

From Data to Knowledge: Translating Functional Genomics Data into Information for Livestock Production
F. M. McCarthy¹, C. Gresham², J. Koltcs³, M.T. Arick², E. Lyons¹, M.W. Vaughn⁴, E.T. Dawson⁴, N. Hopkins¹, S.C. Burgess¹.

¹University of Arizona, ²Mississippi State University, ³Iowa State University, ⁴Texas Advanced Computing Center.

ABSTRACT: New sequencing technologies enable the generation of an increasing number of livestock genomes. However using this data to understand how changes in the genotype affect function is hindered by poor annotation. For example, the Gene Ontology (GO) is routinely used for analyzing functional genomics data, however the GO does not include all key aspects affecting agricultural production (e.g. does not capture information about disease states or tissue expression). Moreover, technologies such as RNASeq identify many novel genes that have no known function. We have integrated new and existing tools to rapidly provide a first pass functional annotation for transcriptome data based upon sequence analysis of conserved protein motifs and sequence homology to better annotate genes. The result of this pipeline is a set of GO and pathway annotations that can be used to determine functional enrichment in a set of differentially expressed genes.

Keywords: bioinformatics; Gene Ontology; functional genomics

Introduction

The development of next generation sequencing techniques for transcription profiling democratized functional genomics studies by enabling researchers to study an even broader range of species instead of focusing on species that have well defined genomes and resources (e.g., microarray platforms) available. However this same technique puts further pressure on the development of annotations and functional modeling tools that can support a much larger range of species and predict functions for novel genes identified by this same technique. These same sequencing approaches also result in long lists of differentially expressed genes that do not *per se* provide useful information about the biological system being studied. Instead, researchers must rely on biological modeling to understand how these gene expression lists provide insights about their biological systems (Cordero et al. (2007); McCarthy et al. (2007)). As a result, while the gap between data and knowledge is closing in for several, well studied species, but fundamental annotation to support functional modeling in a broad range of species is still required.

In the following sections we describe the development of bioinformatics tools that enable livestock researchers to rapidly model their large scale functional genomics data sets, the integrated workflows for adding functional annotation data and how this data is used for functional analyses of differentially expressed gene sets (Figure 1). We also demonstrate the use of these workflows using livestock genomics data and describe how these workflows are being deployed on publicly accessible, high

performance computing environments to provide livestock researchers with the easily accessible computing power they require.

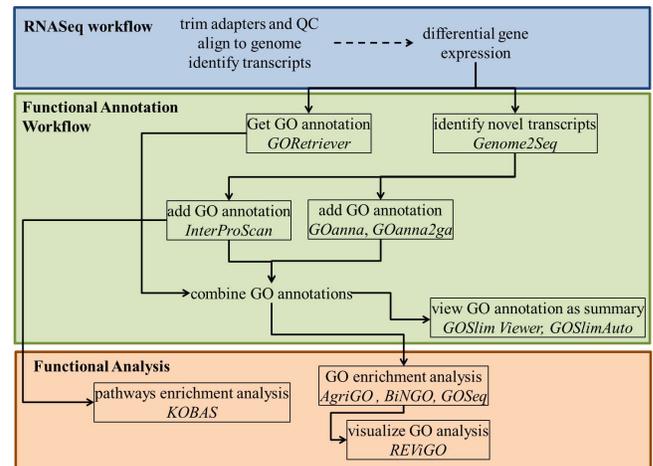


Figure 1: Workflows for functionally modeling RNASeq data sets. Workflows include adding functional annotation and using this data to do functional analysis. Tool names are italicised.

Bioinformatics Workflows and Tools

Genome2Seq Analysis. A typical RNASeq analysis for functional genomics includes (a) initial quality checks and adapter trimming of sequence data; (b) aligning RNASeq reads to an existing genome or assembling transcripts de novo; (c) calculating expression values for transcripts; and (d) identifying transcripts that are differentially expressed between experimental conditions. The end result is a long list of genomic locations representing identified transcripts, along with their expression value. The *Genome2Seq* tool accepts lists of genomic locations and provides a look up service for common livestock genomes that links the genomic location to annotated genes and their GO annotation, or, if the location corresponds to unannotated genes, a fasta file of sequences that can be used for functional annotation. The *Genome2Seq* tool is available via the AgBase database (McCarthy et al. (2009)) and additional genomes can be requested via this resource.

Researchers can easily obtain GO annotations for annotated genes by submitting the list of gene accessions in their differentially expressed list to *GORetriever*, a web-based tool available from the AgBase database. *GORetriever* accepts a list of accessions and returns (a) a GO annotation file for this list; (b) a list of accessions that have no GO (that may be annotated using the functional annotation pipelines described below); and (c) a GO

summary file that can be used as input for *GOSlimViewer* (described below). Since GO annotations are associated with gene products (proteins and RNAs), if researchers submit gene accessions, *GORetriever* also provides mapping information about the original input accession and its protein accession that the GO is associated with; this allows researchers to link their accessions directly with the GO information provided.

Functional Annotation Workflow. An advantage of the transcriptomics approach is that it is able to identify previously unannotated genes (or ‘novel’ genes) and transcript variants, making it a powerful tool for functional genomics for the many commercially important livestock species where genome annotation is poor or incomplete. However the cost of moving to a distributed model for genome annotation is that while many laboratories have the power to identify genes, providing functional information about what these genes do is a less well-defined process. If, as part of gene identification, genes are related to homologous or orthologous genes in related species, then initial gene nomenclature and prediction of function may also be assigned on this basis. However, this approach does not lend itself well to downstream functional analysis.

Our functional annotation pipeline leverages existing functional enrichment tools by providing GO annotation for enrichment analysis. We utilize an existing tool, *InterProScan*, in combination with the *GOanna* tool to rapidly provide ‘breadth’ of annotation (i.e., most gene products have at least some GO annotation) across a wide range of species.

The *InterProScan* tool (Jones, Binns, Chang et al. (2014)) uses an input file of sequences in a fasta format and scans these sequences for short motifs and domains linked to conserved functions. Since this tool is very computationally intensive, it is not suitable for deployment as a web service; instead we have deployed it on the iPlant Discovery Environment (DE) (Goff, Vaughn, McKay, et al. (2011)), where it utilizes Texas Advanced Computing Center (TACC) high performance computing resources. A related DE tool, *InterProScan Parser*, takes the html output of *InterProScan* and maps the InterPro results to GO and pathway accessions, providing summaries as well as a standard GO annotation file format file that can be used in downstream GO enrichment analyses.

The *GOanna* tool uses BLAST to identify homologous genes, searching against specialized databases that contain only GO annotated proteins (McCarthy et al. 2006). If the query sequence matches a sequence in the database, the GO annotations can be transferred to this query and this homologous relationship indicated by assigning the annotation as ‘Inferred from Sequence Alignment’ (ISA) evidence code and a related tool, *GOanna2GA* will convert *GOanna* output to a standard GO annotation file format. Since GO consortium standards require that only annotations assigned on the basis of experimental data are transferred based upon homology or orthology, *GOanna*’s default setting is to only report these annotations. In contrast, *Blast2GO*, a similar tool that assigns GO terms based upon an annotation score calculated from the compilation of multiple, similar GO annotations (Conesa and Goetz, 2008). Since *GOanna* was

developed for use with smaller data sets (of up to 5,000 sequences or a fasta file size of 6Mb), it returns complete BLAST alignment data as an xml file linked to an excel file that contains the GO information. We have updated *GOanna* to handle larger data sets where the user will not want to manually scan individual alignments by presenting users with alignment data as a tab-separated file; this version of *GOanna* has no input file limit and is deployed on the iPlant DE so that it can make use of high performance computing resources. We strongly recommend that users first test a subset of their data against the AgBase version of *GOanna* to identify suitable BLAST parameters and avoid spurious matches with their own data.

It is worth noting that the GO Consortium is actively providing information about GO terms whose functions should be excluded from certain taxa (e.g. mammalian gene products should not be annotated to GO:0019684 photosynthesis). Neither *InterProScan* nor *GOanna* currently apply these taxon restrictions for GO terms to the output data and researchers should check their GO annotation data to ensure that assigned function is taxonomically consistent with their species of interest. We expect that future developments to these tools will address this current limitation.

Combining GO annotations from *InterProScan* and *GOanna* enables researchers to rapidly produce a set of GO annotations for their experimental data set or species of interest. They may choose to summarize their data using the *GOSlimViewer* or *GOSlimAuto* (Davis, Sehgal and Ragan, 2010) tools. *GOSlimViewer* uses pre-computed sets of high order GO terms (‘GO Slim sets’) that describe broad functional categories. A newer tool, *GOSlimAuto* computes the most informative GO terms to describe particular experimental data sets, effectively creating an ‘on the fly’ GO Slim set. Both tools are available via the AgBase database.

Functional Enrichment Analysis. Once the functional annotation workflow has produced a comprehensive set of GO annotations (or pathways annotations which are also produced by the *InterProScan* tool), a common analysis is to identify over-represented terms in the differentially expressed genes/transcripts. There are multiple tools that determine differential expression, including CuffDiff from the Tuxedo suite (Trapnell, Hendrickson, Sauvageau et al. (2014)), DESeq (Anders and Huber, 2010) and edgeR (Robinson, McCarthy, Smyth, 2010), which are available from the iPlant DE. Functional enrichment analysis is done by comparing the set of functional annotations for the differentially expressed transcripts to the functional annotations generated from the complete transcript set.

Once the differentially expressed genes are identified, this gene set can be functionally analyzed to identify GO, pathways and other functional categories that are enriched in this set. There are a large number of tools specifically designed for GO enrichment analysis; some of them even provide a broader range of functional analysis by combining GO and pathways analyses. However, many of these tools are limited in the number of species for which they can do functional analysis, and even fewer consider

technical biases generated by RNASeq data compared to microarray data sets.

The *AgriGO* tool is a web-based tool specifically designed to support analysis of agricultural data, including chicken, cow, pig and horse data sets (Du, Zhou, Ling (2010)). This tool also allows users to upload their own GO annotation sets so that they can include analysis of novel gene data. Another advantage of this tool is that the output data is linked to *REViGO*. *REViGO* is a web-based tool that takes long lists of GO terms and summarizes them by clustering closely related terms (Supek, Bosnjak, Skunca (2011)). This enables researchers to more clearly identify functional changes in their differentially expressed gene lists. We note that this feature is useful for visualizing GO enrichment data, *REViGO* also accepts lists of GO terms generated as the output of any GO enrichment tool.

Another useful tool for GO enrichment analysis is *BiNGO* (Maere, S., Heymans, K. and Kuiper, M. (2005)). *BiNGO* is Java tool that can be run as a plugin or “app” using Cytoscape (Saito R., Smoot M.E., Ono K. et al. (2012)). *BiNGO* allows researchers to upload their own GO annotations and enables analysis of data from a broad range of species, including common pathogens. Since it uses the Cytoscape system, the data is easily visualized and the analysis includes network analysis as well as the GO enrichment analysis.

The *KOBAS* tool (Xie C., Mao X., Huang J. et al. (2011)) also includes a wide range of species, including microbial species, and does functional enrichment based upon GO, pathways and disease databases.

A notable limitation of these functional enrichment tools is that they rely on the assumption that all genes are independent and equally likely to be selected as differentially expressed; this assumption is false for RNASeq experiments. In RNASeq data, expression values are linked to transcript length, and we have more statistical power to identify long genes as being differentially expressed than we do short genes. If this bias is not corrected, functional categories with a preponderance of long genes are more likely to be identified as enriched in an RNASeq data set. Young et al. (2010) proposed a model to correct for this bias and we have adapted this initial work to make it available as a web-based tool, *GOSeq*, on the iPlant DE.

Livestock Analysis Example

To demonstrate the workflows described above, we will use a bovine RNASeq data set collected from fetal pancreatic tissue. After Illumina sequencing, adapter trimmed sequences were mapped to the bovine genome (UMD 3.1) and differentially expressed genes identified using the Tuxedo pipeline and Cufflinks (v2.1.1). This analysis resulted in 914 differentially expressed transcripts, including 137 novel transcripts. Using *Genome2Seq*, we can map 35 of these ‘novel’ transcripts to annotated genes, indicating that they are transcript variants of an annotated gene. The AgBase database *GORetriever* tool identifies 26,760 GO annotations for 748 gene products; 26% of these GO annotations are already annotated using the combined functional annotation workflow.

Functional analysis was done on this data set using both the *AgriGO* and *KOBAS* functional analysis tools. Our initial analysis used only GO annotations provided without the functional annotation workflow outlined above as input for the *AgriGO* Singular Enrichment Analysis (SEA) tool. Using this approach only 392 of the 914 differentially expressed genes had GO annotation. We used the Fisher statistical test and Hochberg multiple correction to identify 391 GO terms with P-values ≤ 0.01 . Among the GO terms identified are very general GO terms such as developmental process, cell death and response to stress. A criticism of using the GO for this type of analysis is that all too often the results in a list of broadly descriptive terms, such as these.

Repeating the analysis with the same statistical parameters but including additional GO annotations provided by the functional annotation workflow identified 14 GO terms with P-values ≤ 0.01 , and these results were visualized using *REViGO* (Figure 2). While fewer GO terms are identified, these enriched functions include peptidase activity and cytokine activity terms that are expected from normal digestive and endocrine pancreatic function. This demonstrates the importance of obtaining a comprehensive set of GO annotations for functional analysis.

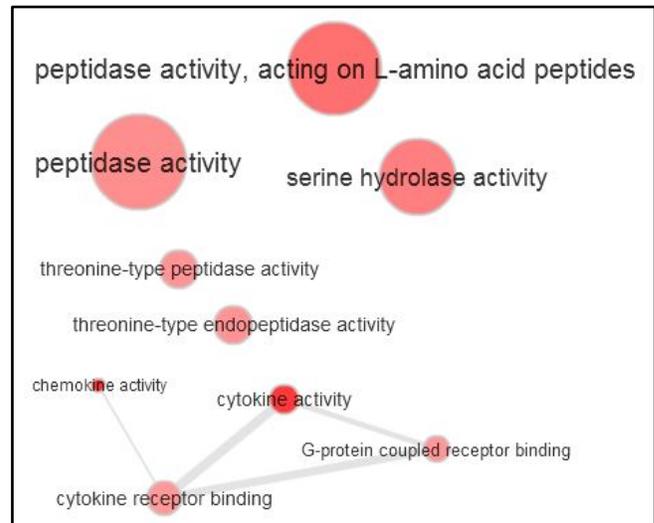


Figure 2: Enriched GO terms from bovine pancreatic RNASeq data. Each GO term is represented by a circle, and related terms as shown as linked. Larger circles indicate a larger proportion of transcripts, compared to the background genome and darker red indicates a lower P-values. Only terms with P-values ≤ 0.01 are reported.

Likewise, *KOBAS* pathways analysis using the same data set (with a Fisher exact test with Benjamini and Hochberg multiple test correction) identifies six differentially expressed pathways (P-values ≤ 0.01), including pancreatic secretion and protein digestion and absorption pathways. This information complements the GO enrichment analysis and together develops a model for genes involved in the developing pancreas.

Conclusions and Future Development

The work presented here demonstrates the utility of functional annotation and analysis workflows to rapidly provide functional modeling for livestock RNASeq data. As RNASeq is applied to an increasingly diverse range of species and more data is produced, providing resources that enable researchers to understand their functional genomic data requires not just mapping sequences to genomes, but also the ability to identify enriched functions from the same data sets.

We are collaborating with iPlant developers to either adapt existing tools and workflows to the iPlant cyberinfrastructure or to develop new tools, as required. Our goal is to enable livestock researchers to do functional modeling on their large, RNASeq data sets in the same computational environment that they use for RNASeq. Future development will include adding additional tools to the iPlant DE and ensuring that links between these tools are seamless, so that the researchers can easily use the output of one tool as the input for the next tool in the workflow.

Literature Cited

- Anders S., Huber W. (2010). *Genome Biol.* 11(10):R106.
- Cordero, F., Botta, M. and Calogero, R. A. (2007). *Brief Funct Genomic Proteomic* 6(4): 265-281.
- Conesa A., Götz S. (2008). *Int J Plant Genomics.* 2008:619832.
- Davis M.J., Sehgal M.S., Ragan M.A. (2010). *BMC Bioinformatics.* 11:498.
- Du Z., Zhou X., Ling Y. et al (2010). *Nucleic Acids Res.* 38(Web Server issue):W64-70.
- Goff S.A., Vaughn M.W., McKay S. et al. (2011). *Front Plant Sci.* 2:34.
- Jones P., Binns D., Chang H.Y., et al (2014). *Bioinformatics* doi:10.1093/bioinformatics/btu031.
- Maere, S., Heymans, K. and Kuiper, M. (2005). *Bioinformatics* 21, 3448-3449.
- McCarthy F.M., Wang N., Magee G.B., et al (2007). *Nucleic Acids Res.* 35(Database issue):D599-603.
- McCarthy, F. M., Bridges S. M. and Burgess S. C. (2007). *Cytogenet Genome Res* 117(1-4): 278-287.
- McCarthy F.M., Gresham C.R., Buza T.J. et al (2010). *Nucleic Acids Res.* 39(Database issue):D497-506.
- Robinson M.D., McCarthy D.J., Smyth G.K. (2010). *Bioinformatics* 26(1):139-140.
- Saito R., Smoot M.E., Ono K. et al (2012) *Nature Methods* 9(11):1069-76.
- Supek F., Bošnjak M., Škunca N., et al (2011). *PLoS One.* 6(7):e21800.
- Trapnell C., Hendrickson D.G., Sauvageau M. (2014). *Nat Biotechnol.* 31(1):46-53.
- Xie C., Mao X., Huang J. et al. (2011). *Nucleic Acids Res.* 39(Web Server issue):W316-22.
- Young M.D., Wakefield M.J., Smyth G.K. et al (2010). *Genome Biology* 11:R14.