, adding GOPubMed numbers for publications that have biological process, molecular function or cellular component terms will overestimate the number of papers that have functional literature, as many papers will be counted more than once. Neither method can effectively account for GO term synonyms, recognize variations in gene product names or account for functional data that may not be mentioned in the title and abstract of an article. Trained biocurators are essential for recognizing and curating experimental data from published literature but cannot keep up with the increasing amount of functional literature without improved tools and resources to support biocuration. However, by capturing the different direct experimental evidence for different species it is possible to extrapolate functional data to other, less well-annotated species. Given the increasing number of organisms to which functional genomics and proteomics analyses is applied, providing quality functional annotations for a diverse range of organisms is a critical research need. By developing a quantitative measure to assess GO quality, we provide a means for researchers to make the most of existing GO annotations and for biocurators to more efficiently focus their GO annotation efforts. The GAQ scripts will be freely distributed via the AgBase website (http://www.agbase.msstate.edu) and users provided with assistance in using or calculating *GAQ* scores to suit their specific needs.

In summary, we demonstrate the utility of the *GAQ* score for assessing GO annotation quality in nine different species that have varying levels of GO annotation and by assessing the improvement in GO annotation for both chicken and mouse based on time since a dedicated GO annotation effort commenced for each species. We also show how the *GAQ* score may be used to assess specific areas of the ontologies and this can also be applied to specific datasets (including microarrays). A quantitative assessment of GO quality will help biocurators to better

direct current GO annotation efforts to specific areas that are important for their organisms' research community and provides researchers with valuable information about their model systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Financial assistance for T.J.B. and F.M.M. is provided by the USDA NRI, MSU Office of Research, MSU Bagley College of Engineering, MSU College of Veterinary Medicine, the MSU Life Science and Biotechnology institute and the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number MISV-329140. Funding to pay the Open Access publication charges for this article was provided by the Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University.

Conflict of interest statement. None declared.

REFERENCES

1. Istrail,S., Sutton,G.G., Florea,L., Halpern,A.L., Mobarry,C.M., Lippert,R., Walenz,B., Shatkay,H., Dew,I., Miller,J.R. *et al.* (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA*, **101**, 1916–1921.
2. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Gregory,T.R., Nicol,J.A., Tamm,H., Kullman,B., Kullman,K., Leitch,I.J., Murray,B.G., Kapraun,D.F., Greilhuber,J. and Bennett,M.D (2007) Eukaryotic genome size databases. *Nucleic Acids Res.*, **35**, D332–D338.
4. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
6. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
7. Lewis,S.E. (2005) Gene ontology: looking backwards and forwards. *Genome Biol.*, **6**, 103.
8. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
9. Biswas,M., O'Rourke,J.F., Camon,E., Fraser,G., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N. *et al.* (2002) Applications of InterPro in protein annotation and genome analysis. *Brief. Bioinform.*, **3**, 285–295.
10. Alterovitz,G., Xiang,M., Mohan,M. and Ramoni,M.F. (2007) GO PaD: the Gene Ontology Partition Database. *Nucleic Acids Res.*, **35**, D322–D327.
11. McCarthy,F.M., Bridges,S.M., Wang,N., Magee,G.B., Williams,W.P., Luthe,D.S. and Burgess,S.C. (2006) AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res.*, **35**, D599–D603.
12. Lu,Z., Cohen,K.B. and Hunter,L. (2006) Finding GeneRIFs via gene ontology annotations. *Pac. Symp. Biocomput.*, **52–63.**
13. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33**, W783–W786.

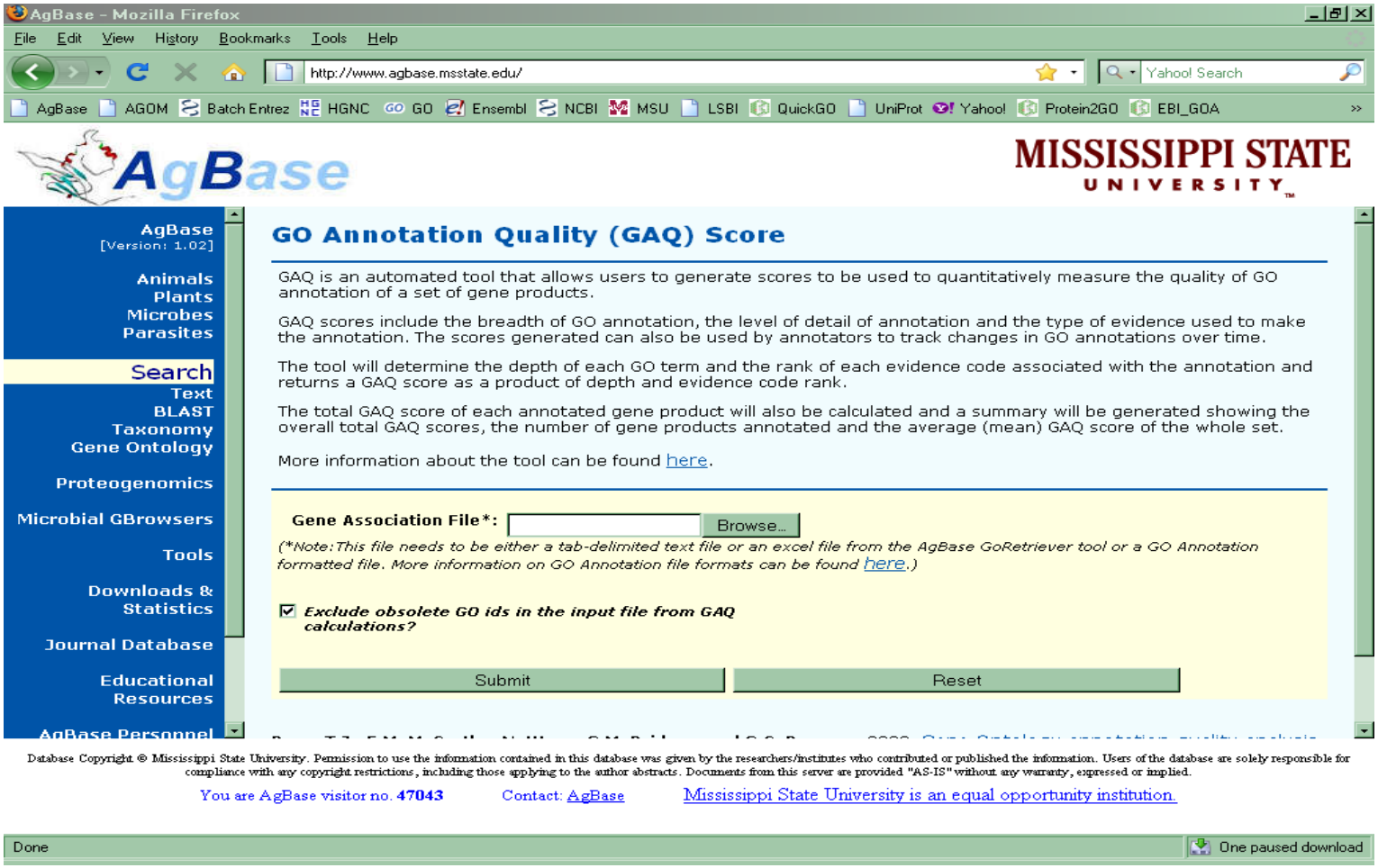

Figure 4.5. Website for calculating GAQ scores: http://www.agbase.msstate.edu/

Table 4.7 Format of an input file to upload when running the *GAQ* program

Column Name	Entry 1	Entry 2	Entry 3
DB	UniProtKB	UniProtKB	GenBank
DB_Object_ID	Q03237	Q5ZKC5	XP_426579
DB_Object_Symbol	MYBL2	YAF2	Rapgef4
Qualifier			
GO ID	GO:0045449	GO:0005634	GO:0030552
GO name	regulation of transcription	nucleus	cAMP binding
DB:Reference(|DB:Reference)	GOA:interpro|GO_REF:0000002	UniProtKB:Q5ZKC5	GenBank:XP_428446
Evidence	IEA	ISS	ISO
With(or)From	InterPro:IPR012287	UniProtKB:Q8IY57	UniProtKB:Q9Z1C7
Aspect	P	C	F
DB_Object_Name	Myb-related protein B	YY1 associated factor 2	similar to cAMP-GEFII
DB_Object_Synonym(|Synonym)	BMYB|IPI00581971	RCJMB04_11m21	Cgef2|Epac2|CAMP-GEFII
DB_Object_Type	protein	protein	Protein
Taxon	taxon:9031	taxon:9031	taxon:9031
Date	20090112	20060524	20090318
Assigned_by	UniProtKB	AgBase	TJB

NOTE: The input file should contain 16 columns matching the format of the gene association file as downloaded from GORetriever at: http://www.agbase.msstate.edu/

Table 4.8 Sample of GAQ output file # 1 showing GAQ score calculated for each GO term associated with the gene product

Gene _ID	GO_ID	GO_name	DB:Reference	With(or)From	Aspect	Evidence	GAQ_score	Date
Q91018	GO:0001889	liver development	PMID:11789987		P	IEP	18	20090211
Q91018	GO:0003677	DNA binding	GO_REF:0000004	SP_KW:KW-0371	F	IEA	8	20090722
Q91018	GO:0005634	nucleus	PMID:11850194		C	IDA	25	20090211
Q91018	GO:0005737	cytoplasm	PMID:11850194		C	IDA	30	20090211
Q5GQ97	GO:0016020	membrane	GO_REF:0000004	SP_KW:KW-0472	C	IEA	8	20090722
Q5GQ97	GO:0016021	integral to membrane	GO_REF:0000004	SP_KW:KW-0812	C	IEA	14	20090722
Q70GM8	GO:0007219	Notch signaling pathway	PMID:14722768	UniProtKB:Q9R229	P	ISS	12	20080604
Q70GM8	GO:0007512	adult heart development	PMID:14722768	UniProtKB:Q9R229	P	ISA	14	20090303
Q90ZG0	GO:0005624	membrane fraction	PMID:11341768	UniProtKB:P26883	C	ISA	12	20090205
Q90ZG0	GO:0032526	response to retinoic acid	PMID:14511757		P	IEP	21	20090204
Q90ZG0	GO:0048185	activin binding	PMID:11341768	UniProtKB:P62942	F	ISA	10	20090205
Q90ZG0	GO:0004871	signal transducer activity	PMID:11341768	UniProtKB:P62942	F	ISA	6	20090205
P21760	GO:0005215	transporter activity	GO_REF:0000002	InterPro:IPR002345	F	IEA	4	20090716
P21760	GO:0005615	extracellular space	PMID:10777107		C	IDA	20	20090610
P21760	GO:0008283	cell proliferation	PMID:12891703		P	IMP	15	20080729
Q9PVN4	GO:0016049	cell growth	GO_REF:0000002	InterPro:IPR003942	P	IEA	10	20090716
Q9PVN4	GO:0040007	growth	GO_REF:0000002	InterPro:IPR001111	P	IEA	4	20090716
Q90998	GO:0003007	heart morphogenesis	PMID:10092230		P	IDA	35	20080816
Q90998	GO:0004872	receptor activity	GO_REF:0000004	SP_KW:KW-0675	F	IEA	8	20090722
P30371	GO:0048666	neuron development	GO_REF:0000024	UniProtKB:P27090	P	ISS	12	20051205
P30371	GO:0031012	extracellular matrix	PMID:10401723	UniProtKB:P61812	C	ISA	8	20090424
P30371	GO:0001654	eye development	UniProtKB:P30371	UniProtKB:P61812	P	ISS	14	20090616

Table 4.9 Sample of GAQ output file # 2 showing summary of GAQ score of individual gene product and the mean GAQ score of the whole set

Gene_product_ID	**GAQ score**
Q91018	81
Q5GQ97	22
Q70GM8	26
Q90ZG0	49
P21760	24
P21760	15
Q9PVN4	14
Q90998	43
P30371	34
SUMMARY	
Total GAQ score	308
Number of annotated gene products	8
Mean GAQ score	**38.5**

NOTE: GAQ scores for each gene product are summed-up and a summary is generated.

CHAPTER 5

FACILITATING FUNCTIONAL ANNOTATION OF CHICKEN MICROARRAY DATA[1]

[1] Reprint from T.J. Buza, R. Kumar, C.R. Gresham, S. C. Burgess, F.M. McCarthy.2009. Facilitating functional annotation of chicken microarray data. BMC Bioinformatics 2009, 10(Suppl 11):S2
This article is available from: http://www.biomedcentral.com/1471-2105/10/S11/S2

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Facilitating functional annotation of chicken microarray data

Teresia J Buza*[1,2], Ranjit Kumar[1,2], Cathy R Gresham[2,5], Shane C Burgess[†1,2,3,4] and Fiona M McCarthy[†1,2]

Address: [1]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA, [2]Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA, [3]Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS 39762, USA, [4]Mississippi Agricultural and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA and [5]Department of Computer Science and Engineering, Bagley College of Engineering, Mississippi State University, Mississippi State, MS 39762, USA

E-mail: Teresia J Buza* - tbuza@cvm.msstate.edu; Ranjit Kumar - rkumar@cvm.msstate.edu; Cathy R Gresham - gresham@cse.msstate.edu; Shane C Burgess - fmccarthy@cvm.msstate.edu; Fiona M McCarthy - burgess@cvm.msstate.edu

*Corresponding author †Equal contributors

from Sixth Annual MCBIOS Conference. Transformational Bioinformatics: Delivering Value from Genomes
Starkville, MS, USA 20–21 February 2009

Published: 08 October 2009

BMC Bioinformatics 2009, **10**(Suppl 11):S2 doi: 10.1186/1471-2105-10-S11-S2

This article is available from: http://www.biomedcentral.com/1471-2105/10/S11/S2

Abstract

Background: Modeling results from chicken microarray studies is challenging for researchers due to little functional annotation associated with these arrays. The Affymetrix GenChip chicken genome array, one of the biggest arrays that serve as a key research tool for the study of chicken functional genomics, is among the few arrays that link gene products to Gene Ontology (GO). However the GO annotation data presented by Affymetrix is incomplete, for example, they do not show references linked to manually annotated functions. In addition, there is no tool that facilitates microarray researchers to directly retrieve functional annotations for their datasets from the annotated arrays. This costs researchers amount of time in searching multiple GO databases for functional information.

Results: We have improved the breadth of functional annotations of the gene products associated with probesets on the Affymetrix chicken genome array by 45% and the quality of annotation by 14%. We have also identified the most significant diseases and disorders, different types of genes, and known drug targets represented on Affymetrix chicken genome array. To facilitate functional annotation of other arrays and microarray experimental datasets we developed an Array GO Mapper (*AGOM*) tool to help researchers to quickly retrieve corresponding functional information for their dataset.

Conclusion: Results from this study will directly facilitate annotation of other chicken arrays and microarray experimental datasets. Researchers will be able to quickly model their microarray dataset into more reliable biological functional information by using *AGOM* tool. The disease, disorders, gene types and drug targets revealed in the study will allow researchers to learn more about how genes function in complex biological systems and may lead to new drug discovery and development of therapies. The GO annotation data generated will be available for public use via AgBase website and will be updated on regular basis.

Background

The development of microarray high-throughput screening platforms for chicken is an important step for gene expression profiling in changes occurring in avian as a response to different challenges and stimuli [1-3]. The chicken research community uses microarrays for a wide range of applications, including gene expression analysis [1,4], exon expression analysis [5-7], novel transcript discovery [8], genotyping [9,10] and resequencing [11,12]. In addition, microarray analysis can also be combined with chromatin immunoprecipitation to perform genome-wide identification of transcription factors and their respective binding sites [13].

According to statistics obtained from "Gallus Expression *in Situ* Hybridization Analysis" (GEISHA; http://geisha.arizona.edu/geisha/microarray.jsp; 03/14/2009), there is already significant resources constructed for the "Whole Genome" Chicken Microarrays. Listed in GEISHA are: 1) Arizona *Gallus gallus* 20.7 K Long Oligo Array, 2) Affymetrix array which cover 32,773 transcripts corresponding to over 28,000 chicken genes, 3) FHCRC Chicken 13 K Array, 4) University of Delaware-Larry Cogburn which produced UD_Liver_3.2 K, UD 7.4 K Metabolic/Somatic Systems, Chicken Neuroendocrine System 5 K and the DEL-MAR 14 K Integrated Systems and 5) ARK Genomics which offers a 1,153 clone chicken embryo array, a 5,000 cDNA chicken immune array, and a 4,800 clone chicken neuroendocrine array. Gene Expression Omnibus (GEO), publicly accessible through the World Wide Web at http://www.ncbi.nlm.nih.gov/geo, is a curated public repository for high-throughput gene expression data [14,15]. Platform is one of central data entities of GEO which contains a list of probes that define what set of molecules may be detected and can easily be browsed, queried and retrieved to fit user's interests [14,16].

Comprehensive annotation of these arrays will benefit chicken researchers, because they will be able to functionally model their expressed dataset to obtain relevant information about their biological system. However, most arrays are not associated to any functional information. The only array that is comprehensively annotated to GO is the Affymetrix chicken GeneChip array [17]. This array is the mostly used for gene expression studies as shown in a survey when the chicken research community was polled in July 2008 http://doodle.com/participation.html?pollId=zwvmhpt5t23tvfv8. The Affymetrix NetAffx database links probesets on Affymetrix GenChip microarrays to GO using data from the GO Consortium [18]. However, the GO evidence codes are not linked to any reference that was used to make functional assertions. This is a challenge to researchers who want to associate their dataset with functional information at the same time showing supporting evidence. For example, use of an experimental evidence code in a GO annotation should be associated with a paper that displays results from a physical characterization of a gene/gene product being annotated. This allows the researcher to access the detailed information that was used to make the GO annotation.

In this study we have re-annotated all gene products associated with probesets on Affymetrix chicken genome array using GO standards. However, the GO describes normal gene or gene product function [19] such that information about which genes are associated with significant diseases and disorders and which are known to be drug targets is not captured using the GO. This type of information would clearly benefit researchers in modeling diseases. We therefore used Ingenuity Pathway Analysis to identify significant diseases, disorders, drug targets and types of gene represented on Affymetrix chicken genome array. Furthermore, we demonstrate how other microarrays can be annotated using the annotations from Affymetrix chicken array.

Results

Initial assessment of structural and functional annotation of chicken array

Most of chicken arrays currently available are linked to either gene or gene products but very few of the arrays are annotated to any functional information (Table 1). The Affymetrix chicken array was chosen for this study because it represents most of genomic elements annotated on chicken genome. Initial assessment of annotation of Affymetrix chicken genome array are shown (Additional file 1). Over 97% of chicken Affymetrix probesets are mapped to 27,852 genes or gene products in total. Other probesets represented on this array are for studying 17 different avian viruses. About 51% of the probesets are associated with GO annotations made for 12,457 genes or gene product.

Functional annotation and GO annotation quality

The GO annotation of Affymetrix chicken probesets does not show any reference supporting the evidence of the annotation as pointed out in methods section. We re-annotated all gene products represented on this array, regardless of their initial annotations, according to GO standards. We were able to increase the number of GO annotations in all three ontologies (Figure 1); re-annotation increased the total GO annotations by 45%, the number of annotated gene products by 10% and the number of probe sets linked to annotated gene

Table 1: Initial assessment of structural and functional annotation of chicken array

Name of Microarray	Size	Cross reference		GO	Evd*
		Gene/EST	Protein		
ARK-Genomics G. gallus 20 K v1.0 (GPL5480)	22,176	+	-	-	-
ARK-Genomics G. gallus 13 K v4.0 (GPL5673)	27,648	+	-	-	-
Affymetrix GenChip® chicken genome array	38,535	+	+	+	+
Chicken 44 K custom Agilent microarray (GPL4993)	42,034	+	+	-	-
Arizona Gallus gallus 20.7 K Oligo Array v1.0 (GPL6049)	21,120	+	-	-	-
FHCRC Chicken 13 K Array (GPL1836)	15,769	+	-	-	-
Custom 4 × 2 K miRNA microarray (#4166) (GPL7472)	1,412	+	-	-	-
Chick Pineal 2004 (GPL1289)	9,056	+	+	-	-
DEL-MAR 14 K Integrated Systems(GPL1731)	19,200	+	-	-	-
Avian Innate Immunity Microarray (AIIM) (GPL1461)	14,877	+	-	-	-
UD 7.4 K Metabolic/Somatic Systems (GPL1737)	7,680	+	-	-	-
UD_Liver_3.2 K (GPL1742)	3,456	+	-	-	-
Chicken_Neuroendocrine_System_5 K (GPL1744)	7,000	+	-	-	-

Different chicken arrays (column 1) have different gene products represented on them (column 2). Column 3 & 4 shows whether the printed transcripts are linked to a gene (G), mRNA (R), EST or protein. GO in column 5 indicates GO functional annotation linked to gene products represented on these arrays and evd (column 6) indicates evidence code supporting the functional information. The (+) or (-) in columns 3 – 6 indicates presence or absence of the parameter in that specific column.
+ or **-** shows that the array is linked or not linked.
*Evidence that support the GO annotation.

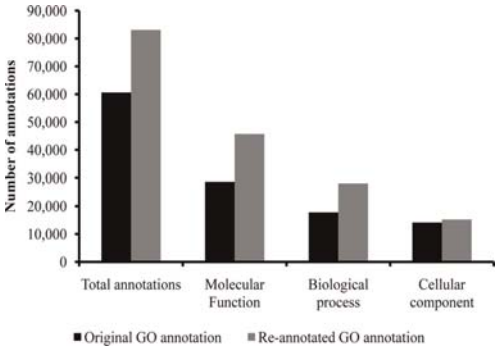

Figure 1
Functional annotation of Affymetrix chicken genome array. Original annotation of Affymetrix chicken array (grey bars) were compared with re-annotated GO (black bars). All biological ontologies show improvements realized from the re-annotation.

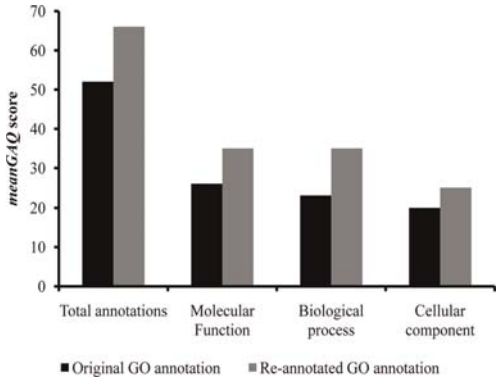

Figure 2
The mean GAQ score of the GO annotation. The mean GAQ scores are calculated for both original (black bar) and re-annotated (grey bar) GO annotations. The mean GAQ score is based only on the unique gene products with GO, not individual the probesets.

products by 13%. Moreover, the quality of the original GO annotations in all three GO ontologies, as determined by GAQ score [20], was improved by the additional annotations (Figure 2). Briefly, the GAQ score quantitatively assess the level of detail provided by the GO annotation and the type of evidence used to make the annotations. The overall mean GAQ score of all annotations regardless of biological ontology, increased from 52 to 66.

Additional functional information was obtained using the Ingenuity Pathway Analysis (IPA) tool to identify the significant biological functions, diseases and disorders that are represented on Affymetrix chicken genome array (Table 2). The most significant diseases and disorders represented on this array are cancer and genetic disorders, respectively. Cell death was identified to be the most significant molecular and cellular function while organismal survival was the most significant

Table 2: Biological functions represented on Affymetrix chicken GenChip® array

Biological Function	Number of Genes	P-value*
Diseases and Disorders		
Cancer	2,298	2.43E-53 – 6.86E-08
Neurologic disease	1,219	4.94E-52 – 6.76E-08
Genetic disorder	1,152	6.69E-37 – 6.69E-37
Cardiovascular disease	583	3.18E-36 – 6.17E-08
Developmental disease	554	6.01E-30 – 6.17E-08
Molecular and Cellular Functions		
Cell death	1,604	1.19E-55 – 6.51E-08
Cellular growth and proliferation	1,774	6.66E-42 – 4.87E-08
Cellular development	1,231	1.00E-35 – 5.68E-08
Gene expression	1,231	2.82E-35 – 2.21E-08
Cellular movement	931	1.89E-32 – 6.78E-08
Physiological System Development and Function		
Organismal survival	718	5.95E-38 – 1.18E-12
Tissue development	920	6.07E-36 – 4.30E-08
Organismal development	879	5.40E-34 – 5.36E-08
Organ development	585	2.33E-33 – 4.90E-08
Tissue morphology	666	2.03E-27 – 1.20E-08

Significant biological functions represented on Affymetrix chicken genome array.
*Based on Fisher's Exact Test P-value ≤ 0.05.

process among the physiological system development. Different types of genes and known drug targets were also identified (Figure 3).

Tool for array GO mapping

Improved functional annotation of Affymetrix chicken array proved to facilitate the annotation of other arrays, such as the Arizona Gallus gallus 20.7 K Oligo Array v1.0 (GPL6049). An Array GO Mapper (*AGOM*) tool developed in this study was able to map Entrez genes, Ensembl genes and GenBank accessions from the Arizona array to Affymetrix annotations in order to retrieve GO annotations. We successfully identified 79% of genes that were common in both arrays (Figure 3), out of which 72% were mapped to GO annotations (Figure 4). The total number of GO annotations generated for Arizona array was 60,846. An example of output generated by *AGOM* is shown on additional file 2 which includes only the first 1,000 gene association lines generated for Arizona chicken array. The mean GAQ score associated with the GO annotations retrieved was 59 and was calculated by summing up all GAQ scores of all 60,846 GO associations and dividing these by the number of annotated gene products. These results provide an initial assessment of GO annotations available for the Arizona chicken array and demonstrates how GO annotations can be transferred to identical transcriptional elements represented on multiple arrays.

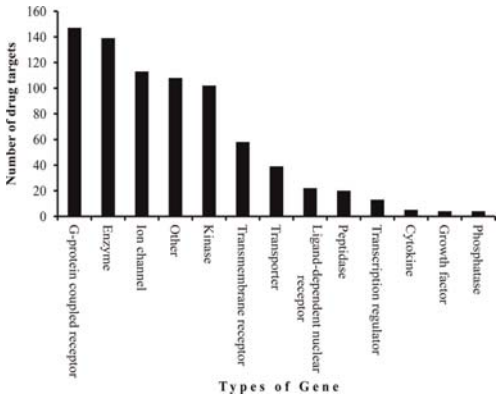

Figure 3
Types of genes and drug targets represented on Affymetrix GenChip® chicken genome array Sample figure title. The probesets matching different types of genes (A) were determined by using Ingenuity Pathway Analysis software. Some probesets were mapped to genes that are considered drug targets (B).

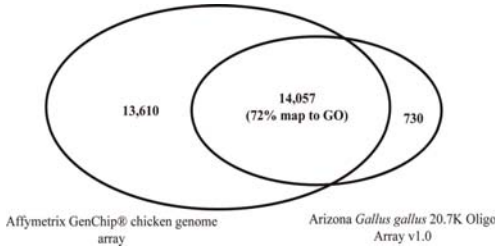

Figure 4
Distribution of genes and gene products represented on Affymetrix and Arizona chicken array.

Discussion

The major challenge that faces microarray researchers is interpretation of hundreds of differentially expressed genes into a biologically relevant context. The Gene Ontology (GO) Consortium provides a controlled vocabulary to annotate the biological knowledge associated with genes or gene products. In order to make the functional interpretation of microarray dataset less challenging, microarray developers can associate their arrays with functional information.

However, most chicken arrays either have no associated GO information or do not follow the GO annotation standards [21]. In this study we have re-annotated and improved the GO annotation of Affymetrix chicken

genome array to facilitate annotation of other chicken arrays and microarray experimental datasets. Further, we developed the Array GO Mapper (*AGOM*) tool to generate GO annotations for chicken arrays with no GO information or for microarray experimental datasets and demonstrated its utility by annotating the Arizona chicken array which had no associated GO information. By implementing *AGOM* researchers will not only obtain functional information for their experimental dataset but will also obtain GAQ scores associated with each GO term retrieved. This will help researchers determine the quality of annotations made to their datasets and also help tracking the improvement made by any additional GO when there are any updates.

We also provided additional functional information not covered by the GO but is associated with the Affymetrix chicken genome array. This additional data broadens the ability of array users to model their datasets, for example infectious disease datasets. The additional information obtained on diseases, disorders and known drug targets represented on this array will provide light to future research in drug and therapy development.

Conclusion

Improved amount and quality of GO annotations of gene products represented on the Affymetrix chicken genome array will help researchers to model their genes of interest to high quality functional information by using AGOM tool. The existing chicken microarray studies can use AGOM and this demonstrates how this tool can enhance functional annotation in these studies. Annotation of microarrays of other species will be included in the future. The top significant diseases and disorders represented on the chicken array correlate well with how the chicken is used as a biomedical model organism to study human diseases and development. The identified gene types and drug targets allows researchers to learn more about how genes function in complex biological systems and may lead to new drug discovery and development of therapies.

Methods

Initial assessment of structural and functional annotation of chicken array

We downloaded 12 chicken array platforms deposited in the NCBI Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/) database (Table 1). Affymetrix GenChip chicken genome array annotations were downloaded from the Affymetrix website http://www.affymetrix.com. In each array we assessed whether the printed transcripts were structurally linked to any gene, EST or protein. Gene Ontology (GO) was used as criteria for initial assessment of functional annotation. The purpose of this assessment was to determine which whole chicken genome arrays could be used as reference for structural and functional annotation of other arrays or experimental datasets. Affymetrix chicken genome array was the only one that had been comprehensively structurally and functionally annotated and was selected for further improvement.

Functional annotation

Further assessment and improvement of GO annotation of the Affymetrix chicken array was necessary. The GO annotations associated with the probe sets on Affymetrix chicken array do not show detail information to support the annotation. For example; were experimental evidence codes are shown there is no any literature referenced to support the annotation. For this reason we decided to re-annotate all gene products linked to the probesets on this array, regardless of their original annotations, in order to provide high quality and standard functional information to the array users. We first used *GORetriever* [22] to download chicken GO annotations for all UniProtKB accessions linked to the probesets. Further annotations for linked gene products with RefSeq number and Ensembl gene identifiers were obtained from AgBase-community databases and Gene Ontology Annotation (GOA) project using an in-house Perl script (*GOMapper.pl*). Additional GO was retrieved by implementing an in-house tool (*ISO.pl*) to transfer the experimental GO annotations from 1:1 chicken-human/mouse/rat orthologs to the corresponding chicken proteins orthologs. The improved GO annotations will be made available publicly via AgBase.

Additional functional information

In addition to the molecular function, biological process and cellular component annotations provided via the GO, other functional information is also useful for researchers wishing to assess the type of biological information represented by transcript printed on an array. For example, researchers will also benefit by knowing which genes on the array are associated with disease and disorders and which are known drug targets. We used Ingenuity Pathways Analysis (IPA) software to determine known drug targets and significant disease and disorders. The Fischer's exact test was used to calculate a P-value determining the probability that the biological functions, diseases or disorders assigned to the array datasets was due to chance alone.

Assessment of GO annotation quality (GAQ)

To assess the improvement made in the re-annotated functional annotations of the Affymetrix chicken array, the *meanGAQ* score for GO initially associated with the array was calculated as previously described [20] and

compared to that calculated for GO after re-annotations. Briefly, the GAQ score takes into account the quality of GO annotations by quantitatively assessing the level of detail provided by the GO annotation and the type of evidence used to make the functional association. Mathematically the ***GAQ*** score of a GO annotation ***(a)*** can be defined as the product of annotation depth in the ontology ***(Dd)*** and the evidence code rank ***(ECR)*** of the annotation, represented as:

$$GAQ(a) = ECR_a \cdot Dd_a$$

When you have a set of gene products ***(S)*** annotated to a number of GO terms ***(A)***, the ***GAQ*** score can be defined as:

$$GAQ(S) = \sum_{a=1}^{A} (ECR_a \cdot Dd_a)$$

In this study we reported the **mean GAQ** score based on number of gene products ***(n)*** that have GO and was calculated as:

$$meanGAQ(S) = GAQ(S) / n$$

Development of Array GO Mapper (AGOM)

AGOM was developed to GO annotate chicken arrays and chicken microarray experimental datasets using improved Affymetrix GO annotations generated in the work described here. The tool is written in Perl and works on both windows and Linux platforms. It requires a tab delimited input file containing the microarray dataset cross references for which the GO annotations are searched. The Affymetrix improved GO data file was used as a database to search from. This database contains 6 cross-reference identifier types, which facilitate mapping between arrays and experimental datasets. *AGOM* works with any type of array (whole genome and specific array platform) and experimental datasets with common identifier(s) between the arrays/datasets and the Affymetrix data. The gene associations are presented in 16 columns according to GO standards (Additional file 3). The depth of a GO term, evidence code rank and GAQ score of individual GO term associated with the Affymetrix GO data are in the last 3 columns of file.

We demonstrated *AGOM* implementation by searching GO annotation for Arizona chicken array (GPL6049) from improved Affymetrix chicken array GO data. The Arizona chicken array was chosen because it has no existing GO associated with its gene products (Table 1). In addition, the Arizona array probes are linked to a variety of identifiers (GenBank accession, Entrez Gene ID and Ensembl ID) that can be used to search the Affymetrix GO data while most of other arrays contain only GenBank accessions (Additional file 3). For example, in this study GenBank accession, Entrez Gene ID and Ensembl ID linked to Arizona array were searched against the improved Affymetrix GO annotations to retrieve corresponding GO records. The output generated from the search includes Arizona array identifiers in the first 5 columns; Oligo_ID (unique ID), GenBank accession, Entrez Gene ID, Ensembl ID and array Spot number. When a match is found the corresponding GO information is added to a tab-delimited output file.

AGOM is available via AgBase (http://www.agbase.msstate.edu/; see under Array annotation) where users can use the tool directly online or can download it as a standalone program. When implementing the tool online, users will be given options to retrieve any data associated with the Affymetrix chicken array (Additional file 3). The script is also available upon request and advice is available by e-mail.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TJB is responsible for the study designing, data generation, data analysis, formulation of workflow for *AGOM* and writing the draft of the manuscript. RK wrote the script for *AGOM* and contributed in manuscript preparation. CRG modified the script, developed *AGOM* webpage and contributed in manuscript preparation. Both FMM and SCB contributed in manuscript preparation and in technical advice. All authors read and approved the final manuscript.

Additional material

Additional file 1
Initial assessment of annotation of Affymetrix chicken genome array. *Additional file descriptions text (including details of how to view the file, if it is in a nonstandard format).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S11-S2-S1.pdf]

Additional file 2

Example of output generated by the Array GO Mapper (AGOM). *Unique identifier for the Arizona array (Oligo_ID) is displayed in column 1. Column 2–4 displays GenBank accessions, Entrez gene ID and Ensembl gene ID used for mapping. The array spot number is in column 5. The name of database and the corresponding gene product in Affymetrix annotations are shown in column 6 & 7. The GO and name of the GO term are displayed in column 8 & 9 with the evidence code for the annotation in column 10. Column 11 shows the aspects of gene ontology either molecular function (F), cellular component (C) or biological process (P). The GO Annotation Quality (GAQ) score for individual GO term is displayed in column 12 and the date the output was generated in column 13.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S11-S2-S2.xls]

Additional file 3

Chicken array platform cross-reference. *Each column represents one array platform showing the identifiers that can be used to search GO annotations from Affymetrix GO data. (+) indicates presence of identifier in the corresponding array platform. (-) indicates absence of identifier in the corresponding array platform.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S11-S2-S3.xls]

Acknowledgements

The project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number MISV-329140.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 11, 2009: Proceedings of the Sixth Annual MCBIOS Conference. Transformational Bioinformatics: Delivering Value from Genomes. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S11.

References

1. Heidari M, Huebner M, Kireev D and Silva RF: **Transcriptional profiling of Marek's disease virus genes during cytolytic and latent infection.** *Virus Genes* 2008, **36(2):**383–392.
2. Sarson AJ, Read LR, Haghighi HR, Lambourne MD, Brisbin JT, Zhou H and Sharif S: **Construction of a microarray specific to the chicken immune system: profiling gene expression in B cells after lipopolysaccharide stimulation.** *Can J Vet Res* 2007, **71(2):**108–118.
3. Smith J, Speed D, Hocking PM, Talbot RT, Degen WG, Schijns VE, Glass EJ and Burt DW: **Development of a chicken 5 K microarray targeted towards immune function.** *BMC Genomics* 2006, **7:**49.
4. Masker K, Golden A, Gaffney CJ, Mazack V, Schwindinger WF, Zhang W, Wang LH, Carey DJ and Sudol M: **Transcriptional profile of Rous Sarcoma Virus transformed chicken embryo fibroblasts reveals new signaling targets of viral-src.** *Virology* 2007, **364(1):**10–20.
5. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A and Blume JE: **Discovery of tissue-specific exons using comprehensive human exon microarrays.** *Genome Biol* 2007, **8(4):**R64.
6. Grigoryev DN, Ma SF, Shimoda LA, Johns RA, Lee B and Garcia JG: **Exon-based mapping of microarray probes: recovering differential gene expression signal in underpowered hypoxia experiment.** *Mol Cell Probes* 2007, **21(2):**134–139.
7. Xing Y, Kapur K and Wong WH: **Probe selection and expression index computation of Affymetrix Exon Arrays.** *PLoS ONE* 2006, **1:**e88.
8. Cao W, Epstein C, Liu H, DeLoughery C, Ge N, Lin J, Diao R, Cao H, Long F and Zhang X, *et al*: **Comparing gene discovery from Affymetrix GeneChip microarrays and Clontech PCR-select cDNA subtraction: a case study.** *BMC Genomics* 2004, **5(1):**26.
9. Butcher LM, Meaburn E, Liu L, Fernandes C, Hill L, Al-Chalabi A, Plomin R, Schalkwyk L and Craig IW: **Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits.** *Behav Genet* 2004, **34(5):**549–555.
10. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I and Schalkwyk LC: **Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs.** *BMC Genomics* 2005, **6(1):**52.
11. Corless CE, Kaczmarski E, Borrow R and Guiver M: **Molecular characterization of Neisseria meningitidis isolates using a resequencing DNA microarray.** *J Mol Diagn* 2008, **10(3):** 265–271.
12. Lebet T, Chiles R, Hsu AP, Mansfield ES, Warrington JA and Puck JM: **Mutations causing severe combined immunodeficiency: detection with a custom resequencing microarray.** *Genet Med* 2008, **10(8):**575–585.
13. Chung HR, Kostka D and Vingron M: **A physical model for tiling array analysis.** *Bioinformatics* 2007, **23(13):**i80–86.
14. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M and Marshall KA, *et al*: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37 Database:** D885–890.
15. Edgar R, Domrachev M and Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30(1):**207–210.
16. Zhu Y, Davis S, Stephens R, Meltzer PS and Chen Y: **GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus.** *Bioinformatics* 2008, **24(23):**2798–2800.
17. **GeneChip® Arrays: Chicken Genome Array.** [http://www.affymetrix.com/technology/index.affx, http://www.affymetrix.com/products/arrays/specific/chicken.affx].
18. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D and Siani-Rose MA: **NetAffx: Affymetrix probesets and annotations.** *Nucleic Acids Res* 2003, **31(1):**82–86.
19. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B and Mungall C, *et al*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32 Database:** D258–261.
20. Buza TJ, McCarthy FM, Wang N, Bridges SM and Burgess SC: **Gene Ontology annotation quality analysis in model eukaryotes.** *Nucleic Acids Res* 2008, **36(2):**e12.
21. Rhee SY, Wood V, Dolinski K and Draghici S: **Use and misuse of the gene ontology annotations.** *Nat Rev Genet* 2008, **9(7):** 509–515.
22. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS and Burgess SC: **AgBase: a unified resource for functional analysis in agriculture.** *Nucleic Acids Res* 2007, **35 Database:** D599–603.

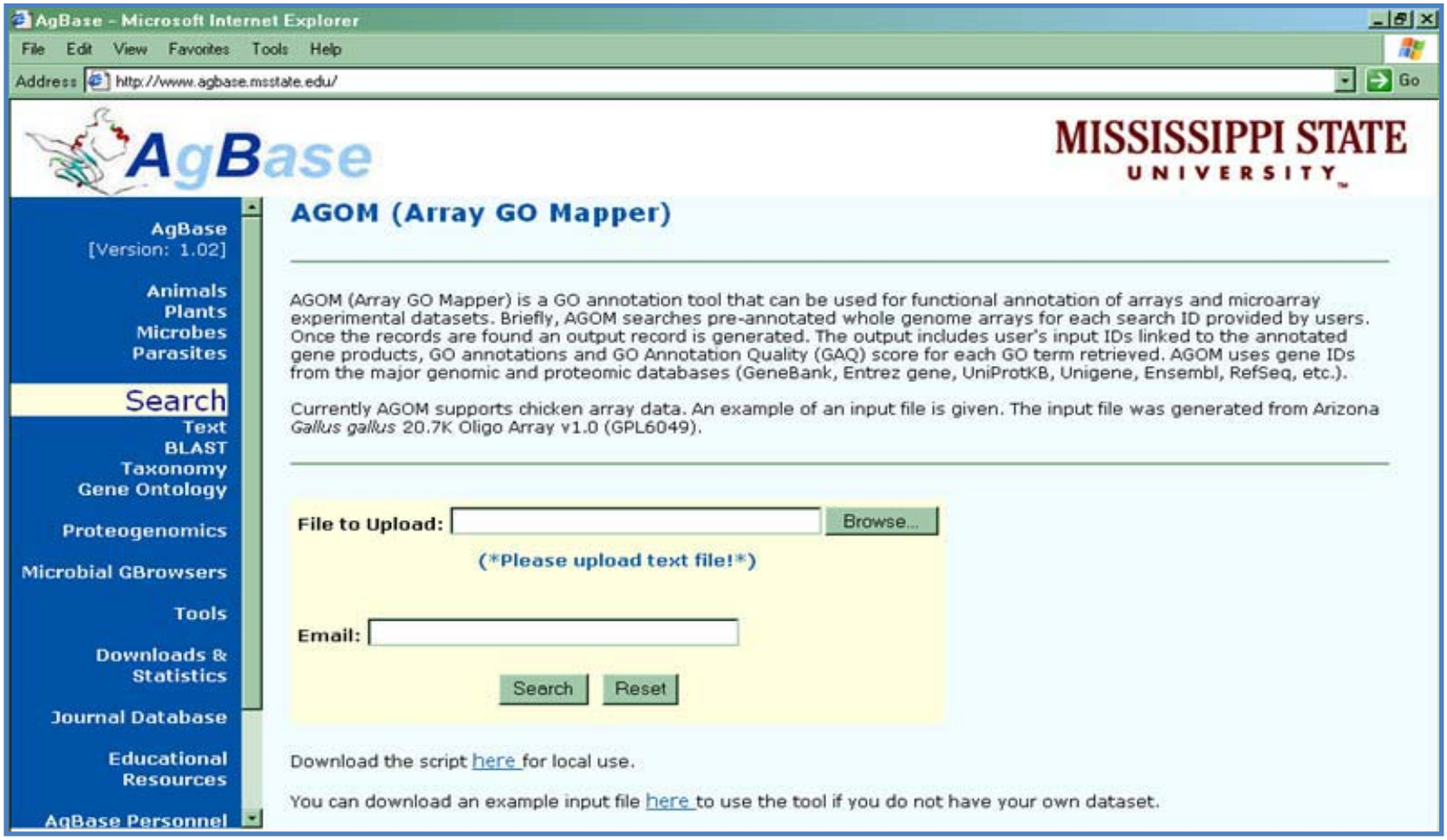

Figure 5 Website for AGOM: http://www.agbase.msstate.edu/

Table 5.3 List of cross reference and gene associations of Affymetrix chicken array for mapping by AGOM

A: Gene product identifiers linked to Affymetrix chicken array probesets				
Column #	Array cross reference	Example 1	Example 2	Example 3
1	Affy_Probeset_ID	Gga.4532.1.S1_at	Gga.1754.1.S1_s_at	GgaAffx.5904.1.S1_at
2	Transcript_ID(Array_Design)	Gga.4532.1	Gga.1754.1	GgaAffx.5904.1
3	Representative public ID:			
	GenBank Accession		BU326744	
	mRNA RefSeq ID	NM_205318		XM_426579
	Ensembl transcript ID			ENSGALT00000015240
4	UniGene_ID	Gga.4532	Gga.1754	Gga.22040
5	Ensembl_gene_ID	ENSGALG00000003503	ENSGALG00000009540	ENSGALG00000009357
6	Entrez_Gene_ID	396258	417790	429022
7	SwissProt_AC	Q03237	Q5ZKC5	
8	RefSeq_Protein_ID	NP_990649	NP_001007851	XP_426579
B: Column names of GO annotation (Adopted AgBase gene association file format)				
9	DB	UniProtKB	UniProtKB	GenBank
10	DB_Object_ID	Q03237	Q5ZKC5	XP_426579
11	DB_Object_Symbol	MYBL2	YAF2	Rapgef4
12	Qualifier			

Table 5.3 (continued)

13	GO_ID	GO:0045449	GO:0005634	GO:0030552
14	GO_name	regulation of transcription	nucleus	cAMP binding
15	DB:Reference(|DB:Reference)	GOA:interpro|GO_REF:0000002	UniProtKB:Q5ZKC5	GenBank:XP_428446
16	Evidence	IEA	ISS	ISO
17	With(or)From	InterPro:IPR012287	UniProtKB:Q8IY57	UniProtKB:Q9Z1C7
18	Aspect	P	C	F
19	DB_Object_Name	Myb-related protein B	YY1 associated factor 2	similar to cAMP-GEFII
20	DB_Object_Synonym	BMYB|IPI00581971	RCJMB04_11m21	Cgef2|Epac2|CAMP-GEFII
21	DB_Object_Type	protein	protein	protein
22	Taxon_ID	taxon:9031	taxon:9031	taxon:9031
23	Date	20090112	20060524	20090318
24	Assigned_by	UniProtKB	AgBase	TJB
25	Submitted_GOA			IDA
C: Gene ontology annotation quality				
26	GO_Depth	6	5	8
27	Evidence_code_rank	2	2	3
28	GAQ_score	12	10	24

Table 5.4 Sample of input file to upload to AGOM for GO annotation of ten gene products linked to Arizona chicken array

OLIGO_ID	GenBank ID	Entrez gene	Ensembl gene	Spot number
RIGG03527	BX933209		ENSGALG00000015746	24.11.17
RIGG03528	BX933190		ENSGALG00000015764	32.11.17
RIGG04490	NM_204158	373965		11.7.16
RIGG05105	CR407421		ENSGALG00000015751	47.19.16
RIGG05209	AF514777	373895		33.19.16
RIGG06460	AF257352	373985	ENSGALG00000008072	18.10.14
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14
RIGG07281	BX932369		ENSGALG00000015763	14.2.13
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3

NOTE: In Arizona chicken array GenBank accessions (column 2), Entrez gene ID (column 3) and Ensembl gene ID (column 4) which are linked to unique OLIGO-ID (column 1) and spot number (column 5) can be used for mapping into Affymetrix chicken array GO database to retrieve GO annotations.

Table 5.5 AGOM output for GO annotation of ten gene products linked to Arizona chicken array

Arizona OLIGO_ID	Arizona GenBank ID	Arizona Entrez gene	Arizona Ensembl gene	Arizona Spot number	DB Object ID	GO ID	Evidence code	Aspect	GAQ score	Date
RIGG03527	BX933209		ENSGALG00000015746	24.11.17					0	20090723
RIGG03528	BX933190		ENSGALG00000015764	32.11.17	Q5ZIR7	GO:0005215	IEA	F	4	20090723
RIGG03528	BX933190		ENSGALG00000015764	32.11.17	Q5ZIR7	GO:0005488	IEA	F	4	20090723
RIGG03528	BX933190		ENSGALG00000015764	32.11.17	Q5ZIR7	GO:0006810	IEA	P	8	20090723
RIGG03528	BX933190		ENSGALG00000015764	32.11.17	Q5ZIR7	GO:0008289	IEA	F	6	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0003779	IEA	F	10	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0005516	IEA	F	8	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0005624	IEA	C	10	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0006936	IEA	P	10	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0017022	IEA	F	10	20090723
RIGG04490	NM_204158	373965		11.7.16	P12957	GO:0030478	IEA	C	20	20090723
RIGG05105	CR407421		ENSGALG00000015751	47.19.16	XP_001231639	GO:0005739	ISO	C	24	20090723
RIGG05105	CR407421		ENSGALG00000015751	47.19.16	XP_001231639	GO:0005743	ISO	C	18	20090723
RIGG05105	CR407421		ENSGALG00000015751	47.19.16	XP_416685	GO:0005739	ISO	C	24	20090723
RIGG05105	CR407421		ENSGALG00000015751	47.19.16	XP_416685	GO:0005743	ISO	C	18	20090723
RIGG05209	AF514777	373895		33.19.16	Q8AYE5	GO:0003677	IEA	F	8	20090723
RIGG05209	AF514777	373895		33.19.16	Q8AYE5	GO:0003700	IEA	F	10	20090723
RIGG05209	AF514777	373895		33.19.16	Q8AYE5	GO:0005634	IEA	C	10	20090723
RIGG05209	AF514777	373895		33.19.16	Q8AYE5	GO:0006355	IEA	P	14	20090723
RIGG05209	AF514777	373895		33.19.16	Q8AYE5	GO:0045449	IEA	P	12	20090723

Table 5.5 (continued)

RIGG06460	AF257352	373985	ENSGALG00000008072	18.10.14	Q90ZD5	GO:0005576	IEA	C	4	20090723
RIGG06460	AF257352	373985	ENSGALG00000008072	18.10.14	Q90ZD5	GO:0007275	IEA	P	6	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	A5A2G0	GO:0000166	IEA	F	6	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	A5A2G0	GO:0003676	IEA	F	6	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0000166	IEA	F	6	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0003676	IEA	F	6	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0003723	IEA	F	8	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0005634	IEA	C	10	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0005737	IEA	C	12	20090723
RIGG06663	NM_001012521	373923	ENSGALG00000008097	5.11.14	Q5F3T7	GO:0006397	IEA	P	14	20090723
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14	Q98TF6	GO:0003735	IEA	F	6	20090723
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14	Q98TF6	GO:0005622	IEA	C	8	20090723
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14	Q98TF6	GO:0005840	IEA	C	16	20090723
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14	Q98TF6	GO:0006412	IEA	P	12	20090723
RIGG07052	AB046396	373936	ENSGALG00000000474	19.13.14	Q98TF6	GO:0030529	IEA	C	12	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0000122	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0001701	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0003677	IEA	F	8	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0003700	IEA	F	10	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0005515	IEA	F	6	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0005634	IEA	C	10	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0006355	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0006357	IEA	P	12	20090723

Table 5.5 (continued)

RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5DWF6	GO:0043565	IEA	F	10	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0000122	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0001701	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0003677	IEA	F	8	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0003700	IEA	F	10	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0005515	IEA	F	6	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0005634	IEA	C	10	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0006355	IEA	P	14	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0006357	IEA	P	12	20090723
RIGG07281	BX932369		ENSGALG00000015763	14.2.13	Q5ZJ46	GO:0043565	IEA	F	10	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0003677	IEA	F	8	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0003700	IEA	F	10	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0004871	IEA	F	6	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0005634	IEA	C	10	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0006350	IEA	P	10	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0006355	IEA	P	14	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0007165	IEA	P	8	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0030528	IEA	F	4	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0045449	IEA	P	12	20090723
RIGG17763	AF246958	373925	ENSGALG00000014096	18.10.3	Q8QGQ7	GO:0048511	IEA	P	4	20090723

NOTE: Unique identifiers for the Arizona chicken array are displayed in column 1 (Oligo ID) & 5 (Spot ID) which are linked to different gene product identifiers (column 2-4) used for mapping to the re-annotated Affymetrix GO database. The GO annotation data in columns 6 – 9 and the GAQ scores in column 10 were retrieved from the Affymetrix chicken GO database and were specified by user. More or less data to be retrieved can be specified by user but must match the list shown on supplementary data 5.2. Column 11 shows the actual date the data is retrieved.

CHAPTER 6

CONCLUSION

Genome annotation is crucial for deriving value from a genome sequence. Generally proteomics offers a fast, relatively cheap and precise method for obtaining a large amount of experimental evidence to assist genome annotation. The value of proteomics in genome annotation, as demonstrated in this study, was to provide a higher level confirmation of protein expression *in vivo*. In this study we used mass spectrometry (MS) data obtained from multiple chicken tissues to confirm *in vivo* expression of electronically predicted proteins. Expression of about 7,811 chicken predicted proteins was confirmed. The results demonstrate the utility of proteome data for genome annotation. Proteomics data can be used to experimentally validate predicted proteins and offers an additional support that genes that code for these proteins are not only transcribed, but also translated. However, a big list of confirmed proteins does not mark the end point for proteomics. Proteins need to be assigned useful biological information.

The most complex component of annotation is linking the genome to biological functions. Functional annotation, a major feature of genome sequence analysis, enables researchers to model their experimental dataset and provide answers to their research questions. Since predicted proteins usually have no functional

information, we transferred Gene Ontology (GO) annotations and standardized gene nomenclature from human and mouse orthologs to the chicken protein that we identified from the proteomics analysis. Using GO we were able to group the confirmed proteins according to their molecular function, involvement in a particular process or subcellular location. As a result we were able to improve the functional annotation of chicken genome by 8%. Improved functional annotation provides researchers with valuable resource for modeling their experimental datasets and answers most of functional genomics questions. As a point of caution the quality of these annotations should be known and maintained.

The GO Annotation Quality (*GAQ*) score developed in this study provides a measure to quantify and assess the overall quality of GO annotations. As demonstrated in nine different species, *GAQ* can be used to assess quantity and quality of functional annotation available for a species. Analysis using *GAQ* scores will enable researchers to determine what species have better GO annotations. Researchers will be able to compare orthologs across species and determine the best annotated orthologs based on higher GAQ scores. This will facilitate the choice of sources of information to be transferred across species whenever deemed necessary. In addition, the *GAQ* score can be used to help biocurators better direct annotation efforts to specific gene products found to have low scores and also to track the improvement of GO annotation over time. *GAQ* scoring can be applied to GO annotations assigned to either proteomics or microarray data in any species.

The last part of this dissertation demonstrates a comprehensive approach that facilitates structural and functional annotation of gene products at the same time showing the *GAQ* scores of each association. The Affymetrix GenChip chicken whole genome array is used as a case study. This array is associated with gene and protein cross references (structural annotation), GO annotations (functional annotation) as well as *GAQ* scores (quality assessment). The number of transcripts, genes and gene products on this array is considered comprehensive because it represents the entire chicken genome. The structural coverage of chicken genome gave us a reason to improve its functional annotation using GO standards and assessing the annotation quality using the *GAQ* scores. We have assigned GO annotations that have been either experimentally verified or computational annotations that have been manually checked or electronically predicted.

We have used the improved annotations of the Affymetrix chicken array as a database that can facilitate annotation of other arrays and experimental datasets from either proteomics or microarray studies. To make this possible, we developed an Array GO Mapper (*AGOM*) tool and demonstrated its implementation by annotating the Arizona chicken array (GPL6049). The Arizona chicken array is linked to GenBank accession, Entrez Gene ID and Ensembl ID. These identifiers were used in the mapping process to retrieve GO annotations from the Affymetrix GO annotation database we developed. In the mapping process over 95% of genes represented on Arizona array were found to be common with the Affymetrix array, and 72% of these genes were mapped to

GO annotations (mean *GAQ* score = 59). Likewise, the existing chicken microarray studies can use *AGOM* to enhance functional annotation in these studies.

In this dissertation, one point worth knowing is that, besides GO, other functional information can as well be useful for researchers wishing to assess the type of biological information represented by transcript printed on an array. In any array knowing which genes are associated with diseases and disorders and which are known drug targets is crucial. This can be achieved through integration of functional annotation from multiple databases such as GO annotations and Ingenuity Pathway analysis knowledge base. In the future, this information can be linked to the Affymetrix functional information to show the importance of chicken as a biomedical model organism to study human diseases, development and any other biomedical issues. If gene types, drug targets and even pathways are comprehensively identified, it will allow researchers to properly design their studies and learn more about how genes function in complex biological systems. Ultimately, this may lead to new drug discovery and development of therapies.

To sum up, the results reported in this dissertation provides a foundation for comprehensive annotation of chicken genome. The methods applied facilitate improvement of both structural and functional annotation of the chicken genome. Conversely, the approach that was used and the tools we developed are simple to implement and are applicable to any species; prokaryotes or eukaryotes, as long as the format suggested is maintained. The broad applicability of this approach will accelerate the genomics knowledge base and understanding of the complex biological system of

poorly annotated and newly sequenced genomes. Ultimately, improved genome annotation will be realized.