

8-1-2011

## Development of computational tools and resources for systems biology of bacterial pathogens

Ranjit Kumar

Follow this and additional works at: <https://scholarsjunction.msstate.edu/td>

---

### Recommended Citation

Kumar, Ranjit, "Development of computational tools and resources for systems biology of bacterial pathogens" (2011). *Theses and Dissertations*. 1453.  
<https://scholarsjunction.msstate.edu/td/1453>

This Dissertation - Open Access is brought to you for free and open access by the Theses and Dissertations at Scholars Junction. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholars Junction. For more information, please contact [scholcomm@msstate.libanswers.com](mailto:scholcomm@msstate.libanswers.com).

DEVELOPMENT OF COMPUTATIONAL TOOLS AND RESOURCES FOR  
SYSTEMS BIOLOGY OF BACTERIAL PATHOGENS

By

Ranjit Kumar

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 Department of Basic Sciences,  
College of Veterinary medicine

Mississippi State, Mississippi

August 2011

DEVELOPMENT OF COMPUTATIONAL TOOLS AND RESOURCES FOR  
SYSTEMS BIOLOGY OF BACTERIAL PATHOGENS

By

Ranjit Kumar

Approved:

---

Bindu Nanduri  
Assistant Professor of CVM Basic Sciences  
(Major Professor and Director of  
Dissertation)

---

Larry A. Hanson  
Professor of CVM Basic Sciences  
(Graduate Coordinator of CVM Basic  
Sciences)

---

Mark L. Lawrence  
Professor of CVM Basic Sciences  
(Co-major Professor)

---

Shane C. Burgess  
Associate Dean for Strategic Initiatives  
and Economic Development and  
CVM Basic Sciences  
Director of Institute for Genomics,  
Biocomputing and Biotechnology  
(Committee member)

---

Susan M. Bridges  
Professor Emeritus of Computer Science  
and Engineering  
(Committee member)

---

Changhe Yuan  
Assistant Professor of Computer Science  
and Engineering  
(Committee member)

---

Kent H. Hoblet  
Dean and Professor of CVM

Name: Ranjit Kumar

Date of Degree: August 6, 2011

Institution: Mississippi State University

Major Field: Veterinary Medical Science

Major Professor: Dr. Bindu Nanduri

Title of Study: DEVELOPMENT OF COMPUTATIONAL TOOLS AND  
RESOURCES FOR SYSTEMS BIOLOGY OF BACTERIAL  
PATHOGENS

Pages in Study: 110

Candidate for Degree of Doctor of Philosophy

Bacterial pathogens are a major cause of diseases in human, agricultural plants and farm animals. Even after decades of research they remain a challenge to health care as they are known to rapidly evolve and develop resistance to the existing drugs. Systems biology is an emerging area of research where all of the components of the system, their interactions, and the dynamics can be studied in a comprehensive, quantitative, and integrative fashion to generate predictive models. When applied to bacterial pathogenesis, systems biology approaches will help identify potential novel molecular targets for drug discovery.

A pre-requisite for conducting systems analysis is the identification of the building blocks of the system i.e. individual components of the system (structural annotation), identification of their functions (functional annotation) and identification of the interactions among the individual components (interaction prediction). In the context of bacterial pathogenesis, it is necessary to identify the host-pathogen interactions. This

dissertation work describes computational resources that enable comprehensive systems level study of host pathogen system to enhance our understanding of bacterial pathogenesis. It specifically focuses on improving the structural and functional annotation of pathogen genomes as well as identifying host-pathogen interactions at a genome scale.

The novel contributions of this dissertation towards systems biology of bacterial pathogens include three computational tools/resources. “TAAPP” (Tiling array analysis pipeline for prokaryotes) is a web based tool for the analysis of whole genome tiling array data for bacterial pathogens. TAAPP helps improve the structural annotation of bacterial genomes. “ISO-IEA” (Inferred from sequence orthology - Inferred from electronic annotation) is a tool that can be used for the functional annotation of any sequenced genome. “HPIDB” (Host pathogen interaction database) is developed with data a mining capability that includes host-pathogen interaction prediction. The new knowledge gained due to the implementation of these tools is the description of the non coding RNA as well as a computationally predicted host-pathogen interaction network for the human respiratory pathogen *Streptococcus pneumoniae*. In summary, the computation tools and resources developed in this dissertation study will enable building systems biology models of bacterial pathogens.

## DEDICATION

I dedicate this dissertation to my father, the late Sri Upendra Prasad, my loving sister Kumari Gunjan, my brother Ranveer Kumar, my mom Chandrawati Devi and my loving girl friend Prachi Matsye.

## ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Nanduri and Dr. Lawrence, my advisors and mentors, for the opportunity to pursue my Ph.D. study under their supervision. Throughout the entire program they provided me constant support, professional guidance, and intellectual freedom.

In addition, I would like to extend my gratitude to my committee member, Dr. Shane Burgess for guidance, support, discussions, and exposure to many interesting research opportunities and challenges. I would also like to thank my committee members Dr. Susan Bridges and Dr. Changhe Yuan, for their generous support and guidance. The advice and recommendations of the committee members have been very useful for the completion of research described in this dissertation.

Special thanks go to my colleagues especially Divyaswetha Peddinti, Shyamesh Kumar, Teresia Buza, Dusan Kunec, Allen Shack and Lakshmi Pillai for providing guidance and support as senior students. I also want to thank my dear friends Ravi Kant, Mitendra Anand, Pradeep Aggrawal, Vikas Sharma, Ritin Sharma and Prachi Matsye for their endless support.

My heartfelt thanks go to my father Late Sri Upendra Prasad, who always supported and encouraged me for higher studies. I also thank all my family members for

their everlasting prayers, sacrifice and support that enabled me to achieve my academic goals.

My sincere gratitude is also extended to the Department of Basic Sciences, College of Veterinary Medicine and National Science Foundation (Mississippi EPSCoR), United States Department of Agriculture for providing financial resources for this research.



## TABLE OF CONTENTS

DEDICATION .....	ii
ACKNOWLEDGEMENTS .....	iii
LIST OF TABLES .....	viii
LIST OF FIGURES .....	ix
CHAPTER	
I. INTRODUCTION .....	1
REFERENCES CITED.....	4
II. REVIEW OF PERTINENT LITERATURE .....	5
<i>Streptococcus pneumoniae</i> TIGR4 – a human respiratory pathogen.....	5
Systems biology of infectious diseases .....	6
Structural annotation of pathogen genomes.....	7
Functional annotation of genomes .....	10
Host-pathogen interaction prediction.....	12
REFERENCES CITED.....	14
III. TAAPP: TILING ARRAY ANALYSIS PIPELINE FOR PROKARYOTES .....	21
Abstract.....	22
Introduction.....	22
Module .....	23
Module 1: TAR generation .....	23
Module 2: feature extraction and annotation .....	24
Sub-module 1: sRNA identifier .....	24
Sub-module 2: antisense identifier.....	25
Sub-module 3: gene expression .....	26
Sub-module 4: operon structure.....	26
Application.....	27

Acknowledgements.....	27
Authors' contributions .....	27
Competing interests .....	27
References.....	27

#### IV. IDENTIFICATION OF NOVEL NON-CODING SMALL RNAS FROM *Streptococcus pneumoniae* TIGR4 USING HIGH-RESOLUTION GENOME TILING ARRAYS.....29

Abstract.....	30
Background.....	30
Results.....	31
Transcriptionally active regions in TIGR4 genome.....	31
Identification and sequence characterization of sRNAs .....	31
Comparative genomics of sRNA sequences .....	37
Computational functional prediction of sRNAs .....	37
Motif and structural analysis of sRNA sequences .....	37
Gene expression profile and identification of operon structures .....	40
Experimental validation of sRNAs .....	41
Discussion.....	42
Conclusions.....	43
Methods.....	43
Isolation of total RNA from <i>S. pneumoniae</i> TIGR4 .....	43
High density genome tiling and hybridization.....	44
Normalization and data analysis .....	44
Analysis of annotated regions of TIGR4 genome.....	44
Gene expression .....	44
Operons .....	45
sRNAs identification, genomic and structural analysis .....	45
qReal-time PCR .....	45
Additional material .....	45
Authors' contributions .....	46
Acknowledgements.....	46
Author Details.....	46
References.....	46

#### V. AUTOMATED PIPELINE FOR ADDING GENE ONTOLOGY ANNOTATION FOR NON MODEL SPECIES.....49

Abstract.....	49
Background.....	50
Results and Discussion .....	53
Testing of chicken proteins annotation using ISO method.....	54
Annotation of chicken predicted proteins using ISO – IEA pipeline .....	55
Annotation of FHCRC chicken array using ISO – IEA pipeline.....	57

Comparison of ISO and IEA annotations .....	58
Conclusions.....	59
Methods.....	60
Implementation of ISO – IEA pipeline.....	60
ISO annotation .....	60
IEA annotation: A wrapper for InterProScan based sequence analysis.....	61
Annotation of chicken predicted proteins and cDNA chicken microarray using ISO – IEA pipeline .....	62
Authors’ contributions .....	63
Acknowledgements.....	63
REFERENCES CITED.....	93
VI. HPIDB -- A UNIFIED RESOURCE FOR HOST-PATHOGEN INTERACTIONS .....	96
Abstract.....	97
Background.....	97
Construction and content .....	98
Utility and discussion.....	99
Using the web interface .....	99
Case study 1 .....	101
Case study 2 .....	101
Case study 3.....	101
Conclusions.....	101
Availability and requirements.....	102
Additional material .....	102
List of abbreviations used .....	102
Authors’ contributions .....	102
Competing interests .....	102
Acknowledgements.....	102
Author details.....	102
References.....	102
VII. CONCLUSION.....	103
REFERENCES CITED.....	108

## LIST OF TABLES

4.1.	<i>S. pneumoniae</i> TIGR4 sRNAs, their genome location, additional features and comparative genomics .....	33
4.2.	Comparison of <i>S. pneumoniae</i> TIGR4 operons identified by tiling arrays with Streptococcus operons described in literature.....	41
5.1.	ISO annotation result for 86 test genes. Annotation colored in black are derived from human, red from mouse and green are from rat orthologs .....	70
5.2.	Ranking for GO evidence code used for calculating GAQ score .....	92
6.1.	Summary of representative host and pathogen species in HPIDB .....	100

## LIST OF FIGURES

2.1.	The paradigm of systems biology .....	13
3.1.	Flow chart of tiling array analysis and annotation pipeline steps .....	24
3.2.	Web interface of TAAPP modules and sub-modules .....	25
3.3.	Snapshot of a short region of <i>S. pneumoniae</i> TIGR4 genome visualized in Genome Browser .....	26
4.1.	Tiling array data analysis workflow .....	32
4.2.	<i>S. pneumoniae</i> TIGR4 sRNA SN1 visualized in the genome browser .....	36
4.3.	Sequence motifs identified in sRNAs by MEME .....	38
4.4.	<i>S. pneumoniae</i> TIGR4 genes expressed in different TIGR protein families (TIGRFAMs) .....	39
5.1.	ISO-IEA pipeline .....	64
5.2.	Orthologs distribution for chicken predicted proteins .....	65
5.3.	ISO – IEA annotation results for chicken predicted proteins .....	66
5.4.	Overall improvement in GO annotation for the chicken proteome .....	67
5.5.	Distribution of probes for annotation in FHCRC chicken 13K array .....	68
5.6.	Comparison of ISO and IEA annotations .....	69
6.1.	Workflow for the construction of HPIDB .....	99
6.2.	Distribution of PPIs from various databases present in HPIDB .....	101

## CHAPTER I

### INTRODUCTION

Bacterial pathogens are major cause of diseases in human (pneumoniae, tetanus, typhoid fever, diphtheria, syphilis, cholera, food borne illness, leprosy, tuberculosis), agricultural plants (leaf spot, fire blight, wilts etc), and farm animals (Johne's disease, mastitis, salmonella and anthrax). Even after decades of research, bacterial pathogens remain a challenge to health care. They rapidly evolve and develop resistance to the existing drugs (1). In addition, there is a steep decline in the approval of new anti-bacterial drugs (1). Therefore, there is a need to increase our understanding of bacterial pathogenesis for the identification of novel targets for prophylactic and therapeutic intervention strategies. While reductionist approaches study one gene at a time to determine its biological significance, systems biology approaches to study infectious diseases have the potential to expedite drug discovery process. Systems biology is an emerging area where all of the components of the system, their interactions, and the dynamics can be studied in a comprehensive, quantitative, and integrative fashion.

Systems level analysis is facilitated by the recent advances in genome scale high throughput technologies like genome sequencing, microarrays and next generation sequencing. A pre-requisite to conduct systems level analyses is the description of all the

building blocks i.e. functional elements of the system. Beyond this initial identification of the components of the system (structural annotation), it is necessary to describe biological function (functional annotation) to the components. Subsequently, the interactions between the components to achieve a specific goal (interaction prediction) are determined. Understanding the regulatory circuits in the system help generate predictive models of the system. These models in turn help in understanding the behavior and dynamics of the system.

There are gaps in the existing knowledge of bacterial pathogens that need to be addressed for conducting meaningful systems analyses. Recent studies have shown that many components of bacterial genomes (small proteins, small non-coding RNAs, riboswitches and other regulatory elements) are not identified using current computational methods highlighting the need for complementary experimental approaches (2-5). Some of these missing elements have housekeeping functions and are important for virulence. Furthermore, the biological function of many genes (around 30-40% of predicted genes) is either not known or classified as “hypothetical” (6,7). Functional annotation of these genes is crucial for systems level modeling. Beyond the identification and description of the functions for the building blocks, it is important to determine interactions among these elements. For comprehensive understanding of bacterial pathogenesis, it is necessary to study the interactions between the pathogen and the host. Resources for host-pathogen interactions are limited. There are no computational tools that enable prediction of host-pathogen interactions at the genome level.

This dissertation contributes to all of the above stated aspects of host-pathogen systems biology. It specifically focuses on improving the structural and functional annotation of pathogen genomes as well as identifying host-pathogen interactions at a genome scale. Experimental methods such as tiling arrays can improve structural annotation of bacterial genomes. However, existing tiling array data analysis methods are predominantly tailored to eukaryotic genomes and cannot be readily applied to bacterial pathogens. We developed a computational web based tool (TAAPP) for prokaryotes. TAAPP was used to improve the structural annotation of *S. pneumoniae* TIGR4 genome. We also developed a computational tool for adding Gene Ontology based functional annotations to gene products. The tool performs orthology based annotation transfer (where available) as well as conserved sequence features like motifs, and functional domains. Identification of host-pathogen interactions is crucial for understanding bacterial pathogenesis. We developed a novel integrated database of host-pathogen interactions for searching, mining and analyzing these crucial inter-species interactions. The database allows the users to transfer existing interactions to new species of interest based on homology. In summary, the computation tools and resources developed in this dissertation study will enable building systems biology models of bacterial pathogens.



## REFERENCES CITED

1. Boucher, H.W., Talbot, G.H., Bradley, J.S., Edwards, J.E., Gilbert, D., Rice, L.B., Scheld, M., Spellberg, B. and Bartlett, J. (2009) Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin Infect Dis*, **48**, 1-12.
2. Guell, M., van Noort, V., Yus, E., Chen, W.H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kuhner, S. *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science (New York, N.Y)*, **326**, 1268-1271.
3. Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C. and Vogel, J. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet*, **4**, e1000163.
4. Liu, J.M., Livny, J., Lawrence, M.S., Kimball, M.D., Waldor, M.K. and Camilli, A. (2009) Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic acids research*, **37**, e46.
5. Livny, J. and Waldor, M.K. (2007) Identification of small RNAs in diverse bacterial species. *Curr Opin Microbiol*, **10**, 96-101.
6. Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet*, **15**, 132-133.
7. Green, M.L. and Karp, P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res*, **33**, 4035-4039.

CHAPTER II  
REVIEW OF PERTINENT LITERATURE

***Streptococcus pneumoniae* TIGR4 – a human respiratory pathogen**

*S. pneumoniae*, a gram-positive human pathogen, is the most common cause of community-acquired pneumonia and a leading cause of meningitis, sinusitis, chronic bronchitis, and otitis media (1). Pneumococci cause approximately 63,000 invasive infections and 6,100 deaths every year in the United States alone (2). Through a combination of virulence-factor activity and the ability to escape the initial barriers of the host immune response, this organism can spread from the upper respiratory tract to the sterile regions of the lower respiratory tract, which ultimately leads to pneumonia. The complete genome of a capsular serotype 4 isolate of *S. pneumoniae* TIGR4 strain (GenBank accession number AE005672) was sequenced in 2001 by the random shotgun sequencing strategy (3). The sample for this clinical isolate was taken from the blood of a 30-year-old male patient and found to be highly invasive and virulent in the mouse model of infection (4). The genome consists of a single circular chromosome of 2,160,837 base pairs (bp) with a G + C content of 39.7%. There are a total of 2236 genes predicted by automated gene prediction methods of which 1440 (64%) were assigned a biological role. No small RNAs were reported in the genome.

## **Systems biology of infectious diseases**

Systems biology is based on the philosophy that biological systems have “emergent properties” that can only be described by studying all the components and their interactions in a holistic manner (5,6). Systems biology provides a way to study the complex interactions between large number of genes, proteins and other genomic elements at a systems level (7-9). A general process for building a systems level model is divided into several steps (Figure 1). It begins with the identification of the building blocks i.e. components of the system (structural annotation), determining the function of the components (functional annotation) as well as the interactions among the components (interaction prediction). Deciphering the regulatory relationships among the components allows the development of predictive models (10). Genome sequencing and high-throughput omics methods (transcriptomics, proteomics, metabolomics, lipidomics, etc.) record genome response to different perturbations and are often instrumental in constructing and refining the predictive models in an iterative process (11).

Recent studies emphasize the importance of high throughput techniques such as transcriptomics (12), proteomics (13), metabolomics (14) and lipidomics (15) in host-pathogen systems biology. These approaches allow the researchers to capture the dynamic behavior of the components of the system during infection. For example, using genome wide yeast two-hybrid assays, 173 interactions were identified between Epstein-Barr virus and human proteins (16). Microarray based miRNA profiling uncovered specific miRNA signatures which correlated with CD4+ T-cell counts in HIV infected

individuals (17). The transcriptome analysis (RNA-Seq) of *Neisseria gonorrhoeae*, a human pathogen, in anaerobic condition, revealed many novel transcriptional regulators and induction of new small RNAs (18). These studies clearly demonstrate that systems analysis of high throughput data can generate predictive computational models and hypothesis (9,19). Acknowledging the possible impact of systems biology approaches in infectious disease research, the National Institute of Allergy and Infectious Diseases (NIAID) has sponsored the systems biology program for specific pathogens. The pathogens of interest to NIAID include *Mycobacterium tuberculosis*, *H5N1 avian influenza virus*, *Staphylococcus aureus*, *Salmonella enterica* and *Yersinia pestis* (20). Various data integration tools and methods are being developed for host-pathogen systems biology (21,22). However, these resources are currently limited for bacterial pathogens. The integrated approach of studying the host and pathogen at a systems level will increase our current understanding of bacterial pathogenesis and will help in translational research (11).

### **Structural annotation of pathogen genomes**

Genome annotation is a multi-level process that includes prediction of not just protein coding genes, but also pseudogenes, promoter regions, repeat elements, regulatory elements like small non coding RNAs, riboswitches and other genomic features of biological significance. A significant component of structural annotation for any genome is the prediction of its protein coding genes. Once a genome is sequenced it undergoes computational gene prediction for initial structural annotation. The two major

computational approaches for gene prediction are extrinsic and intrinsic approaches (23,24). The extrinsic approach is based on evolutionary conservation of protein-coding regions in the genome sequences. The intrinsic approaches (ab initio methods) use sequence properties like nucleotide composition to predict the location of genes and are more commonly used. For example gene prediction methods like Glimmer (25) or GenMark (26) use Hidden Markov models (intrinsic approach) which are based on the training set consisting of well annotated genes. An evaluation of commonly used gene prediction programs Fgenesh (27), GlimmerHMM (28), and GeneMark.hmm (29) against cDNA verified reference genes of rice genome showed similar exon prediction accuracy with sensitivity around 78% and specificity between 72% - 76% (30). However, these programs may not predict all the exons for a particular gene, which leads to decreased accuracy for the gene prediction with the sensitivity between 22% - 25% and specificity between 15% - 21% (30). The gene prediction accuracy increases in bacterial species due to their relatively simple gene structure. In bacterial species the current gene prediction methods reach to a sensitivity  $> 90\%$  and specificity  $> 85\%$  (31). Bacterial gene prediction programs face difficulties due to the absence of intron elements (unlike higher eukaryotic species), multiple start codons (six), which result in the prediction of overlapping open reading frames (ORFs). Identifying the correct coding frame becomes a difficult task (32). The prediction programs utilize a user defined minimum length cutoff to filter short ORFs, which may lead to incorrect identification of small genes. Apart from gene prediction, when it comes to the structural annotation of other elements like small RNA prediction, the accuracy of bacterial computational prediction decreases. For

example, bioinformatics analysis predicted 40 sRNAs in *S. pneumoniae* D39 strain of which only nine were validated by Northern blotting (33,34). A comparison of sRNA prediction programs reveals the fact that different algorithms identify a different set of sRNAs for the same genome possibly due to the differences in the training set and algorithm parameters (35). Difficulties in small RNA prediction programs are due to very low sRNA sequence conservation across other species (36,37), missing protein coding frame and the limited accuracy of transcriptional signal prediction programs (like promoter prediction and rho-independent terminator prediction).

The issues with computational methods for bacterial genome annotation demonstrate the need for alternative experimental methods to improve the structural annotation of genome. Common experimental methods include high throughput transcriptomics (RNA profile) and proteomics (protein profile) methods to characterize novel elements in the genome (38-42). The benefit of using experimental RNA based methods is that it can account for events like transcriptional errors and RNA editing where the RNA transcript differs from the DNA template (43,44). There are different types of RNA editing events like adenosine deamination to inosine which is recognized by translational machinery as guanosine (A-G) or cytidine is edited to uridine (44). These RNA editing events not only lead to change in single amino acid but sometimes cause changes in a large portion of the protein due to frameshift (44,45). High throughput experimental methods usually validate and improve the existing annotation. Tiling array expression analysis of the intergenic regions of *E. coli* and *Mycobacterium leprae* indicates the expression of small non coding RNAs (38,39). A recent study identified 27

sRNAs in *Caulobacter crescentus* using tiling array approach (40). Using parallel sequencing, a large number of putative sRNAs were reported in *Vibrio cholerae* (46). Immuno precipitation with Hfq (sRNA binding protein) antibody followed by deep sequencing identified 64 sRNAs in *Salmonella Typhimurium* (47). Proteomic methods are also becoming popular where peptide matches are used to identify and validate the existing annotation and scan intergenic regions to discover novel protein fragments (41,42). Studies have shown that sRNAs are involved in various housekeeping activities, regulatory roles and virulence (48). Specifically, they are known to perform regulatory roles in sugar metabolism (49,50), iron homeostasis (51) and cell surface composition and virulence (52,53). To date, no experimental studies have been performed to identify the non coding RNAs and other regulatory elements in respiratory pathogen *Streptococcus pneumoniae* TIGR4 using global transcriptomics. It is highly probable that pneumococcal genome has structural elements which are still un-characterized.

### **Functional annotation of genomes**

Assigning biological function to the genomic elements i.e. functional annotation is important for understanding the underlying biology. The Gene Ontology (GO) project (54) provides a controlled vocabulary for functional annotation. The GO describes three attributes of gene products: molecular function, biological process and cellular component (54). The use of standard vocabulary enables the user to perform GO based functional analysis of high throughput datasets. Gold standard functional annotations are derived from experimental methods (where individual functional assays are described)

and through literature curation by trained biocurators. Annotation to the GO involves providing information about the gene product being annotated, its attributed function and the evidence for associating the function with this gene product (55). Literature curation for annotation is time and labor consuming. Therefore, automated GO annotation pipelines are required for providing GO rapidly, while maintaining the quality of annotations.

Automated GO annotation tools described in the literature (56), are mostly based on sequence similarity searches. However, transfer of function based on orthology (57) is the best way to provide GO annotation when there is no functional literature available for the gene product of interest. Orthologs or orthologous genes are genes in different species that arose from a common ancestor and are assumed to be functionally equivalent. The function of orthologs proteins is usually conserved even when their sequence or structure changes in due course of evolution (58). Annotation transfer can be confounded by the presence of paralogs and gene gain or loss, that could result in inaccurate predicted functions (59). Many a times the sequence similarity also exists in small fragment of the gene (representing similar motif or domain) which may not entirely represent the gene with same function. Estimates of the error rate of curated bacterial genome sequence protein and gene-name annotations lie between 6.8% and 8% and majority of the errors are accounted for functional predictions which are made on the low sequence identity, which is not sufficient to accurately pinpoint the function (60,61). There were no tools available to perform high throughput GO annotation using ortholog information. The availability of such a tool will greatly help in functional analysis at systems level.



## **Host-pathogen interaction prediction**

In addition to identification of structural and functional components of the system, it is important to know how these components interact with each other in the system. Proteins are considered to be the work horses of the cell and they interact with other proteins to carry out biological functions such as signal transduction, protein transport, immune response etc. Protein-protein Interactions (PPIs) can be classified into two main categories: "Intra-species PPI", where two proteins from the same species interact with each other and "Inter-species PPI" where two proteins from two different species interact. Host-pathogen protein-protein interactions (HPIs) are a subset of inter-species interactions and are relevant to studying bacterial pathogenesis.

Although a number of databases are described in literature that store known experimental PPIs (62-64), only a few databases contain HPIs (56,65-67). To create a useable set of HPIs for any analysis, users have to access multiple databases followed by manual curation that requires a lot of programming and data processing. Apart from the limited availability of experimental HPIs, very few computational approaches are reported that predict HPIs. For example, protein domain profiles of existing intra-species PPIs were used to predict the interaction between human and plasmodium proteins (68). In another study, existing intra-species PPIs were used to identify orthologous interactions (interologs), which were used to predict inter-species interactions (69,70). Both of these computational studies use intra-species PPIs to predict inter-species interactions. They also do not provide a tool for predicting HPIs. The interologs based

system for HPI prediction can be improved by using a set of known inter-species PPIs (HPIs) instead of intra-species interactions. However, the current limitation is the unavailability of a centralized database which stores all the experimentally known inter-species interactions (HPIs). Although a few efforts have been made towards developing dedicated host-pathogen interaction databases the existing resources are still limited in scope or confined to limited number of species (71-74). Developing a unified resource that integrates HPIs from multiple databases into a single, non-redundant set for data mining purposes will be critical for host-pathogen systems biology.

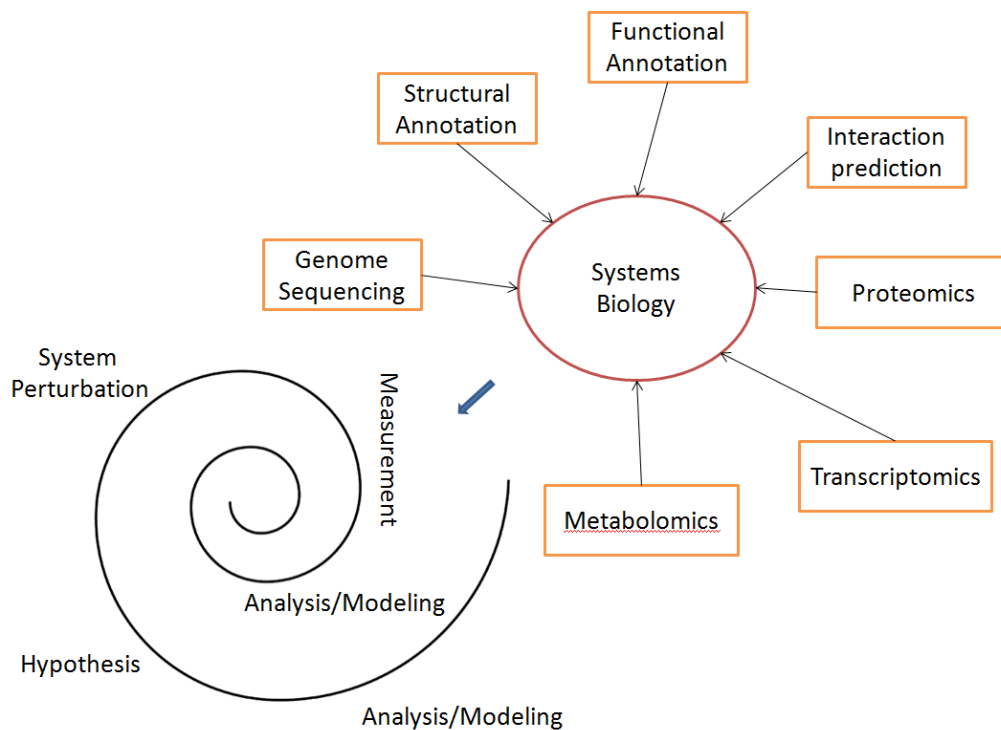


Figure 2.1 The paradigm of systems biology. (A) The building blocks of systems biology are shown in boxes. (B) Once the initial system model is ready it undergoes an iterative process of data analysis, modeling, perturbation to continually refine the model and use for systems level analysis.

## REFERENCES CITED

1. Bridy-Pappas, A.E., Margolis, M.B., Center, K.J. and Isaacman, D.J. (2005) Streptococcus pneumoniae: description of the pathogen, disease epidemiology, treatment, and prevention. *Pharmacotherapy*, **25**, 1193-1212.
2. Schuchat, A., Hilger, T., Zell, E., Farley, M.M., Reingold, A., Harrison, L., Lefkowitz, L., Danila, R., Stefonek, K., Barrett, N. *et al.* (2001) Active bacterial core surveillance of the emerging infections program network. *Emerg Infect Dis*, **7**, 92-99.
3. Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J. *et al.* (2001) Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science (New York, N.Y.)*, **293**, 498-506.
4. Aaberge, I.S., Eng, J., Lermark, G. and Lovik, M. (1995) Virulence of Streptococcus pneumoniae in mice: a standardized method for preparation and frozen storage of the experimental bacterial inoculum. *Microb Pathog*, **18**, 141-152.
5. Csete, M.E. and Doyle, J.C. (2002) Reverse engineering of biological complexity. *Science (New York, N.Y.)*, **295**, 1664-1669.
6. Zak, D.E. and Aderem, A. (2009) Systems biology of innate immunity. *Immunol Rev*, **227**, 264-282.
7. Forst, C.V. (2006) Host-pathogen systems biology. *Drug discovery today*, **11**, 220-227.
8. Aderem, A., Adkins, J.N., Ansong, C., Galagan, J., Kaiser, S., Korth, M.J., Law, G.L., McDermott, J.G., Proll, S.C., Rosenberger, C. *et al.* (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *MBio*, **2**.
9. Peng, X., Chan, E.Y., Li, Y., Diamond, D.L., Korth, M.J. and Katze, M.G. (2009) Virus-host interactions: from systems biology to translational research. *Current opinion in microbiology*, **12**, 432-438.
10. Zhang, W., Li, F. and Nie, L. (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology (Reading, England)*, **156**, 287-301.

11. Bumann, D. (2009) System-level analysis of Salmonella metabolism during infection. *Current opinion in microbiology*, **12**, 559-567.
12. Sturdevant, D.E., Virtaneva, K., Martens, C., Bozinov, D., Ogundare, O., Castro, N., Kanakabandi, K., Beare, P.A., Omsland, A., Carlson, J.H. *et al.* (2010) Host-microbe interaction systems biology: lifecycle transcriptomics and comparative genomics. *Future Microbiol*, **5**, 205-219.
13. Bumann, D. (2010) Pathogen proteomes during infection: A basis for infection research and novel control strategies. *J Proteomics*, **73**, 2267-2276.
14. Han, J., Antunes, L.C., Finlay, B.B. and Borchers, C.H. (2010) Metabolomics: towards understanding host-microbe interactions. *Future Microbiol*, **5**, 153-161.
15. van der Meer-Janssen, Y.P., van Galen, J., Batenburg, J.J. and Helms, J.B. (2010) Lipids in host-pathogen interactions: pathogens exploit the complexity of the host cell lipidome. *Prog Lipid Res*, **49**, 1-26.
16. Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, D.E. *et al.* (2007) Epstein-Barr virus and virus human protein interaction maps. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7606-7611.
17. Houzet, L., Yeung, M.L., de Lame, V., Desai, D., Smith, S.M. and Jeang, K.T. (2008) MicroRNA profile changes in human immunodeficiency virus type 1 (HIV-1) seropositive individuals. *Retrovirology*, **5**, 118.
18. Isabella, V.M. and Clark, V.L. (2011) Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*. *BMC genomics*, **12**, 51.
19. Fruh, K., Finlay, B. and McFadden, G. (2010) On the road to systems biology of host-pathogen interactions. *Future Microbiol*, **5**, 131-133.
20. Aderem, A., Adkins, J.N., Ansong, C., Galagan, J., Kaiser, S., Korth, M.J., Law, G.L., McDermott, J.G., Proll, S.C., Rosenberger, C. *et al.* (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *MBio*, **2**, e00325-00310.
21. McGarvey, P.B., Huang, H., Mazumder, R., Zhang, J., Chen, Y., Zhang, C., Cammer, S., Will, R., Odle, M., Sobral, B. *et al.* (2009) Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets. *PloS one*, **4**, e7162.

22. Kozhenkov, S., Sedova, M., Dubinina, Y., Gupta, A., Ray, A., Ponomarenko, J. and Baitaluk, M. (2011) BiologicalNetworks--tools enabling the integration of multi-scale data for the host-pathogen studies. *BMC systems biology*, **5**, 7.
23. Borodovsky, M., Rudd, K.E. and Koonin, E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res*, **22**, 4756-4767.
24. Overbeek, R., Bartels, D., Vonstein, V. and Meyer, F. (2007) Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev*, **107**, 3431-3447.
25. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*, **26**, 544-548.
26. Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, **29**, 2607-2618.
27. Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome research*, **10**, 516-522.
28. Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*, **20**, 2878-2879.
29. Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, **26**, 1107-1115.
30. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome biology*, **9**, R7.
31. McHardy, A.C., Goesmann, A., Puhler, A. and Meyer, F. (2004) Development of joint application strategies for two microbial gene finders. *Bioinformatics (Oxford, England)*, **20**, 1622-1631.
32. Palleja, A., Harrington, E.D. and Bork, P. (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC genomics*, **9**, 335.

33. Livny, J., Brenic, A., Lory, S. and Waldor, M.K. (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res*, **34**, 3484-3493.
34. Tsui, H.C., Mukherjee, D., Ray, V.A., Sham, L.T., Feig, A.L. and Winkler, M.E. (2009) Identification and Characterization of Non-Coding Small RNAs in *Streptococcus pneumoniae* Serotype 2 Strain D39. *Journal of bacteriology*.
35. Sridhar, J., Sambaturu, N., Sabarinathan, R., Ou, H.Y., Deng, Z., Sekar, K., Rafi, Z.A. and Rajakumar, K. (2010) sRNAscanner: a computational tool for intergenomic small RNA detection in bacterial genomes. *PloS one*, **5**, e11970.
36. Kulkarni, R.V. and Kulkarni, P.R. (2007) Computational approaches for the discovery of bacterial small RNAs. *Methods*, **43**, 131-139.
37. Backofen, R. and Hess, W.R. Computational prediction of sRNAs and their targets in bacteria. *RNA biology*, **7**.
38. Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E. and Rosenow, C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res*, **30**, 3732-3738.
39. Akama, T., Suzuki, K., Tanigawa, K., Kawashima, A., Wu, H., Nakata, N., Osana, Y., Sakakibara, Y. and Ishii, N. (2009) Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. *Journal of bacteriology*, **191**, 3321-3327.
40. Landt, S.G., Abeliuk, E., McGrath, P.T., Lesley, J.A., McAdams, H.H. and Shapiro, L. (2008) Small non-coding RNAs in *Caulobacter crescentus*. *Molecular microbiology*, **68**, 600-614.
41. Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M., Nicora, C.D., Lipton, M.S. and Levy, S.B. (2009) Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. *PloS one*, **4**, e8455.
42. Lamontagne, J., Beland, M., Forest, A., Cote-Martin, A., Nassif, N., Tomaki, F., Moriyon, I., Moreno, E. and Paramithiotis, E. (2010) Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome. *BMC genomics*, **11**, 300.

43. Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M. and Cheung, V.G. (2011) Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science (New York, N.Y.)*
44. Knoop, V. (2011) When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci*, **68**, 567-586.
45. Sharma, V., Firth, A.E., Antonov, I., Fayet, O., Atkins, J.F., Borodovsky, M. and Baranov, P.V. (2011) A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. *Mol Biol Evol*.
46. Liu, J.M., Livny, J., Lawrence, M.S., Kimball, M.D., Waldor, M.K. and Camilli, A. (2009) Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res*, **37**, e46.
47. Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C. and Vogel, J. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet*, **4**, e1000163.
48. Livny, J. and Waldor, M.K. (2007) Identification of small RNAs in diverse bacterial species. *Curr Opin Microbiol*, **10**, 96-101.
49. Gorke, B. and Vogel, J. (2008) Noncoding RNA control of the making and breaking of sugars. *Genes & development*, **22**, 2914-2925.
50. Weilbacher, T., Suzuki, K., Dubey, A.K., Wang, X., Gudapaty, S., Morozov, I., Baker, C.S., Georgellis, D., Babitzke, P. and Romeo, T. (2003) A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Mol Microbiol*, **48**, 657-670.
51. Vasil, M.L. (2007) How we learnt about iron acquisition in *Pseudomonas aeruginosa*: a series of very fortunate events. *Biometals*, **20**, 587-601.
52. Gottesman, S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*, **21**, 399-404.
53. Geissmann, T., Possedko, M., Huntzinger, E., Fechter, P., Ehresmann, C. and Romby, P. (2006) Regulatory RNAs as mediators of virulence gene expression in bacteria. *Handbook of experimental pharmacology*, 9-43.
54. 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.
55. Hill, D.P., Smith, B., McAndrews-Hill, M.S. and Blake, J.A. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC bioinformatics*, **9 Suppl 5**, S2.
  56. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, **37**, D619-622.
  57. Buza, T.J., McCarthy, F.M. and Burgess, S.C. (2007) Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC genomics*, **8**, 425.
  58. Fernandez, A. and Lynch, M. (2011) Non-adaptive origins of interactome complexity. *Nature*, **474**, 502-505.
  59. Gabaldon, T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, **9**, 235.
  60. Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet*, **15**, 132-133.
  61. Green, M.L. and Karp, P.D. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res*, **33**, 4035-4039.
  62. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W. and Stumpflen, V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, **34**, D436-441.
  63. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. *et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res*, **37**, D767-772.
  64. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**, D449-451.
  65. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J. *et al.* (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, **38**, D525-531.



66. Gilbert, D. (2005) Biomolecular interaction network database. *Brief Bioinform*, **6**, 194-198.
67. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002) MINT: a Molecular INTERaction database. *FEBS letters*, **513**, 135-140.
68. Dyer, M.D., Murali, T.M. and Sobral, B.W. (2007) Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics (Oxford, England)*, **23**, i159-166.
69. Kim, J.G., Park, D., Kim, B.C., Cho, S.W., Kim, Y.T., Park, Y.J., Cho, H.J., Park, H., Kim, K.B., Yoon, K.O. *et al.* (2008) Predicting the interactome of *Xanthomonas oryzae pathovar oryzae* for target selection and DB service. *BMC Bioinformatics*, **9**, 41.
70. Lee, S.A., Chan, C.H., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y. and Huang, C.Y. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, **9 Suppl 12**, S11.
71. Driscoll, T., Dyer, M.D., Murali, T.M. and Sobral, B.W. (2009) PIG--the pathogen interaction gateway. *Nucleic Acids Res*, **37**, D647-650.
72. Winnenburg, R., Urban, M., Beacham, A., Baldwin, T.K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K.E. and Kohler, J. (2008) PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res*, **36**, D572-576.
73. Navratil, V., de Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., Andre, P., Lotteau, V. and Rabourdin-Combe, C. (2009) VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res*, **37**, D661-668.
74. Zhang, C., Crasta, O., Cammer, S., Will, R., Kenyon, R., Sullivan, D., Yu, Q., Sun, W., Jha, R., Liu, D. *et al.* (2008) An emerging cyberinfrastructure for biodefense pathogen and pathogen-host data. *Nucleic Acids Res*, **36**, D884-891.

## CHAPTER III

### TAAPP: TILING ARRAY ANALYSIS PIPELINE FOR PROKARYOTES<sup>1</sup>

---

<sup>1</sup> Reprint from Kumar R, Burgess SC, Lawrence ML, Nanduri B. Genomics Proteomics Bioinformatics.

2011. TAAPP: Tiling Array Analysis Pipeline for Prokaryotes. This article is available from:

<http://www.ncbi.nlm.nih.gov/pubmed/21641563>

Application Note

## TAAPP: Tiling Array Analysis Pipeline for Prokaryotes

Ranjit Kumar<sup>1,2\*</sup>, Shane C. Burgess<sup>1,2,3</sup>, Mark L. Lawrence<sup>1,2</sup>, and Bindu Nanduri<sup>1,2</sup>

<sup>1</sup>College of Veterinary Medicine, Mississippi State University, Mississippi 39762, USA;

<sup>2</sup>Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi 39762, USA;

<sup>3</sup>Mississippi Agriculture and Forestry Experiment Station, Mississippi State University, Mississippi 39762, USA.

Genomics Proteomics Bioinformatics 2011 Apr; 9(1-2): 56-62 DOI: 10.1016/S1672-0229(11)60008-9

Received: Oct 1, 2010; Accepted: Jan 11, 2011

### Abstract

High-density tiling arrays provide closer view of transcription than regular microarrays and can also be used for annotating functional elements in genomes. The identified transcripts usually have a complex overlapping architecture when compared to the existing genome annotation. Therefore, there is a need for customized tiling array data analysis tools. Since most of the initial tiling arrays were conducted in eukaryotes, data analysis methods are well suited for eukaryotic genomes. For using whole-genome tiling arrays to identify previously unknown transcriptional elements like small RNA and antisense RNA in prokaryotes, existing data analysis tools need to be tailored for prokaryotic genome architecture. Furthermore, automation of such custom data analysis workflow is necessary for biologists to apply this powerful platform for knowledge discovery. Here we describe TAAPP, a web-based package that consists of two modules for prokaryotic tiling array data analysis. The transcript generation module works on normalized data to generate transcriptionally active regions (TARs). The feature extraction and annotation module then maps TARs to existing genome annotation. This module further categorizes the transcription profile into potential novel non-coding RNA, antisense RNA, gene expression and operon structures. The implemented workflow is microarray platform independent and is presented as a web-based service. The web interface is freely available for academic use at <http://lims.lsbj.mafes.msstate.edu/TAAPP-HTML/>.

**Key words:** transcriptomics, small RNA, operon, prokaryotes, tiling arrays

### Introduction

Genomic tiling arrays (overlapping oligonucleotide probes tiled across both strands of genome sequence) provide an unbiased view of genome expression, and have been used to generate transcriptional maps in eukaryotic genomes describing small RNAs (sRNAs), antisense expression, 5' and 3' untranslated regions

(UTRs) (1-3). There is increasing appreciation for the significant role that sRNAs play in bacterial adaptation to stress and pathogenesis (4-6). Computational methods are used for identifying sRNAs, but they still need biological validation (7, 8). Due to the smaller size of prokaryotic genome, tiling arrays are now being used for whole-genome analysis to detect novel transcripts in bacteria (9-12). Generally, computational tools that automate tiling array data analysis are based on two color arrays (13, 14), and are tailored for eukaryotic genomes. Recently, new tools that focus on prokaryotic genome architecture for probe de-

\*Corresponding author.

E-mail: [rkumar@cvm.msstate.edu](mailto:rkumar@cvm.msstate.edu)

© 2011 Beijing Institute of Genomics. All rights reserved.

sign and normalization procedures were described (15, 16). However, these tools and other described analysis workflows stop with the identification of transcriptionally active regions (TARs); the end user with little or no computational skills are left with difficult task of mapping these TARs back to the genome and performing feature extraction for knowledge discovery.

Here we describe, for the first time, a computational pipeline named TAAPP (implemented in Perl), which is tailored for prokaryotic tiling array data, and consists of two modules: the first module handles normalized data from single color arrays, identifies expressed probes and then joins them to generate TARs; the second module maps these identified TARs back to the existing genome annotation, facilitating identification of sRNA elements, gene expression, operon structures and antisense RNA. sRNA elements can be identified in the non-coding area of genome where no annotation is available on either strand whereas antisense RNA is usually identified on the opposite strand of any annotated gene/RNA. The design of TAAPP into two separate modules allows data from two color tiling arrays (after analysis into differentially expressed TARs) to be mapped onto the genome directly using the second module. We applied TAAPP to analyze transcriptome of *Streptococcus pneumoniae* TIGR4 genome using custom high-density tiling arrays. The web interface is freely available for academic use at <http://lims.lsbj.mafes.msstate.edu/TAAPP-HTML/>.

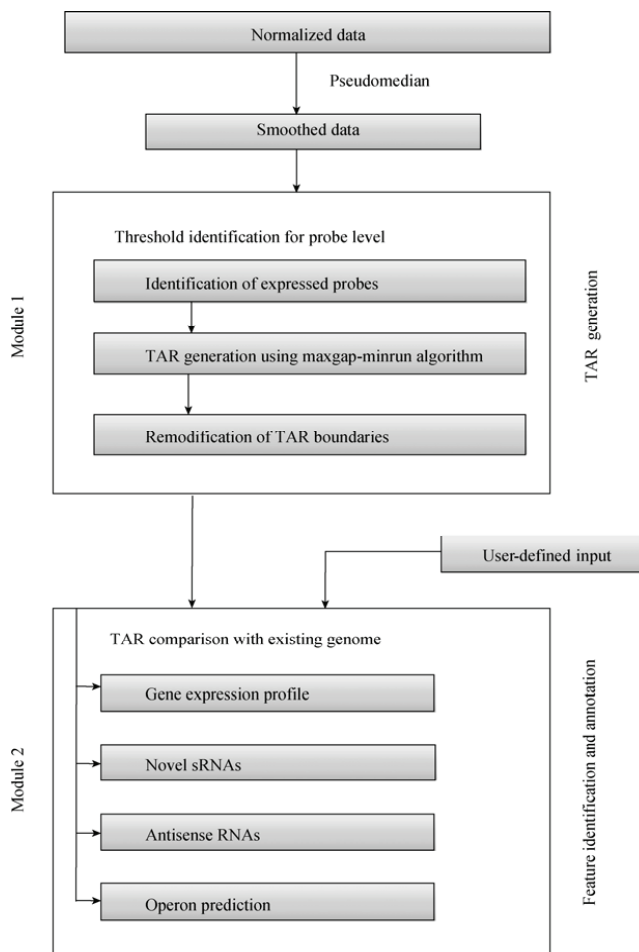
## Module

The software consists of two modules. The first module identifies the expressed regions and the second module compares it with existing genome annotation to identify gene expression pattern and novel elements. The flow chart presented in **Figure 1** shows the various steps involved in data analysis.

### Module 1: TAR generation

The TAR generation module accepts normalized probe-level data as tab-delimited text file (making the pipeline microarray platform independent) (**Figure 2**). For classifying the probes as expressed, this module requires the user to input probe intensity cutoff value

or supply the files with positive and negative controls for automated calculation. This value is often determined based on the distribution of normalized intensity values for negative and positive control probes on the array and varies with array design (2). A lower cutoff value is associated with higher false positive rates of identification and *vice versa*. In the absence of experimental controls, user can use the top 90 intensity percentile as a cutoff value (17). To minimize sequence-based effects on probe intensity, a pseudomedian filter is applied, which takes adjacent probe intensities into account and provides smoothing to the data. A pseudomedian filter works by calculating median of all possible pairwise averages in a sliding window and assigning it to the probe at the center (18). The sliding window is then shifted to the next probe and the process is continued for the complete genome sequence. Probes with intensity greater than the cutoff value are classified as expressed probes and consecutive expressed probes are further joined using maxgap-minrun algorithm (2) to generate TARs. The maxgap parameter allows certain number of probes (one or two probes) to be below the cutoff while still being incorporated into the TAR, whereas the minrun parameter defines at least a certain length of the transcript to be considered as TAR (discarding small length transcripts). To accurately identify genes in the densely packed prokaryotic genomes (marked by short intergenic regions), the maxgap parameter is set to zero for the intergenic region. This helps to differentiate transcript of two consecutive genes, which are usually separated by very short intergenic region, if they are not expressed as an operon. Due to the smoothing of dataset generated by pseudomedian filter, slight errors are introduced in the identification of transcript boundaries (start and end). Therefore, we implemented a new step that remodified transcript boundaries using average intensity values (data before pseudomedian calculation). Remodification is conducted by either elongating or shortening transcript ends until the average intensity value of the probe is greater than or equal to the threshold cutoff. Remodified transcripts are again processed using maxgap-minrun method to generate TARs. Expression data for both strands are processed separately to generate TARs.



**Figure 1** Flow chart of tiling array analysis and annotation pipeline steps.

**Figure 3** shows the transcriptome snapshot of a short region of *S. pneumoniae* TIGR4 genome visualized in Genome Browser, during various steps of data analysis.

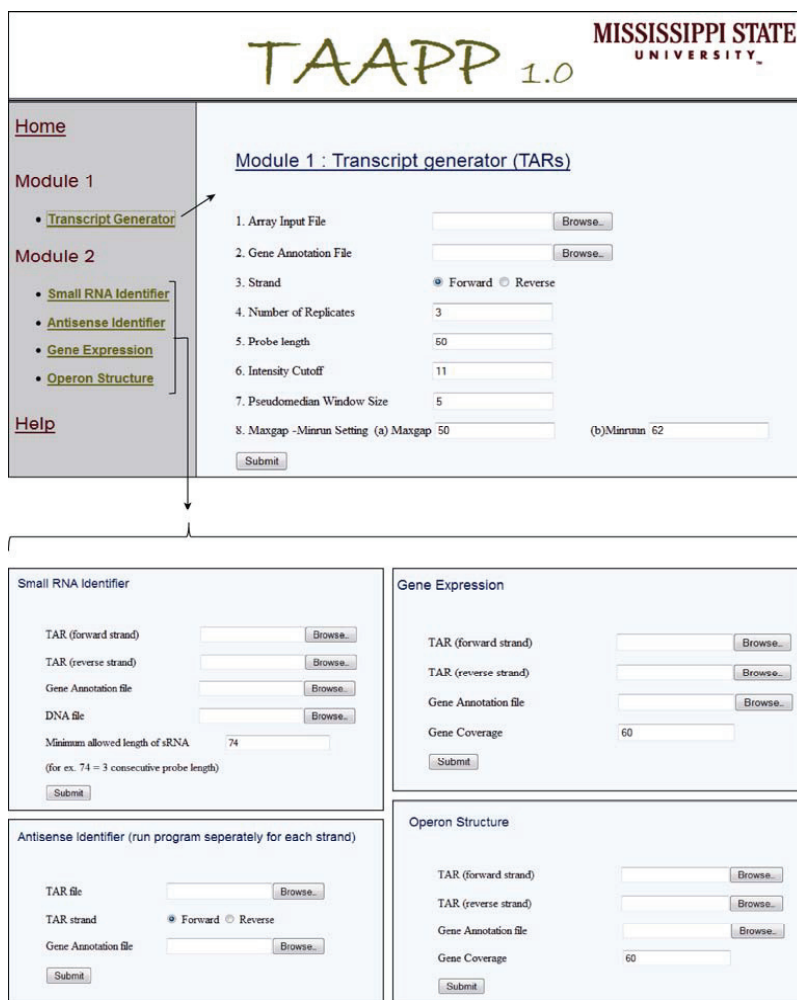
### Module 2: feature extraction and annotation

This module maps the identified TARs generated from module 1 (or any other tiling analysis workflow) with the existing genome annotation. Mapping TARs to annotated open reading frames (ORFs) helps identify the basal transcription of the genome under experi-

mental conditions. On the other hand, TARs identified outside the ORF boundaries are potential novel expressed regions missed by the initial annotation. Module 2 is further divided into four separate sub-modules.

#### Sub-module 1: sRNA identifier

To identify sRNAs, TARs were mapped onto the intergenic regions of the *S. pneumoniae*. Intergenic regions within operons, small 5' and 3' UTR of mRNAs, and non-unique regions (mobile genetic elements and repetitive regions) of the genome were excluded for



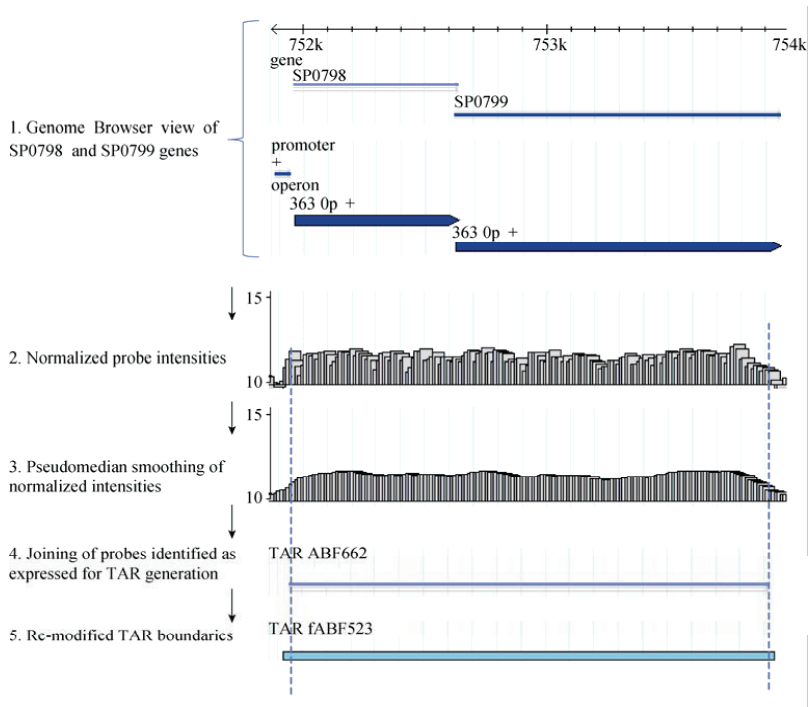
**Figure 2** Web interface of TAAPP modules and sub-modules.

this analysis. Transcripts expressed (greater than the specified minimum length) from the intergenic regions were classified as novel sRNA. The results for sRNAs include the start and end coordinates along with the DNA sequence.

**Sub-module 2: antisense identifier**

This sub-module generates a list of TARs (called antisense RNAs) that are found on the non-coding strand of a gene. The antisense RNAs for genes show

different kinds of expression patterns. For example, a gene might have many antisense RNAs or an antisense RNA may overlap the whole gene. Apart from listing all the genes that had detectable antisense RNA, the module classifies them into four different categories—5DASH overlap (antisense transcript overlapping 5'-end of gene), 3DASH overlap (antisense transcript overlapping 3'-end of gene), PART (antisense transcript as a small part located between gene ends), OVERLAP (antisense transcript fully overlapping the gene). Earlier studies have shown that 5'/3' antisense



**Figure 3** Snapshot of a short region of *S. pneumoniae* TIGR4 genome visualized in Genome Browser. Track 1 shows the operon region containing two genes SP0798 and SP0799 along with the predicted promoter. Tracks 2 to 5 show the probe intensity corresponding to the region depicted in Track 1 at various steps of tiling array data analysis.

overlaps are likely to be involved in regulatory activities (19).

### Sub-module 3: gene expression

Due to experimental variations, probes for a given genomic region may not be always expressed. Therefore, a gene region may be represented as a mixed set of expressed and non-expressed probes. A gene is considered as expressed if it has relatively higher proportion of expressed probes. The default cutoff value is taken as 70%, which represents the proportion of probes classified as expressed ( $P < 0.001$  in a binomial test) (20). The program generates a list of expressed genes based on the default selection criteria.

### Sub-module 4: operon structure

Since tiling arrays measure expression in the intergenic regions of the genomes, they can be used to identify operon structures in bacteria. Two or more consecutive genes are considered to be part of an operon, if they fulfill the following criteria: (1) they are expressed; (2) they are transcribed in the same direction; and (3) the intergenic region between the genes is identified as a single expressed transcript that overlaps the genes in both directions. Overlapping pairs of genes are joined together to identify large operon structures.

TAAPP is implemented in Perl. The software is available as a web server, so it does not need any special software installation. The two TAAPP modules are independent of each other and their simple in-

put/output format makes them suitable for any microarray platform. An extensive help file with sample input dataset is provided online.

## Application

Whole-genome tiling arrays are used to study transcriptional pattern in eukaryotes as well as prokaryotic species. Many conventional tiling array analysis programs exist for the design and analysis of tiling array datasets, but most of them were developed for eukaryotic genomes (13). The majority of these programs do not work for single color tiling arrays or customized tiling arrays. Very few software tools were described in literature for prokaryotic tiling array data (15, 16). However, these tools mainly focus on tiling array probe design and data normalization. To our knowledge, there is no software tool for prokaryotes, which performs transcript comparison with genome annotation and helps in the identification of novel features. In prokaryotes, tiling arrays can also be used to identify operon structures in bacteria, which is not possible in eukaryotic genomes.

Here we described a set of programs tailored for prokaryotic genome architecture that identifies expressed transcripts from normalized data and performs feature extraction. We implemented TAAPP on a custom *S. pneumoniae* TIGR4 single color Nimblegen tiling array dataset (Roche NimbleGen, Madison, USA), obtained from Gene Expression Omnibus database at NCBI (GSE12636). Initial data processing was done using NMPP module, which is used for preprocessing of Nimblegen specific microarray chips (21). Normalized data were used as the input for TAAPP. The TAR generation module identified 1,324 TARs in the forward (+) strand and 1,190 TARs in the reverse (-) strand with default settings. The feature identification module identified a set of 50 novel non-coding sRNAs in the intergenic regions. In total, 994 genes were expressed out of 2,015 annotated genes. The operon identifier sub-module identified 202 operon structures, consisting of 520 genes. These results for sRNA identification and operon prediction along with more analyses and RT-PCR validation were published in a separate manuscript (22). A descriptive help file is also provided with sample input

and output files, along with instructions for executing and interpreting the results of the two modules.

TAAPP automates the analysis of prokaryotic tiling array datasets and is provided as an easy-to-use web interface. The future work includes addition of confidence scores to identified novel regions and inclusion of features (like promoter and terminator) to identified transcriptional elements. Another possible improvement could be the modification of module 1 to facilitate the input of deep sequencing data.

## Acknowledgements

This project was partially supported by a grant from the National Science foundation of USA (Mississippi EPSCoR-0903787). We acknowledge the Institute for Genomics, Biocomputing and Biotechnology (IGBB), Mississippi State University for assistance with the article-processing charges for the manuscript. We thank Tony Arick and IGBB, Mississippi State University, for hosting the TAAPP web server.

## Authors' contributions

All authors contributed to the development and design of TAAPP. RK wrote all the scripts for the implementation and drafted the manuscript. BN, MLL and SCB edited the draft manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors have declared that no competing interests exist.

## References

- 1 He, H., et al. 2007. Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.* 17: 1471-1477.
- 2 Kampa, D., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14: 331-342.
- 3 Yamada, K., et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302: 842-846.
- 4 Narberhaus, F. and Vogel, J. 2009. Regulatory RNAs in



- prokaryotes: here, there and everywhere. *Mol. Microbiol.* 74: 261-269.
- 5 Romby, P., et al. 2006. The role of RNAs in the regulation of virulence-gene expression. *Curr. Opin. Microbiol.* 9: 229-236.
  - 6 Toledo-Arana, A., et al. 2007. Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.* 10: 182-188.
  - 7 Livny, J. and Waldor, M.K. 2007. Identification of small RNAs in diverse bacterial species. *Curr. Opin. Microbiol.* 10: 96-101.
  - 8 Kulkarni, R.V. and Kulkarni, P.R. 2007. Computational approaches for the discovery of bacterial small RNAs. *Methods* 43: 131-139.
  - 9 Akama, T., et al. 2009. Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. *J. Bacteriol.* 191: 3321-3327.
  - 10 Miyakoshi, M., et al. 2009. High-resolution mapping of plasmid transcriptomes in different host bacteria. *BMC Genomics* 10: 12.
  - 11 Toledo-Arana, A., et al. 2009. The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* 459: 950-956.
  - 12 Tsui, H.C., et al. 2010. Identification and characterization of noncoding small RNAs in *Streptococcus pneumoniae* serotype 2 strain D39. *J. Bacteriol.* 192: 264-279.
  - 13 Liu, X.S. 2007. Getting started in tiling microarray analysis. *PLoS Comput. Biol.* 3: 1842-1844.
  - 14 Zhang, Z.D., et al. 2007. Telescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol.* 8: R81.
  - 15 Phillippy, A.M., et al. 2009. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics* 10: 293.
  - 16 Thomassen, G.O., et al. 2009. Custom design and analysis of high-density oligonucleotide bacterial tiling microarrays. *PLoS One* 4: e5943.
  - 17 Bertone, P., et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306: 2242-2246.
  - 18 Royce, T.E., et al. 2007. An efficient pseudomedian filter for tiling microarrays. *BMC Bioinformatics* 8: 186.
  - 19 Brantl, S. 2007. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr. Opin. Microbiol.* 10: 102-109.
  - 20 David, L., et al. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA* 103: 5320-5325.
  - 21 Wang, X., et al. 2006. NMPP: a user-customized NimbleGen microarray data processing pipeline. *Bioinformatics* 22: 2955-2957.
  - 22 Kumar, R., et al. 2010. Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. *BMC Genomics* 11: 350.

CHAPTER IV  
IDENTIFICATION OF NOVEL NON-CODING SMALL RNAS FROM  
*Streptococcus pneumoniae* TIGR4 USING HIGH-RESOLUTION  
GENOME TILING ARRAYS<sup>1</sup>

---

<sup>1</sup> Reprint from Kumar R, Shah P, Swiatlo E, Burgess SC, Lawrence ML, Nanduri B. BMC Genomics. 2010. Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. This article is available from:  
<http://www.ncbi.nlm.nih.gov/pubmed/20525227>

RESEARCH ARTICLE

Open Access

# Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays

Ranjit Kumar<sup>†1,2</sup>, Pratik Shah<sup>†6,7</sup>, Edwin Swiatlo<sup>†5</sup>, Shane C Burgess<sup>†1,2,3,4</sup>, Mark L Lawrence<sup>†1,2</sup> and Bindu Nanduri<sup>†1,2</sup>

## Abstract

**Background:** The identification of non-coding transcripts in human, mouse, and *Escherichia coli* has revealed their widespread occurrence and functional importance in both eukaryotic and prokaryotic life. In prokaryotes, studies have shown that non-coding transcripts participate in a broad range of cellular functions like gene regulation, stress and virulence. However, very little is known about non-coding transcripts in *Streptococcus pneumoniae* (pneumococcus), an obligate human respiratory pathogen responsible for significant worldwide morbidity and mortality. Tiling microarrays enable genome wide mRNA profiling as well as identification of novel transcripts at a high-resolution.

**Results:** Here, we describe a high-resolution transcription map of the *S. pneumoniae* clinical isolate TIGR4 using genomic tiling arrays. Our results indicate that approximately 66% of the genome is expressed under our experimental conditions. We identified a total of 50 non-coding small RNAs (sRNAs) from the intergenic regions, of which 36 had no predicted function. Half of the identified sRNA sequences were found to be unique to *S. pneumoniae* genome. We identified eight overrepresented sequence motifs among sRNA sequences that correspond to sRNAs in different functional categories. Tiling arrays also identified approximately 202 operon structures in the genome.

**Conclusions:** In summary, the pneumococcal operon structures and novel sRNAs identified in this study enhance our understanding of the complexity and extent of the pneumococcal 'expressed' genome. Furthermore, the results of this study open up new avenues of research for understanding the complex RNA regulatory network governing *S. pneumoniae* physiology and virulence.

## Background

The emerging regulatory roles of RNA in prokaryotic and eukaryotic organisms are expanding the central dogma of molecular biology. While the full spectrum of cellular functions regulated by small non-coding RNA (called sRNA in prokaryotes) are yet to be established, work is going on to identify and study the role of non-coding regulatory RNAs in biological systems. In bacteria alone, more than 150 sRNAs are described [1]. The majority were identified in *E. coli*, and their functional characterization showed that they perform regulatory roles in sugar metabolism [2-4], iron homeostasis [5] and cell surface composition. In bacteria, sRNA also mediates post-

transcriptional gene regulation, which can be important in virulence [6,7]. Large-scale identification of sRNAs is a necessary step towards understanding their functions in normal bacterial physiology and virulence.

*S. pneumoniae*, a Gram-positive human pathogen, is the most common cause of community-acquired pneumonia and a leading cause of meningitis, sinusitis, chronic bronchitis, and otitis media [8]. Pneumococci cause approximately 63,000 invasive infections and 6,100 deaths every year in the United States alone [9]. There is a precedent for sRNA involvement in pneumococcal physiology and virulence. Investigation of the CiaRH regulon in *S. pneumoniae* strain R6 using classic molecular biology and genetic approaches resulted in the identification of 15 promoters which are regulated by CiaRH, of which five encodes sRNAs [10]. This two component regulatory system CiaRH is involved in maintaining cell integrity, competence and virulence. Expression of these sRNAs

\* Correspondence: b.nanduri@cvm.msstate.edu

<sup>1</sup> Department of Basic sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA

<sup>†</sup> Contributed equally

Full list of author information is available at the end of the article



BioMed Central

© 2010 Kumar et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

was confirmed by northern blots, and analysis of sRNA mutants showed that two of these sRNAs were important for stationary phase autolysis. Two sRNAs identified by experimental approaches in *Streptococcus pneumoniae* strain D39 had demonstrated cis-acting effects on the transcription of adjacent genes [11]. Thus there is a need for increased identification of non-coding functional elements in the pneumococcal genome.

A number of computational as well as experimental approaches have been described for identifying sRNAs in bacteria [12]. Computational methods usually rely upon sRNA conservation in closely related species [12,13] and are often limited to accuracy of transcriptional signal prediction programs (like promoter prediction and rho-independent terminator prediction). Although computational prediction of sRNAs in *S. pneumoniae* TIGR4 using program sRNAPredict2 [14] resulted in a list of 63 sRNAs, only nine were validated by Northern blotting in *S. pneumoniae* D39 strain [15]. This lack of agreement between computational prediction and experimental validation necessitates experimental approaches. Experimental methods for sRNA identification include genetic and molecular biology approaches [6,16,17]. Nowadays, genomic tiling arrays and RNA-seq methods are commonly used for genome-wide transcriptome analysis in bacteria [18]. Expression in the intergenic regions of *E. coli* and *Mycobacterium leprae* were identified using tiling arrays, suggesting the likely expression of small non-coding RNAs [19,20]. A recent study identified 27 sRNAs in *Caulobacter crescentus* using tiling array approach [21]. Using parallel sequencing, a large number of putative sRNAs were reported in *Vibrio cholerae* [22]. Immunoprecipitation with Hfq (sRNA binding protein) antibody followed by deep sequencing identified 64 sRNAs in *Salmonella Typhimurium* [23]. A total of 14 sRNAs identified by molecular biology techniques are described in *S. pneumoniae* (strains R6 and D39). To date, global experimental approaches for sRNA identification in the *Streptococcus pneumoniae* have not been reported. Here we describe a genomic tiling array approach for comprehensive identification of sRNAs in *S. pneumoniae* serotype 4 clinical isolate TIGR4. We used whole genome tiling arrays for these analyses because they offer an unbiased view of transcription at the genome level. Another reason was absence of Hfq protein in *S. pneumoniae* which eliminates the possibility of immunoprecipitation based identification of sRNAs.

*S. pneumoniae* TIGR4 genomic tiling arrays identified 50 novel sRNAs in genome, thirteen of which were validated by qRT-PCR. Computational analysis for predicting the function of TIGR4 sRNAs was conducted using Rfam database searches, BLAST searches and sequence motif analysis. Tiling arrays also identified 202 operon structures expressed in TIGR4. Overall, our results pro-

vide new insights towards understanding the complex regulatory network of the pneumococcus and underscore the importance of genomic features present in non-coding regions.

## Results

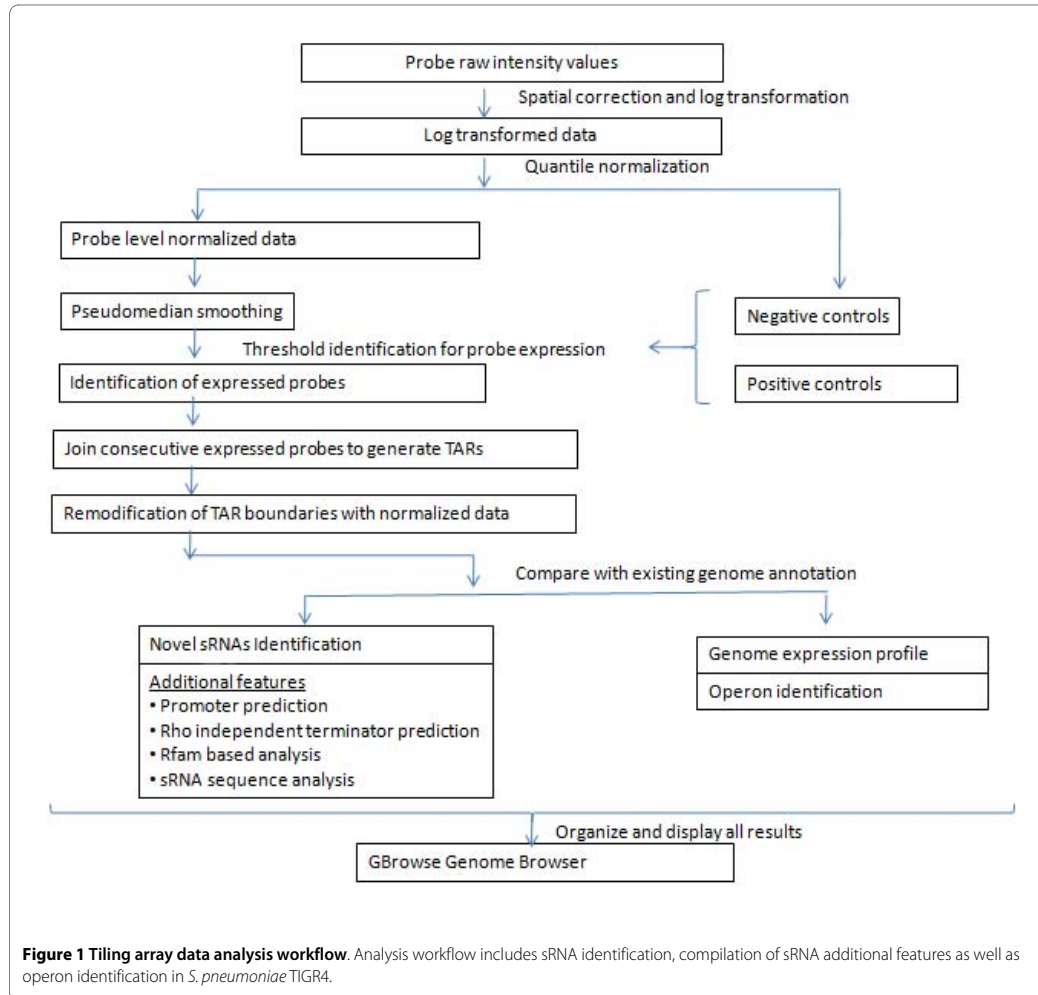
### Transcriptionally active regions in TIGR4 genome

A fundamental aspect unique to tiling array data analysis workflow was defining the baseline for the identification of expressed regions of the genome. Fluorescence intensities of spiked positive and/or negative control probes included in the array design are often used for identifying a probe level threshold for expression. RNA for the tiling array experiment was isolated from *S. pneumoniae* strain TIGR4 [24] during mid-log phase (OD<sub>600</sub> nm, 0.4-0.5). To derive a baseline for expression in our tiling experiment, we used random probes (~20,000) spotted on the array as negative controls. For positive controls, we utilized *S. pneumoniae* TIGR4 proteome data and selected 35 proteins known to be expressed under identical growth conditions [25]. To minimize sequence based effects on probe intensity, we took adjacent probe intensities into account and applied a pseudomedian filter [26]. The threshold for probe expression was set as 11.0 based on the distribution of the intensities of positive and negative control probes, pseudomedian filter setting, and the accuracy of transcript boundary detection. This threshold intensity had an associated FPR (false positive rate) of 1.63% (Additional file 1). Therefore, probes with intensity values  $\geq 11.0$  were considered to be expressed.

Consecutive expressed probes were joined together for the generation of transcriptionally active regions (TARs). We identified 2514 TARs in the TIGR4 genome, of which 1324 were found on the nominal forward (+) strand and 1190 were identified on the nominal reverse (-) strand. The genome size of *S. pneumoniae* is 2.2 Mb (2,160,837 bp), of which 88.2% is annotated as genes [24] and rest 11.8% as intergenic region. Overall, our results show that 68% of the annotated regions (that constitutes 50% genes) of the genome are expressed during mid-log phase. In addition, approximately 55% of the intergenic region was expressed which includes sRNAs, UTR regions of mRNAs, and intergenic region within operons. High level of transcription was detected in the repetitive regions present inside the intergenic regions, which were excluded from further analysis. Figure 1, shows the important steps involved in tiling array data analysis.

### Identification and sequence characterization of sRNAs

Novel non-protein coding sRNAs were identified from the intergenic region of *S. pneumoniae* TIGR4 genome. Our results identified expression in more than 55% of the intergenic region. We excluded intergenic region within operons, small UTRs (untranslated extensions of mRNA)



and repetitive regions (including insertion sequences and highly conserved mobile repeat sequences like BOX [27] and RUP [28] elements) from our analysis for identifying novel sRNA. Here we report for the first time, identification of 50 sRNAs (Table 1, sRNA SN1 in Figure 2A) in the genome. The majority of the identified sRNAs were shorter than 200 nucleotides (length range 74 - 480 nucleotides). Since our tiling array design with overlapping probes arranged at 12 bp intervals does not provide a single nucleotide resolution, we cannot accurately identify the exact transcription start/end site for sRNA. As such the start and end for sRNA in Table 1 refer to the boundaries of transcriptionally active region (putative sRNAs) and in most cases a promoter is predicted within 25 bp of transcript start site (Additional file 2). The over-

lap between the 50 sRNAs identified in this study and the 63 computationally predicted sRNAs [14] reported for *S. pneumoniae* TIGR4 is very small. Only 8 sRNAs are shared between these two datasets of which four were validated by Northern blotting [11]. A comparison of computationally predicted [14] and experimentally verified sRNAs [10,11] is available (Additional file 3). Five of the sRNAs (SN1, SN5, SN6, SN7 and SN35) were found to be homologs (BLAST identity > 98%, coverage = 100%) of the previously described sRNAs (ccnC, ccnA, ccnB, ccnD and ccnE respectively) from *S. pneumoniae* R6 strain [10]. The identification of all of the five previously identified pneumococcal sRNAs in our study, though not expected *a priori*, nevertheless strengthens our workflow. We utilized these five sRNAs as a benchmark dataset for

Table 1: *S. pneumoniae* TIGR4 sRNAs, their genome location, additional features and comparative genomics.

ID	Start	End	Length (nt)	Strand	Promoter (transcription start site)	Rho independent terminator	Flanking genes Left Right	Rfam prediction	Conservation across other genomes
SN1	24145	24254	110	+	Y	Y	SP0019(+)	SP0020(+)	$\alpha\beta$
SN2	40243	40508	266	+	Y	Y	SP0041(+)	SP0042(+)	$\alpha$
SN3	116167	116372	206	+	Y	-	SP0114(-)	SP0115(+)	$\alpha$
SN4	171543	171712	170	-	Y	Y	SP0178(-)	SP0179(+)	FMN(Cis-reg,riboswitch) $\alpha\beta\gamma$
SN5	228604	228713	110	+	Y	Y	SP0256(+)	SP0257(+)	$\alpha\beta$
SN6	230748	230916	171	+	Y	-	SP0257(+)	SP0258(-)	$\alpha\beta$
SN7	233177	233262	93	+	Y	Y	SP0260(+)	SP0261(+)	$\alpha\beta$
SN8	350572	351050	479	+	Y	Y	SP0372(+)	SP0373(+)	RNaseP_bact_b $\alpha\beta\gamma$
SN9	414094	414215	122	+	Y	-	SP0439(+)	SP0440(+)	$\alpha$
SN10	467128	467294	172	+	Y	-	SP0486(+)	SP0487(-)	FMN(Cis-reg,riboswitch) $\alpha\beta\gamma$
SN11	623211	623332	122	+	Y	Y	SP0649(-)	SP0650(-)	$\alpha$
SN12	667995	668092	98	+	Y	Y	SP0700(-)	SP0701(+)	PyrR(Cis-reg) $\alpha\beta$
SN13	681801	681922	122	+	Y	-	SP0715(+)	SP0716(+)	TPP(Cis-reg,riboswitch) $\alpha\beta\gamma$
SN14	783289	783434	146	+	Y	-	SP0834(+)	SP0835(+)	$\alpha\beta$

**Table 1: *S. pneumoniae* TIGR4 sRNAs, their genome location, additional features and comparative genomics. (Continued)**

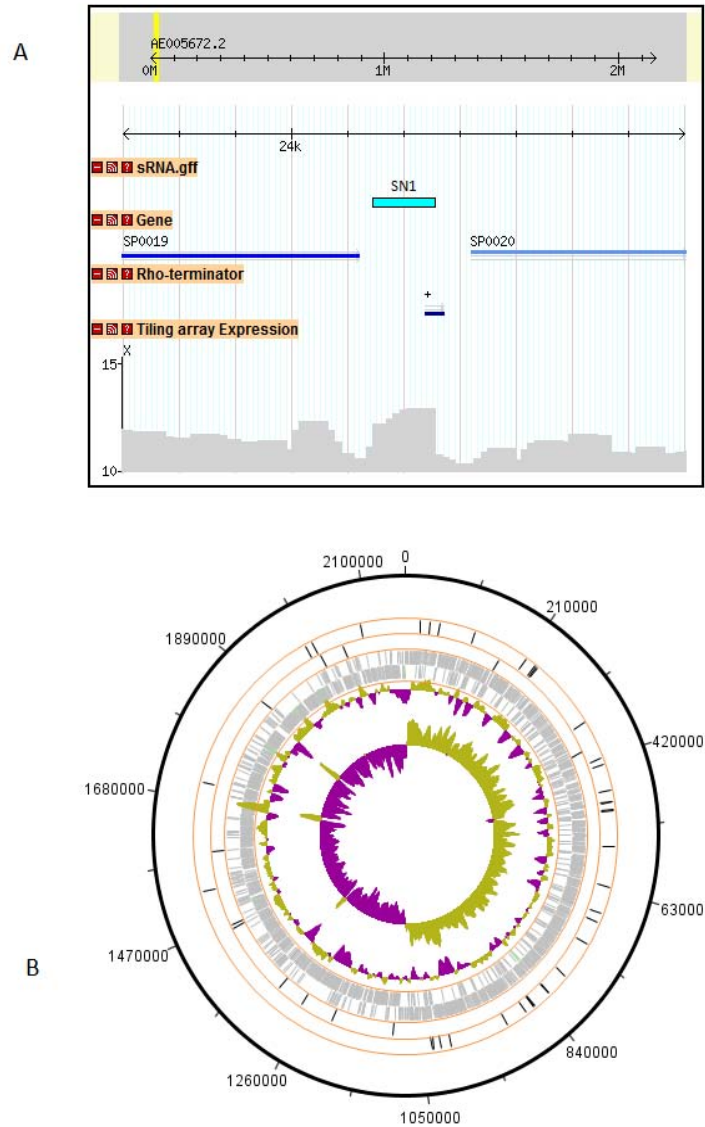
SN15	821508	821581	74	+	Y	-	SP0873(+)	SP0874(-)		$\alpha$
SN16	821892	822301	410	+	Y	Y	SP0873(+)	SP0874(-)	tmRNA	$\alpha\beta\zeta$
SN17	853100	853586	487	+	Y	-	SP0897(+)	SP0898(-)		$\alpha$
SN18	854355	854559	205	+	Y	-	SP0898(-)	SP0899(+)		$\alpha$
SN19	855530	855627	100	+	Y	-	SP0899(+)	SP0900(-)		$\alpha$
SN20	869478	869791	318	+	Y	Y	SP0915(-)	SP0916(+)		$\alpha\beta$
SN21	1005291	1005532	242	+	Y	Y	SP1068(+)	SP1069(+)	T-box(Cis-reg)	$\alpha\beta$
SN22	1033894	1034015	125	+	Y	-	SP1100(+)	SP1101(-)		$\alpha$
SN23	1324023	1324276	256	-	Y	-	SP1400(-)	SP1401(+)		$\alpha$
SN24	1529942	1530039	98	+	Y	-	SP1629(+)	SP1630(+)	T-box(Cis-reg)	$\alpha$
SN25	1592924	1593285	362	-	Y	-	SP1691(-)	SP1692(+)		$\alpha$
SN26	1989967	1990063	97	+	-	-	SP2078(+)	SP2079(-)		$\alpha$
SN27	2086051	2086304	277	+	Y	Y	SP2168(+)	SP2169(-)		$\alpha$
SN28	485360	485540	181	+	Y	-	SP0502(+)	SP0503(-)		$\alpha$
SN29	497140	497360	221	+	Y	-	SP0516(+)	SP0517(+)		$\alpha$
SN30	499750	499970	231	+	Y	-	SP0518(+)	SP0519(+)		$\alpha\beta$
SN31	2000722	2001113	392	+	Y	-	SP2092(+)	SP2093(+)		$\alpha$
SN32	1022430	1022539	121	+	Y	-	SP1086(+)	SP1087(+)		$\alpha$
SN33	392134	392231	105	-	Y	-	SP0411(-)	SP0412(-)		$\alpha\beta$

**Table 1: *S. pneumoniae* TIGR4 sRNAs, their genome location, additional features and comparative genomics. (Continued)**

SN34	1706645	1706890	246	+	Y	-	SP1790(+)	Spt11(+)	6S	$\alpha\beta$
SN35	209748	209905	158	+	Y	Y	SP0239(+)	SP0240(+)		$\alpha\beta$
SN36	423848	423992	145	+	Y	Y	SP0451(+)	SP0452(-)		$\alpha$
SN37	557778	557971	194	+	Y	Y	SP0587(-)	SP0588(+)		$\alpha\beta$
SN38	485578	485759	182	+	Y	-	SP0502(+)	SP0503(-)		$\alpha$
SN39	721337	721446	110	+	Y	-	SP0761(+)	SP0762(+)		$\alpha\beta$
SN40	907168	907301	134	+	Y	Y	SP0958(+)	SP0959(+)	L20_leader(Cis-reg)	$\alpha\beta$
SN41	1037030	1037185	166	+	Y	-	SP1104(+)	SP1105(+)	L21_leader(Cis-reg)	$\alpha\beta$
SN42	1214232	1214365	137	-	Y	Y	SP1278(-)	SP1279(-)	PyrR(Cis-reg)	$\alpha$
SN43	1275596	1275742	147	-	Y	Y	SP1355(-)	SP1356(-)	L10_leader(Cis-reg)	$\alpha\beta\text{‡}$
SN44	1460966	1461207	242	-	Y	Y	SP1551(-)	SP1552(+)	yypP-ykoY(Cis-reg)	$\alpha\beta$
SN45	2005540	2005697	181	-	Y	Y	SP2097(-)	SP2098(-)		$\alpha$
SN46	2048539	2048648	112	-	Y	-	SP2136(-)	SP2137(+)		$\alpha$
SN47	56069	56190	122	+	-	-	SP0051(+)	SP0052(+)		$\alpha\beta$
SN48	1102915	1103083	169	-	-	-	SP1166(-)	SP1167(-)		$\alpha$
SN49	1455217	1455362	146	-	Y	-	SP1547(-)	SP1548(-)		$\alpha$
SN50	1874532	1874844	313	-	-	-	SP1966(-)	SP1967(-)		$\alpha\beta$

sRNA sequences conserved in;  $\alpha$ -different *Streptococcus pneumoniae* strains like CGSP14, G54, Hungary19A-6, R6, D39.  $\beta$ -different species of *Streptococcus* like *S. mitis*, *S. gordonii*, *S. sanguinis* SK36.  $\text{‡}$ -other species outside *Streptococcaceae* (for example *Lactobacillus*, *Clostridium*, and *Bacillus*). The start and end represents the boundaries of identified TAR (transcriptionally active region) which is a potential sRNA region.





**Figure 2 A *S. pneumoniae* TIGR4 sRNA SN1 visualized in the genome browser.** The sRNA and the additional features are shown as different tracks in the genome browser. sRNA track (in blue) shows the presence of small RNA SN1. Tiling array expression track indicates the higher level of expression in the sRNA SN1 region (located in-between genes SP0019 and SP0020) relative to the intensity threshold cutoff (11.0). Rho-independent terminator track shows a predicted terminator near the 3' end of sRNA. **B. Circular representation of *S. pneumoniae* TIGR4 genome depicting open reading frames and sRNAs.** The outermost track (solid black circle, track one) is TIGR4 genome. With reference to track one, moving inward, tracks two and three represent sRNAs in the forward and reverse strand respectively. Tracks four and five (gray) shows the presence of genes on the forward and reverse strand respectively. Track six is the GC plot and the seventh (innermost track) shows the GC skew of the genome.

evaluating the results of our computational analyses of sRNA sequences.

The expression of sRNAs showed a strong bias towards the forward strand (38 sRNAs) relative to the reverse strand (12 sRNAs) even though the distribution of protein coding genes in TIGR4 is almost equal for both strands of DNA. We found that the TIGR4 genome has gene orientation bias, a common feature of low-GC (Gram-positive) organisms. Approximately, half of the total genes were located to the right of the origin of replication, of which 79% are transcribed in the same direction as DNA replication and vice versa [24] (Figure 2B). Since two thirds of the identified sRNAs were located to the right of origin of replication, the majority of the sRNAs in our study were expressed in the forward strand.

Transcription is usually facilitated by promoter sequences located in the 5' upstream region on same strand of DNA. Earlier comparative genomics studies also have reported the presence of rho-independent transcription terminators as evidence for the identification of sRNA [29]. Both promoter and rho-independent terminators were also experimentally identified in the five homologs of previously identified pneumococcal sRNAs from R6 strain [10]. The results of computational analysis for promoter/terminator showed that most of the sRNAs had a predicted promoter within 25 nt upstream of the TAR start site. In some cases more than one promoter was predicted in the upstream region of sRNA sequence. Rho-independent transcription terminators were predicted for 20 sRNAs within 25 bp downstream of transcription end site. The predicted promoter sequence with transcription start site and terminator sequences for sRNAs are present (Additional file 2). We also evaluated the potential protein coding capacity of sRNAs by translating the sequences in all three open reading frames. Our results indicate that two sRNAs (SN48 and SN50) encoding regions have the potential to code smaller proteins. Further analysis of the DNA sequence in these regions using "FGENESB" gene prediction tool <http://www.softberry.com> identified the presence of smaller ORF (open reading frame). We did not find any predicted promoter sequences in the upstream regions of these two sRNAs, suggesting they may constitute part of an operon. Further analysis revealed that SN48 is indeed located in a four gene operon (SP1166 to SP1169). BLAST based sequence searches against non-redundant protein database at NCBI did not identify any matches for these two sRNAs in other genomes suggesting that these potential novel genes are currently unique to *S. pneumoniae* TIGR4. While SN48 and SN50 could encode proteins, in absence of experimental validation of ORF, it is not possible to rule out their functional involvement as a sRNA. Therefore we included SN48 and SN50 in our sRNA list (Table 1).

#### Comparative genomics of sRNA sequences

The average GC content of sRNAs ( $35\% \pm 5\%$ ) was slightly less than the average GC content of the TIGR4 genome (39.7%). BLAST analysis of sRNA sequences against the non-redundant nucleotide database at NCBI revealed that all sRNA sequences were highly conserved (coverage ~ 100%, identity > 97%) within other pneumococcal strains (including CGSP14, G54, Hungary19A-6, R6, and D39; Table 1). But only 25 is found to be conserved in closely related species of *Streptococcus* (for example *S. mitis*, *S. gordonii*, and *S. sanguinis* SK36) [30]. However, these sRNAs were not conserved in other species of *Streptococcus* like *S. pyogenes*, *S. mutans*, or *S. bovis*. This lack of sRNA sequence conservation at the genus level indicates that these sRNAs might have been acquired during pneumococcal evolution. Six sRNA sequences were found to be conserved in other species outside *Streptococcaceae* (for example *Lactobacillus*, *Clostridium*, and *Bacillus*) and are known to be involved in various regulatory functions.

#### Computational functional prediction of sRNAs

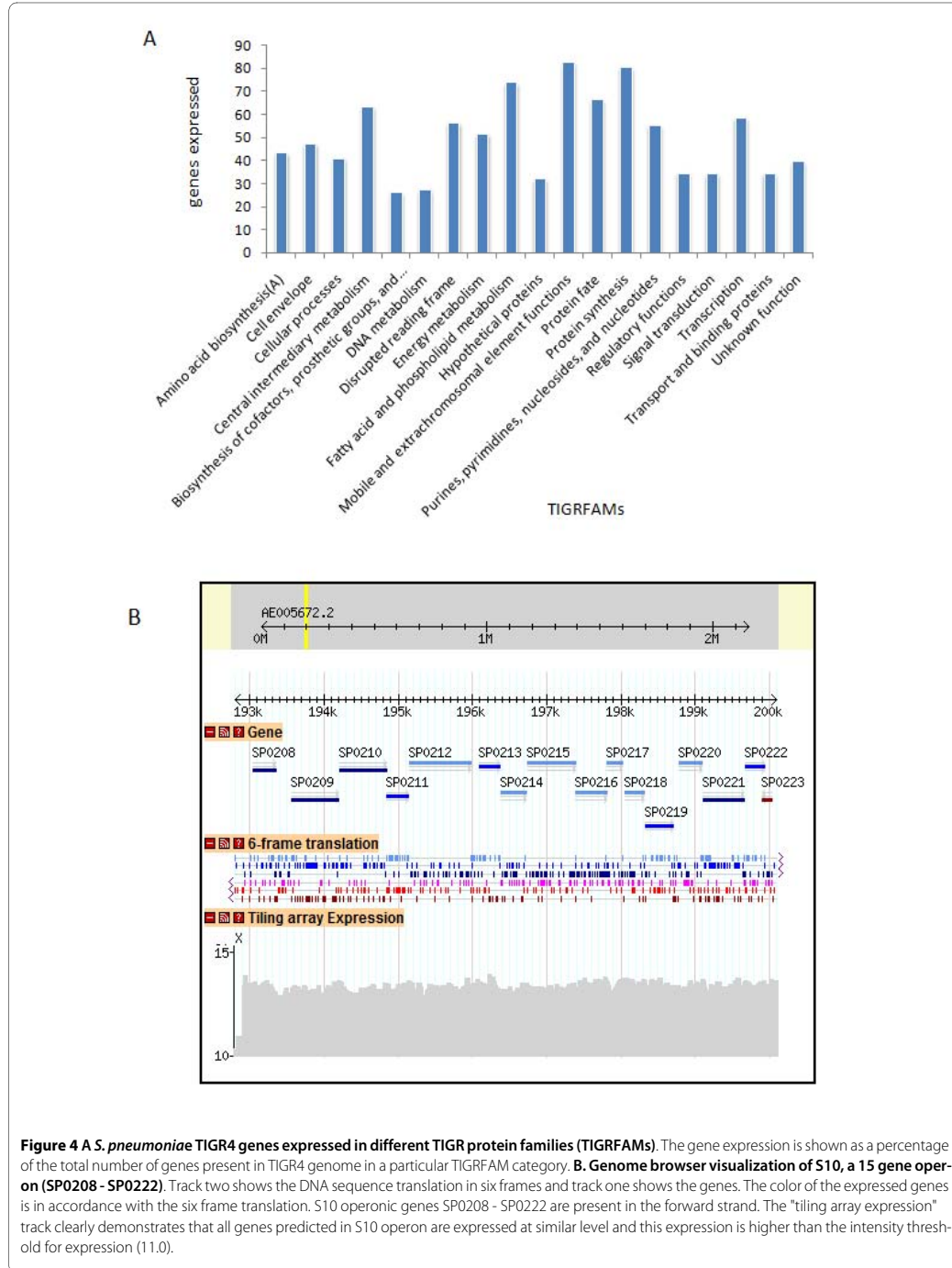
sRNAs can be functionally characterized as either cis- or trans- regulators based on the location of their target genes. The Rfam database [31] is a collection of non-coding RNA families represented by multiple sequence alignments and profile stochastic context-free grammars. We searched all TIGR4 sRNA sequences against the Rfam database to determine their putative functions. We found that some of the pneumococcal sRNAs we identified were homologs to well characterized sRNAs in other genomes. The identified functional categories include FMN riboswitches, TPP riboswitch, PyrR family, Tbox leader elements, r-protein leader autoregulatory structure, putative endoribonuclease (RNaseP\_bact\_b), tmRNA, and 6S (Table 1; description of individual categories is available at Rfam). With Rfam database searches we could assign putative functions to 14 sRNAs, 11 of which were predicted to be cis-regulators. Three of the cis-sRNAs were predicted riboswitches that could directly bind a small target molecule. For 36 sRNAs we could not predict function using computational methods. These sRNAs likely represent a novel set of non-coding sRNAs in pneumococci.

#### Motif and structural analysis of sRNA sequences

To identify sequence characteristics among the pneumococcal sRNAs, we searched for the presence of overrepresented sequence motifs using MEME SUITE [32]. A sequence motif is a nucleotide sequence pattern that is widespread and has, or is predicted to have, structural or biological significance. All sRNA sequences were used for motif prediction, and the top 8 motifs (present in total 22 sRNAs) were selected based on high score, length (> 15

Motif	Weblogo	sRNA having motif	P-value range	Rfam predicted functions of sRNA
M1		SN1, SN5, SN6, SN7, SN35	2.16e-12 to 4.27e-11	Homologs of CiaR regulated sRNA of R6 strain
M2		SN1, SN5, SN6, SN7, SN35	1.06e-18 to 5.74e-16	Homologs of CiaR regulated sRNA of R6 strain
M3		SN4, SN10, SN20	4.90e-22 to 9.32e-17	SN4 and SN10 are FMN riboswitch
M4		SN14, SN30, SN37	2.47e-14 to 1.88e-13	None
M5		SN21, SN24, SN33	4.84e-20 to 3.35e-19	SN21 and SN24 are cis-regulated T-box
M6		SN12, SN42	1.60e-19 to 1.74e-18	SN12 and SN42 are cis-regulated PyrR
M7		SN17, SN32, SN38	4.11e-19 to 2.01e-17	None
M8		SN16, SN29, SN49	1.38e-11 to 1.21e-10	SN16 is tmRNA

**Figure 3 Sequence motifs identified in sRNAs by MEME.** Overrepresented sequence motifs among non-aligned sRNA sequences were identified by MEME. Rfam annotation for sRNAs are shown where available.



**Figure 4** *S. pneumoniae* TIGR4 genes expressed in different TIGR protein families (TIGRFAMs). The gene expression is shown as a percentage of the total number of genes present in TIGR4 genome in a particular TIGRFAM category. **B. Genome browser visualization of S10, a 15 gene operon (SP0208 - SP0222).** Track two shows the DNA sequence translation in six frames and track one shows the genes. The color of the expressed genes is in accordance with the six frame translation. S10 operonic genes SP0208 - SP0222 are present in the forward strand. The "tiling array expression" track clearly demonstrates that all genes predicted in S10 operon are expressed at similar level and this expression is higher than the intensity threshold for expression (11.0).

nucleotides), and p-value ( $< 1e^{-10}$ ). Our results indicate that sRNAs predicted to have similar functions share common motif sequences (Figure 3). All members of motif group M1 and M2 were functionally similar, the five sRNAs which were homologs of CiaRH regulated sRNA in *S. pneumoniae* R6 strain. Similarly, members of motif group M3, M5 and M6 share similar functions.

We also investigated secondary structure of motif sequences based on MFOLD predicted sRNA structures [33]. Our results showed that in most motif groups, the sRNA sequences had similar motif structures (Additional file 4). Motif M1 always forms a partial stem loop-like structure in all five sRNAs (SN1, SN5, SN6, SN7, and SN35), while motif M2 forms a large unpaired segment. Motif M2 in SN7 and SN35 assumes a partial stem loop structure while a large portion of the sequence still remains unpaired. Motifs M3, M4, M5, and M7 form stem loop structures in corresponding sRNA sequences. Motifs M6 and M8 includes two stem loop structures along with the unpaired region between them. The 28 sRNAs that had no detectable sequence motifs could represent a set of diverse sequences having different mode of action.

Searching these motifs in motif database using TOM-TOM [34] results in identification of motif M6 associated with pyrR (transcriptional attenuator and uracil phosphoribosyltransferase activity) regulated function, similar to sRNAs (SN12 and SN42) predicted function. Motif M6 was identified to be a part of antiterminator binding region in regulatory protein, PyrR, where it regulates the transcription of pyr operon by attenuation mechanism [35-37]. We also analyzed two motifs M3, M5 that were present in the sRNAs whose functions are well described in literature. Motif M3 was found to be a part of aptamer structure (the region binding to small molecules) of FMN riboswitches [38,39]. Motif M5 was found to be present in the conserved part of the specifier loop of T-box regulated genes [40,41]. T-box antitermination is considered as one of the main mechanisms to regulate gene expression in amino acid metabolism in gram-positive bacteria. The other described motifs could represent important novel structural or functional regions to be investigated.

#### Gene expression profile and identification of operon structures

Our results indicate that ~50% of *S. pneumoniae* TIGR4 genes were expressed during mid-log growth phase. We characterized the set of expressed genes, which represent basal transcriptional activity under our growth condition, using TIGRFAMs (Figure 4A). The expressed genes are involved in fundamental biological processes such as transcription, protein synthesis, protein fate, and cell division (Additional file 5). Processes such as fatty acid and phospholipid metabolism were also represented in

the expression profile. We found that approximately 40% of the expressed genes were involved in processes mediating DNA metabolism, regulatory functions, and signal transduction. Almost all genes with mobile and extra-chromosomal functions were expressed. Genes encoding surface proteins, proteins involved in acquiring nutrients, and transporters were also expressed [24]. Interestingly, one third of the annotated hypothetical genes (97) and around half of the genes annotated as disrupted reading frames (52 out of 92) were expressed.

In bacteria genes involved in carrying out similar function are often organized into operon structures. Identifying operon structures is critical for understanding coordinated regulation of bacterial transcriptome. Identifying transcriptional units can also help in assigning function to hypothetical genes when present in an operon of known function [42]. Tiling arrays efficiently identify co-expressed genes and transcription units at a genomic scale. We identified co-expression for 520 pairs of TIGR4 genes (Additional file 6) that were transcribed together and constituted minimal operons. By joining consecutive overlapping pairs of co-expressed genes, we identified 202 distinct transcription units/operons (size varied between two to fifteen genes; Additional file 7).

The operons identified in this study were compared to previously described pneumococcal operons (Table 2). The *vic*, *man*, *atp*, and *marMP* operons identified by tiling arrays concur with previously described operon structures [43-46]. In *S. pneumoniae* R6, *marMP* operon is considered to have three genes (SP2108-SP2110 in *S. pneumoniae* TIGR4). In contrast, our results identified only two genes as transcription unit (SP2109-2110). Our data clearly shows that the expression of SP2108 is higher than SP2109 - SP2110 (Additional file 8), suggests that SP2108 is either expressed as an independent transcription unit or there exists a possibility of overlapping transcripts among these three genes. We did not identify *murMN*, *phg*, and *comCDE* operon expression, suggesting that these genes may not be required for mid-log growth phase. Lack of expression of competence related genes is expected as THB medium used for propagating *S. pneumoniae* does not support competence.

Comparing our experimentally identified co-expressed genes with computationally predicted operons using "DOOR" [47] showed that there was an approximately 63% overlap between both datasets (291 gene pairs, excluding rRNA and tRNA; Additional file 9). Thus, our dataset experimentally validates 291 DOOR gene-pair predictions. Tiling array expression analysis also identified 229 additional co-expressed gene-pairs that were not predicted by DOOR, which may help in refining the boundaries of identified transcriptional units with greater accuracy. For example, DOOR predicted the S10 operon (coding for ribosomal proteins) in TIGR4 as a 14 gene

**Table 2: Comparison of *S. pneumoniae* TIGR4 operons identified by tiling arrays with Streptococcus operons described in literature.**

Operon name	Experimental prediction	Tiling array predictions	Literature reference (PUBMED ID)
murMN	SP0615-SP0616	-	10759563
vic	SP1225-SP1226-SP1227	SP1225-SP1226-SP1227	12379689
MiaR reg MarMP (3 operon)	SP2106-SP2107	SP2106-SP2107	11278784
	SP2108-SP2109-SP2110	SP2108	
	SP2111-SP2112	SP2109-SP2110	
		SP2111-SP2112	
phg	SP1043-SP1044-SP1045	-	15271918
TrmD	SP0776-SP0777-SP0778- SP0779-SP0780	SP0778-SP0779-SP0780	15060037
ComCDE	SP2235-SP2236-SP2237	-	9352904
luxS & Dcw3	SP0340	SP0340	16436421
	SP0334-SP0335-SP0336- SP0337	SP0336-SP0337	
man	SP0282-SP0283-SP0284	SP0282-SP0283-SP0284	12486041
atp	SP1507-SP1508-SP1509- SP1510-SP1511-SP1512- SP1513-SP1514	SP1507-SP1508-SP1509- SP1510-SP1511-SP1512- SP1513-SP1514	15576803

operon (SP0209-SP0222). However, tiling analysis indicated that the S10 operon has 15 genes (SP0208-SP0222) and included SP0208 (Figure 4B). Interestingly, we found that *Bacillus subtilis* S10 operon structure is similar to our experimentally derived pneumococcal S10 operon structure (fifteen genes, including a SP0208 homolog) [48]. One possible reason for the exclusion of SP0208 by DOOR could be the relatively large 217 bp intergenic region between SP0208 and SP0209. In another example, tiling expression identified rplK-rplA (SP0630-SP0631) genes as part of single transcriptional unit, but DOOR failed to identify this unit possibly due to the presence of a large 207 bp intergenic region between rplK and rplA.

Proteins encoded by genes in the same operon often have related function or are in the same biological pathway. Therefore, putative function may be assigned to

hypothetical genes when located in an operon of known function [42]. In our operon dataset, approximately 20% (147) of the genes encode hypothetical proteins. In operon 8, a three gene operon (SP0077 - SP0079), two genes encode Trk family of potassium uptake proteins, and one gene (SP0077) encodes a hypothetical protein. Therefore, it is possible that SP0077 may be a member of the Trk transporter protein family. In another three gene operon (SP0904-SP0906), all genes encode hypothetical proteins; it is possible these proteins have similar as yet un-assigned functions.

#### Experimental validation of sRNAs

Expression of 14 sRNAs identified by genomic tiling expression analysis was analyzed by qRT-PCR. The sRNAs selected for validation included 5 sRNAs identified in *S. pneumoniae* R6 strain and 9 novel TIGR4

sRNAs identified in the current study. Statistical t-tests were performed for each sRNA between the  $C_t$  value for the (reverse strand) vs the  $C_t$  value of the background (no primer) to determine if there was significantly higher expression than background. Another t-test was conducted for each sRNA between the  $C_t$  value for the reverse strand vs the  $C_t$  value for the forward strand to determine if there was significant expression from the sense strand. At  $p \leq 0.05$ , for 13 sRNAs we found significantly higher expression (lower  $C_t$  value) for the coding strand specific qRT-PCR compared to the non-coding strand and background (no primer) (Additional file 10). The  $p$ -value of sRNA SN24 was not significant at  $p < 0.05$ . Three of the validated sRNAs (SN4, SN12 and SN16) had available annotations (Table 1). Although validated, no functional information was predicted for sRNAs SN2, SN11, SN22 and SN27. All five sRNAs whose homologs were present in *S. pneumoniae* R6 strain were also positively validated. Overall, qRT-PCR validations were successful for thirteen out of fourteen sRNAs.

## Discussion

Tiling array analysis is widely used in eukaryotes to study transcriptional complexity and identifying non-coding transcripts [49-52]. Recent studies in *Mycobacterium leprae* and *E. coli* described whole genome tiling array approach for sRNA identification [20]. Parallel sequencing technology was used for sRNA identification in *Salmonella* [23] and *Vibrio cholerae* [22]. Individual experimental studies [10,15] altogether identified 14 sRNAs in two different strains of *S. pneumoniae* (D39 and R6 strain). To our knowledge, this is the first study to report the use of whole genome tiling arrays for experimental identification of sRNAs at a global scale in *S. pneumoniae*. The tiling array analysis method described here is a combination of the methods described by others [49,52], but tailored for prokaryotic genomes. Hfq protein plays a central role in sRNA function in *E. coli*, facilitating the pairing of sRNA with its mRNA target [53]. One experimental approach for sRNA identification in bacteria could be the co-immunoprecipitation of sRNA using Hfq antibodies [16]. However, *S. pneumoniae* TIGR4 genome does not code for Hfq protein which precludes applying this method to TIGR4 genome. Therefore, tiling array approach described in this study is a pragmatic experimental approach for identifying sRNAs. Identifying the sRNA repertoire of TIGR4 is the first step towards understanding the sRNA regulatory network of this human pathogen.

The transcriptome map generated in this study identified expression in two thirds of TIGR4 genome. Tiling array analyses of *E. coli* and yeast reported expression of 87% and 90% of the genome respectively [50,54]. Com-

pared to these studies, TIGR4 genome expression in this study was relatively in lower proportion (68%). Possible reasons for this lower expression could be the growth conditions and/or the stringent intensity cutoff used for identification of expressed regions. We choose a stringent intensity cutoff (11.0) to maintain a low false positive rate (1.63%) for identifying sRNAs, which are usually short in length (50-200 bp).

As a result, we report for the first time genome-wide identification of 50 novel sRNAs in pneumococcus using tiling arrays. Additional features, such as presence of a promoter and rho-independent terminator, were computationally predicted for identified sRNAs. Almost half of the identified sRNAs showed the presence of a rho-independent terminator. As speculated by others [29,55], our analysis indicates that identification of rho-independent terminator sequence is the strongest determinant for the identification of sRNA. Furthermore, the identification of rho-independent terminator downstream from sRNA sequences helped us in differentiating the sRNAs from the 5' untranslated extensions of genes. However, it is possible that some sRNAs may be associated with a rho-independent terminator and thus would not be identified in our search.

Comparative genomics of sRNA sequences revealed that only six sRNA sequences involved in various regulatory activities were conserved beyond *Streptococcaceae* (example *Lactobacillus*, *Clostridium*, *Bacillus* (Table 1). The evolutionary tree of Streptococcus family [30] indicates that *S. mitis*, *S. gordonii*, *S. sanguinis* SK36 are phylogenetically closer to *S. pneumoniae* than other species (like *S. pyrogens*, *S. mutans* or *S. bovis*) which explains the conservation of 25 sRNAs in *S. mitis*, *S. gordonii*, and *S. sanguinis* SK36), but not present in other species like *S. pyogenes*, *S. mutans*, or *S. bovis*. It also indicates that sRNA prediction algorithms that rely on comparative genomics need to first account for the observed low sequence conservation of sRNAs among different species [13]. Our results suggest that computational methods which rely on comparative genomics to find sRNAs need to focus on carefully selected closely related species. The 50 sRNAs identified in this study along with their comparative genomics could serve as a training dataset for further computational sRNA predictions in pneumococcus, particularly for the identification of sRNAs which are not expressed under our experimental conditions. At last, we speculate that computational prediction of *Streptococcus* sRNAs using comparative genomics with *S. mitis*, *S. gordonii*, and *S. sanguinis* SK36 will identify new as yet undescribed sRNAs.

Exploring the sequence characteristics of sRNAs described in this study showed that sRNAs predicted to have similar biological function share common sequence motif. We identified 8 sequence motifs, of which five

were identified in TIGR4 for the first time. Members of the motif group without predicted function could have similar structural or functional properties. For example SN20 had motif M3 and might function as a FMN switch similar to SN4 and SN10, which also contain this motif. Likewise, sRNAs present in motif group M4 could be predicted to have similar yet undefined function. Structural analysis of motif (Additional file 4) suggests that they mainly form two kinds of structure in sRNAs; firstly, the whole motif forms a stem loop structure (like motif M5) and secondly, the motif is present as two stem loop structures including the unpaired region between them (like motif M6). Furthermore, motifs present in sRNAs with similar function also formed a conserved secondary structure (for example, motifs M1, M2, and M5). We speculate that (SN32 and SN38), (SN16 and SN29), (SN21, SN24 and SN33), (SN14 and SN37) contains similar motif structure and might share similar yet unknown structure/function. This structural conservation of motifs also suggests that motif regions of sRNA could be structurally or functionally important regions and can be used as targets for mutational studies to decipher function.

The accuracy of computational operon prediction in bacteria is 85-91% in terms of specificity and sensitivity for predicting operonic gene pairs (pairs of consecutive genes that are part of the same operon) in *E. coli* and *B. subtilis*, respectively [56]. However, the sensitivity of prediction drops to as low as 50% when predicting transcription units with more than one gene [56]. Two examples were discussed in results where the computational prediction failed to identify a gene pair as a part of an operon due to the presence of a large intergenic region between them. The accuracy of computational operon prediction algorithms also decreases when performing predictions for newly sequenced genomes for which no training dataset is available. Based on tiling array analysis, we generated 520 gene pairs that were co-expressed and identified 202 transcription units in *S. pneumoniae* TIGR4. Our results clearly demonstrate the effective use of tiling arrays for operon identification at a whole genome scale. An obvious limitation to the tiling array approach is the inability to identify operons whose genes are not expressed in the experimental growth condition. Nevertheless, our results demonstrate that combining operons identified by tiling with computational prediction greatly improves operon identification in genomes, as speculated by other researchers [57]. The operons identified in this study, though not comprehensive, still represent a validated dataset of approximately 202 operons.

Around 8% of the *S. pneumoniae* TIGR4 genome is repetitive in nature. It includes sequences (> 50 bp) that are present at multiple locations in the genome, such as mobile genetic elements, small dispersed repeats like RUP and BOX elements, and other repetitive regions.

Although these regions were excluded for identifying sRNA, we detected a high level of transcription in these repetitive regions from both sense and antisense strands. Because it is not possible to identify the actual origin of transcription with tiling arrays, future experiments designed to analyze the transcriptional activity in these repeat regions are warranted. In view of recent findings where sRNAs are involved in repressing expression of toxic proteins [58] and are present in multiple copies, we speculate that these repetitive regions may be involved in various regulatory activities within the cell.

In conclusion, our combinatorial approach of experimental identification of sRNAs on a genome scale using tiling arrays in conjunction with computational analyses of sRNAs in *S. pneumoniae* TIGR4 has resulted in the description of 50 sRNAs in this clinically relevant strain. Our result forms the initial framework for understanding sRNA-based regulation of *S. pneumoniae* gene expression.

## Conclusions

Here we have demonstrated the utility of tiling arrays to study whole genome transcription in prokaryotes. The analysis of high-resolution transcription map of the *S. pneumoniae* clinical isolate TIGR4 results in identification of 50 novel sRNAs. Bioinformatics sequence based searches helped to predict function of 14 sRNAs. Comparative genomics shows that half of the identified sRNA sequences are unique to *S. pneumoniae* genome. We identified eight overrepresented sequence motifs among sRNA sequences that correspond to different functional categories. We identified 202 operon structures in the genome, further validated by available experimental identifications. Overall, this work elucidated pneumococcal operon structures and identified previously undiscovered sRNAs, which will enhance our understanding of the complexity and extent of the pneumococcal 'expressed' genome. Also, this work opens up new avenues for understanding the complex RNA regulatory network governing *S. pneumoniae* physiology and virulence.

## Methods

### Isolation of total RNA from *S. pneumoniae* TIGR4

*S. pneumoniae* strain TIGR4 [24] was grown in Todd-Hewitt broth supplemented with 0.5% yeast extract (THY). Cells were harvested during mid-log phase ( $OD_{600}$  nm, 0.4-0.5) of growth by centrifugation from two biological replicates. The harvested pellets were washed twice in sterile phosphate-buffered saline (PBS; pH 7.4) and stored at -80°C. RNA was purified from frozen bacterial pellets using Qiagen RNeasy kit <http://www.qiagen.com/> following the manufacturer's protocol. Isolated RNA was treated with DNase, and the purity was checked by performing a one-step RT-PCR using primers specific



for 16 S rRNA in the presence or absence of reverse transcriptase. RT-PCR performed in the presence reverse transcriptase in the reactions resulted in the amplification of the desired PCR product. In contrast, no PCR product was generated when reverse transcriptase was excluded from the reaction mix, confirming that the isolated RNA did not have genomic DNA. RNA concentration and quality were determined by using Agilent Bioanalyzer (Agilent, Foster City, CA). Purified RNA was stored in nuclease free water at -80°C. One microgram of total RNA was used by Nimblegen systems (Roche NimbleGen, Inc. Madison, WI) for labeling and hybridization.

#### High density genome tiling and hybridization

High density oligonucleotide microarrays from Nimblegen Systems that incorporate "Maskless Array Synthesis" [59] technology for designing probes were used to study the expression of TIGR4 genome. The tiling array was designed based on the TIGR4 genome sequence (obtained from Genbank, accession number [NC\\_003028](#)). Probes of 50 nucleotide length were designed in an overlapping fashion at 12 bp intervals for both strands across the entire genome, resulting in a total of 359,366 probes. Twenty thousand random probes were included for measuring non-specific hybridizations. Labeling of cDNA with Cy3, hybridization, and scanning were conducted by Nimblegen Systems (detailed protocol available at <http://www.nimblegen.com/products/lit/lit.html>) and Nimblegen provided resulting raw fluorescence intensity values.

#### Normalization and data analysis

Spatial effects (uneven washing or scanning) were removed from the fluorescence intensity data using a global distance-weighted smoothing algorithm for correction available in the NimbleGen Microarray Data Processing Pipeline (NMPP) [60]. NMPP output was log transformed for further analysis. Quantile normalization was performed using the Affy package available in R language <http://pbil.univ-lyon1.fr/library/affy/html/normalize.quantiles.html> to remove systematic errors (biases) from the replicate slides and to generate identical intensity distribution for both chips [61]. The correlation coefficient between the intensities of the two chips was  $r^2 \geq 0.90$ .

Although a number of methods are described in the literature for tiling array data analysis [62-65], most were not readily applicable to our dataset because of our single color array design. Furthermore, the existing methods are not tailored for prokaryotic genomes. Therefore, for processing our TIGR4 tiling array data, we modified Kampa et al. method [52] as described below:

1. Instead of using PM (positive-match) - MM (mismatch) intensities, we used PM probe intensities only.

2. Pseudomedian filter (which takes adjacent probe intensities into account) was used to adjust for sequence based variation at the probe level and provide an initial smoothing of the raw probe intensity values [26]. Pseudomedian (Hodges-Lehman estimator) for each probe was calculated with a sliding window size of 11 probes (170 bp).

3. To identify the transcribed regions of the genome, we considered a probe to be expressed when its pseudomedian intensity was found to be higher than a threshold value. The threshold value was determined on the basis of distribution of positive and negative control probe intensities, pseudomedian filter setting, accuracy of transcript boundary detection, and the associated false positive rate.

4. To identify TARs (transcriptionally active regions), consecutive expressed (transcribed) probes were joined together using maxgap-minrun method [52]. The maxgap parameter allows certain number of probes (one or two probes) to be below the cutoff while still being incorporated into the TAR, whereas the minrun parameter requires at least a certain length of the TAR to be considered further. To account for the densely packed prokaryotic genomes (shorter intergenic regions), the maxgap feature was not applied in the intergenic regions. We used a minrun value of 74 (at least 3 consecutive probes) for sRNA detection.

5. The pseudomedian filter can result in slightly erroneous identification of transcript boundaries (start and end). Therefore, we implemented a new step that re-modified transcript boundaries using normalized average raw intensity values. Re-modification was conducted by either elongating or shortening transcript ends until the average raw intensity values of the probe (not pseudomedian value) was greater than or equal to the threshold cutoff. Overlapping transcripts were then joined together for TAR generation.

All of the above analytical steps were performed using in-house PERL scripts. Steps one, four, and five were modifications of the Kampa method and are specific to our analysis. The tiling array data from this study have been submitted to Gene Expression Omnibus under accession no. GSE12636.

#### Analysis of annotated regions of TIGR4 genome

##### Gene expression

Identified TARs were mapped to the current annotation of *S. pneumoniae* TIGR4 genome [24]. We found that each gene was represented by a mixed set of expressed and non expressed probes. Genes that had a significantly higher proportion of expressed probes in a binomial test [50] were considered to be expressed ( $p < .001$ , which results in at least 70% gene length coverage by TAR). This set of expressed genes represented the basal transcription of TIGR4. Functional analysis of the expressed genes was

conducted based on "TIGRFAMs" <http://www.tigr.org/TIGRFAMs/index.shtml>.

#### Operons

Because tiling arrays measure expression in the intergenic regions of annotated genomes, they can be used to identify and predict operon structures in bacteria. Two or more consecutive genes were considered to be part of an operon, if they fulfilled the following criteria: (a) they are expressed, (b) they are transcribed in same direction, and (c) the intergenic region between the genes was identified as a single expressed transcript that overlapped the genes in both directions. Overlapping pairs of genes are joined together to identify large operon structures.

#### sRNAs identification, genomic and structural analysis

To identify small RNAs, TARs were mapped to intergenic regions of the *S. pneumoniae* chromosome. Intergenic regions within operons, small 5' and 3' untranslated extensions (UTR) of mRNAs, and non-unique regions (mobile genetic elements and repetitive regions) of the genome were excluded. Only sRNAs that were identified at a minimum length of 74 bp (3 consecutive probes) were considered. Additional features for sRNAs such as promoters and transcription terminators were predicted computationally to add confidence in their identification. Bacterial promoter prediction was done using the "Neural Network Promoter Prediction" program [http://www.fruitfly.org/seq\\_tools/promoter.html](http://www.fruitfly.org/seq_tools/promoter.html) [66]. Putative sRNA sequences including 50 base pair upstream region were utilized for promoter prediction. Rho-independent transcription terminators were identified using program TransTermHP [67]. The putative sRNA sequence with 50 base pair downstream region is included for terminator prediction. UTR regions of length less than 100 bp are discarded. Variation in transcriptional intensity, presence of promoter and presence of rho-independent terminators are used as evidences to identify structural regulatory elements located inside the leader sequences. A circular *S. pneumoniae* TIGR4 genome map along with genes and sRNAs was generated using DNAPlotter [68]. All sRNA sequences were searched against Rfam database [31] for functional annotation. BLASTN searches were performed against non redundant nucleotide database at NCBI to determine sRNA sequence conservation among other genomes. MEME [69] was used for the identification of motifs in non-aligned sRNA sequences, where a motif is a sequence pattern that occurs repeatedly in a group of nucleotide sequences. Selected motifs were searched for their presence against the preexisting motif database using TOMTOM [34]. Sequence logos for predicted motifs were generated by WebLogo [70]. sRNA secondary structures were predicted using MFOLD [33]. The sRNAs, along with additional features, were mapped onto the TIGR4 genome in Genome Browser "GBrowse"

[71] <http://gbrowse.lsbj.mafes.msstate.edu/cgi-bin/gbrowse/TIGR4/> for visualization, analysis, and web based accessibility.

#### qReal-time PCR

Expressions of 13 sRNAs were validated by complementary quantitative Reverse Transcription - Polymerase Chain Reaction (qRT-PCR). PCR primers were designed (Additional file 11) using Primer3 [72] with at least one GC clamp on the 3' end. The same RNA used for tiling array labeling and hybridization was used as the template for qRT-PCR. All reverse transcription (RT) and subsequent PCR reactions were done in parallel and in triplicate. For each sRNA, three different RT reactions were set up. To measure possible expression of each complementary DNA strand, two strand-specific RT reactions were done; each reaction used only one strand-specific primer (forward or reverse). The third RT reaction was conducted in the absence of primers (to account for primer independent cDNA synthesis). After the RT step, both primers were added to all three reactions to complete the PCR step. RT-PCR was performed with 10 ng *S. pneumoniae* RNA using the Platinum<sup>®</sup> SYBR<sup>®</sup> Green One-Step qRT-PCR Kit (Invitrogen Corporation, Carlsbad, CA) as described [73]. Briefly, strand-specific RT reaction was conducted at 50°C for 10 min, 95°C for 5 min, and 0°C for 5 min. At this stage, the PCR primers were added to the reaction, and amplification and detection of specific PCR products was accomplished using the iCycler iQ Real-Time PCR Detection System (Bio-Rad Laboratories, Inc., Hercules, CA) with the following cycle profile: 95°C for 5 min, followed by 45 identical cycles at 95°C for 15 s and 60°C for 1 min. Melt curve analysis used 95°C for 1 min and 55°C for 1 min, followed by 80 cycles of 55°C for 10 s. The  $C_t$  (threshold cycle) values from all three RT-PCR reactions in triplicate were analyzed to detect sRNAs expression (Additional file 10).

#### Additional material

**Additional file 1 Determination of intensity threshold for probe expression.** Distribution of the intensities for positive and negative control probes was used to determine the threshold cutoff for probe level expression.

**Additional file 2 Genomic features of identified sRNAs.** *S. pneumoniae* TIGR4 sRNAs and their DNA sequences are shown with the transcription start sites (TSS, bold) predicted by "Neural Network Promoter Prediction". For sRNAs that were also identified in strain R6 (SN1, SN5, SN6, SN7 and SN35) the experimental (\*) TSS are shown.

**Additional file 3 Comparison of sRNAs from different studies.** Comparison of sRNAs identified in this study with previously described sRNAs (using computational and experimental approaches) in *S. pneumoniae*.

**Additional file 4 sRNA secondary structure prediction.** Predicted secondary structure of sRNAs using MFOLD. Motif regions are colored.

**Additional file 5 Gene expression profile.** *S. pneumoniae* TIGR4 genes identified as expressed in the present study, their associated TIGRFAM roles, sub roles and functions (where available).

**Additional file 6 List of co-expressed genes.** Pairs of co-expressed genes in *S. pneumoniae* TIGR4 identified by genomic tiling arrays.  
**Additional file 7 List of identified transcription units.** Transcription units identified by joining co-expressed gene pairs in *S. pneumoniae* TIGR4.  
**Additional file 8 GBrowse visualization of a transcription unit.** Genome browser visualization of genes SP2108 - SP2110. The tracks shown include translation in all six frames and tiling array expression. All three genes are present in the forward strand. The "tiling array expression" track clearly shows high level of expression for SP2108 compared to SP2109-SP2110.  
**Additional file 9 Comparison of co-expressed gene pairs.** Comparison of co-expressed gene pairs identified by tiling arrays with the results of computational operon prediction program "DOOR".  
**Additional file 10 qRT-PCR validations of sRNAs.** qRT-PCR validations of *S. pneumoniae* TIGR4 sRNAs carried out in triplicate.  
**Additional file 11 Primer sequences for qRT-PCR.** DNA sequences of primers used for qRT-PCR for validation for sRNAs in *S. pneumoniae* TIGR4.

#### Authors' contributions

RK designed the analysis workflow with BN, wrote all the scripts required for the analysis, carried out data analysis and wrote the initial draft of this manuscript. PS prepared the RNA samples for tiling array analysis and helped in data interpretation and initial draft preparation. ES, SB and ML and BN conceived and designed this collaborative study, helped with data analysis and interpretation. ES, SB and ML and BN helped draft the final version of the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

This project was partially supported by a grant from the National Science Foundation (Mississippi EPSCoR-0903787). We acknowledge Allen Shack and Dusan Kunec from College of Veterinary Medicine, Mississippi State University for technical help in conducting RT-PCR. We acknowledge Tony Arick and Life Sciences and Biotechnology Institute, Mississippi State University for hosting *S. pneumoniae* GBrowse. We acknowledge the Department of Basic Sciences, College of Veterinary Medicine, and the Life Sciences and Biotechnology Institute, Mississippi State University for assistance with the article-processing charges of this manuscript. Approved for publication as Journal Article by the Mississippi Agricultural and Forestry Experiment Station, Mississippi State University.

#### Author Details

<sup>1</sup>Department of Basic sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA, <sup>2</sup>Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA, <sup>3</sup>Mississippi Agriculture and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA, <sup>4</sup>MSU Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS 39762, USA, <sup>5</sup>Research Service (151), Veterans Affairs Medical Center, Jackson, MS 39216, USA, <sup>6</sup>Department of Microbiology, University of Mississippi Medical Center, Jackson, MS 39216 USA and <sup>7</sup>Departments of Molecular Biology and Microbiology and Molecular Genetics, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

Received: 10 November 2009 Accepted: 3 June 2010

Published: 3 June 2010

#### References

1. Livny J, Waldor MK: Identification of small RNAs in diverse bacterial species. *Curr Opin Microbiol* 2007, **10**(2):96-101.
2. Vanderpool CK, Gottesman S: Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Molecular microbiology* 2004, **54**(4):1076-1089.
3. Gorke B, Vogel J: Noncoding RNA control of the making and breaking of sugars. *Genes Dev* 2008, **22**(21):2914-2925.
4. Weillbacher T, Suzuki K, Dubey AK, Wang X, Gudapaty S, Morozov I, Baker CS, Georgellis D, Babitzke P, Romeo T: A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Molecular microbiology* 2003, **48**(3):657-670.
5. Vasil ML: How we learnt about iron acquisition in *Pseudomonas aeruginosa*: a series of very fortunate events. *Biometals* 2007, **20**(3-4):587-601.
6. Gottesman S: Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* 2005, **21**(7):399-404.
7. Geissmann T, Possedko M, Huntzinger E, Fechter P, Ehresmann C, Romby P: Regulatory RNAs as mediators of virulence gene expression in bacteria. *Handb Exp Pharmacol* 2006:9-43.
8. Bridy-Pappas AE, Margolis MB, Center KJ, Isaacman DJ: Streptococcus pneumoniae: description of the pathogen, disease epidemiology, treatment, and prevention. *Pharmacotherapy* 2005, **25**(9):1193-1212.
9. Schuchat A, Hilger T, Zell E, Farley MM, Reingold A, Harrison L, Lefkowitz L, Danila R, Stefonek K, Barrett N, et al.: Active bacterial core surveillance of the emerging infections program network. *Emerg Infect Dis* 2001, **7**(1):92-99.
10. Halfmann A, Kovacs M, Hakenbeck R, Bruckner R: Identification of the genes directly controlled by the response regulator CiaR in *Streptococcus pneumoniae*: five out of 15 promoters drive expression of small non-coding RNAs. *Molecular microbiology* 2007, **66**(1):110-126.
11. Tsui HC, Mukherjee D, Ray VA, Sham LT, Feig AL, Winkler ME: Identification and characterization of noncoding small RNAs in *Streptococcus pneumoniae* serotype 2 strain D39. *Journal of bacteriology* 2010, **192**(1):264-279.
12. Backofen R, Hess WR: Computational prediction of sRNAs and their targets in bacteria. *RNA biology* 2010, **7**(1):33-42.
13. Kulkarni RV, Kulkarni PR: Computational approaches for the discovery of bacterial small RNAs. *Methods* 2007, **43**(2):131-139.
14. Livny J, Brencic A, Lory S, Waldor MK: Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic acids research* 2006, **34**(12):3484-3493.
15. Tsui HC, Mukherjee D, Ray VA, Sham LT, Feig AL, Winkler ME: Identification and Characterization of Non-Coding Small RNAs in *Streptococcus pneumoniae* Serotype 2 Strain D39. *Journal of bacteriology* 2010, **192**(1):264-279.
16. Altuvia S: Identification of bacterial small non-coding RNAs: experimental approaches. *Curr Opin Microbiol* 2007, **10**(3):257-261.
17. Vogel J, Sharma CM: How to find small non-coding RNAs in bacteria. *Biol Chem* 2005, **386**(12):1219-1238.
18. Sorek R, Cossart P: Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature reviews* 2010, **11**(1):9-16.
19. Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C: Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research* 2002, **30**(17):3732-3738.
20. Akama T, Suzuki K, Tanigawa K, Kawashima A, Wu H, Nakata N, Osana Y, Sakakibara Y, Ishii N: Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. *Journal of bacteriology* 2009, **191**(10):3321-3327.
21. Landt SG, Abeliuk E, McGrath PT, Lesley JA, McAdams HH, Shapiro L: Small non-coding RNAs in *Caulobacter crescentus*. *Molecular microbiology* 2008, **68**(3):600-614.
22. Liu JM, Livny J, Lawrence MS, Kimball MD, Waldor MK, Camilli A: Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic acids research* 2009, **37**(6):e46.
23. Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J: Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 2008, **4**(8):e1000-163.
24. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, et al.: Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* 2001, **293**(5529):498-506.
25. Nanduri B, Shah P, Ramkumar M, Allen EB, Swiatlo E, Burgess SC, Lawrence ML: Quantitative analysis of *Streptococcus pneumoniae* TIGR4 response to in vitro iron restriction by 2-D LC ESI MS/MS. *Proteomics* 2008, **8**(10):2104-2114.

26. Royce TE, Carriero NJ, Gerstein MB: **An efficient pseudomedian filter for tiling microarrays.** *BMC Bioinformatics* 2007, **8**:186.
27. Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R, et al.: **A highly conserved repeated DNA element located in the chromosome of Streptococcus pneumoniae.** *Nucleic acids research* 1992, **20**(13):3479-3483.
28. Oggioni MR, Claverly JP: **Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae.** *Microbiology (Reading, England)* 1999, **145**(Pt 10):2647-2653.
29. Livny J, Teonadi H, Livny M, Waldor MK: **High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs.** *PLoS ONE* 2008, **3**(9):e3197.
30. Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T: **Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus Streptococcus.** *Int J Syst Bacteriol* 1995, **45**(2):406-408.
31. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic acids research* 2005:D121-124.
32. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic acids research* 2009:W202-208.
33. Mathews DH, Turner DH, Zuker M: **RNA secondary structure prediction.** *Curr Protoc Nucleic Acid Chem* 2007, **Chapter 11**:1-12.
34. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs.** *Genome Biol* 2007, **8**(2):R24.
35. Bonner ER, D'Elia JN, Billips BK, Switzer RL: **Molecular recognition of pyr mRNA by the Bacillus subtilis attenuation regulatory protein PyrR.** *Nucleic acids research* 2001, **29**(23):4851-4865.
36. Lu Y, Turner RJ, Switzer RL: **Function of RNA secondary structures in transcriptional attenuation of the Bacillus subtilis pyr operon.** *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(25):14462-14467.
37. Tomchick DR, Turner RJ, Switzer RL, Smith JL: **Adaptation of an enzyme to regulatory function: structure of Bacillus subtilis PyrR, a pyr RNA-binding attenuation protein and uracil phosphoribosyltransferase.** *Structure* 1998, **6**(3):337-350.
38. Winkler WC, Breaker RR: **Regulation of bacterial gene expression by riboswitches.** *Annual review of microbiology* 2005, **59**:487-517.
39. Winkler WC, Cohen-Chalalmish S, Breaker RR: **An mRNA structure that controls gene expression by binding FMN.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(25):15908-15913.
40. Vitreschak AG, Mironov AA, Lyubetsky VA, Gelfand MS: **Comparative genomic analysis of T-box regulatory systems in bacteria.** In *RNA Volume 14*. Issue 4 New York, NY; 2008:717-735.
41. Wels M, Groot Kormelink T, Kleerebezem M, Siezen RJ, Francke C: **An in silico analysis of T-box regulated genes and T-box evolution in prokaryotes, with emphasis on prediction of substrate specificity of transporters.** *BMC genomics* 2008, **9**:330.
42. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23**(9):324-328.
43. Wagner C, Saizieu Ad A, Schonfeld HJ, Kamber M, Lange R, Thompson CJ, Page MG: **Genetic analysis and functional characterization of the Streptococcus pneumoniae vic operon.** *Infection and immunity* 2002, **70**(11):6121-6128.
44. Nieto C, Puyet A, Espinosa M: **MalR-mediated regulation of the Streptococcus pneumoniae malMP operon at promoter PM. Influence of a proximal divergent promoter region and competition between MalR and RNA polymerase proteins.** *The Journal of biological chemistry* 2001, **276**(18):14946-14954.
45. Mascher T, Zahner D, Merai M, Balmelle N, de Saizieu AB, Hakenbeck R: **The Streptococcus pneumoniae cia regulon: CiaR target sites and transcription profile analysis.** *Journal of bacteriology* 2003, **185**(1):60-70.
46. Kuhnert WL, Zheng G, Faustoferri RC, Quivey RG Jr: **The F-ATPase operon promoter of Streptococcus mutans is transcriptionally regulated in response to external pH.** *Journal of bacteriology* 2004, **186**(24):8524-8528.
47. Mao F, Dam P, Chou J, Olman V, Xu Y: **DOOR: a database for prokaryotic operons.** *Nucleic acids research* 2009:D459-463.
48. Li X, Lindahl L, Sha Y, Zengel JM: **Analysis of the Bacillus subtilis S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase S10-spc-alpha cluster.** *Journal of bacteriology* 1997, **179**(22):7046-7054.
49. He H, Wang J, Liu T, Liu XS, Li T, Wang Y, Qian Z, Zheng H, Zhu X, Wu T, et al.: **Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray.** *Genome Res* 2007, **17**(10):1471-1477.
50. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(14):5320-5325.
51. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al.: **Empirical analysis of transcriptional activity in the Arabidopsis genome.** *Science* 2003, **302**(5646):842-846.
52. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al.: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome research* 2004, **14**(3):331-342.
53. Aiba H: **Mechanism of RNA silencing by Hfq-binding small RNAs.** *Curr Opin Microbiol* 2007, **10**(2):134-139.
54. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: **RNA expression analysis using a 30 base pair resolution Escherichia coli genome array.** *Nat Biotechnol* 2000, **18**(12):1262-1268.
55. Saito S, Kakeshita H, Nakamura K: **Novel small RNA-encoding genes in the intergenic regions of Bacillus subtilis.** *Gene* 2009, **428**(1-2):2-8.
56. Dam P, Olman V, Harris K, Su Z, Xu Y: **Operon prediction using both genome-specific and general genomic information.** *Nucleic acids research* 2007, **35**(1):288-298.
57. Brouwer RW, Kuipers OP, van Hijum SA: **The relative value of operon predictions.** *Brief Bioinform* 2008, **9**(5):367-375.
58. Fozo EM, Hemm MR, Storz G: **Small toxic proteins and the antisense RNAs that repress them.** *Microbiol Mol Biol Rev* 2008, **72**(4):579-589. Table of Contents.
59. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, et al.: **Gene expression analysis using oligonucleotide arrays produced by maskless photolithography.** *Genome research* 2002, **12**(11):1749-1755.
60. Wang X, He H, Li L, Chen R, Deng XW, Li S: **NMPP: a user-customized NimbleGen microarray data processing pipeline.** *Bioinformatics* 2006, **22**(23):2955-2957.
61. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**(2):185-193.
62. Ji H, Wong WH: **TileMap: create chromosomal map of tiling array hybridizations.** *Bioinformatics (Oxford, England)* 2005, **21**(18):3629-3636.
63. Halasz G, van Batenburg MF, Perusse J, Hua S, Lu XJ, White KP, Bussemaker HJ: **Detecting transcriptionally active regions using genomic tiling arrays.** *Genome biology* 2006, **7**(7):R59.
64. Zhang ZD, Rozowsky J, Lam HY, Du J, Snyder M, Gerstein M: **Telescope: online analysis pipeline for high-density tiling microarray data.** *Genome biology* 2007, **8**(5):R81.
65. Liu XS: **Getting started in tiling microarray analysis.** *PLoS computational biology* 2007, **3**(10):1842-1844.
66. Reese MG: **Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome.** *Comput Chem* 2001, **26**(1):51-56.
67. Kingsford CL, Ayanbule K, Salzberg SL: **Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.** *Genome Biol* 2007, **8**(2):R22.
68. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J: **DNAPlotter: circular and linear interactive genome visualization.** *Bioinformatics* 2009, **25**(1):119-120.
69. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
70. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14**(6):1188-1190.

71. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, *et al.*: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12**(10):1599-1610.
72. Rozen S, Skaletsky H: **Primer3 on the www for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
73. Kunec D, Nanduri B, Burgess SC: **Experimental annotation of channel catfish virus by probabilistic proteogenomic mapping.** *Proteomics* 2009, **9**(10):2634-2647.

doi: 10.1186/1471-2164-11-350

**Cite this article as:** Kumar *et al.*, Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays *BMC Genomics* 2010, **11**:350

Submit your next manuscript to BioMed Central  
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



CHAPTER V  
AUTOMATED PIPELINE FOR ADDING GENE ONTOLOGY  
ANNOTATION FOR NON MODEL SPECIES

**Abstract**

The Gene Ontology (GO) project provides a controlled vocabulary to facilitate high-quality functional gene annotation for many species. This enables the user to perform GO based functional analysis of high throughput microarray and proteomic datasets. Apart from gene or protein annotation, high throughput microarray experiments have large number of Expressed Sequence Tag (EST) sequences for which very limited GO is available. Furthermore, in spite of being a valuable resource for functional modeling, detailed species specific GO annotations are limited to few model species (<http://www.geneontology.org/GO.people.shtml>) which hinders functional genomics and hypothesis generation in non model species.

Here we describe an automated pipeline (the ISO-IEA pipeline) that can be used to generate GO for a diverse range of species. The pipeline transfers the available high quality, experimental based GO annotations of orthologous proteins from closely related species to the species being studied. The pipeline adds GO annotation in a two step process. First, for a given protein in a species of interest, we identify orthologs from

species that already have experimentally derived GO and use this to transfer the GO annotation to the original gene product. By relying on orthology rather than sequence alignment, we take advantage of conserved function from predicted orthologs. Second, in the absence of orthologous proteins in GOA (Gene Ontology Annotation project at EBI, UK) database, we use its sequence to search against InterPro database, identify functional motifs and assign GO based on it. The workflow inputs a set of UniProt/Ensembl/IPI/RefSeq identifiers and generates GO in a gene association file format that can be directly used in many GO based functional analysis tools.

The ISO - IEA pipeline provides rapid, automated and high quality sequence based GO annotation for any given species. The pipeline increases the GO coverage, while maintaining functional annotation quality for both EST and protein sequences, which facilitates modelling of high throughput data and generation of testable hypothesis.

## **Background**

Functional genomics has emerged as a major tool for genomic research but requires systems level modelling of biological functions. Functional analysis is an important step in data analysis and largely depends on the available functional annotation of the genome. GO (1) has now become the most widespread *de facto* standard for functional annotation. The GO is a directed acyclic graph (or DAG) consisting of defined terms and the relationships between them that describe three attributes of gene products: Molecular Function, Biological Process and Cellular Component (1). Annotation to the GO provides information about the gene product, its attributed function and the evidence

for associating it with the function (2). There are two broad types of GO evidence codes: direct experimental codes (the evidence codes used for biocuration of published literature) and indirect evidence codes. Indirect evidence codes include function prediction based on sequence known as “inferred from sequence orthology” (ISO) where functional conservation is inferred for predicted orthologs, and “inferred from electronic annotation” (IEA), which includes function predicted based on functional motifs and domains or keywords from curated databases such as SwissProt, etc (3). It should be noted that there are methods for providing IEA annotations other than based upon functional motifs and domains, but here we have used IEA because it works on raw DNA or protein sequence. Note that while the IEA annotations are not individually reviewed by biocurators, they are based on mapping files that are continually reviewed and updated (3).

A large number of tools are available that depend on GO for high throughput functional genomics data analysis (4). While the gold standard for providing GO annotation is expert biocuration (5) of experimental literature, this process is very slow (6) and is not available for broad range of species currently being investigated using functional genomics approaches. Therefore, we need automated GO annotation pipelines for providing GO rapidly, while maintaining the quality of annotations.

Automated GO annotation tools described in the literature (7), are mostly based on sequence similarity searches. However, transfer of function based on orthology (8) is the best way to provide GO annotation when there is no functional literature available for the gene product of interest. Orthologs or orthologous genes are genes in different



species that arose from a common ancestor and are assumed to be functionally equivalent. Therefore, functional annotation transfer based on orthologs is much more reliable compared to general BLAST based sequence similarity searches (9).

Here, we describe a new pipeline which performs annotation based on both sequence orthology (ISO evidence) and computational analysis of functional motifs (IEA evidence). The value of this ISO-IEA pipeline is to provide a platform for rapidly assigning breadth of GO coverage for the many species that do not have a focused effort to provide detailed literature biocuration. While the European Bioinformatics Institute (EBI) GOA Project provides IEA based GO annotation for all species represented in the UniProtKB database (3), many proteins are not found in the UniProtKB database (10) and microarrays for many non-model organisms are based upon ESTs sequences. While there are already tools that can attribute GO to gene products on the basis of BLAST searches (11,12), our method assigns GO first based upon orthology, or if this is not possible, based upon conserved functional motifs. Since orthologous genes emerge from a single ancestor, they are presumed to have conserved function and we believe this is a much more precise way to provide functional annotation than relying on BLAST searches. This pipeline is suitable for both EST or protein sequences and we demonstrate the utility of the pipeline by providing GO annotation for a dataset of computationally predicted proteins that have no experimental literature available and array that contains probes based upon EST sequences. We also quantitatively assess the GO annotation provided by both ISO and IEA.

GO annotations for a genome or an array can be provided using three different approaches. The first step is literature curation, where experimental results are used for functional annotation of known proteins. It provides highest quality of annotation, but is very slow, needs experimental evidence and involves biocurator expertise. The second step is annotation of proteins which are predicted to exist, but are not validated through experimental methods. This set of predicted proteins can only be annotated through sequence based features like orthology or sequence motifs. For these predicted proteins, the ISO-IEA pipeline can provide quick and automated high quality GO annotation. The third step encompasses annotation of the EST sequences for which no protein sequence information exists and thus they can only be annotated to IEA based on sequence searches. ISO-IEA pipeline that can annotate based on sequence orthology or protein functional domains helps provide high quality, broad level annotation for any species. Providing rapid GO annotation will help researchers derive value from their high throughput functional genomics datasets.

## **Results and Discussion**

The pipeline we developed, called ISO-IEA consists of two parts (Fig. 1). The ISO pipeline identifies orthologs in related species that already have experimentally derived GO available and use this to assign function to the original gene product. It is worthwhile noting that the ISO annotations attributed using this method are only as accurate as the orthology predictions; there are multiple resources that predict orthology (13) and these resources use different approaches to determine orthology. In this

manuscript we use Ensembl orthology 1:1 ortholog predictions (where 1:1 orthologs refers to strict ortholog pairs where only one copy of the gene is present in each species) but the ISO pipeline can also accept orthologs predicted from other databases or can be user defined. The IEA pipeline searches against the InterPro database (14), and assigns GO based on the identified functional motifs. Since ISO pipeline provides GO annotations based on experimental evidence codes, we expect it to provide more detailed annotations than the IEA pipeline. In contrast, IEA pipeline rapidly provides a large quantity of GO terms, i.e. more breadth or coverage.

### ***Testing of chicken proteins annotation using ISO method***

For testing the ISO pipeline, chicken genes for which experimental GO annotation is available are used. A total of 148 genes were downloaded from (GO annotation with evidence code IEA, ISS were removed) EBI GOA database. Ensembl ids were successfully mapped on 86 genes. These 86 genes had 215 GO terms defined in chicken genome and we call it experimental set. ISO pipeline is used to annotate these 86 genes based on human, mouse and rat orthologs. The pipeline produced a total of 844 GO terms for 86 genes from human (429), mouse (452) and rat (76) orthologs (Table 5.1).

Comparing both datasets we found that the ISO pipeline produced broad set of GO annotation for test genes. Average DAG depth per GO terms is found to be higher for ISO annotation (5.85) with respect to experimental chicken annotation (5.4). Looking into specific details we found that 84 GO terms out of 215 of experimental set were found in ISO annotation. Also, for half of the test genes (41 genes out of 86) at least one GO

term was common in both dataset. ISO annotations exactly matched the available experimental GO annotation for many genes, for ex. Q9IA88 (6 GO terms), P54519 (6 GO terms), P83038 (8 GO terms), P56517 (6 GO terms) representing the sensitivity of the results.

### ***Annotation of chicken predicted proteins using ISO – IEA pipeline***

The National Center for Biotechnology Information (NCBI) Genome annotation pipeline (<http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml>) combines *ab initio* predictions with sequence homology based upon RefSeq transcript alignments to produce “predicted” genes (and proteins) based upon sequence similarity with known genes from other species. These predicted gene products initially have no functional literature available to provide experimental based GO annotation but are likely to have recognizable orthologs in better GO annotated species. To test the ISO-IEA pipeline, 14,404 chicken predicted proteins were downloaded from GOA EBI website. EBI already had electronic annotation available for 6907 proteins with 21176 GO terms, which mostly contains annotation with evidence code “IEA”. All the proteins were ISO annotated using 1:1 orthologs from human, mouse and rat. Ensembl ids were mapped for 8338 proteins and their orthologs were identified using Ensembl 52. To maintain high quality of annotation, only gene products having experimental GO evidence codes in orthologous genomes were used for chicken protein annotation and assigned the evidence code “ISO”. As expected, since the phylogenetic distance from chicken for all three species is similar, they had had similar numbers of orthologs: human (6313), mouse (6265) and rat (5822)

(Fig. 2). It is notable that although 78% of the chicken dataset had 1:1 orthologs, very few of these orthologs had experimentally derived GO annotation, even in mouse and human, which are among the founding species of the GO Consortium (15). The largest numbers of GO annotations transferred to chicken proteins were from human and mouse. This is expected because the number of experimental based annotations is 79754 for mouse and 62431 for human, compared to 12980 for rat (as of 20<sup>th</sup> May 2009; based upon EBI GOA Project figures from <http://www.ebi.ac.uk/GOA/>). The ISO pipeline resulted in annotation of 4257 chicken predicted proteins with 24553 GO terms.

Next, 8,834 predicted proteins that could not be annotated to GO through ISO pipeline were submitted to the IEA pipeline. The IEA pipeline uses sequence features from 13 different databases associated with the InterPro database (14) to identify protein motifs and domains. These motifs and domains are linked to appropriate GO terms using an InterPro2GO mapping file provided by the EBI GOA Project and updated by GO Consortium biocurators on a monthly basis (<http://www.geneontology.org/external2go/interpro2go>). The IEA pipeline provided an additional 12,692 annotations for 4328 predicted proteins. Altogether, using ISO-IEA pipeline we were able to provide GO annotation for 60% of the chicken predicted proteins (8486) (Fig. 3). The ISO-IEA pipeline provides “no data” (ND) annotations for any products that do not have GO annotations. We do this because, when the input is ESTs or “predicted” proteins that do not have any literature available for manually biocuration, if no GO can be assigned using ISO or IEA it is standard practice to assign an ND code to indicate that there is simply no functional data available for these gene

products. This breadth for functional annotation will help researchers to model high throughput functional genomics data. Moreover, in chicken these predicted proteins had 21176 annotations for 6907 genes and the ISO-IEA pipeline increased the total number of chicken predicted GO annotations by 218%, i.e. from 21176 to 46282. This increase is reflected in all three categories of GO; biological process, cellular component and molecular function (Fig. 4).

#### ***Annotation of FHCRC chicken array using ISO – IEA pipeline***

ISO – IEA pipeline is also well suited for annotation of microarrays. Here we used the pipeline for annotation of FHCRC Chicken 13K Array (16), multi-tissue cDNA microarray with 13,007 features. Around 671 elements were removed as their entries were no longer available in NCBI. All probe elements were mapped to 11869 unique protein/EST ids. This list is further classified by accession type into three categories which can be annotated using different approaches (Fig. 5). In category “A” 32% of ids matched known manually verified proteins, for which some GO annotations were available in public databases and could be further annotated by literature curation. Category “B” had around 25% (3048) of the array ids that matched to predicted proteins that can only be annotated based on ISO and IEA pipeline. The ISO pipeline provided annotation for 1403 of these predicted proteins, and IEA added annotations to 636 additional proteins. Using ISO-IEA we were able to annotate 70% of the predicted proteins represented on this array. Category “C” represents the largest proportion, containing 43% sequences on the array that represent EST sequences for which no

gene/protein mapping is currently available. These ESTs can only be annotated using IEA pipeline. IEA analysis helped to add GO to 2799 of the EST sequences represented on this array (55%). In total, using ISO – IEA pipeline we were able to rapidly annotate 72% of the FHCRC chicken array.

### ***Comparison of ISO and IEA annotations***

Manually curated experimental GO annotations are of high quality and represent annotations to more specific terms, but are limited in number. Around 98% of the available GO annotations are IEA annotations ([http://www.ebi.ac.uk/GOA/uniprot\\_release.html](http://www.ebi.ac.uk/GOA/uniprot_release.html)) which are based on sequence features and represent very broad functions such as ‘protein binding’ and ‘enzyme binding’. However the structure of the Gene Ontology is based on a directed acyclic graph (DAG), 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 (child node) is an indicator of the level of functional detail captured in the annotation. The GO Annotation Quality (GAQ) Score measures DAG depth and other quantitative measures of GO annotation quality (17). Here, we compared the DAG depth and GAQ score between ISO and IEA annotation for 2886 chicken predicted proteins for which IEA annotations are available from EBI GOA database and these proteins were also annotated using ISO pipeline to assess the overall quality for each of these methods. In calculating the GAQ Scores, we assigned experimental evidence codes a rank of 3-5, the ISS evidence codes (ISS, ISO, ISA and ISM) a rank of 2-3 and the IEA evidence code a rank of 2 (Table 5.2).

We calculated the DAG depth and GAQ score for all annotations generated for chicken predicted proteins. We found that the average DAG depth and average GAQ score is higher for proteins having ISO annotation compared to IEA annotation (Fig. 6). These results are consistent with our hypothesis that ISO annotation provides more detailed GO annotations, and highlights the importance of literature biocuration to underpin the transfer of GO annotations from one species to another. In contrast, IEA annotation is rapid and provides greater breadth of GO annotations. Therefore, it is always preferable to use ISO pipeline prior to IEA, to get the best possible GO annotations available when there is no biocurated GO annotation.

## **Conclusions**

The ISO - IEA pipeline provides rapid, automated and high quality sequence based GO for any given species. The pipeline increases the GO coverage, while maintaining functional annotation quality for both EST and protein sequences. It generates the output as a standard gene association file format which can be directly used by various GO based data analysis tools, facilitating modelling of high throughput data and generation of testable biological hypotheses. Moreover we note that this method relies on the availability of both predicted orthologous genes and experimental based GO annotation for these orthologous genes.



## Methods

### *Implementation of ISO – IEA pipeline*

The ISO-IEA pipeline was written in perl language and requires a standard perl installation. It is tested on both windows and Linux platforms. All the Perl scripts along with a sample dataset and the help file [will](#) be made available from the AgBase tools website (<http://www.agbase.msstate.edu/ISOIEA.html>).

### *ISO annotation*

For ISO annotations, users are required to select the genome which is phylogenetically closer and have the GO annotations available. The quality and accuracy of ISO annotations will depend on how accurate we define the orthologs. User can use orthologs from databases like Inparanoid (18), Homologene (19), OrthoMCL (20) or Ensembl (21). If no orthologous information is available for a particular genome in these databases, the user can upload results of their own orthologous predictions using any of the available prediction tools. In order to facilitate communication among various databases, we determined equivalent accession from multiple databases (“ID mapping”) using Ensembl database. For example, we map input gene or protein accessions to their orthologous gene product accessions and then to UniProtKB accessions. This ID mapping system provides more flexibility for users who can upload their own id mapping file, when the pipeline is unable to map ids. For a given protein in a species of interest, once an ortholog is identified and have experimentally-derived GO available, we use this to

assign function to the original gene product. Latest GO annotations were downloaded from EBI (3). We do not transfer annotations which have evidence codes derived from sequence based predictions (for example IEA, ISS, ISO, ISM, ISA, ND) (22). Qualified-annotations which refer to annotations with a qualifier (e.g. “NOT” qualified annotations) are not included in the annotation transfer process. If an ortholog is not available for the protein sequence, the sequence is processed by IEA pipeline for sequence based annotation. The tool produces the output in Gene association file format which can be easily used with other data analysis programs. The produced annotations have the evidence code “ISO” and include the reference of the protein from which the annotation is derived.

#### ***IEA annotation: A wrapper for InterProScan based sequence analysis***

We have written a wrapper which formats and validates user provided EST/protein sequences and scans the InterPro database (14). The InterPro database is a collection of 13 different protein recognition methods combined into one unified resource. InterProScan installation is a prerequisite for the IEA pipeline. InterProScan can take EST/nucleotide sequences and translate them in all six frames for possible protein sequences and searches these protein sequences for functional motifs. IEA pipeline scans the user provided input protein/EST sequence for any formatting errors or duplicates, creates a corrected input file and starts the InterPro database searches. The InterPro output is analysed separately for protein and EST sequences, sorted and reformatted into a gene association file, useful for GO related analysis. IEA pipeline calls

the latest InterPro2GO mapping file to ensure that the IEA annotations provided are continually reviewed and updated. ISO and IEA pipeline are combined into an automated pipeline where sequences having no ISO annotation are inputs for IEA pipeline. The ISO and IEA pipelines can also be used separately if required.

### ***Annotation of chicken predicted proteins and cDNA chicken microarray using ISO – IEA pipeline***

To test the pipeline on a dataset which was likely to have orthologs available, we downloaded 14,404 predicted chicken gene products from NCBI. For another larger dataset containing a mixture of genes likely to have orthologs and EST sequences, we used the Fred Hutchinson Cancer Research Centre (FHCRC) 13K chicken cDNA array (16) which has 13007 distinct features. The complete array dataset was downloaded from the NCBI Gene Expression Omnibus (GEO) database. The data on this array was mapped to different public database accessions using the ArrayIDer tool (23). Both the dataset are processed with ISO and IEA pipeline for annotation. This allowed us to determine which of the elements represented on the array could be matched to genes while which were represented by ESTs that do not currently map to the chicken genome. Where we were able to identify a corresponding gene we used both ISO and then IEA while ESTs sequences were GO annotated using IEA.

Human, rat and mouse genomes were used for identifying orthologs to annotate the chicken datasets. Orthologous predictions were downloaded from Ensembl (21). User given ID, Ensembl orthologs ID and Uniprot ID were mapped using Biomart services

(24). The latest GOA files for human, rat and mouse were downloaded from EBI (3) and used to provide GO when a 1:1 orthologous gene in either human, mouse or rat was identified for the chicken gene product. This information was output as a gene association file containing the GO annotation information and a list of accessions that had either no 1:1 orthologs or no GO. The second file was used to obtain sequence in a fasta file and this data was input into the IEA pipeline. The chicken predicted protein annotation results are also used to compare the DAG depth and GAQ score (explained in results section).

### **Authors' contributions**

RK developed all components of the pipeline and did extensive testing of the pipeline. KJ developed the initial prototype of ISO pipeline that was modified by RK. RK drafted the manuscript. TJB assisted with development of the ISO pipeline and ISO annotation. FM and BN initiated the project and assisted with manuscript preparation. All authors read and approved the final manuscript.

### **Acknowledgements**

This project was partially supported by a grant from the National Science Foundation (EPS-0556308-06040293), the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number MISV-329140 and USDA Microbial Genomes Program. The authors acknowledge Bryce Magee and Prashanti Manda for their help with data processing for IEA pipeline.

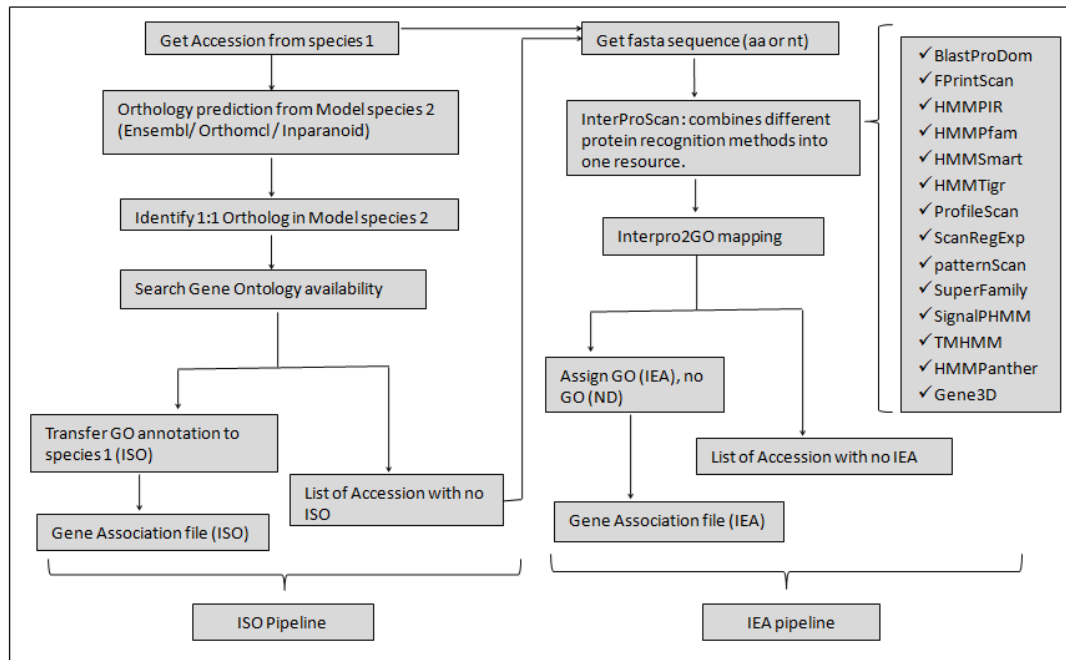


Figure 5.1 ISO-IEA pipeline. The strategy used to by the ISO - IEA pipeline is shown in a stepwise manner. Note that both the ISO and IEA sections of the pipeline can also be run independently.

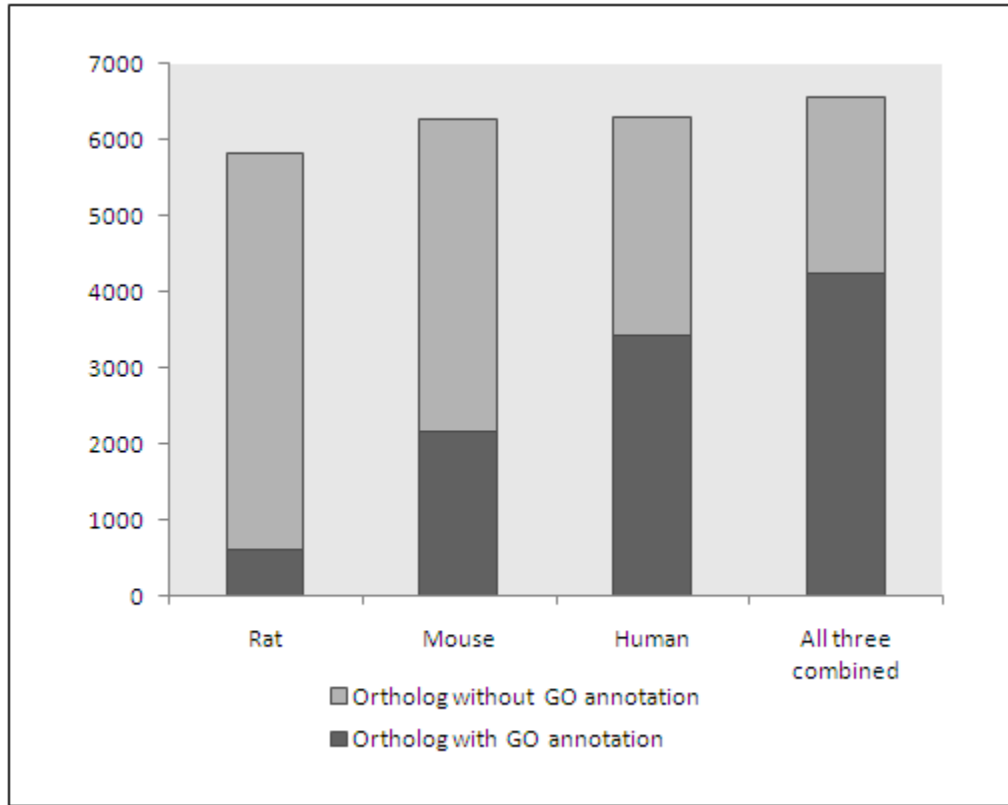


Figure 5.2 Orthologs distribution for chicken predicted proteins. 14,404 chicken predicted proteins from NCBI were downloaded and GO annotated using the ISO pipeline. The number of chicken proteins with a 1:1 ortholog in rat, mouse and human is shown with the number of genes having GO annotations. Fourth bar “all three combined” represent orthologs and GO annotation combined from all three resources human, mouse and rat.

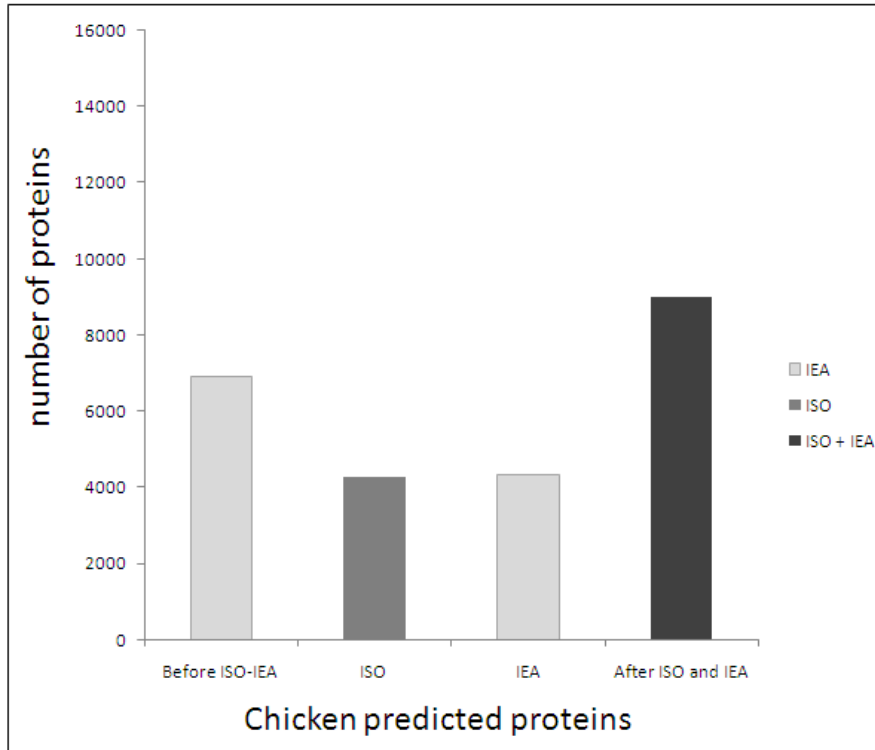


Figure 5.3 ISO – IEA annotation results for chicken predicted proteins. Chicken predicted proteins from the NCBI database were GO annotated using the ISO-IEA annotation pipeline and the results are shown. Before running the pipeline IEA annotations were available for 6907 proteins. Bars named “ISO” refers to GO annotations transferred from mammalian to chicken based upon 1:1 orthology while “IEA” refers to GO annotation based upon analysis of functional motifs form sequence. The fourth bar “After ISO and IEA” includes all the available information after running the pipeline. This method is found capable of providing GO annotation for 60% of the predicted proteins.

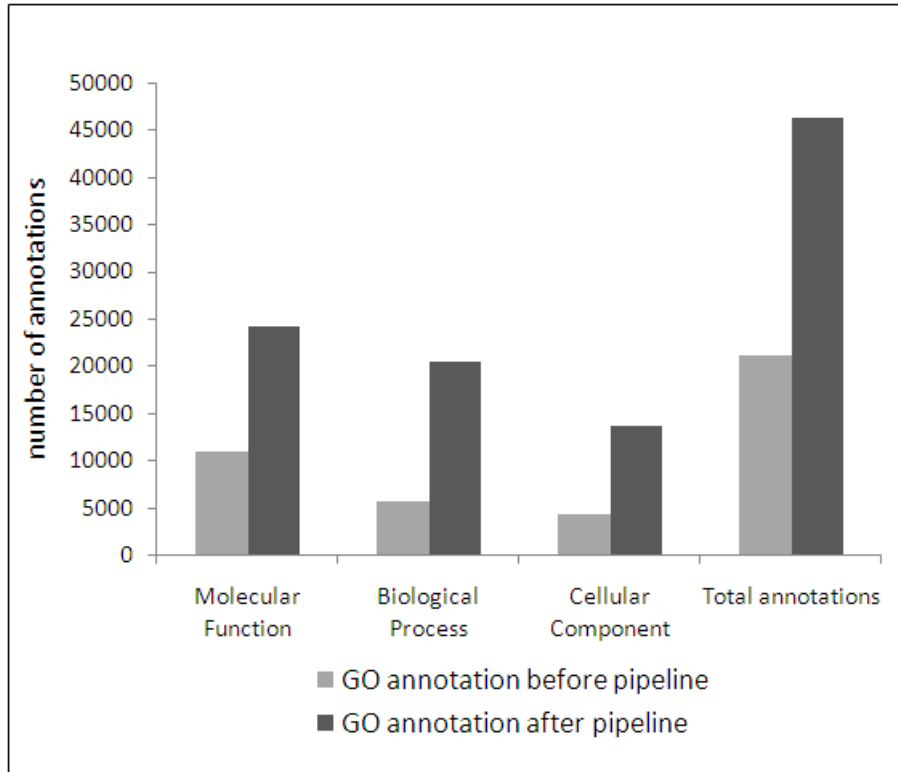


Figure 5.4 Overall improvement in GO annotation for the chicken proteome. The total GO annotations initially available for chicken predicted protein were 21176. Using the ISO-IEA pipeline to add GO annotation to chicken predicted proteins increased the overall GO annotation for chicken proteins to 46282. The increase in GO annotation is shown here for Molecular Function, Biological Process and Cellular Component as well as the total GO annotation.



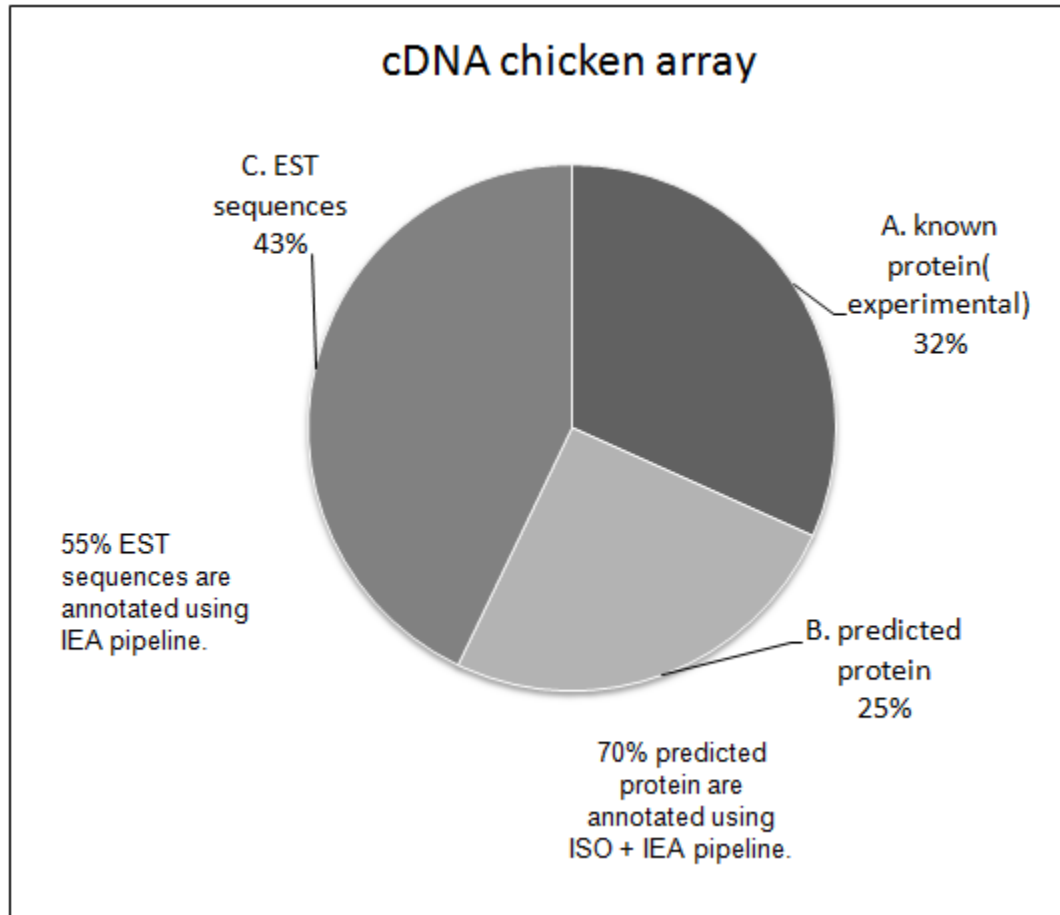


Figure 5.5 Distribution of probes for annotation in FHCRC chicken 13K array. To demonstrate the utility of the ISO-IEA pipeline for rapidly providing GO annotation, we GO annotated the gene products represented on the in FHCRC chicken 13K array. These gene products were divided into three groups for GO annotation by different approaches. Group A is a set of known proteins, group B is a set of predicted proteins and group C belongs to EST sequences for which no corresponding protein is found. For group A literature review can be used to provide more detailed GO annotation, while group B is annotated using ISO and IEA pipeline and group C is annotated using IEA pipeline, since there are no identifiable orthologs for these gene products.

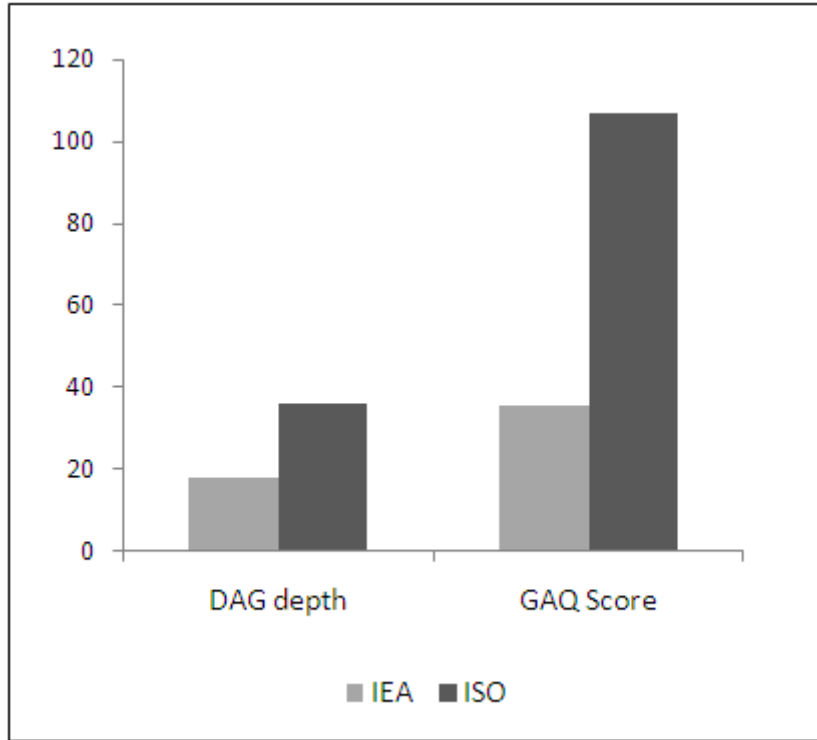


Figure 5.6 Comparison of ISO and IEA annotations. Since transferring GO annotations from orthologous genes that are already GO annotated relies on experimentally derived GO annotation, we expect that ISO annotations are likely to be more detailed than those obtained from IEA annotation. To test the quality of the GO annotations provided by these two methods we calculated the DAG depth and GAQ Score for GO annotations from the chicken predicted dataset.

Table 5.1

ISO annotation result for 86 test genes. Annotation colored in black are derived from human, red from mouse and green are from rat orthologs.

<b>Ensembl id</b>	<b>GO id</b>	<b>Evidence code</b>	<b>Attribute</b>	<b>Go Depth</b>
ENSGALP00000000170	GO:0005102	ISO	F	4
ENSGALP00000000170	GO:0019903	ISO	F	6
ENSGALP00000000365	GO:0005183	ISO	F	6
ENSGALP00000000365	GO:0005625	ISO	C	5
ENSGALP00000000365	GO:0007165	ISO	P	4
ENSGALP00000000365	GO:0007267	ISO	P	4
ENSGALP00000000365	GO:0007275	ISO	P	3
ENSGALP00000000365	GO:0008285	ISO	P	5
ENSGALP00000000678	GO:0000723	ISO	P	7
ENSGALP00000000678	GO:0000781	ISO	C	6
ENSGALP00000000678	GO:0001516	ISO	P	9
ENSGALP00000000678	GO:0003720	ISO	F	8
ENSGALP00000000678	GO:0005697	ISO	C	6
ENSGALP00000000678	GO:0007165	ISO	P	4
ENSGALP00000000678	GO:0050220	ISO	F	5
ENSGALP00000000678	GO:0051082	ISO	F	4
ENSGALP00000000680	GO:0001822	ISO	P	6
ENSGALP00000000680	GO:0001974	ISO	P	6
ENSGALP00000000680	GO:0002005	ISO	P	7
ENSGALP00000000680	GO:0002019	ISO	P	8
ENSGALP00000000680	GO:0003081	ISO	P	6
ENSGALP00000000680	GO:0003779	ISO	F	5
ENSGALP00000000680	GO:0005515	ISO	F	3
ENSGALP00000000680	GO:0005615	ISO	C	4
ENSGALP00000000680	GO:0005624	ISO	C	5
ENSGALP00000000680	GO:0005768	ISO	C	8
ENSGALP00000000680	GO:0005886	ISO	C	5
ENSGALP00000000680	GO:0008144	ISO	F	3
ENSGALP00000000680	GO:0008237	ISO	F	5
ENSGALP00000000680	GO:0008241	ISO	F	6
ENSGALP00000000680	GO:0009897	ISO	C	7
ENSGALP00000000680	GO:0019229	ISO	P	8
ENSGALP00000000680	GO:0031404	ISO	F	5

Table 5.1 (continued)

ENSGALP0000000680	GO:0031711	ISO	F	6
ENSGALP0000000680	GO:0032943	ISO	P	4
ENSGALP0000000680	GO:0042312	ISO	P	8
ENSGALP0000000680	GO:0042447	ISO	P	5
ENSGALP0000000680	GO:0043171	ISO	P	5
ENSGALP0000000680	GO:0050482	ISO	P	6
ENSGALP0000000680	GO:0060218	ISO	P	6
ENSGALP0000000926	GO:0005886	ISO	C	5
ENSGALP0000000926	GO:0005887	ISO	C	8
ENSGALP0000000926	GO:0007155	ISO	P	3
ENSGALP00000002808	GO:0000122	ISO	P	7
ENSGALP00000002808	GO:0003700	ISO	F	5
ENSGALP00000002808	GO:0003702	ISO	F	3
ENSGALP00000002808	GO:0003714	ISO	F	4
ENSGALP00000002808	GO:0005515	ISO	F	3
ENSGALP00000002808	GO:0005634	ISO	C	5
ENSGALP00000002808	GO:0008270	ISO	F	6
ENSGALP00000002808	GO:0045893	ISO	P	8
ENSGALP00000002861	GO:0005886	ISO	C	5
ENSGALP00000002861	GO:0005887	ISO	C	8
ENSGALP00000002861	GO:0006928	ISO	P	4
ENSGALP00000002861	GO:0016023	ISO	C	5
ENSGALP00000003019	GO:0000012	ISO	P	7
ENSGALP00000003019	GO:0000785	ISO	C	6
ENSGALP00000003019	GO:0003682	ISO	F	3
ENSGALP00000003019	GO:0003684	ISO	F	5
ENSGALP00000003019	GO:0003690	ISO	F	6
ENSGALP00000003019	GO:0003725	ISO	F	5
ENSGALP00000003019	GO:0005654	ISO	C	6
ENSGALP00000003019	GO:0005730	ISO	C	6
ENSGALP00000003019	GO:0006302	ISO	P	7
ENSGALP00000003019	GO:0006974	ISO	P	4
ENSGALP00000003019	GO:0008967	ISO	F	7
ENSGALP00000003019	GO:0031647	ISO	P	6
ENSGALP00000003019	GO:0033699	ISO	F	10
ENSGALP00000003019	GO:0042542	ISO	P	6
ENSGALP00000003019	GO:0046403	ISO	F	8
ENSGALP00000003019	GO:0047485	ISO	F	4
ENSGALP00000004141	GO:0000118	ISO	C	5

Table 5.1 (continued)

ENSGALP00000004141	GO:0004407	ISO	F	6
ENSGALP00000004141	GO:0005634	ISO	C	5
ENSGALP00000004141	GO:0005737	ISO	C	6
ENSGALP00000004141	GO:0006916	ISO	P	7
ENSGALP00000004141	GO:0008134	ISO	F	4
ENSGALP00000004141	GO:0016568	ISO	P	7
ENSGALP00000004174	GO:0005515	ISO	F	3
ENSGALP00000004174	GO:0005634	ISO	C	5
ENSGALP00000004174	GO:0005737	ISO	C	6
ENSGALP00000004174	GO:0005768	ISO	C	8
ENSGALP00000004174	GO:0005829	ISO	C	8
ENSGALP00000004174	GO:0006357	ISO	P	6
ENSGALP00000004174	GO:0007242	ISO	P	5
ENSGALP00000004174	GO:0030036	ISO	P	7
ENSGALP00000004174	GO:0042169	ISO	F	5
ENSGALP00000005166	GO:0003677	ISO	F	4
ENSGALP00000005166	GO:0003723	ISO	F	4
ENSGALP00000005166	GO:0005070	ISO	F	5
ENSGALP00000005166	GO:0005515	ISO	F	3
ENSGALP00000005166	GO:0005634	ISO	C	5
ENSGALP00000005166	GO:0006397	ISO	P	7
ENSGALP00000005166	GO:0007050	ISO	P	6
ENSGALP00000005166	GO:0007166	ISO	P	5
ENSGALP00000005166	GO:0008283	ISO	P	3
ENSGALP00000005166	GO:0016020	ISO	C	4
ENSGALP00000005212	GO:0000118	ISO	C	5
ENSGALP00000005212	GO:0003700	ISO	F	5
ENSGALP00000005212	GO:0004407	ISO	F	6
ENSGALP00000005212	GO:0005634	ISO	C	5
ENSGALP00000005212	GO:0005737	ISO	C	6
ENSGALP00000005212	GO:0006916	ISO	P	7
ENSGALP00000005212	GO:0008134	ISO	F	4
ENSGALP00000005212	GO:0016568	ISO	P	7
ENSGALP00000005212	GO:0019899	ISO	F	4
ENSGALP00000005212	GO:0042802	ISO	F	4
ENSGALP00000005255	GO:0003713	ISO	F	4
ENSGALP00000005436	GO:0000775	ISO	C	6
ENSGALP00000005436	GO:0005515	ISO	F	3
ENSGALP00000005436	GO:0007059	ISO	P	3

Table 5.1 (continued)

ENSGALP0000006645	GO:0003700	ISO	F	5
ENSGALP0000006811	GO:0000118	ISO	C	5
ENSGALP0000006811	GO:0004407	ISO	F	6
ENSGALP0000006811	GO:0005634	ISO	C	5
ENSGALP0000006811	GO:0005737	ISO	C	6
ENSGALP0000006811	GO:0006954	ISO	P	5
ENSGALP0000006811	GO:0007049	ISO	P	3
ENSGALP0000006811	GO:0007275	ISO	P	3
ENSGALP0000006811	GO:0007399	ISO	P	5
ENSGALP0000006811	GO:0008134	ISO	F	4
ENSGALP0000006811	GO:0016564	ISO	F	3
ENSGALP0000006811	GO:0016568	ISO	P	7
ENSGALP0000006811	GO:0030183	ISO	P	6
ENSGALP0000006811	GO:0045843	ISO	P	6
ENSGALP0000008131	GO:0006928	ISO	P	4
ENSGALP0000008131	GO:0007155	ISO	P	3
ENSGALP0000008131	GO:0030055	ISO	C	8
ENSGALP0000008131	GO:0030336	ISO	P	6
ENSGALP0000008131	GO:0043234	ISO	C	3
ENSGALP0000008131	GO:0043297	ISO	P	7
ENSGALP0000008131	GO:0045294	ISO	F	4
ENSGALP00000010292	GO:0007418	ISO	P	7
ENSGALP00000010951	GO:0003700	ISO	F	5
ENSGALP00000010951	GO:0005737	ISO	C	6
ENSGALP00000010951	GO:0006406	ISO	P	7
ENSGALP00000010951	GO:0007049	ISO	P	3
ENSGALP00000010951	GO:0008134	ISO	F	4
ENSGALP00000010951	GO:0008190	ISO	F	5
ENSGALP00000010951	GO:0010552	ISO	P	12
ENSGALP00000010951	GO:0010553	ISO	P	12
ENSGALP00000010951	GO:0016525	ISO	P	5
ENSGALP00000010951	GO:0030948	ISO	P	7
ENSGALP00000012176	GO:0000910	ISO	P	4
ENSGALP00000012176	GO:0005515	ISO	F	3
ENSGALP00000012176	GO:0005721	ISO	C	8
ENSGALP00000012176	GO:0005819	ISO	C	7
ENSGALP00000012176	GO:0007059	ISO	P	3
ENSGALP00000012176	GO:0043234	ISO	C	3
ENSGALP00000012728	GO:0003700	ISO	F	5

Table 5.1 (continued)

ENSGALP00000012728	GO:0005515	ISO	F	3
ENSGALP00000012728	GO:0043123	ISO	P	7
ENSGALP00000014143	GO:0005515	ISO	F	3
ENSGALP00000014143	GO:0005634	ISO	C	5
ENSGALP00000014143	GO:0005737	ISO	C	6
ENSGALP00000014143	GO:0005829	ISO	C	8
ENSGALP00000014143	GO:0006915	ISO	P	6
ENSGALP00000014143	GO:0007569	ISO	P	4
ENSGALP00000014143	GO:0016481	ISO	P	7
ENSGALP00000014143	GO:0045786	ISO	P	5
ENSGALP00000014363	GO:0003700	ISO	F	5
ENSGALP00000014363	GO:0005634	ISO	C	5
ENSGALP00000014363	GO:0006325	ISO	P	6
ENSGALP00000014363	GO:0006355	ISO	P	7
ENSGALP00000015070	GO:0000228	ISO	C	6
ENSGALP00000015070	GO:0003700	ISO	F	5
ENSGALP00000015070	GO:0006355	ISO	P	7
ENSGALP00000015070	GO:0007275	ISO	P	3
ENSGALP00000016266	GO:0005515	ISO	F	3
ENSGALP00000016343	GO:0005634	ISO	C	5
ENSGALP00000016343	GO:0005737	ISO	C	6
ENSGALP00000016343	GO:0005829	ISO	C	8
ENSGALP00000016343	GO:0006915	ISO	P	6
ENSGALP00000016343	GO:0007253	ISO	P	13
ENSGALP00000016343	GO:0008139	ISO	F	5
ENSGALP00000016343	GO:0010552	ISO	P	12
ENSGALP00000016343	GO:0010745	ISO	P	
ENSGALP00000016343	GO:0031625	ISO	F	5
ENSGALP00000016343	GO:0032270	ISO	P	6
ENSGALP00000016343	GO:0032376	ISO	P	9
ENSGALP00000016343	GO:0033256	ISO	C	6
ENSGALP00000016343	GO:0042345	ISO	P	12
ENSGALP00000016343	GO:0042802	ISO	F	4
ENSGALP00000016343	GO:0043392	ISO	P	6
ENSGALP00000016343	GO:0045833	ISO	P	5
ENSGALP00000016343	GO:0051059	ISO	F	5
ENSGALP00000016475	GO:0046658	ISO	C	8
ENSGALP00000016502	GO:0003700	ISO	F	5
ENSGALP00000016502	GO:0007530	ISO	P	4

Table 5.1 (continued)

ENSGALP00000016502	GO:0008584	ISO	P	7
ENSGALP00000017170	GO:0004672	ISO	F	6
ENSGALP00000017170	GO:0005524	ISO	F	7
ENSGALP00000017370	GO:0004888	ISO	F	5
ENSGALP00000017370	GO:0005515	ISO	F	3
ENSGALP00000017370	GO:0005886	ISO	C	5
ENSGALP00000017370	GO:0005901	ISO	C	7
ENSGALP00000017370	GO:0006508	ISO	P	6
ENSGALP00000017370	GO:0006629	ISO	P	4
ENSGALP00000017370	GO:0006897	ISO	P	6
ENSGALP00000017370	GO:0007165	ISO	P	4
ENSGALP00000017370	GO:0016021	ISO	C	7
ENSGALP00000017370	GO:0019221	ISO	P	6
ENSGALP00000017370	GO:0030229	ISO	F	6
ENSGALP00000017370	GO:0034187	ISO	F	5
ENSGALP00000018493	GO:0005113	ISO	F	5
ENSGALP00000018493	GO:0005576	ISO	C	2
ENSGALP00000018493	GO:0007267	ISO	P	4
ENSGALP00000018493	GO:0015485	ISO	F	6
ENSGALP00000019120	GO:0005634	ISO	C	5
ENSGALP00000019120	GO:0005737	ISO	C	6
ENSGALP00000019120	GO:0006986	ISO	P	5
ENSGALP00000019120	GO:0007140	ISO	P	7
ENSGALP00000019120	GO:0007286	ISO	P	6
ENSGALP00000019120	GO:0009986	ISO	C	4
ENSGALP00000019120	GO:0051082	ISO	F	4
ENSGALP00000019451	GO:0007165	ISO	P	4
ENSGALP00000019477	GO:0003702	ISO	F	3
ENSGALP00000019477	GO:0006357	ISO	P	6
ENSGALP00000020036	GO:0006366	ISO	P	5
ENSGALP00000020036	GO:0007567	ISO	P	3
ENSGALP00000020572	GO:0004872	ISO	F	4
ENSGALP00000020572	GO:0005887	ISO	C	8
ENSGALP00000020572	GO:0007165	ISO	P	4
ENSGALP00000020572	GO:0008283	ISO	P	3
ENSGALP00000020720	GO:0005509	ISO	F	5
ENSGALP00000021000	GO:0005515	ISO	F	3
ENSGALP00000021000	GO:0006417	ISO	P	6
ENSGALP00000021000	GO:0016049	ISO	P	5



Table 5.1 (continued)

ENSGALP00000021000	GO:0016202	ISO	P	8
ENSGALP00000023309	GO:0005178	ISO	F	5
ENSGALP00000023309	GO:0005730	ISO	C	6
ENSGALP00000023309	GO:0005856	ISO	C	5
ENSGALP00000023309	GO:0005884	ISO	C	7
ENSGALP00000023309	GO:0005925	ISO	C	10
ENSGALP00000023309	GO:0007155	ISO	P	3
ENSGALP00000023309	GO:0008307	ISO	F	3
ENSGALP00000023309	GO:0030035	ISO	P	8
ENSGALP00000023309	GO:0030175	ISO	C	5
ENSGALP00000023309	GO:0031143	ISO	C	5
ENSGALP00000023309	GO:0031432	ISO	F	5
ENSGALP00000023309	GO:0042802	ISO	F	4
ENSGALP00000023309	GO:0042981	ISO	P	7
ENSGALP00000023309	GO:0043197	ISO	C	7
ENSGALP00000023309	GO:0046983	ISO	F	4
ENSGALP00000023309	GO:0048041	ISO	P	6
ENSGALP00000023309	GO:0051289	ISO	P	8
ENSGALP00000023309	GO:0051370	ISO	F	5
ENSGALP00000023309	GO:0051374	ISO	F	6
ENSGALP00000023309	GO:0070080	ISO	F	
ENSGALP00000023335	GO:0003712	ISO	F	5
ENSGALP00000023335	GO:0005634	ISO	C	5
ENSGALP00000023626	GO:0001725	ISO	C	7
ENSGALP00000023626	GO:0005515	ISO	F	3
ENSGALP00000023626	GO:0005886	ISO	C	5
ENSGALP00000023626	GO:0005887	ISO	C	8
ENSGALP00000023626	GO:0005913	ISO	C	9
ENSGALP00000023626	GO:0005925	ISO	C	10
ENSGALP00000023626	GO:0007155	ISO	P	3
ENSGALP00000023626	GO:0007165	ISO	P	4
ENSGALP00000023626	GO:0007267	ISO	P	4
ENSGALP00000023838	GO:0000122	ISO	P	7
ENSGALP00000023838	GO:0003700	ISO	F	5
ENSGALP00000023838	GO:0003714	ISO	F	4
ENSGALP00000023838	GO:0005515	ISO	F	3
ENSGALP00000023838	GO:0007275	ISO	P	3
ENSGALP00000023851	GO:0000070	ISO	P	7
ENSGALP00000023851	GO:0000775	ISO	C	6

Table 5.1 (continued)

ENSGALP00000023851	GO:0005515	ISO	F	3
ENSGALP00000023851	GO:0005634	ISO	C	5
ENSGALP00000023851	GO:0007049	ISO	P	3
ENSGALP00000023851	GO:0007051	ISO	P	8
ENSGALP00000023851	GO:0048015	ISO	P	7
ENSGALP00000023972	GO:0003682	ISO	F	3
ENSGALP00000023972	GO:0005634	ISO	C	5
ENSGALP00000023972	GO:0006281	ISO	P	6
ENSGALP00000023972	GO:0006334	ISO	P	6
ENSGALP00000023972	GO:0006345	ISO	P	9
ENSGALP00000023972	GO:0016585	ISO	C	8
ENSGALP00000023972	GO:0042393	ISO	F	4
ENSGALP00000024133	GO:0000118	ISO	C	5
ENSGALP00000024133	GO:0003700	ISO	F	5
ENSGALP00000024133	GO:0004407	ISO	F	6
ENSGALP00000024133	GO:0005634	ISO	C	5
ENSGALP00000024133	GO:0005737	ISO	C	6
ENSGALP00000024133	GO:0006355	ISO	P	7
ENSGALP00000024133	GO:0008134	ISO	F	4
ENSGALP00000024133	GO:0016568	ISO	P	7
ENSGALP00000024133	GO:0019899	ISO	F	4
ENSGALP00000025107	GO:0005634	ISO	C	5
ENSGALP00000025107	GO:0016564	ISO	F	3
ENSGALP00000025107	GO:0045892	ISO	P	8
ENSGALP00000026069	GO:0000287	ISO	F	5
ENSGALP00000026069	GO:0004674	ISO	F	7
ENSGALP00000026069	GO:0005515	ISO	F	3
ENSGALP00000026069	GO:0005524	ISO	F	7
ENSGALP00000026069	GO:0005634	ISO	C	5
ENSGALP00000026069	GO:0005737	ISO	C	6
ENSGALP00000026069	GO:0006468	ISO	P	8
ENSGALP00000026069	GO:0007243	ISO	P	6
ENSGALP00000026160	GO:0000050	ISO	P	6
ENSGALP00000026160	GO:0004585	ISO	F	6
ENSGALP00000026160	GO:0005759	ISO	C	6
ENSGALP00000026160	GO:0006526	ISO	P	8
ENSGALP00000026160	GO:0009348	ISO	C	8
ENSGALP00000026395	GO:0005509	ISO	F	5
ENSGALP00000026395	GO:0005515	ISO	F	3

Table 5.1 (continued)

ENSGALP00000027736	GO:0003707	ISO	F	6
ENSGALP00000027736	GO:0005102	ISO	F	4
ENSGALP00000027736	GO:0007165	ISO	P	4
ENSGALP00000027736	GO:0007267	ISO	P	4
ENSGALP00000027736	GO:0047485	ISO	F	4
ENSGALP00000031021	GO:0001726	ISO	C	5
ENSGALP00000031021	GO:0005515	ISO	F	3
ENSGALP00000031021	GO:0005516	ISO	F	4
ENSGALP00000031021	GO:0005634	ISO	C	5
ENSGALP00000031021	GO:0005654	ISO	C	6
ENSGALP00000031021	GO:0005737	ISO	C	6
ENSGALP00000031021	GO:0005794	ISO	C	8
ENSGALP00000031021	GO:0006897	ISO	P	6
ENSGALP00000031021	GO:0016461	ISO	C	5
ENSGALP00000031021	GO:0016591	ISO	C	8
ENSGALP00000031021	GO:0030048	ISO	P	7
ENSGALP00000031021	GO:0030330	ISO	P	7
ENSGALP00000031021	GO:0031941	ISO	C	7
ENSGALP00000031021	GO:0031965	ISO	C	6
ENSGALP00000031021	GO:0045944	ISO	P	7
ENSGALP00000031021	GO:0048471	ISO	C	8
ENSGALP00000031021	GO:0051015	ISO	F	6
ENSGALP00000031021	GO:0051046	ISO	P	4
ENSGALP00000031021	GO:0060001	ISO	F	4
ENSGALP00000032226	GO:0005515	ISO	F	3
ENSGALP00000032226	GO:0005634	ISO	C	5
ENSGALP00000032226	GO:0043515	ISO	F	3
ENSGALP00000032226	GO:0051383	ISO	P	6
ENSGALP00000034039	GO:0005041	ISO	F	6
ENSGALP00000034039	GO:0005624	ISO	C	5
ENSGALP00000034039	GO:0005886	ISO	C	5
ENSGALP00000034039	GO:0007165	ISO	P	4
ENSGALP00000034039	GO:0007399	ISO	P	5
ENSGALP00000034039	GO:0007613	ISO	P	5
ENSGALP00000034039	GO:0030229	ISO	F	6
ENSGALP00000034039	GO:0034447	ISO	P	
ENSGALP00000035264	GO:0005515	ISO	F	3
ENSGALP00000035264	GO:0045944	ISO	P	7
ENSGALP00000035339	GO:0001525	ISO	P	6

Table 5.1 (continued)

ENSGALP00000035339	GO:0005215	ISO	F	2
ENSGALP00000035339	GO:0005515	ISO	F	3
ENSGALP00000035339	GO:0005739	ISO	C	8
ENSGALP00000035339	GO:0005753	ISO	C	7
ENSGALP00000035339	GO:0005754	ISO	C	9
ENSGALP00000035339	GO:0005886	ISO	C	5
ENSGALP00000035339	GO:0006091	ISO	P	3
ENSGALP00000035339	GO:0009986	ISO	C	4
ENSGALP00000035339	GO:0015992	ISO	P	8
ENSGALP00000035339	GO:0042288	ISO	F	6
ENSGALP00000035339	GO:0042645	ISO	C	7
ENSGALP00000035339	GO:0043499	ISO	F	4
ENSGALP00000035339	GO:0046961	ISO	F	10
ENSGALP00000035339	GO:0051453	ISO	P	9
ENSGALP00000036053	GO:0005887	ISO	C	8
ENSGALP00000036053	GO:0005922	ISO	C	10
ENSGALP00000036053	GO:0006810	ISO	P	4
ENSGALP00000036053	GO:0007601	ISO	P	7
ENSGALP00000036053	GO:0015267	ISO	F	5
ENSGALP00000036647	GO:0000082	ISO	P	6
ENSGALP00000036647	GO:0001501	ISO	P	6
ENSGALP00000036647	GO:0001541	ISO	P	8
ENSGALP00000036647	GO:0004871	ISO	F	3
ENSGALP00000036647	GO:0005125	ISO	F	5
ENSGALP00000036647	GO:0005179	ISO	F	5
ENSGALP00000036647	GO:0005576	ISO	C	2
ENSGALP00000036647	GO:0006917	ISO	P	7
ENSGALP00000036647	GO:0006952	ISO	P	3
ENSGALP00000036647	GO:0007050	ISO	P	6
ENSGALP00000036647	GO:0007166	ISO	P	5
ENSGALP00000036647	GO:0007267	ISO	P	4
ENSGALP00000036647	GO:0007399	ISO	P	5
ENSGALP00000036647	GO:0008083	ISO	F	5
ENSGALP00000036647	GO:0009605	ISO	P	3
ENSGALP00000036647	GO:0030154	ISO	P	4
ENSGALP00000036647	GO:0030308	ISO	P	6
ENSGALP00000036647	GO:0032925	ISO	P	9
ENSGALP00000036647	GO:0042326	ISO	P	8
ENSGALP00000036647	GO:0042541	ISO	P	7

Table 5.1 (continued)

ENSGALP00000036647	GO:0042802	ISO	F	4
ENSGALP00000036647	GO:0043509	ISO	C	6
ENSGALP00000036647	GO:0043512	ISO	C	6
ENSGALP00000036647	GO:0045077	ISO	P	8
ENSGALP00000036647	GO:0045578	ISO	P	8
ENSGALP00000036647	GO:0045648	ISO	P	9
ENSGALP00000036647	GO:0045650	ISO	P	10
ENSGALP00000036647	GO:0045786	ISO	P	5
ENSGALP00000036647	GO:0045944	ISO	P	7
ENSGALP00000036647	GO:0046881	ISO	P	9
ENSGALP00000036647	GO:0046882	ISO	P	9
ENSGALP00000036647	GO:0048184	ISO	F	4
ENSGALP00000037503	GO:0000922	ISO	C	7
ENSGALP00000037503	GO:0005509	ISO	F	5
ENSGALP00000037503	GO:0005737	ISO	C	6
ENSGALP00000037503	GO:0005813	ISO	C	9
ENSGALP00000037503	GO:0005829	ISO	C	8
ENSGALP00000037503	GO:0005876	ISO	C	7
ENSGALP00000037503	GO:0005886	ISO	C	5
ENSGALP00000037503	GO:0007186	ISO	P	6
ENSGALP00000037503	GO:0019904	ISO	F	4
ENSGALP00000037503	GO:0031432	ISO	F	5
ENSGALP00000037503	GO:0031997	ISO	F	5
ENSGALP00000037503	GO:0032465	ISO	P	5
ENSGALP00000037503	GO:0051592	ISO	P	6
ENSGALP00000039165	GO:0003700	ISO	F	5
ENSGALP00000039165	GO:0003714	ISO	F	4
ENSGALP00000039165	GO:0005515	ISO	F	3
ENSGALP00000039165	GO:0005634	ISO	C	5
ENSGALP00000039165	GO:0006357	ISO	P	6
ENSGALP00000039165	GO:0009653	ISO	P	4
ENSGALP00000039447	GO:0005737	ISO	C	6
ENSGALP00000039447	GO:0005856	ISO	C	5
ENSGALP00000039447	GO:0006446	ISO	P	5
ENSGALP00000039447	GO:0006916	ISO	P	7
ENSGALP00000039447	GO:0006928	ISO	P	4
ENSGALP00000039447	GO:0006986	ISO	P	5
ENSGALP00000039447	GO:0009986	ISO	C	4
ENSGALP00000039447	GO:0042802	ISO	F	4

Table 5.1 (continued)

ENSGALP0000000170	GO:0005737	ISO	C	6
ENSGALP0000000170	GO:0043025	ISO	C	4
ENSGALP0000000365	GO:0010468	ISO	P	4
ENSGALP0000000365	GO:0045471	ISO	P	5
ENSGALP0000000365	GO:0048545	ISO	P	5
ENSGALP0000000680	GO:0008217	ISO	P	6
ENSGALP0000000926	GO:0001764	ISO	P	6
ENSGALP0000000926	GO:0009986	ISO	C	4
ENSGALP0000000926	GO:0030424	ISO	C	6
ENSGALP0000001715	GO:0005515	ISO	F	3
ENSGALP0000001715	GO:0005667	ISO	C	8
ENSGALP0000001715	GO:0043565	ISO	F	5
ENSGALP0000001715	GO:0045944	ISO	P	7
ENSGALP0000002808	GO:0006306	ISO	P	8
ENSGALP0000002808	GO:0006349	ISO	P	5
ENSGALP0000002808	GO:0009048	ISO	P	6
ENSGALP0000002808	GO:0010216	ISO	P	6
ENSGALP0000002861	GO:0005737	ISO	C	6
ENSGALP0000002861	GO:0007010	ISO	P	5
ENSGALP0000002861	GO:0007194	ISO	P	7
ENSGALP0000002861	GO:0008104	ISO	P	4
ENSGALP0000002861	GO:0008360	ISO	P	7
ENSGALP0000002861	GO:0030818	ISO	P	10
ENSGALP0000002861	GO:0031750	ISO	F	7
ENSGALP00000004141	GO:0003677	ISO	F	4
ENSGALP00000004141	GO:0007346	ISO	P	5
ENSGALP00000004174	GO:0005070	ISO	F	5
ENSGALP00000004174	GO:0045309	ISO	F	5
ENSGALP00000005166	GO:0016481	ISO	P	7
ENSGALP00000005166	GO:0016564	ISO	F	3
ENSGALP00000005166	GO:0046831	ISO	P	9
ENSGALP00000005255	GO:0002053	ISO	P	6
ENSGALP00000005255	GO:0005634	ISO	C	5
ENSGALP00000005255	GO:0030326	ISO	P	7
ENSGALP00000005255	GO:0042472	ISO	P	7
ENSGALP00000005255	GO:0042474	ISO	P	7
ENSGALP00000005255	GO:0045880	ISO	P	6
ENSGALP00000005255	GO:0048701	ISO	P	7
ENSGALP00000005255	GO:0048844	ISO	P	6

Table 5.1 (continued)

ENSGALP0000005255	GO:0051216	ISO	P	7
ENSGALP0000005255	GO:0060021	ISO	P	4
ENSGALP0000006117	GO:0016477	ISO	P	4
ENSGALP0000006117	GO:0018108	ISO	P	9
ENSGALP0000006117	GO:0030900	ISO	P	8
ENSGALP0000006645	GO:0005515	ISO	F	3
ENSGALP0000006645	GO:0005634	ISO	C	5
ENSGALP0000006645	GO:0006355	ISO	P	7
ENSGALP0000006811	GO:0000122	ISO	P	7
ENSGALP0000006811	GO:0001501	ISO	P	6
ENSGALP0000006811	GO:0003677	ISO	F	4
ENSGALP0000006811	GO:0005515	ISO	F	3
ENSGALP0000006811	GO:0008285	ISO	P	5
ENSGALP0000008131	GO:0005515	ISO	F	3
ENSGALP0000008131	GO:0005886	ISO	C	5
ENSGALP0000008131	GO:0005911	ISO	C	8
ENSGALP0000008131	GO:0005912	ISO	C	8
ENSGALP0000008131	GO:0005916	ISO	C	10
ENSGALP0000008131	GO:0005925	ISO	C	10
ENSGALP0000008131	GO:0030032	ISO	P	8
ENSGALP0000008131	GO:0030334	ISO	P	5
ENSGALP0000008131	GO:0043034	ISO	C	10
ENSGALP00000010292	GO:0001569	ISO	P	7
ENSGALP00000010292	GO:0001570	ISO	P	6
ENSGALP00000010292	GO:0001656	ISO	P	7
ENSGALP00000010292	GO:0001658	ISO	P	7
ENSGALP00000010292	GO:0001708	ISO	P	4
ENSGALP00000010292	GO:0001755	ISO	P	6
ENSGALP00000010292	GO:0001947	ISO	P	6
ENSGALP00000010292	GO:0002052	ISO	P	6
ENSGALP00000010292	GO:0002053	ISO	P	6
ENSGALP00000010292	GO:0002076	ISO	P	6
ENSGALP00000010292	GO:0005113	ISO	F	5
ENSGALP00000010292	GO:0005515	ISO	F	3
ENSGALP00000010292	GO:0005615	ISO	C	4
ENSGALP00000010292	GO:0007165	ISO	P	4
ENSGALP00000010292	GO:0007228	ISO	P	7
ENSGALP00000010292	GO:0007267	ISO	P	4
ENSGALP00000010292	GO:0007368	ISO	P	7

Table 5.1 (continued)

ENSGALP00000010292	GO:0007389	ISO	P	4
ENSGALP00000010292	GO:0007411	ISO	P	9
ENSGALP00000010292	GO:0007435	ISO	P	8
ENSGALP00000010292	GO:0007442	ISO	P	6
ENSGALP00000010292	GO:0007507	ISO	P	6
ENSGALP00000010292	GO:0007596	ISO	P	6
ENSGALP00000010292	GO:0008209	ISO	P	6
ENSGALP00000010292	GO:0009952	ISO	P	6
ENSGALP00000010292	GO:0009986	ISO	C	4
ENSGALP00000010292	GO:0014003	ISO	P	8
ENSGALP00000010292	GO:0021513	ISO	P	6
ENSGALP00000010292	GO:0021904	ISO	P	6
ENSGALP00000010292	GO:0021938	ISO	P	7
ENSGALP00000010292	GO:0021940	ISO	P	6
ENSGALP00000010292	GO:0021978	ISO	P	6
ENSGALP00000010292	GO:0030162	ISO	P	7
ENSGALP00000010292	GO:0030336	ISO	P	6
ENSGALP00000010292	GO:0030539	ISO	P	7
ENSGALP00000010292	GO:0030850	ISO	P	7
ENSGALP00000010292	GO:0030900	ISO	P	8
ENSGALP00000010292	GO:0030901	ISO	P	8
ENSGALP00000010292	GO:0030902	ISO	P	8
ENSGALP00000010292	GO:0031016	ISO	P	6
ENSGALP00000010292	GO:0031069	ISO	P	8
ENSGALP00000010292	GO:0032435	ISO	P	8
ENSGALP00000010292	GO:0042130	ISO	P	8
ENSGALP00000010292	GO:0042307	ISO	P	11
ENSGALP00000010292	GO:0042475	ISO	P	6
ENSGALP00000010292	GO:0042733	ISO	P	8
ENSGALP00000010292	GO:0043010	ISO	P	8
ENSGALP00000010292	GO:0043237	ISO	F	5
ENSGALP00000010292	GO:0045121	ISO	C	7
ENSGALP00000010292	GO:0045445	ISO	P	6
ENSGALP00000010292	GO:0045449	ISO	P	6
ENSGALP00000010292	GO:0045596	ISO	P	5
ENSGALP00000010292	GO:0045944	ISO	P	7
ENSGALP00000010292	GO:0046639	ISO	P	9
ENSGALP00000010292	GO:0048546	ISO	P	5
ENSGALP00000010292	GO:0048568	ISO	P	6



Table 5.1 (continued)

ENSGALP00000010292	GO:0048589	ISO	P	3
ENSGALP00000010292	GO:0048598	ISO	P	5
ENSGALP00000010292	GO:0048663	ISO	P	7
ENSGALP00000010292	GO:0048706	ISO	P	7
ENSGALP00000010292	GO:0048714	ISO	P	9
ENSGALP00000010292	GO:0048859	ISO	P	5
ENSGALP00000010292	GO:0051146	ISO	P	8
ENSGALP00000010292	GO:0060020	ISO	P	6
ENSGALP00000010292	GO:0060438	ISO	P	
ENSGALP00000010292	GO:0060441	ISO	P	
ENSGALP00000010292	GO:0060458	ISO	P	
ENSGALP00000010951	GO:0001889	ISO	P	6
ENSGALP00000010951	GO:0005634	ISO	C	5
ENSGALP00000010951	GO:0009952	ISO	P	6
ENSGALP00000010951	GO:0016564	ISO	F	3
ENSGALP00000010951	GO:0030177	ISO	P	7
ENSGALP00000010951	GO:0030878	ISO	P	7
ENSGALP00000010951	GO:0030900	ISO	P	8
ENSGALP00000010951	GO:0035050	ISO	P	7
ENSGALP00000010951	GO:0042127	ISO	P	4
ENSGALP00000011799	GO:0001701	ISO	P	7
ENSGALP00000011799	GO:0030534	ISO	P	4
ENSGALP00000011799	GO:0042127	ISO	P	4
ENSGALP00000011799	GO:0045604	ISO	P	6
ENSGALP00000012176	GO:0000801	ISO	C	8
ENSGALP00000012176	GO:0030496	ISO	C	4
ENSGALP00000012728	GO:0001816	ISO	P	3
ENSGALP00000012728	GO:0005634	ISO	C	5
ENSGALP00000012728	GO:0045084	ISO	P	8
ENSGALP00000012728	GO:0045893	ISO	P	8
ENSGALP00000013605	GO:0004981	ISO	F	9
ENSGALP00000013605	GO:0005624	ISO	C	5
ENSGALP00000013605	GO:0007197	ISO	P	10
ENSGALP00000014363	GO:0000122	ISO	P	7
ENSGALP00000014363	GO:0001708	ISO	P	4
ENSGALP00000014363	GO:0002052	ISO	P	6
ENSGALP00000014363	GO:0003682	ISO	F	3
ENSGALP00000014363	GO:0005515	ISO	F	3
ENSGALP00000014363	GO:0005667	ISO	C	8

Table 5.1 (continued)

ENSGALP00000014363	GO:0005737	ISO	C	6
ENSGALP00000014363	GO:0019827	ISO	P	6
ENSGALP00000014363	GO:0021879	ISO	P	7
ENSGALP00000014363	GO:0021984	ISO	P	7
ENSGALP00000014363	GO:0021987	ISO	P	6
ENSGALP00000014363	GO:0030178	ISO	P	7
ENSGALP00000014363	GO:0030539	ISO	P	7
ENSGALP00000014363	GO:0030910	ISO	P	8
ENSGALP00000014363	GO:0032526	ISO	P	7
ENSGALP00000014363	GO:0042221	ISO	P	3
ENSGALP00000014363	GO:0042472	ISO	P	7
ENSGALP00000014363	GO:0043565	ISO	F	5
ENSGALP00000014363	GO:0043586	ISO	P	7
ENSGALP00000014363	GO:0045665	ISO	P	8
ENSGALP00000014363	GO:0045666	ISO	P	8
ENSGALP00000014363	GO:0045668	ISO	P	7
ENSGALP00000014363	GO:0045747	ISO	P	6
ENSGALP00000014363	GO:0045944	ISO	P	7
ENSGALP00000014363	GO:0046148	ISO	P	5
ENSGALP00000014363	GO:0048568	ISO	P	6
ENSGALP00000014363	GO:0048663	ISO	P	7
ENSGALP00000014363	GO:0048852	ISO	P	5
ENSGALP00000014363	GO:0050910	ISO	P	8
ENSGALP00000014363	GO:0050973	ISO	P	7
ENSGALP00000014363	GO:0060042	ISO	P	7
ENSGALP00000014363	GO:0060235	ISO	P	11
ENSGALP00000015070	GO:0001501	ISO	P	6
ENSGALP00000015070	GO:0005515	ISO	F	3
ENSGALP00000015070	GO:0005667	ISO	C	8
ENSGALP00000015070	GO:0007389	ISO	P	4
ENSGALP00000015070	GO:0042733	ISO	P	8
ENSGALP00000015910	GO:0001709	ISO	P	4
ENSGALP00000015910	GO:0001945	ISO	P	7
ENSGALP00000015910	GO:0002088	ISO	P	9
ENSGALP00000015910	GO:0003705	ISO	F	4
ENSGALP00000015910	GO:0005634	ISO	C	5
ENSGALP00000015910	GO:0005737	ISO	C	6
ENSGALP00000015910	GO:0008285	ISO	P	5
ENSGALP00000015910	GO:0030240	ISO	P	9

Table 5.1 (continued)

ENSGALP00000015910	GO:0045446	ISO	P	5
ENSGALP00000015910	GO:0046619	ISO	P	7
ENSGALP00000015910	GO:0048845	ISO	P	6
ENSGALP00000015910	GO:0055005	ISO	P	7
ENSGALP00000015910	GO:0055009	ISO	P	7
ENSGALP00000015910	GO:0055010	ISO	P	7
ENSGALP00000015910	GO:0060214	ISO	P	7
ENSGALP00000015910	GO:0060298	ISO	P	
ENSGALP00000015910	GO:0060412	ISO	P	9
ENSGALP00000015910	GO:0060414	ISO	P	
ENSGALP00000015910	GO:0060421	ISO	P	
ENSGALP00000015910	GO:0070309	ISO	P	
ENSGALP00000016343	GO:0000060	ISO	P	10
ENSGALP00000016343	GO:0031663	ISO	P	6
ENSGALP00000016343	GO:0032495	ISO	P	6
ENSGALP00000016343	GO:0032496	ISO	P	5
ENSGALP00000016343	GO:0034142	ISO	P	8
ENSGALP00000016343	GO:0042127	ISO	P	4
ENSGALP00000016343	GO:0043330	ISO	P	5
ENSGALP00000016343	GO:0045638	ISO	P	8
ENSGALP00000016343	GO:0045746	ISO	P	6
ENSGALP00000016343	GO:0070427	ISO	P	
ENSGALP00000016343	GO:0070431	ISO	P	
ENSGALP00000017370	GO:0005041	ISO	F	6
ENSGALP00000017370	GO:0005615	ISO	C	4
ENSGALP00000017370	GO:0021766	ISO	P	6
ENSGALP00000017370	GO:0021819	ISO	P	10
ENSGALP00000017370	GO:0045860	ISO	P	8
ENSGALP00000018798	GO:0007612	ISO	P	5
ENSGALP00000018798	GO:0008021	ISO	C	8
ENSGALP00000018798	GO:0048488	ISO	P	7
ENSGALP00000019120	GO:0005515	ISO	F	3
ENSGALP00000019120	GO:0005739	ISO	C	8
ENSGALP00000019477	GO:0003690	ISO	F	6
ENSGALP00000019477	GO:0005634	ISO	C	5
ENSGALP00000019477	GO:0016563	ISO	F	3
ENSGALP00000019477	GO:0045944	ISO	P	7
ENSGALP00000019505	GO:0007283	ISO	P	6
ENSGALP00000020036	GO:0001701	ISO	P	7

Table 5.1 (continued)

ENSGALP00000020036	GO:0045604	ISO	P	6
ENSGALP00000020572	GO:0001841	ISO	P	8
ENSGALP00000020572	GO:0005113	ISO	F	5
ENSGALP00000020572	GO:0005515	ISO	F	3
ENSGALP00000020572	GO:0005576	ISO	C	2
ENSGALP00000020572	GO:0008158	ISO	F	6
ENSGALP00000020572	GO:0008201	ISO	F	6
ENSGALP00000020572	GO:0008270	ISO	F	6
ENSGALP00000020572	GO:0008544	ISO	P	6
ENSGALP00000020572	GO:0008589	ISO	P	7
ENSGALP00000020572	GO:0009887	ISO	P	6
ENSGALP00000020572	GO:0009953	ISO	P	6
ENSGALP00000020572	GO:0015485	ISO	F	6
ENSGALP00000020572	GO:0016485	ISO	P	8
ENSGALP00000020572	GO:0030326	ISO	P	7
ENSGALP00000020572	GO:0030879	ISO	P	7
ENSGALP00000020572	GO:0040015	ISO	P	5
ENSGALP00000020572	GO:0042593	ISO	P	7
ENSGALP00000020572	GO:0043616	ISO	P	4
ENSGALP00000020572	GO:0050680	ISO	P	6
ENSGALP00000020720	GO:0030424	ISO	C	6
ENSGALP00000020720	GO:0030425	ISO	C	6
ENSGALP00000020720	GO:0050806	ISO	P	7
ENSGALP00000022313	GO:0005515	ISO	F	3
ENSGALP00000022313	GO:0005634	ISO	C	5
ENSGALP00000022313	GO:0031558	ISO	P	9
ENSGALP00000022313	GO:0046686	ISO	P	6
ENSGALP00000023309	GO:0005865	ISO	C	11
ENSGALP00000023309	GO:0006936	ISO	P	5
ENSGALP00000023309	GO:0030018	ISO	C	12
ENSGALP00000023309	GO:0030375	ISO	F	6
ENSGALP00000023309	GO:0030674	ISO	F	4
ENSGALP00000023309	GO:0042803	ISO	F	5
ENSGALP00000023309	GO:0051015	ISO	F	6
ENSGALP00000023838	GO:0001843	ISO	P	7
ENSGALP00000023838	GO:0003682	ISO	F	3
ENSGALP00000023838	GO:0007368	ISO	P	7
ENSGALP00000023838	GO:0008285	ISO	P	5
ENSGALP00000023838	GO:0009953	ISO	P	6

Table 5.1 (continued)

ENSGALP00000023838	GO:0016566	ISO	F	4
ENSGALP00000023838	GO:0045666	ISO	P	8
ENSGALP00000023838	GO:0048146	ISO	P	6
ENSGALP00000023838	GO:0048387	ISO	P	7
ENSGALP00000023838	GO:0060041	ISO	P	9
ENSGALP00000024133	GO:0000792	ISO	C	7
ENSGALP00000024133	GO:0005657	ISO	C	6
ENSGALP00000024133	GO:0016575	ISO	P	9
ENSGALP00000024294	GO:0005886	ISO	C	5
ENSGALP00000024294	GO:0010001	ISO	P	6
ENSGALP00000024294	GO:0016198	ISO	P	6
ENSGALP00000024294	GO:0030424	ISO	C	6
ENSGALP00000024294	GO:0045165	ISO	P	5
ENSGALP00000026069	GO:0007346	ISO	P	5
ENSGALP00000026069	GO:0045595	ISO	P	5
ENSGALP00000026160	GO:0005739	ISO	C	8
ENSGALP00000026160	GO:0005743	ISO	C	6
ENSGALP00000026395	GO:0005634	ISO	C	5
ENSGALP00000026395	GO:0005737	ISO	C	6
ENSGALP00000026395	GO:0005829	ISO	C	8
ENSGALP00000026395	GO:0007626	ISO	P	4
ENSGALP00000027736	GO:0001542	ISO	P	8
ENSGALP00000027736	GO:0002070	ISO	P	5
ENSGALP00000027736	GO:0005634	ISO	C	5
ENSGALP00000027736	GO:0006355	ISO	P	7
ENSGALP00000027736	GO:0030879	ISO	P	7
ENSGALP00000027736	GO:0050678	ISO	P	5
ENSGALP00000027736	GO:0050847	ISO	P	8
ENSGALP00000034039	GO:0005515	ISO	F	3
ENSGALP00000034039	GO:0005615	ISO	C	4
ENSGALP00000034039	GO:0045860	ISO	P	8
ENSGALP00000035264	GO:0000122	ISO	P	7
ENSGALP00000035264	GO:0001829	ISO	P	6
ENSGALP00000035264	GO:0001892	ISO	P	8
ENSGALP00000035264	GO:0003700	ISO	F	5
ENSGALP00000035264	GO:0005634	ISO	C	5
ENSGALP00000035264	GO:0005667	ISO	C	8
ENSGALP00000035264	GO:0006350	ISO	P	5
ENSGALP00000035339	GO:0005743	ISO	C	6

Table 5.1 (continued)

ENSGALP00000035339	GO:0006629	ISO	P	4
ENSGALP00000035339	GO:0006933	ISO	P	6
ENSGALP00000036053	GO:0002088	ISO	P	9
ENSGALP00000037355	GO:0001890	ISO	P	6
ENSGALP00000037355	GO:0005622	ISO	C	4
ENSGALP00000037355	GO:0005737	ISO	C	6
ENSGALP00000037355	GO:0005739	ISO	C	8
ENSGALP00000037503	GO:0007049	ISO	P	3
ENSGALP00000037503	GO:0043388	ISO	P	6
ENSGALP00000038276	GO:0000146	ISO	F	3
ENSGALP00000038276	GO:0001750	ISO	C	7
ENSGALP00000038276	GO:0005509	ISO	F	5
ENSGALP00000038276	GO:0005516	ISO	F	4
ENSGALP00000038276	GO:0005737	ISO	C	6
ENSGALP00000038276	GO:0005794	ISO	C	8
ENSGALP00000038276	GO:0005882	ISO	C	7
ENSGALP00000038276	GO:0006887	ISO	P	6
ENSGALP00000038276	GO:0007268	ISO	P	6
ENSGALP00000038276	GO:0007601	ISO	P	7
ENSGALP00000038276	GO:0016459	ISO	C	7
ENSGALP00000038276	GO:0030050	ISO	P	8
ENSGALP00000038276	GO:0030073	ISO	P	8
ENSGALP00000038276	GO:0030141	ISO	C	6
ENSGALP00000038276	GO:0030318	ISO	P	6
ENSGALP00000038276	GO:0031585	ISO	P	5
ENSGALP00000038276	GO:0031987	ISO	P	5
ENSGALP00000038276	GO:0032252	ISO	P	7
ENSGALP00000038276	GO:0032400	ISO	P	5
ENSGALP00000038276	GO:0032402	ISO	P	8
ENSGALP00000038276	GO:0042438	ISO	P	6
ENSGALP00000038276	GO:0042470	ISO	C	7
ENSGALP00000038276	GO:0042476	ISO	P	5
ENSGALP00000038276	GO:0042552	ISO	P	5
ENSGALP00000038276	GO:0042640	ISO	P	6
ENSGALP00000038276	GO:0042641	ISO	C	7
ENSGALP00000038276	GO:0042759	ISO	P	7
ENSGALP00000038276	GO:0043025	ISO	C	4
ENSGALP00000038276	GO:0048066	ISO	P	3
ENSGALP00000038276	GO:0050808	ISO	P	5

Table 5.1 (continued)

ENSGALP00000038276	GO:0051010	ISO	F	7
ENSGALP00000038276	GO:0051643	ISO	P	6
ENSGALP00000039165	GO:0000790	ISO	C	7
ENSGALP00000039165	GO:0007369	ISO	P	5
ENSGALP00000039447	GO:0005622	ISO	C	4
ENSGALP00000039447	GO:0005625	ISO	C	5
ENSGALP00000039447	GO:0005626	ISO	C	5
ENSGALP00000039447	GO:0005634	ISO	C	5
ENSGALP00000039447	GO:0005886	ISO	C	5
ENSGALP00000039447	GO:0030018	ISO	C	12
ENSGALP00000039447	GO:0043292	ISO	C	8
ENSGALP00000000170	GO:0005080	ISO	F	7
ENSGALP00000000170	GO:0007205	ISO	P	10
ENSGALP00000000365	GO:0005615	ISO	C	4
ENSGALP00000000365	GO:0006916	ISO	P	7
ENSGALP00000000365	GO:0007565	ISO	P	3
ENSGALP00000000365	GO:0030728	ISO	P	6
ENSGALP00000000365	GO:0031960	ISO	P	6
ENSGALP00000000365	GO:0042698	ISO	P	5
ENSGALP00000000680	GO:0014910	ISO	P	7
ENSGALP00000000926	GO:0007160	ISO	P	5
ENSGALP00000000926	GO:0007409	ISO	P	8
ENSGALP00000001715	GO:0006916	ISO	P	7
ENSGALP00000001715	GO:0016563	ISO	F	3
ENSGALP00000004141	GO:0000785	ISO	C	6
ENSGALP00000004141	GO:0042493	ISO	P	4
ENSGALP00000013605	GO:0005085	ISO	F	4
ENSGALP00000013605	GO:0005515	ISO	F	3
ENSGALP00000013605	GO:0014069	ISO	C	6
ENSGALP00000013605	GO:0032279	ISO	C	3
ENSGALP00000013605	GO:0043025	ISO	C	4
ENSGALP00000013605	GO:0043679	ISO	C	6
ENSGALP00000018798	GO:0008022	ISO	F	4
ENSGALP00000018798	GO:0043087	ISO	P	6
ENSGALP00000018798	GO:0046982	ISO	F	5
ENSGALP00000019451	GO:0001662	ISO	P	5
ENSGALP00000019451	GO:0001836	ISO	P	8
ENSGALP00000019451	GO:0005615	ISO	C	4
ENSGALP00000019451	GO:0006919	ISO	P	8

Table 5.1 (continued)

ENSGALP00000019451	GO:0007205	ISO	P	10
ENSGALP00000019451	GO:0030425	ISO	C	6
ENSGALP00000019451	GO:0032099	ISO	P	9
ENSGALP00000019451	GO:0032461	ISO	P	7
ENSGALP00000019451	GO:0043065	ISO	P	6
ENSGALP00000019451	GO:0043194	ISO	C	7
ENSGALP00000019451	GO:0043195	ISO	C	7
ENSGALP00000019451	GO:0043203	ISO	C	5
ENSGALP00000019451	GO:0043204	ISO	C	5
ENSGALP00000019451	GO:0050731	ISO	P	9
ENSGALP00000019451	GO:0051901	ISO	P	6
ENSGALP00000019451	GO:0051930	ISO	P	7
ENSGALP00000020720	GO:0005515	ISO	F	3
ENSGALP00000020720	GO:0005829	ISO	C	8
ENSGALP00000020720	GO:0008427	ISO	F	6
ENSGALP00000020720	GO:0045921	ISO	P	7
ENSGALP00000020720	GO:0048015	ISO	P	7
ENSGALP00000021000	GO:0006584	ISO	P	6
ENSGALP00000021000	GO:0021707	ISO	P	6
ENSGALP00000021000	GO:0030424	ISO	C	6
ENSGALP00000021000	GO:0043403	ISO	P	6
ENSGALP00000021000	GO:0051146	ISO	P	8
ENSGALP00000022313	GO:0044445	ISO	C	9
ENSGALP00000026069	GO:0000122	ISO	P	7
ENSGALP00000026069	GO:0005829	ISO	C	8
ENSGALP00000026069	GO:0016564	ISO	F	3
ENSGALP00000026395	GO:0007611	ISO	P	4
ENSGALP00000026395	GO:0042359	ISO	P	6
ENSGALP00000026395	GO:0048167	ISO	P	5
ENSGALP00000032013	GO:0006942	ISO	P	7
ENSGALP00000032013	GO:0060048	ISO	P	7
ENSGALP00000034039	GO:0001666	ISO	P	4
ENSGALP00000034039	GO:0005634	ISO	C	5
ENSGALP00000034039	GO:0007166	ISO	P	5
ENSGALP00000034039	GO:0007584	ISO	P	6
ENSGALP00000034039	GO:0009725	ISO	P	4
ENSGALP00000034039	GO:0030296	ISO	F	6
ENSGALP00000034039	GO:0032496	ISO	P	5
ENSGALP00000034039	GO:0032869	ISO	P	7



Table 5.1 (continued)

ENSGALP00000035264	GO:0003690	ISO	F	6
ENSGALP00000036647	GO:0046982	ISO	F	5
ENSGALP00000039447	GO:0000502	ISO	C	6
ENSGALP00000039447	GO:0043130	ISO	F	5

Table 5.2

Ranking for GO evidence code used for calculating GAQ score

Evidence code	Rank
IDA	5
IGI	5
IMP	5
IPI	5
IC	4
TAS	4
IEP	3
ISS	2
RCA	3
IGC	3
IEA	2
NAS	2
NR	1
ND	0
ISO	3
EXP	5
ISA	2
ISM	2

## REFERENCES CITED

1. 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.
2. Hill, D.P., Smith, B., McAndrews-Hill, M.S. and Blake, J.A. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9 Suppl 5, S2.
3. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. and Apweiler, R. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, 37, D396-403.
4. Lee, S.A., Chan, C.H., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y. and Huang, C.Y. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC bioinformatics*, 9 Suppl 12, S11.
5. Jones, C.E., Brown, A.L. and Baumann, U. (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8, 170.
6. Baumgartner, W.A., Jr., Cohen, K.B., Fox, L.M., Acquaaah-Mensah, G. and Hunter, L. (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23, i41-48.
7. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37, D619-622.
8. Buza, T.J., McCarthy, F.M. and Burgess, S.C. (2007) Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC genomics*, 8, 425.
9. Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39, 309-338.
10. McCarthy, F.M., Bridges, S.M., Wang, N., Magee, G.B., Williams, W.P., Luthe, D.S. and Burgess, S.C. (2007) AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res*, 35, D599-603.

11. Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, 21, 3674-3676.
12. McCarthy, F.M., Wang, N., Magee, G.B., Nanduri, B., Lawrence, M.L., Camon, E.B., Barrell, D.G., Hill, D.P., Dolan, M.E., Williams, W.P. *et al.* (2006) AgBase: a functional genomics resource for agriculture. *BMC Genomics*, 7, 229.
13. Altenhoff, A.M. and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology*, 5, e1000262.
14. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*, 37, D211-215.
15. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25, 1251-1255.
16. Burnside, J., Neiman, P., Tang, J., Basom, R., Talbot, R., Aronszajn, M., Burt, D. and Delrow, J. (2005) Development of a cDNA array for chicken gene expression analysis. *BMC Genomics*, 6, 13.
17. Buza, T.J., McCarthy, F.M., Wang, N., Bridges, S.M. and Burgess, S.C. (2008) Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic acids research*, 36, e12.
18. O'Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33, D476-480.
19. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35, D5-12.
20. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34, D363-368.

21. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res*, 35, D610-617.
22. McGarvey, P.B., Huang, H., Mazumder, R., Zhang, J., Chen, Y., Zhang, C., Cammer, S., Will, R., Odle, M., Sobral, B. *et al.* (2009) Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets. *PloS one*, 4, e7162.
23. van den Berg, B.H., Konieczka, J.H., McCarthy, F.M. and Burgess, S.C. (2009) ArrayIDER: automated structural re-annotation pipeline for DNA microarrays. *BMC Bioinformatics*, 10, 30.
24. Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart--biological queries made easy. *BMC Genomics*, 10, 22.

CHAPTER VI  
HPIDB--A UNIFIED RESOURCE FOR HOST-PATHOGEN  
INTERACTIONS<sup>1</sup>

---

<sup>1</sup> Reprint from Kumar R, Nanduri B. BMC Bioinformatics. 2010. HPIDB--a unified resource for host-pathogen interactions. This article is available from: <http://www.ncbi.nlm.nih.gov/pubmed/20946599>

PROCEEDINGS

Open Access

## HPIDB - a unified resource for host-pathogen interactions

Ranjit Kumar<sup>1,2\*</sup>, Bindu Nanduri<sup>1,2</sup>

From Seventh Annual MCBIOS Conference. Bioinformatics: Systems, Biology, Informatics and Computation Jonesboro, AR, USA. 19-20 February 2010

### Abstract

**Background:** Protein-protein interactions (PPIs) play a crucial role in initiating infection in a host-pathogen system. Identification of these PPIs is important for understanding the underlying biological mechanism of infection and identifying putative drug targets. Database resources for studying host-pathogen systems are scarce and are either host specific or dedicated to specific pathogens.

**Results:** Here we describe "HPIDB" a host-pathogen PPI database, which will serve as a unified resource for host-pathogen interactions. Specifically, HPIDB integrates experimental PPIs from several public databases into a single, non-redundant web accessible resource. The database can be searched with a variety of options such as sequence identifiers, symbol, taxonomy, publication, author, or interaction type. The output is provided in a tab delimited text file format that is compatible with Cytoscape, an open source resource for PPI visualization. HPIDB allows the user to search protein sequences using BLASTP to retrieve homologous host/pathogen sequences. For high-throughput analysis, the user can search multiple protein sequences at a time using BLASTP and obtain results in tabular and sequence alignment formats. The taxonomic categorization of proteins (bacterial, viral, fungi, etc.) involved in PPI enables the user to perform category specific BLASTP searches. In addition, a new tool is introduced, which allows searching for homologous host-pathogen interactions in the HPIDB database.

**Conclusions:** HPIDB is a unified, comprehensive resource for host-pathogen PPIs. The user interface provides new features and tools helpful for studying host-pathogen interactions. HPIDB can be accessed at <http://agbase.msstate.edu/hpi/main.html>.

### Background

Proteins are the work horses of living organisms; they interact with other proteins to carry out most of the biological functions such as signal transduction, protein transport, immune response and other essential functions. PPIs can be classified into two main categories: "Intra-species PPI," where two proteins from the same species interact with each other, and "Inter-species PPI," where two proteins from two different species interact. Host-pathogen protein-protein interactions (HPIs) that play a vital role in initiating infection are a subset of inter-species interactions. Identification and study of

HPIs is critical for understanding molecular mechanisms of infection and subsequent development of drug targets.

Although a number of databases that store PPIs are described in the literature [1-3], only a few databases contain inter-species interactions [4-7]. Thus resources for studying host-pathogen interactions are very limited and users have to access multiple databases followed by manual curation to get the desired set of HPIs. Although there are few efforts toward developing dedicated host-pathogen interaction databases but the existing resources are limited in scope or confined to a limited number of species. The PIG (pathogen Interaction gateway) database provides a collection of HPIs from different resources, but is limited to only one host species, i.e. "human" [8]. Also, the search options in the

\* Correspondence: [rkumar@cvm.msstate.edu](mailto:rkumar@cvm.msstate.edu)

<sup>1</sup>College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA

Full list of author information is available at the end of the article

PIG database are limited to gene identifiers, and BLASTP alignment results are not displayed for sequence searches, so the user cannot evaluate the quality of the alignment. Furthermore, BLASTP cannot be performed in batch mode (multiple sequences at a time), making it difficult to apply for modeling high throughput datasets. Another database "Phi-base" catalogues information about experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens, but does not provide any PPI information [9]. VirhostNet database is dedicated for only virus related PPIs [10]. Other pathogen specific databases have also been reported [11]. Apart from limited availability of experimental HPIs, very few computational approaches have been reported for predicting HPIs. Protein domain profiles of existing intra-species PPIs were used to predict the interaction between human and plasmodium proteins [12]. In another study, existing intra-species PPIs were used to identify orthologous interactions (interologs), which were then used to predict inter-species interactions [13,14]. Both of these computational studies use intra-species PPIs to predict inter-species interactions. Furthermore, they do not provide any web based tool for predicting HPI. In another approach, experimentally identified PPIs are used to search for homologous PPIs to transfer annotations to a new species [15], but the provided tool is limited to predicting intra-species interactions, and has not been applied to predict HPIs.

Here we describe HPIDB, a unified resource that integrates HPIs from multiple resources into a single, non-redundant set in a user friendly web accessible format. The user interface provides multiple options for querying the database content and facilitates BLASTP [16] based sequence searches. It also provides a web based tool which searches for existing homologous HPIs in the HPIDB, which can be used to transfer HPIs to other species.

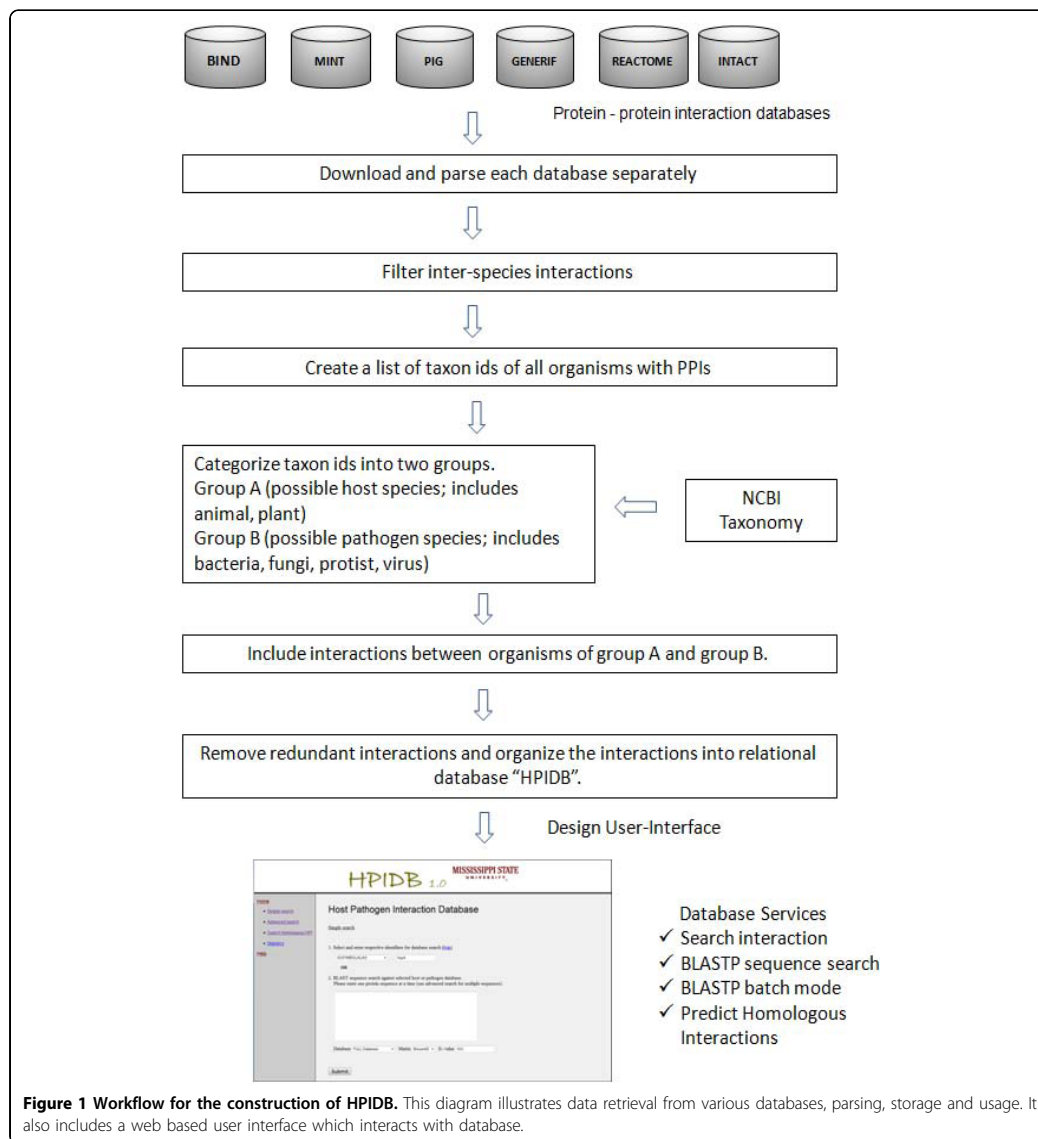
### Construction and content

The HPIDB database is implemented using MYSQL 5.1, with the user interface and web server designed using CGI and Perl. Figure 1 illustrates the workflow of HPIDB and shows data retrieval from various public resources, parsing, storage and specific usage. PPIs from various resources were collected into a single repository. Individual scripts were written to download and parse the data from each PPI database into one unified format. The database schema is available on the website (<http://agbase.msstate.edu/hpi/main.html>). Only inter-species interactions were selected from the repository. For proper classification of PPI into HPI, we classified the taxon ids present in the inter-species interactions into two groups: group A had the taxonomic ids for

host species (includes human, plant, animals, etc.) and group B had the taxonomic ids of pathogenic species (includes bacteria, virus, fungi, protists, etc.). All PPIs where a protein from "group A" interacts with a protein from "group B" are selected and considered as possible Host pathogen interactions (HPIs). To eliminate the possibility of redundant HPIs, all entries were converted into UniProt accession and duplicate entries were eliminated. Where UniProt accession conversion was not successful, duplicate PPI entries were removed based on the protein sequence identity. All the identified interactions were organized into a relational database with additional features like synonym, taxon id, sequence, function, interaction type, experimental information used to identify PPI, and literature information (PubMed id and author information). A web based user interface was designed to query the database using various identifiers, perform BLASTP based protein sequence searches and provide a tool for searching homologous interactions.

All the protein sequences in the database are grouped into major taxonomic groups like plant, animal, bacteria, virus, fungi, and protist. BLASTP sequence alignment functionality was added to the database to search against similar protein sequences. Scripts were written to perform the BLASTP sequence searches in batch mode (search multiple protein sequences at a time) and process the results. The taxonomic classification of protein sequences is integrated with BLASTP, to search only a particular group of sequence databases such as bacteria, virus, animal, all pathogen, all host, etc. Taking advantage of this taxonomic classification of proteins into host and pathogen, we designed a "Search Homologous HPIs" module within the HPIDB. This tool enables the user to search for homologous HPIs in the database for a given set of host and pathogen protein sequences. Internally, this tool is executed in three steps:

1. Input user provided host protein sequences (A) in FASTA format, conduct BLASTP searches against all host proteins in HPIDB and output homologous host protein (HA).
2. Input user provided pathogen protein sequences (B) in FASTA format, conduct BLASTP searches against all pathogen protein in HPIDB and output homologous pathogen protein (HB).
3. Combine the results from step 1 and step 2; any interactions found between HA and HB in HPIDB database are called homologous host-pathogen interactions for proteins A-B. This module provides the user with a set of homologous interactions for further analysis and wherever possible, the results obtained in this module (for example, HA-HB) can be used to transfer annotation to a new species (A-B).



Currently HPIDB contains 22,841 interactions between 49 host and 319 pathogen species. Table 1 shows the prominent set of host and pathogen species represented in HPIDB.

#### Utility and discussion

The database can be accessed using the web interface, which is divided into three separate modules based on specific user needs. Alternatively the whole database can

be downloaded from the website in tab delimited file format.

#### Using the web interface

The web interface is divided into three separate modules:

1. "Simple search" is used to search the database based on user defined identifiers like UniProt id,



**Table 1 Summary of representative host and pathogen species in HPIDB**

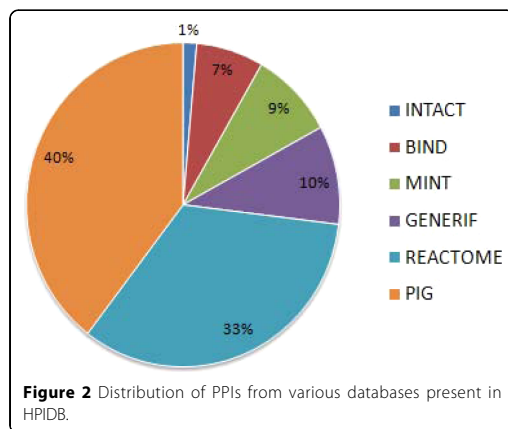
Summary of Host Pathogen PPI stored in HPIDB		
Taxon id	Name	Number of PPI
<b>Host</b>		
9606	<i>Homo sapiens</i>	22386
10090	<i>Mus musculus</i>	147
3702	<i>Arabidopsis thaliana</i>	99
10116	<i>Rattus norvegicus</i>	53
9913	<i>Bos taurus</i>	30
9031	<i>Gallus gallus</i>	19
<b>Pathogen</b>		
1392	<i>Bacillus anthracis</i>	6965
11676	Human immunodeficiency virus 1	3723
119856	<i>Francisella tularensis</i> subsp. <i>tularensis</i>	1341
10376	Human herpesvirus 4	354
11685	Human immunodeficiency virus type 1 (ARV2/SF2 ISOLATE)	344
11696	HIV-1 M:B_MN	341
11689	Human immunodeficiency virus type 1 (ELI ISOLATE)	340
362651	Human immunodeficiency virus type 1 (YU-2 isolate)	340
11688	Human immunodeficiency virus type 1 (JRCSF ISOLATE)	338
11697	Human immunodeficiency virus type 1 (MAL ISOLATE)	338
11701	Human immunodeficiency virus type 1 (RF/HAT ISOLATE)	338
4932	<i>Saccharomyces cerevisiae</i>	337
11678	Human immunodeficiency virus type 1 BH10	320
211044	Influenza A virus (A/Puerto Rico/8/34(H1N1))	303
11683	Human immunodeficiency virus type 1 (Z2/CDC-Z34 ISOLATE)	296
11699	Human immunodeficiency virus type 1 (OYI ISOLATE)	294
11686	Human immunodeficiency virus type 1 (BRU ISOLATE)	292
333284	Hepatitis C virus (isolate Con1)	283

alias name, symbol, taxonomy id, interaction type, literature information (like PubMed id, author, etc.). The results include a set of host-pathogen PPIs along with additional taxonomic categorization in tabular format. A search result with all information about all PPIs is available for download in tab delimited text format, which can be further used in other programs like Cytoscape [17] for network construction and visualization. Protein information is also hyperlinked to other databases for access to available functional annotation, Gene Ontology, and PubMed references. The BLASTP interface is provided, which can be used to determine if a similar protein is involved in the HPI. The user can adjust the BLASTP search parameters and database category (otherwise, default values are provided). The BLASTP search results are returned in both tabular format (for quick analysis) and standard output format (with pair-wise alignment) for user convenience. The results in tabular format are further referenced back to the entries in original database.

2. "Advanced BLAST search" provides the ability to perform BLASTP sequence searches in batch mode. Users can provide more than one protein sequence at a time in FASTA format. Apart from the features provided in a simple BLAST query, users have the option to either get the top hit result for each query or get multiple hits below a user specified E-value.

3. "Search Homologous HPIs" is used to search for homologous HPIs in the HPIDB. For a given set of host and pathogen proteins, first the program tries to identify similar host and pathogenic proteins (based on BLASTP results) in the database. If the identified homologs were involved in HPI interaction in HPIDB, it would be called a homologous HPI. This tool can also be used for only host or pathogen sequences to search homologous host/pathogen proteins and their interacting partners.

The user interface includes a statistics page which summarizes the interactions present in the database (Figure 2). A help file is included, which explains the database schema and the workflow for using the tools



with the sample input and output files. More databases can be easily added to HPIDB and it will be updated every three months. In the future, we plan to extend the homologous HPI prediction and combine it with the protein domain profiles from the HPIDB proteins to develop a computational HPI prediction tool.

Here we describe three case studies that demonstrate the utility of HPIDB to researchers in achieving their objective:

#### Case study 1

**A researcher is studying a particular bacterial species and its related strains that cause infection in humans and animals. In order to identify the host specificity as well as the source for varying infectivity of the bacterial strains of interest, he wants to get a list of all host-pathogen PPIs available for each strain.**

Instead of searching various databases and filtering inter-species PPIs from them individually, this researcher can search the HPIDB using the “simple search” feature and use the taxon ids of all the bacterial species in the search field (one by one) to get the desired PPI dataset.

#### Case study 2

**A researcher has sequenced a new bacterial genome and wants to identify proteins in the genome that are similar to known bacterial proteins involved in host-pathogen interactions.** Existing PPI/HPI resources do not provide sequence searches for a particular taxonomic category. An advantage that HPIDB has over other comparable databases is that it provides the categorization of pathogen protein sequences into categories like bacteria, fungi, protist, virus, etc. and host protein sequences into categories such as animal, plant, etc. based on taxonomy. The researcher can use

the “Advanced BLAST search” option to perform a BLASTP search in batch mode with all protein sequences from the genome against the bacterial protein database.

#### Case study 3

**A user is studying pneumonia, a disease caused by the human pathogen *Streptococcus pneumoniae*. In the absence of any experimental PPIs between human and *S.pneumoniae*, the user needs to identify putative HPIs based on similar homologous interactions (BLASTP E-value <  $10^{-20}$ ) present in the database to generate a testable hypothesis.** Currently, there is no web based tool available that enables the user to search for homologous HPIs. *S. pneumoniae* proteins sequences (2105) were downloaded from NCBI ([ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus\\_pneumoniae\\_TIGR4/NC\\_003028.faa](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_pneumoniae_TIGR4/NC_003028.faa)). In the HPIDB, the “Search Homologous HPIs” can be selected to identify homologous HPIs. Here option A is selected as HPIDB already has human proteins in the database and no further predictions for homologs host proteins are desired. “Form A” should be used which inputs the pathogen protein sequences in FASTA format, the BLASTP parameters can be set to have an E-value <  $10^{-20}$  and the “bacterial proteins” should be selected as database. When we conducted this search, we identified 2001 HPIs between 492 pathogen proteins and 1153 host proteins (mostly human). The dataset can be used further to transfer the homologous interactions and to predict new interactions between human and *S.pneumoniae*. For example, the predicted interactions include previously known virulence factors of *S. pneumoniae* [18] like 3 different capsule proteins (SP0350, SP0357, SP0360), trigger factor (SP0400), exoenzyme enolase (SP1128), pneumolysin (SP1923), Streptococcal lipoprotein rotamase (2012) and serine protease (SP2239) (Additional File 1). Using the output from HPIDB in Cytoscape, one can start exploring the interaction network of all virulence proteins mentioned above with human proteins (Additional File 2).

#### Conclusions

We developed a new host-pathogen protein-protein interaction database “HPIDB” which will serve as a unified and comprehensive resource for HPIs. The user interface provides multiple options to search the database. HPIDB allows high throughput sequence searches in which the user can submit multiple protein sequences at a time and search against a selected taxonomic category. HPIDB also includes a tool that can search for homologous HPIs in the database for user provided sequences. All these features of HPIDB will be helpful for studying host-pathogen interactions.

## Availability and requirements

Project name: HPIDB

Project home page: <http://agbase.msstate.edu/hpi/main.html>

Restrictions for use by non-academics: none

## Additional material

**Additional File 1 : Title: List of homologous HPis for *S. pneumoniae* and human proteins.**Description: Selected homologous HPis identified during Case study 3 for *S. pneumoniae* and human proteins.

**Additional File 2: Title: Cytoscape visualization of homologous HPis in additional File 1.**Description: HPi network for selected *S. pneumoniae* and human proteins visualized using Cytoscape.

## List of abbreviations used

PPI: (protein-protein interaction); PPIs: (protein-protein interactions) ; HPi: (host-pathogen interaction) ; HPis: (host-pathogen interactions); HPIDB: (host-pathogen interaction database)

## Authors' contributions

RK contributed to the design of HPIDB, wrote all of the scripts for the database construction and implementation, and wrote the draft of this manuscript. BN conceived this study, contributed to the design of HPIDB, helped to analyze and interpret the results, and helped to draft the manuscript. All authors read and approved the final manuscript.

## Competing interests

All authors have declared that there are no competing interests.

## Acknowledgements

This project was partially supported by a grant from the National Science Foundation (Mississippi EPSCoR-0903787). We acknowledge Cathy Gresham for her help in implementing the web server at AgBase (<http://www.agbase.msstate.edu>). We thank Dr. Pruett, Dr. Perkins and Ms. Pillai for their help in reviewing this manuscript.

This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 6, 2010: Proceedings of the Seventh Annual MCBIOS Conference. Bioinformatics: Systems, Biology, Informatics and Computation. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/11?issue=S6>.

## Author details

<sup>1</sup>College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA. <sup>2</sup>Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA.

Published: 7 October 2010

## References

- Guldener U, Munsterkott M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Res* 2006, **34**(Database issue):D436-441.
- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: **Human Protein Reference Database-2009 update.** *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.
- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
- Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuerhann M, Ghanbarian AT, Kerrien S, Khadake J, et al: **The IntAct molecular interaction database in 2010.** *Nucleic Acids Res* 2009, **38**(Database issue):D525-531.

- Gilbert D: **Biomolecular interaction network database.** *Brief Bioinform* 2005, **6**(2):194-198.
- Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, et al: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic Acids Res* 2009, **37**(Database issue):D619-622.
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INteraction database.** *FEBS Lett* 2002, **513**(1):135-140.
- Driscoll T, Dyer MD, Murali TM, Sobral BW: **PIG—the pathogen interaction gateway.** *Nucleic Acids Res* 2009, **37**(Database issue):D647-650.
- Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M, Hansen H, Rawlings C, Hammond-Kosack KE, Kohler J: **PHI-base update: additions to the pathogen host interaction database.** *Nucleic Acids Res* 2008, **36**(Database issue):D572-576.
- Navratil V, de Chassey B, Meyniel L, Delmotte S, Gautier C, Andre P, Lotteau V, Rabourdin-Combe C: **VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks.** *Nucleic Acids Res* 2009, **37**(Database issue):D661-668.
- Zhang C, Crasta O, Cammer S, Will R, Kenyon R, Sullivan D, Yu Q, Sun W, Jha R, Liu D, et al: **An emerging cyberinfrastructure for biodefense pathogen and pathogen-host data.** *Nucleic Acids Res* 2008, **36**(Database issue):D884-891.
- Dyer MD, Murali TM, Sobral BW: **Computational prediction of host-pathogen protein-protein interactions.** *Bioinformatics* 2007, **23**(13): i159-166.
- Kim JG, Park D, Kim BC, Cho SW, Kim YT, Park YJ, Cho HJ, Park H, Kim KB, Yoon KO, et al: **Predicting the interactome of *Xanthomonas oryzae* pathovar *oryzae* for target selection and DB service.** *BMC Bioinformatics* 2008, **9**:41.
- Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CY: **Ortholog-based protein-protein interaction prediction and its application to inter-species interactions.** *BMC Bioinformatics* 2008, **9**(Suppl 12):S11.
- Chen CC, Lin CY, Lo YS, Yang JM: **PPISearch: a web server for searching homologous protein-protein interactions across multiple species.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W369-375.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
- Killocoyne S, Carter GW, Smith J, Boyle J: **Cytoscape: a community-based framework for network modeling.** *Methods Mol Biol* 2009, **563**:219-239.
- Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: **VFDB: a reference database for bacterial virulence factors.** *Nucleic Acids Res* 2005, **33**(Database issue):D325-328.

doi:10.1186/1471-2105-11-S6-S16

Cite this article as: Kumar and Nanduri: HPIDB - a unified resource for host-pathogen interactions. *BMC Bioinformatics* 2010 **11**(Suppl 6):S16.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## CHAPTER VII

### CONCLUSION

Bacterial pathogens cause a variety of diseases in human, animals and other plant species. Due to the increased prevalence of antimicrobial drug resistant strains in conjunction with the decreased approval of new drugs, bacterial pathogens pose a major threat to health. Systems biology approaches will increase our understanding of bacterial pathogenesis for subsequent interpretation, prediction and identification of molecular targets for treatment and intervention.

This dissertation is particularly focused on the identification of the building blocks of the system i.e. structural annotation which is essential for systems biology of bacterial pathogens. The major contribution is towards the development of computational tools and resources for tiling array data analysis and feature identification which can be applied to any bacterial pathogen (where genome sequence is available) in a high throughput fashion. First pass structural annotation for genomes is conducted in the genome sequencing projects that use automated gene prediction algorithms to identify all the elements of the system. However, novel structural elements like small non coding RNAs, riboswitches, small genes and other regulatory regions are often missed by the computational gene prediction or other small RNA prediction programs (1-4). To

complement and improve the structural annotation, many experimental approaches like tiling array, RNA sequencing, proteogenomic mapping can be applied to identify previously unknown elements in the genome (2,3,5-9). We used genomic tiling arrays for improving the structural annotation of *S. pneumoniae* TIGR4 genome. Since there were no software tools available that were dedicated to bacterial tiling array analysis, we first developed a computational program called “TAAPP” (Tiling array analysis and annotation pipeline for prokaryotes). “TAAPP” is a web based package and performs data normalization and feature identification and categorization (small RNA, antisense RNA, operons). We used “TAAPP” for tiling array data analysis in *S. pneumoniae* and generated a high-resolution whole genome transcriptional map. We identified around 50 non-coding small RNAs (34 novel sRNAs and 2 novel proteins) and 202 operon structures (consisting of 512 proteins) and improved the structural annotation of this respiratory pathogen. The disadvantage of our tiling array based map of *S. pneumoniae* is that it was generated with RNA expressed at a single experimental growth condition. Therefore to maximize the tiling array coverage RNA samples should be analyzed from different experimental growth and stress conditions. The tiling array map of *S. pneumoniae* is at 12bp resolution. Adopting experimental techniques like RNA-Seq can generate a single nucleotide resolution map that will be more accurate compared to the tiling array (10). The RNA-Seq method is free from any probe design and hybridization bias and can be used to find novel structural elements arising from transcriptional errors and RNA editing events (11-14). Likewise, improvements can be made to the “TAAPP” program to include RNA-Seq datasets. sRNA identified in this study can be used as a

training set to improve the accuracy of computational sRNA prediction programs like sRNAPredict2, sRNAScanner etc (15-17).

Having the structural information is just one step towards conducting systems biology. The next logical step is to identify functional relevance of the components of the system. Availability of controlled vocabularies like the Gene Ontology (GO) that describes biological function is important for functional analysis of the high throughput data (18). Tools that enable automated GO annotation of high throughput datasets expedite biological discovery. The available automated GO annotation tools (19) are mostly based on sequence similarity searches. However, transfer of function based on orthology (20) is the pragmatic approach to provide GO annotation when there is no functional literature available for the gene product (21). The biological function of orthologous proteins is expected to be conserved even when the sequence or tertiary structure changes in due course of evolution (22). This dissertation work developed an automated GO annotation method called “ISO-IEA” to provide functional annotation to gene products from any species in a high throughput manner. The program first transfers the available high quality, experimental based GO annotations of 1:1 orthologous proteins from closely related species known as ISO (Inferred from Sequence Orthology) method. In the absence of orthologous proteins or their experimental annotation, the program uses protein sequence to search against InterPro database (23) to identify functional motifs and assign GO known as IEA (Inferred from Electronic Annotation) method. Using chicken predicted proteins we demonstrated that we were able to increase GO annotations of chicken by 25% using the “ISO-IEA”. A point worth noting is that the

accuracy of the ISO method is dependent on the accuracy of orthology prediction programs. Events such as gene loss, gene duplication and domain rearrangement (24) during protein evolution increase the difficulty of ortholog detection and also could result in orthologous proteins with different biological functions. A logical extension of “ISO-IEA” would be to include domain conservation and positional orthology as a criterion to prevent annotation transfer errors (25).

After developing computation tools for structural and functional annotation of the genome, this dissertation also addresses the prediction of interactions between the components of the host and the pathogen systems. Identification of host-pathogen protein-protein interaction (PPI) is important for understanding the underlying biological mechanism of infection. However, the databases and resources available for studying host-pathogen PPI are scarce and are either host specific or dedicated to specific pathogens. In addition, there is no resource available for predicting host-pathogen interaction. This dissertation work first designed and developed a host-pathogen PPI database "HPIDB" that will serve as a unified resource for searching and analyzing host-pathogen interactions. The database has 22,841 interactions between 49 host and 319 pathogen species. It also enables transfer of existing homologous HPI to new species of interest. In HPIDB the host-pathogen interaction prediction is based on the homology between the protein which have known experimental interaction and proteins for which interactions are being predicted. Increasing the stringency of homolog prediction can enhance the accuracy of the predictions. However, doing so will decrease the total number of predictions. In future, more probabilistic models based on domain

conservation can be added to assign probabilities to predicted interactions. Another logical extension of this work is towards the development of host commensalism interaction database (26). New high throughput experimental methods such as proteomics, transcriptomics etc can also be used in future to identify new host-pathogen and host-commensalism interactions (27).

In conclusion, this dissertation developed computational resources for the structural and functional annotation of genomes as well as the computational prediction of interspecies interactions that are at the heart of host-pathogen systems biology. The computation tools and resources developed will enhance the knowledgebase of infectious disease systems biology.



## REFERENCES CITED

1. Guell, M., van Noort, V., Yus, E., Chen, W.H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kuhner, S. *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science (New York, N.Y.)*, **326**, 1268-1271.
2. Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C. and Vogel, J. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet*, **4**, e1000163.
3. Liu, J.M., Livny, J., Lawrence, M.S., Kimball, M.D., Waldor, M.K. and Camilli, A. (2009) Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic acids research*, **37**, e46.
4. Livny, J. and Waldor, M.K. (2007) Identification of small RNAs in diverse bacterial species. *Curr Opin Microbiol*, **10**, 96-101.
5. Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E. and Rosenow, C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research*, **30**, 3732-3738.
6. Akama, T., Suzuki, K., Tanigawa, K., Kawashima, A., Wu, H., Nakata, N., Osana, Y., Sakakibara, Y. and Ishii, N. (2009) Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. *Journal of bacteriology*, **191**, 3321-3327.
7. Landt, S.G., Abeliuk, E., McGrath, P.T., Lesley, J.A., McAdams, H.H. and Shapiro, L. (2008) Small non-coding RNAs in *Caulobacter crescentus*. *Molecular microbiology*, **68**, 600-614.
8. Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M., Nicora, C.D., Lipton, M.S. and Levy, S.B. (2009) Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. *PloS one*, **4**, e8455.
9. Lamontagne, J., Beland, M., Forest, A., Cote-Martin, A., Nassif, N., Tomaki, F., Moriyon, I., Moreno, E. and Paramithiotis, E. (2010) Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome. *BMC genomics*, **11**, 300.

10. Wilhelm, B.T. and Landry, J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.*
11. Liu, X.S. (2007) Getting started in tiling microarray analysis. *PLoS computational biology*, **3**, 1842-1844.
12. Zhang, Z.D., Rozowsky, J., Lam, H.Y., Du, J., Snyder, M. and Gerstein, M. (2007) Telescope: online analysis pipeline for high-density tiling microarray data. *Genome biology*, **8**, R81.
13. Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M. and Cheung, V.G. (2011) Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science (New York, N.Y.*
14. Knoop, V. (2011) When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci*, **68**, 567-586.
15. Kulkarni, R.V. and Kulkarni, P.R. (2007) Computational approaches for the discovery of bacterial small RNAs. *Methods (San Diego, Calif.*, **43**, 131-139.
16. Livny, J., Brenic, A., Lory, S. and Waldor, M.K. (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res*, **34**, 3484-3493.
17. Sridhar, J., Sambaturu, N., Sabarinathan, R., Ou, H.Y., Deng, Z., Sekar, K., Rafi, Z.A. and Rajakumar, K. (2010) sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PloS one*, **5**, e11970.
18. 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. *Nature genetics*, **25**, 25-29.
19. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, **37**, D619-622.
20. Buza, T.J., McCarthy, F.M. and Burgess, S.C. (2007) Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC genomics*, **8**, 425.

21. Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, **39**, 309-338.
22. Fernandez, A. and Lynch, M. (2011) Non-adaptive origins of interactome complexity. *Nature*, **474**, 502-505.
23. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211-215.
24. Kuzniar, A., van Ham, R.C., Pongor, S. and Leunissen, J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*, **24**, 539-551.
25. Dewey, C.N. (2011) Positional orthology: putting genomic evolutionary relationships into context. *Briefings in bioinformatics*.
26. Casadevall, A. and Pirofski, L.A. (2000) Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect Immun*, **68**, 6511-6518.
27. Kint, G., Fierro, C., Marchal, K., Vanderleyden, J. and De Keersmaecker, S.C. (2010) Integration of 'omics' data: does it lead to new insights into host-microbe interactions? *Future Microbiol*, **5**, 313-328.