

## Proteomics applied to disease in poultry

*B. Nanduri<sup>1†</sup>, F. M. McCarthy<sup>1†</sup>, S.M. Bridges<sup>2</sup>, A. Corzo<sup>3</sup>, M. D. Koter<sup>1</sup>, J.J. Buza<sup>1</sup>, B. van Den Berg<sup>1</sup>, N. Wang<sup>2</sup> & S. C. Burgess<sup>1\*</sup>*

<sup>1</sup>Department of Basic Sciences, College of Veterinary Medicine

<sup>2</sup>Department of Computer Science, College of Engineering

<sup>3</sup>Department of Poultry Science, College of Agriculture

Mississippi State University, Box 6100, MS 39762-6100, USA

Tel: (+ 1) 662 325 1239, Fax: (+ 1) 662 325 1031,\*Email: [burgess@cvm.msstate.edu](mailto:burgess@cvm.msstate.edu)

† = Both authors jointly contributed to this manuscript.

### Summary

Like humans, many other biomedical research species, and some agricultural plants before it, the chicken now has a sequenced genome. The chicken was the first agricultural animal to have its genome sequenced. All fields of chicken research may now choose to utilize techniques in “post-genome biology.” Two post-genomic areas of exponential growth in all fields of biology are proteomics and functional gene product annotation databases. Proteomics allows the high-throughput measurement of thousands of proteins at once. Functional gene product annotation databases are one fundamental tool for modeling proteomics data (as well as data from the more familiar cDNA micro-array experiments). This review provides a primer on proteomics and Gene Ontology (GO) functional gene product annotation as it applies to the chicken. We summarize the current state of chicken proteomics research. It discusses “ChickGO” a publicly accessible database for GO functional chicken gene product annotations that is accessible via the AgBase website (<http://www.agbase.msstate.edu/>). We also review our work on ChickGO and AgBase, chicken proteogenomics, modeling B cell function in health and T cell function after Marek’s disease virus transformation, identifying pathogen-host interactions in the context of a disease resistance and susceptibility to *Salmonella enteritidis*, antibiotic effects on pathogens and the use of proteomics for rapid biomarker detection from chicken plasma or serum using nutritional models.

### Keywords:

proteome, mass spectrometry, MuDPIT, functional genomics, functional annotation.

## Introduction

The poultry community has recently entered the “post genome” era. The millions of dollars spent to sequence the chicken genome are considered an investment that will improve poultry production and, more or less directly, human wellbeing. For the first time, complete functional genomics approaches (including proteomics) are theoretically available for poultry researchers. The proteome of a cell is the protein complement encoded by the genome and “proteomics” is the study of the proteome (Wilkins et al. 1996) and proteomics is growing exponentially (Fig. 1A). The proteome links the DNA sequence, its transcription into mRNA and the resultant phenotype. Proteomic studies with chicken, the first livestock species to have its genome sequenced, provide unique opportunities and challenges in understanding normal and disease physiology.

Proteins are the smallest subunits of most bio-molecular “machines” and, as such, are the basic building blocks of life. Protein amounts are often inferred from transcriptome studies (measurement of mRNA levels). However, there is very little, if any, correlation between mRNA levels and the corresponding protein expression levels (Goodlet 2003). In contrast proteomics directly measures. Like the other functional genomics disciplines, proteomics can be applied to solving very many biological problems. However, unlike these other disciplines, proteomics is, relatively new, immature and still evolving.(Pardanani et al. 2002; Patterson and Aebersold 2003). The proteome is very fluid and context-dependent. Proteomes vary over time and with the environment. Measurements in proteomics include protein identification, quantification, have to address the issues of the dynamic range of proteins in a cell, stoichiometrics, structure (including post translational modifications [PTMs]) and interacting partners.

Proteomics delivers enormous datasets. These datasets first need to be analyzed and then compiled in ways that humans can visualize them. This compilation and visualization can be called “modeling”. Many modeling methods have been tried and certainly the various statistical clustering methods available for transcriptome data may be applied to proteomic data. However, methods that actually allow biological systems to be modeled, and then “tinkered with” (following paradigms from engineering sciences) are becoming popular. These methods are grouped under the term “systems biology”. Critically underpinning systems biology are databases that contain descriptions of the functions of gene products, which are called functional-annotations.

Gene functional-annotation using the Gene Ontology (GO) is now the accepted standard for functional-annotation and the use of GO is growing exponentially (Lewis 2005) (Fig. 1B). Ontologies are simply tightly controlled vocabularies. The GO facilitates the use of the genome information to understand organism function. GO functional-annotations are maintained in species-specific databases, and are as fundamental for doing post genomic biology as are the genome sequences themselves (Lewis 2005).

This review introduces proteomics, GO functional genome annotation and how we integrate the two as a pre-requisite for making biological (i.e. functional) sense of our large proteomics datasets. Proteomics is a valuable tool for biomarker identification from blood and this particular application of proteomics in poultry will be described. The exponential growth of both proteomics and the use of GO (Fig. 1A-C) means that it is practically impossible to include every reference and references are restricted to seminal publications, representative examples or detailed reviews. I do attempt to mention all other poultry researchers using proteomics and any omissions are unintentional and I apologize in advance for any omissions. Much of the proteomics and GO work that we, and others, currently do is to establish the underlying skills and fundamental principles for our systems before applying these to biological questions. The biggest barriers in proteomics are capital and experiment

costs. Together these issues mean that doing proteomics in livestock and their pathogens is even more challenging than in the better funded biomedical species.

The proteome, proteomics and proteomics techniques

Proteomics is driven by three main factors. 1. Complete annotated genome sequences allow rapid protein identification from mass spectrometry data. 2. High throughput technologies to measure proteins from biological samples now exist. 3. Databases, computer algorithms and modeling that integrate genomic and proteomic data are now available. The total number of proteins produced by a eukaryote genome's ~25,000 genes is in the vicinity of 500,000. Alternate mRNA splicing, co- and post- translational modifications contribute most to the disparity between the numbers of genes and proteins. There are more than 200 separate documented protein modifications in vertebrates. More than one of these modifications routinely occurs on most proteins (Gooley and Packer 1997). Furthermore protein localization in the cell determines its function. Many proteins (e.g. transcription factors, signaling molecules and many cytokines) are stored in one place and translocate to other places to cause a change in function and organism phenotype. In short, biology is very complicated.

Proteomics can be divided into three broad areas: 1. *Expression proteomics* is the identification and quantification of proteins, usually between control and test scenarios (Refs) and is conceptually equivalent to cDNA microarrays. 2. *Interactomics* aims to identify protein-protein interactions (REFs). 3. *Structural proteomics* aims to predict the three-dimensional structures of proteins on a genome-wide scale and will not be discussed further as it is beyond the scope of this review.

Expression proteomics is based on electrophoretic or non-electrophoretic systems (or sometimes both). A detailed review of these systems for poultry is given in ((Burgess 2004)). Briefly, electrophoretic-proteomics is done by separating complex mixtures in one (1-D) or two dimensions (2-D) usually using sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE). The aim is to deconvolute a complex protein/peptide mixture by taking advantage of the physico-chemical properties of the proteins/peptides. 2-D SDS PAGE (and recent variations on this theme) is the electrophoresis-based technology most commonly associated with proteomics.

Non-electrophoretic proteomics methods avoid the direct use of electricity and gels for deconvoluting protein mixtures prior to identification. Instead, deconvolution is primarily based on multi-dimensional high performance liquid chromatography (HPLC). Commonly, the words "high performance" are omitted and simply "LC" is used. The LC may be done either off line or directly inline with a mass spectrometer fitted with an electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) source. Regardless of how proteins/peptides are separated in expression proteomics, they must be identified. Identification of proteins by non-gel based proteomics is often referred to as multi dimensional protein identification technology (MudPIT) and entails endopeptidase digestion, mass spectrometry, associated computer algorithms and the databases provided by the genome projects. A detailed description of protein identification is beyond the scope of this review but is given by (Burgess 2004).

Interactomics data sets are derived using a fishing principle. Baits may be monoclonal or polyclonal antibodies and/or tagged proteins that are used to precipitate targets. The immunoprecipitated complexes are again identified by endopeptidase digestion, mass spectrometry etc as above. Yeast-two-hybrid systems follow the same fishing principle but are based initially on molecular constructs. A yeast-two-hybrid system has been used to great effect in the chicken to elucidate MDV-host protein interactions (Liu and Cheng 2003).

There are many practical problems to be overcome before proteomics is as easy to use as, for example, restriction enzymes, the polymerase chain reaction (PCR) or DNA sequencers. However, like these technologies, proteomics will become more widely used. Proteomics suffers from relatively poor sensitivity and these sensitivity issues are generally approached in the wet laboratory by pre-analytical fractionation or at the level of improving mass spectrometry. Chicken proteomics research has contributed to this effort (McCarthy\* et al. 2005).

Regardless of the methods used, gathering proteomics data is currently challenging. Genomic sequence data is linear and directly obtained and transcriptomics data has few dimensions. In contrast, proteomics datasets are non-linear and rely on many preprocessing stages, much human intervention and interpretation. All of these steps require understanding how proteomics technologies work and their limitations. Superficially, at least, proteomics is expensive. However, the methods described above provide extremely large protein datasets more cheaply per unit of data than earlier methods. Having said this, it is extremely difficult to find people who understand proteomics to work in poultry systems.

### *The chicken genome*

Although not absolutely essential, a sequenced genome greatly facilitates proteomics. Ostensibly, the chicken is well placed because its “genome is sequenced”. The first critical step after a genome is sequenced is to rigorously structurally- and functionally-annotate that genome sequence. Genome structural-annotation is demarcating functional nucleotide sequences in genomes and is done first by the sequencing centers using computational methods.

The chicken genome assembly captures 98% of the sequences. However, at least 1.4 Mbase pairs of sequence is in the wrong place and ~17% of the genome (190 Gbase pairs) is unassigned to any chromosome and grouped as the “unassigned chromosome” (Ch.UN) (Hillier et al. 2004; Schmutz and Grimwood 2004). Electronic genome structural-annotation methods produce false positive and false negative predictions and commonly misclassify pseudogenes as functional (Eyras et al. 2005). After electronic genome structural-annotation, manual genome structural-annotation and manual ongoing curation of these annotations is a priority (Searle et al. 2004; Ashurst et al. 2005). Manual genome curation relies heavily on experimental evidence such as industrial-scale reverse-transcription PCR which was used in conjunction with multiple computational “gene finders” to greatly enhance the structural-annotation of the chicken genome (Eyras et al. 2005).

Genome functional-annotation is the assignment of functions to each identified functional element in the genome and is commonly done by using the unified Gene Ontology (GO;(Ashburner et al. 2000); described below). Interpretation of functional genomics experiments done in chicken before the completion of genome sequencing was hindered due to the lack of genome functional annotation. Even now however, the functional-annotation of the chicken genome is its embryonic stages. Over half of the genes in the chicken genome are not even definitively known to be translated into proteins. Such genes are electronically-predicted by homology or by *ab initio* prediction algorithms (Eyras et al. 2005). Genome annotation from existing information is always limited by the information available. Even the very well characterized founding GO consortium member’s genomes (yeast, fruitfly and mouse) are still works-in-progress (Lewis 2005). Although the chicken genome sequences will benefit from advances in human and mouse genomics, they need separate annotation and this task will be an even greater challenge because the smaller size of the research community for these species. Novel paradigms for annotating livestock genomes are needed. For livestock, at least initially, complete genome annotation may be less important than sound functional-annotation of gene products defining economically-important phenotypes.

Current statistics (Table 1) for selected genome databases (human, mouse, rat and cow) and their annotation reveal fundamental problems for genomic structural- and functional-annotation in the chicken. Solving these problems will require different paradigms to those used for other genomes. The chicken genome will always have a low build number and yet it has comparable numbers of known genes (UniGene) to human and mouse and comparable numbers of gene predictions. This means that the genome sequences are relatively poorly compiled (Schmutz and Grimwood 2004). Furthermore, compared to human and mouse, the chicken has 10-fold less ESTs. These statistics, combined with smaller funding pools for manual genome structural-annotation, suggest that the human and mouse paradigm for genome manual structural-annotation is not practicable in the chicken (Eyras et al. 2005).

Gene Ontology, genome functional-annotation, AgBase and ChickGO.

The GO database (Gene Ontology Consortium 2001) is the central repository of all functional gene annotations and it is a critical functional genomics resource. A “GO Consortium” exists and this initially included only three model organisms (Gene Ontology Consortium 2001; Lewis 2005): the *Saccharomyces* Genome database (SGD); FlyBase, the *Drosophila* genome database; and the Mouse Genome database (MGD). However, every species has unique gene products and even gene products that share sequence homology can have different functions in different species (i.e. they can be paralogous). Because organisms, share relatively few “core gene orthologs” with identical functions (Hillier et al. 2004) the GO Consortium has grown to now include members representing invertebrates, plants, a nematode, mold and the Zebrafish. However, until we released AgBase (spearheaded by ChickGO) no livestock species had GO representation.

The GO project provides the basis for the design, development, and implementation of publicly available, expertly curated databases containing comprehensive genome wide gene product functional-annotations. Gene products in every organism are functionally-annotated using three controlled structured vocabularies (the ontologies) representing “biological process”, “molecular function” and “cellular component”. Molecular function describes activities at the molecular level and do not specify where or when, or in what context, the activity occurs. A biological process is accomplished by one or more ordered assemblies of molecular functions. To distinguish between a biological process and a molecular function a process must have more than one distinct step. A biological process is not equivalent to a pathway. Cellular component is a cell component with the proviso that this component is part of some larger object, which may be an anatomical structure or a gene product group (Ashburner et al. 2000; Ashburner and Lewis 2002).

Each GO Consortium member group uses GO to functionally-annotate their species, incorporates these annotations into their own databases and shares functional-annotation files with the GO database. GO annotation has been minimally used in chicken (Fig. 1C). In part this is because of smaller numbers of livestock researchers, but also using GO annotation in livestock first requires researchers to functionally-annotate their own data. At the protein level, the numbers of chicken proteins in the non-redundant protein database (NRPD) are an order of magnitude fewer than human and mouse. The chicken has 10-fold less proteins in the Uniprot database than the human and mouse. These statistics translate to very poor GO functional-annotation.

Any method demonstrating protein kinetics, functions, in-vivo interactions, localization, PTMs, quantity or structure are valuable for functional-annotation (Ge et al. 2003). In addition, protein separation based on sub-cellular fractionation defines a protein's cellular location and can be used to help infer its function. Observation of proteins by mass spectrometry provides direct molecular evidence for the existence of a protein *in-vivo* and

lends more confidence than does inference from genomic sequence alone (Jaffe et al. 2004; Jaffe et al. 2004). In addition, pathogen genomes are of fundamental relevance to the proteomics of immunity and disease. It is axiomatic that well annotated pathogen genome sequences be available, in addition to the chicken genome sequence, if proteomics approaches are to be applied to deciphering chicken – pathogen interactions.

Based on our needs for modeling chicken proteomics data, we established chicken GO databases (ChickGO) that can be accessed from the AgBase web site (<http://www.agbase.msstate.edu>). The AgBase structure is modeled after, and its establishment has been greatly assisted by, the established GO databases. AgBase houses targeted GO annotation databases for agricultural species

ChickGO exemplifies the fundamental difference between AgBase and all other GO databases. We are currently annotating chicken gene products based on our proteomics datasets and our modeling needs. The better funded biomedical species are comprehensively annotating the whole genome. Currently, only 17% of chicken gene products have any GO annotation at all and more than 99% of these proteins are ascribed function without any direct experimental evidence. ChickGO has three goals:

Working with GOA, ChickGO provides freely available GO functional annotations for chicken gene products (GOA Chicken 3.0 has just been released).

ChickGO will aid structural annotation of the chicken genome using proteomics approaches. We have now confirmed the *in vitro* expression of 8% of the 77 600 *ab initio* ORFs frames predicted by Ensembl and these have been submitted to the PRIDE database.

We also develop and support computational tools for functional genomics and modeling, including tools for analysis of GO and proteomics data. These tools are biased towards the agricultural communities but also have general and broad utility.

ChickGO accepts annotations from the research community and will provide an intermediary service for agricultural researchers who wish to learn more about functional genomics and ontologies. ChickGO has served as a model for CowGO and for GO annotation of other agriculturally important genomes with small research communities and limited funds.

### Chicken Proteogenomics.

Proteomics can accurately determine genome structure, including the boundaries and enumeration of functional open reading frames (ORFs) and verify unknown ORFs that cannot be well established on the basis of homology. The correlation of mass spectral proteomics data to the genome structure is called “proteogenomic mapping”. “Proteogenomics” is the use of proteomics for genome annotation and has been used to further structurally and functionally-annotate viruses, bacteria and eukaryotes (Jaffe et al. 2004; Jaffe et al. 2004). Because the resources simply do not exist for chicken to follow the human and mouse models of continued and intensive (manual) genome structural-annotation, we have developed a method to carry out proteogenomics for structural genome-annotation in the chicken. We have used this method on the chicken “unassigned chromosome” (ChrUN) first.

ChrUN encodes known and novel predicted genes, and may also contain unknown open reading frames (ORFs). We have developed “peptide identification sequence tags” (PISTs), which we use to compare the chicken genome to the much better annotated genomes to identify unknown ORFs and to verify ORFs that may be predicted but that cannot be well established based on primary nucleotide sequence homology (Jaffe et al., 2004). We used proteins isolated from bursal B-cells and our DDF MuDPIT technique (McCarthy et al., 2005). The proteins identified from an avian subset of the non-redundant protein database were compared with a database of the ChrUN nucleotide sequence translated in all six open reading frames. Upon removal of matching peptides from both datasets, we found that 392 peptides from ChrUN were not present in the AVDb. The position of these peptides was

mapped on the ChrUN. The peptides were manually extended in the C-terminal direction, either until the end of the readable sequence or until a stop codon was reached. These PISTs were then searched against the entire non-redundant protein database. Our first PIST (PIST-1) serves as a good example of structural annotation using proteogenomics. PIST-1 has no chicken homolog in the NRPD but matches “human PREDICTED: similar to a RIKEN cDNA E130306M17” (a mouse EST). PIST-1 shares no homology with the original mouse RIKEN cDNA or to any known chicken proteins. However, by tracing the homology between the human and chicken protein we were able to identify and artificial concatenation in the ChrUN DNA sequence.

#### Proteomics and disease in chickens

It would be ideal if this review could discuss a lot about the use of proteomics in poultry to identify gene products associated with disease. However, because proteomics is such a new field little has yet been published. Certainly, there have been some seminal papers including yeast-two-hybrid work defining a protein involved in Marek’s disease resistance (Liu and Cheng 2003) and chicken proteomics by 2-D SDS PAGE and mass spectrometry methods (Doherty et al. 2004; Doherty et al. 2004; McLean et al. 2004; McLean et al. 2004; Niikura et al. 2004; Beynon 2005; Beynon and Pratt 2005; Doherty et al. 2005; Hayter et al. 2005). Others have investigated the peptides bound to chicken MHC class I that may be involved in the perennial questions surrounding resistance to Marek’s disease (Haeri et al. 2005). We and others have used proteomics to identify host (Burgess 2004; Scott et al. 2005) and pathogen (Carrillo et al. 2004; de Venevelles et al. 2004) genes involved in chicken diseases. Add turkey.

I will summarize here three current experimental paradigms that my laboratory is taking to use proteomics to understand disease in chickens. The first is to develop methods to model B and T cell function in chicken health and disease. We are currently modeling B cell function in the bursa of Fabricius to provide baseline values for further B cell developmental work and also T cells transformed by Marek’s disease herpesvirus. In both cases we are using a fractionation technique and quantitative proteomics methods as well as GO to allow us to place proteins in cellular locations to model function.

The second paradigm is to take very global approaches to identify pathogen-host interactions in the context of a disease resistance/susceptibility model (*Salmonella enteritidis* [SE] infection). This work is a collaboration with the Lamont Laboratory at Iowa State University. Classical crossing between chickens susceptible and resistant for SE was done and the offspring were then infected with SE. Spleens were isolated from the infected chicks with the upper and lower 10th percentile of bacterial loads 1 week post infection. A major challenge here is that proteomics is normally done on fresh tissues and cells. Before we could do the proteomics experiments, we had to design a method for doing proteomics on frozen spleens. Using separate control spleens, we have now adapted our DDF MudPIT method (McCarthy\* et al. 2005) to frozen tissues, modeled the resulting chicken spleen proteomes and identified the most abundant proteins. We have also developed a novel HPLC method for depleting these most abundant proteins to improve proteome coverage.

Thirdly we are using proteomics to tackle pathogens directly. Most of this work has been done in livestock pathogens other than those of the chicken. We have previously shown a mechanism for a sub-minimum inhibitory concentration (MIC) effects (as would be found in feed additive antibiotics) on leukotoxin in a bovine respiratory pathogen using a novel protein quantification method (Nanduri et al. 2005). We are now, investigating responses to sub-MIC antibiotic concentrations in *Pasturella multocida*. We identified proteins with significantly altered expression, some of which could be predicted based on the antibiotic mechanism of action, but many of which were previously unpredicted

Serum biomarkers in nutritional health.

This final section deals with the unique case of proteomics for rapid biomarker detection from plasma or serum. Of all of the “omics technologies, proteomics is unique in its ability to deal with differential gene expression in plasma or serum. Plasma and serum are traditional and readily accessible sources of real-time (non-destructive) disease biomarkers. We have an on going interest in proteomics and nutrition (Corzo et al. 2004; Corzo et al. 2004; Corzo et al. 2005; Corzo et al. 2005) and we have used a chicken model to investigate changes in serum proteins under the specific nutritional challenge of tryptophan deficiency. We optimized and refined the proteomics technique of  $^{16/18}\text{O}$  labeling and quantification to suit our serum proteomics requirements. We have also used the Spinach protein Rubisco as an internal standard (Koter et al. 2005). This work aimed to identify biomarkers of tryptophan deficiency. Of the 4162 proteins identified; 90 proteins were decreased and 47 increased in tryptophan deficiency. In addition to potential biomarkers of nutritional deficiency these proteins shed light on the molecular mechanisms of tryptophan deficiency. We have manually annotated all of these proteins and are now modeling the impact of tryptophan deficiency on cell physiology in the chicken.

### **Conclusion**

It is still very early days for all post genomic research in the chicken, but this is especially true for proteomics, gene product functional-annotation and complex modeling of biological systems. It is essential that researches have the training and time to become familiar with the new tools and their nuances. It is also essential that appropriate investments are made in people, equipment, databases and computational biology that these investments keep in mind the unique goals of poultry production systems. These investments will be challenging and will require novel paradigms compared with those of the better established, primarily biomedical, systems.

### **Acknowledgements**

The authors acknowledge the technical assistance of LA Shack, T. Harris and T Pechan in preparing data for this manuscript.

## References

- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet 25(1): 25-9.
- Ashburner, M. and S. Lewis (2002). "On ontologies for biologists: the Gene Ontology--untangling the web." Novartis Found Symp 247: 66-80; discussion 80-3, 84-90, 244-52.
- Ashurst, J. L., C. K. Chen, J. G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming and T. Hubbard (2005). "The Vertebrate Genome Annotation (Vega) database." Nucleic Acids Res 33(Database issue): D459-65.
- Beynon, R. J. (2005). "The dynamics of the proteome: strategies for measuring protein turnover on a proteome-wide scale." Brief Funct Genomic Proteomic 3(4): 382-90.
- Beynon, R. J. and J. M. Pratt (2005). "Metabolic labeling of proteins for proteomics." Mol Cell Proteomics 4(7): 857-72.
- Burgess, S. C. (2004). "Proteomics in the chicken: tools for understanding immune responses to avian diseases." Poultry Science 83(4): 552-73.
- Carrillo, C. D., E. Taboada, J. H. Nash, P. Lanthier, J. Kelly, P. C. Lau, R. Verhulp, O. Mykytczuk, J. Sy, W. A. Findlay, K. Amoako, S. Gomis, P. Willson, J. W. Austin, A. Potter, L. Babiuk, B. Allan and C. M. Szymanski (2004). "Genome-wide expression analyses of *Campylobacter jejuni* NCTC11168 reveals coordinate regulation of motility and virulence by *flhA*." J Biol Chem 279(19): 20327-38.
- Corzo, A., M. T. Kidd and S. C. Burgess (2004). "Whole-plasma MALDI-TOF proteomics for identification of biomarkers of nutritional status in the chicken." Journal of Animal and Veterinary Advances 3: 522-526.
- Corzo, A., M. T. Kidd, W. A. Dozier, L. A. Shack and S. C. Burgess (2005). "Protein expression of breast muscle in chickens in response to diets deficient or adequate in dietary methionine." Journal of Nutrition submitted.
- Corzo, A., M. T. Kidd, M. D. Koter and S. C. Burgess (2005). "Assessment of dietary amino acid scarcity on growth and blood plasma proteome status of broiler chickens." Poult Sci 84(3): 419-25.
- Corzo, A., M. T. Kidd, G. T. Pharr and S. C. Burgess (2004). "Initial mapping for the chicken blood plasma proteome." International Journal of Poultry Science 3(3): 157-162.
- de Venevelles, P., J. F. Chich, W. Faigle, D. Loew, M. Labbe, F. Girard-Misguich and P. Pery (2004). "Towards a reference map of *Eimeria tenella* sporozoite proteins by two-dimensional electrophoresis and mass spectrometry." Int J Parasitol 34(12): 1321-31.
- Doherty, M. K., L. McClean, I. Edwards, H. McCormack, L. McTeir, C. Whitehead, S. J. Gaskell and R. J. Beynon (2004). "Protein turnover in chicken skeletal muscle: understanding protein dynamics on a proteome-wide scale." Br Poult Sci 45 Suppl 1: S27-8.
- Doherty, M. K., L. McLean, J. R. Hayter, J. M. Pratt, D. H. Robertson, A. El-Shafei, S. J. Gaskell and R. J. Beynon (2004). "The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain." Proteomics 4(7): 2082-93.
- Doherty, M. K., C. Whitehead, H. McCormack, S. J. Gaskell and R. J. Beynon (2005). "Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates." Proteomics 5(2): 522-33.
- Eyras, E., A. Reymond, R. Castelo, J. M. Bye, F. Camara, P. Flicek, E. J. Huckle, G. Parra, D. D. Shteynberg, C. Wyss, J. Rogers, S. E. Antonarakis, E. Birney, R. Guigo and M. R. Brent (2005). "Gene finding in the chicken genome." BMC Bioinformatics 6(1): 131.
- Ge, H., A. J. Walhout and M. Vidal (2003). "Integrating 'omic' information: a bridge between genomics and systems biology." Trends Genet 19(10): 551-60.

- Gene Ontology Consortium (2001). "Creating the gene ontology resource: design and implementation." Genome Res 11(8): 1425-33.
- Goodlet, D. (2003). Correlation of mRNA and protein expression. Analysing gene expression : a handbook of methods: possibilities and pitfalls. S. Lorkowski and P. Cullen. Weinheim ; New York, Wiley-VCH. 1: 58-63.
- Gooley, A. and N. Packer (1997). Post-translational modifications. Proteome research : new frontiers in functional genomics. Principles and practice. M. R. Wilkins, K. L. Williams, R. D. Appel and D. F. Hochstrasser. Berlin ; New York, Springer: 65-91.
- Haeri, M., L. R. Read, B. N. Wilkie and S. Sharif (2005). "Identification of peptides associated with chicken major histocompatibility complex class II molecules of B21 and B19 haplotypes." Immunogenetics 56(11): 854-9.
- Hayter, J. R., M. K. Doherty, C. Whitehead, H. McCormack, S. J. Gaskell and R. J. Beynon (2005). "The subunit structure and dynamics of the 20S proteasome in chicken skeletal muscle." Mol Cell Proteomics.
- Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S. P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobsson, S. Kerje, L. Andersson, R. P. Crooijmans, J. Aerts, J. J. van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren, I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. Brown, J. Rogers, J. M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K. Wong, J. Wang, B. Liu, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. Dejong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betran, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis and R. K. Wilson (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature 432(7018): 695-716.
- Jaffe, J. D., H. C. Berg and G. M. Church (2004). "Proteogenomic mapping as a complementary method to perform genome annotation." Proteomics 4(1): 59-77.
- Jaffe, J. D., N. Stange-Thomann, C. Smith, D. DeCaprio, S. Fisher, J. Butler, S. Calvo, T. Elkins, M. G. FitzGerald, N. Hafez, C. D. Kodira, J. Major, S. Wang, J. Wilkinson, R. Nicol, C. Nusbaum, B. Birren, H. C. Berg and G. M. Church (2004). "The complete genome and proteome of *Mycoplasma mobile*." Genome Res 14(8): 1447-61.
- Koter, M. D., S. M. Bridges, A. Corzo, T. Cummings, M. T. Kidd, T. Pechan and S. C. Burgess\* (2005). "Quantitative Serum Proteomics using 18O labeling and a Rubisco

- standard." Proceedings of the 53rd American Society for Mass Spectrometry Conference San Antonio, TX, USA.
- Lewis, S. E. (2005). "Gene Ontology: looking backwards and forwards." Genome Biol 6(1): 103.
- Liu, H. C. and H. H. Cheng (2003). "Genetic mapping of the chicken stem cell antigen 2 (SCA2) gene to chromosome 2 via PCR primer mutagenesis." Anim Genet 34(2): 158-60.
- McCarthy\*, F. M., S. C. Burgess\*, B. H. J. van den Berg, M. D. Koter and G. T. Pharr (2005). "Differential detergent fractionation for non-electrophoretic eukaryote cell proteomics." Journal of Proteome Research 4(2): 316-324.
- McLean, L., M. K. Doherty, D. C. Deeming and R. J. Beynon (2004). "The nature of the subcutaneous gel in chick hatchlings: a proteomics approach." Br Poult Sci 45 Suppl 1: S37.
- McLean, L., M. K. Doherty, D. C. Deeming and R. J. Beynon (2004). "A proteome analysis of the subcutaneous gel in avian hatchlings." Mol Cell Proteomics 3(3): 250-6.
- Nanduri, B., M. L. Lawrence, S. Vanguri and S. C. Burgess (2005). "Proteomic analysis using an unfinished bacterial genome: the effects of sub-minimum inhibitory concentrations of antibiotics on *Mannheimia haemolytica* virulence factor expression." Proteomics 5, In press(18).
- Niikura, M., H. C. Liu, J. B. Dodgson and H. H. Cheng (2004). "A comprehensive screen for chicken proteins that interact with proteins unique to virulent strains of Marek's disease virus." Poult Sci 83(7): 1117-23.
- Pardanani, A., E. D. Wieben, T. C. Spelsberg and A. Tefferi (2002). "Primer on medical genomics. Part IV: Expression proteomics." Mayo Clin Proc 77(11): 1185-96.
- Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." Nat Genet 33 Suppl: 311-23.
- Schmutz, J. and J. Grimwood (2004). "Genomes: fowl sequence." Nature 432(7018): 679-80.
- Scott, T. R., A. R. Messersmith, W. J. McCrary, J. L. Herlong and S. C. Burgess (2005). "Hematopoietic Prostaglandin D2 Synthase in the Chicken Harderian Gland." Veterinary Immunology and Immunopathology. In press.
- Searle, S. M., J. Gilbert, V. Iyer and M. Clamp (2004). "The otter annotation system." Genome Res 14(5): 963-70.
- Wilkins, M. R., J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser and K. L. Williams (1996). "Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it." Biotechnol Genet Eng Rev 13: 19-50.

## Legends

### Figure 1

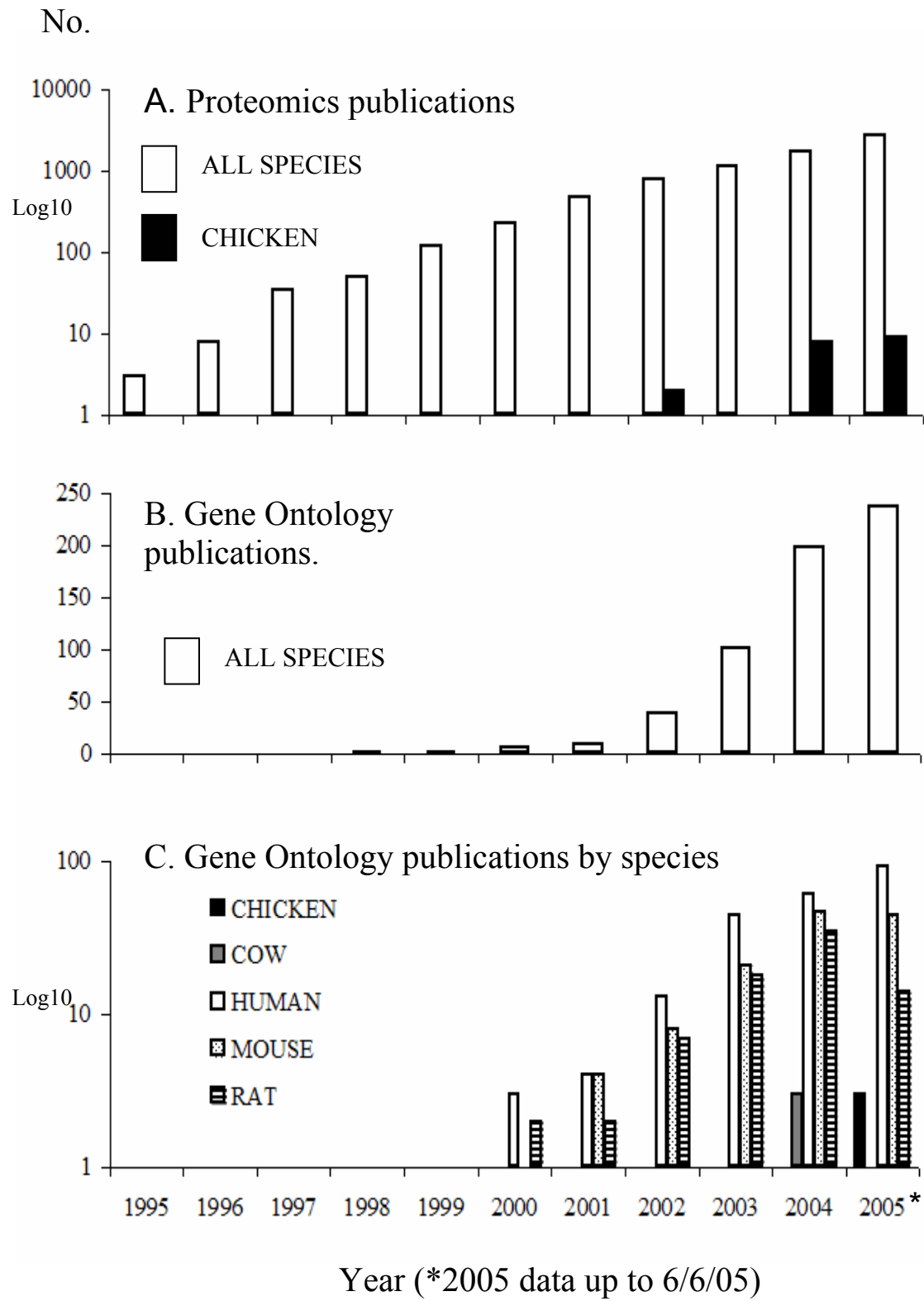
Indications of the growth of proteomics and Gene Ontology (GO) is the exponential increases in the total numbers of proteomics publications in Pubmed published since 1995 when the term proteomics was first coined (A; search term: [proteome OR proteomic OR proteomics] in the title or abstract) and Gene Ontology publications in Pubmed published since 1998 when this term was first coined (B; search term: ["Gene Ontology"] in the title or abstract). The publication data can be further analyzed by species (C). The use of GO annotation has become the accepted standard in humans and mouse and its use in these species is growing exponentially. Furthermore, the rat genome sequence was published only 8 months prior to the chicken genome sequence, and yet it has GO Consortium representation. The rat genome thus has a committed and comprehensive genome annotation effort. As a consequence, like human and mouse, there is exponential growth in rat publications using GO. In contrast GO annotation has been minimally used in chicken (and cow; until now both species had no GO consortium representation). In part this is because of smaller numbers of livestock researchers, but, in addition, using GO annotation in livestock first requires that researchers functionally-annotate their own data. \* all data as at 6/6/05.

### Table 1.

Functional- and structural- genome annotation is always limited by the information available. Even the mouse and human genomes are works-in-progress. Some current statistics for the chicken genome, their associated databases and their annotation, compared with the human, mouse and rat genomes, are given here. These statistics reveal fundamental problems for genomic structural- and functional-annotation in the chicken and two other representative live stock species (cow and catfish). The chicken, cow and catfish genomes all have low build numbers and yet have comparable numbers of known genes (UniGene) to human and mouse and, so far for the chicken at least, comparable numbers of gene predictions. This means that the genome sequences are poorly compiled (Schmutz and Grimwood 2004). Furthermore, compared to human and mouse, the chicken and cow have 10-fold less, and the catfish 100-fold less ESTs. In terms of genome functional-annotation, over half of the genes in the chicken genome are not even definitively known to be translated into proteins; they are electronically-predicted by homology or by ab initio gene prediction algorithms (Eyras et al. 2005). At the protein level, the numbers of chicken proteins in the non-redundant protein database (NRPD) are an order of magnitude fewer than human and mouse. The chicken has 10-fold less proteins in the Uniprot database than the human and mouse. All of these statistics would be predicted to result in very poor GO functional-annotation. However, and on first inspection, the chicken has similar numbers of GO terms associated with gene products (GO associations) compared to human and mouse. But this is misleading. Total numbers of GO associations are not the same as the numbers of proteins which are annotated; total numbers of GO annotations are the total numbers of entries in GO database fields. To date (6/6/05) only 17% of chicken genes have any functional-annotation at all. Worse, only 1.7% of these have evidence codes other than IEA (Inferred from Electronic annotation; the least reliable of all of the GO evidence codes). In comparison ~40% of human and ~70% of mouse gene products have evidence codes other than IEA. The rat genome statistics are much more similar to the livestock genome statistics than to the human and mouse genomes statistics. The rat has a low genome build number and relatively few ESTs or proteins in the NRPD and

UniProt. Like the chicken and cow, the percentage of "predicted" proteins in the rat genome is an order of magnitude higher than human and mouse. However, the rat genome is actively being GO-annotated and, because of this, it has the second lowest percentage of IEA annotations; it is bettered only by the mouse.

Fig 1



| Species             | Build #  | # Autosomes | Genome size |
|---------------------|----------|-------------|-------------|
| Homo sapiens        | 35.1     | 22          | 3.3 Gbp     |
| Mus Musculus        | 33.1     | 19          | 2.9 Gbp     |
| Rattus norvegicus   | 3.4      | 20          | 2.8 Gbp     |
| Gallus gallus       | 1.1      | 32          | 1.05 Gbp    |
| Bos taurus          | 1.1      | 29          | 2.34 Gbp    |
| Ictalurus punctatus | not done | 29          | ~1.0 Gbp    |

| Species             | Genes (UniGene) | Ensembl ab initio gene predictions | ESTs      |
|---------------------|-----------------|------------------------------------|-----------|
| Homo sapiens        | 53,032          | 24,194                             | 6,125,573 |
| Mus Musculus        | 45,659          | 28,069                             | 4,391,843 |
| Rattus norvegicus   | 38,108          | 23,400                             | 659,016   |
| Gallus gallus       | 21,427          | 77,600                             | 543,146   |
| Bos taurus          | 29,189          | not available                      | 618,066   |
| Ictalurus punctatus | not available   | n/a                                | 45,082    |

| Species             | # Proteins (nrpd) | Proteins (Uniprot) | % "Predicted" proteins (NRPD) |
|---------------------|-------------------|--------------------|-------------------------------|
| Homo sapiens        | 258,521           | 63,568             | 2                             |
| Mus Musculus        | 137,327           | 47,475             | 6                             |
| Rattus norvegicus   | 50,750            | 12,999             | 34                            |
| Gallus gallus       | 29,429            | 6,485              | 53                            |
| Bos taurus          | 50,389            | 4,793              | 71                            |
| Ictalurus punctatus | 570               | 0                  | n/a                           |

| Species             | All GO associations | Non-IEA associations | % IEA associations |
|---------------------|---------------------|----------------------|--------------------|
| Homo sapiens        | 62,332              | 23463                | 62.4               |
| Mus Musculus        | 39,132              | 26937                | 31.2               |
| Rattus norvegicus   | 16,912              | 9950                 | 41.2               |
| Gallus gallus       | 24,250              | 422                  | 98.3               |
| Bos taurus          | 21,211              | 904                  | 95.7               |
| Ictalurus punctatus | 0                   | n/a                  | n/a                |

| Species             | Publications | Publications/ORF | Publications/protein (nrpd) |
|---------------------|--------------|------------------|-----------------------------|
| Homo sapiens        | 8980893      | 169              | 35                          |
| Mus Musculus        | 780803       | 17               | 6                           |
| Rattus norvegicus   | 1163349      | 31               | 23                          |
| Gallus gallus       | 92921        | 4                | 3                           |
| Bos taurus          | 228377       | 8                | 5                           |
| Ictalurus punctatus | 1139         | n/a              | 2                           |