ANEXdb: An Integrated Animal Annotation and Microarray EXpression Database

A.S. Leaflet R2556

Oliver Couture, PhD candidate, Interdepartmental Genetics;
Keith Callenberg, undergraduate research assistant, NIH-NSF-BBSI Summer Institute Fellow;
Neeraj Koul, graduate research assistant, Department of Computer Science;
Sushain Pandit, graduate research assistant, Department of Computer Science;
Remy Younes, undergraduate research assistant, Department of Computer Science;
Zhi-Liang Hu, associate scientist;
Jack Dekkers, professor, Department of Animal Science;
James Reecy, professor, Department of Animal Science;
Vasant Honavar, professor, Department of Computer Science;
Christopher Tuggle, professor, Department of Animal Science

Summary and Implications

All publicly available porcine expressed sequences were assembled to create longer, fuller transcripts for annotation purposes. The longer sequences were then used as queries in sequence alignment and comparison software to transfer functional annotation from their homologues in other species. In addition to transferred annotation, sequence variation was also predicted from the assembly. This information can then be used with expression data from high through-put expression measures, such as microarrays, to more fully understand the underlying mechanisms of biological processes. Both kinds of data, expression and annotation, are housed together and available at www.anexdb.org.

ANEXdb allows scientists to store and analyze gene expression data in the form of microarray experiments. Currently **Affymetrix** is the only platform supported. The ANEXdb database schema is **available here for expression** and **here for annotation**. Please refer to the **user manual** for latest submission guidelines.

Introduction

High throughput expression methods, such as microarrays, generate a tremendous amount of data. To fully exploit this data, the transcripts that the microarray tests represent must annotated as to which gene they are, as well as having the raw data easy to manage. Currently very few species other than the human and mouse have good direct annotation of their transcripts. Therefore, most species must rely on sequence homology to one of these well-annotated species for their own annotation. To more fully utilize sequence homology, it is important to have as complete a sequence as possible; requiring an assembly of the shorter EST sequences into longer consensus sequences. By using more complete sequence information, homologous gene family members can be better separated from each other.

It is also important to easily manipulate and utilize the data generated through high throughput methods, as well as package experimental data for easy data sharing. To accomplish this, the Animal ANotation and microarray EXpression database (ANEXdb; www.anexdb.org) along with a web-based application was constructed.

Materials and Methods

ANEXdb is based on a MySQL database, using an online user interface. Users can use the web pages to upload their expression data. Once uploaded, ANEXdb will adjust the expression values depending on the background (both MAS5 and RMA calculations using Bioconductor in R). After this calculation is made, ANEXdb will make the data available for download, providing individual chip downloads, or bulk downloads (one of just the expression data, the other a zip file for submission to GEO).

Approximately 2.5 million sequences were downloaded from NCBI's public databases (dbEST, TRACE, and dbCore). These were first cleaned using seqclean, then clustered and assembled using TGICL. Using BLASTN or BLASTX where appropriate, the resulting assembly was aligned to RefSeq RNA and Protein, as well as Pfam (using either an E-value ≤ 1e-10 or E-value ≤ 1e-10 as a cutoff, respectively). Functional annotation from these alignments was extracted. Exonerate was used to align the sequences against human chromosomes (est2genome model, ≥60% query sequence in alignment, ≥1 HSP with a minimum score of 100). BLASTN was used to align the Affymetrix Porcine Target sequences against the assembly (using E-value ≤ 1e-5 as a cutoff). SNP data was also predicted for the consensus sequences, with frequency being calculated using only the number of unique sequences at each base pair position.

Results and Discussion

After cleaning, ~2.3 million sequences remained for the assembly, which resulted in 140,087 consensus and 103,888 singleton sequences. See Table 1 for full annotation results. In addition, 94% of the Affymetrix target sequences aligned to at least one sequence, and of these 80% have a link to a

RefSeq, providing information for the vast majority of the platform.

The IPA also covers the human portion of RefSeq as well as the mouse RefSeq does. This indicates that, while there are likely overlaps between genes in both databases (i.e. a single IPA accounting for two or more human RefSeqs), the IPA is as complete as mouse RefSeq, relative to human expressed sequences (see Figure 1).

A total of 45,009 consensus sequences contain a total of 2,025,897 SNPs. However, this number requires just a single read of the minor allele. When increasing the stringency to needing at least three reads of the minor allele, the numbers drop to 12,887 consensus sequences containing 202,383 SNPs.

Acknowledgments

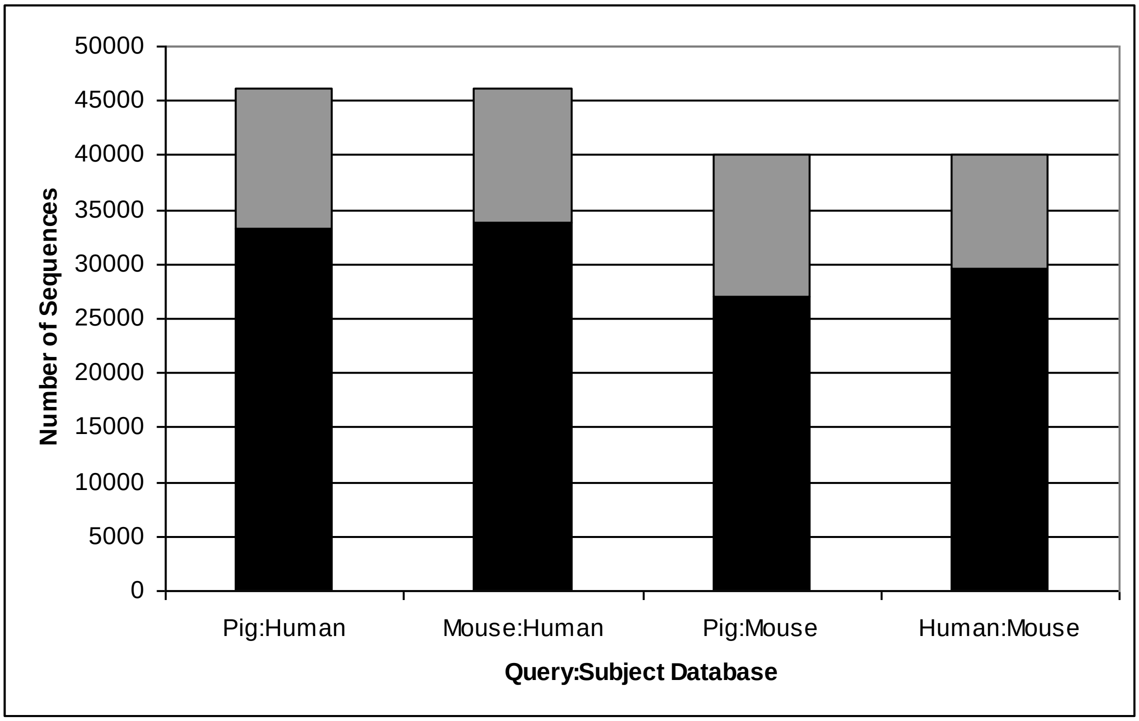
We would like to thank the USDA CSREES-NRI-2005-3560415618 and the ISU Center for Integrative Animal Genomics and ISU Center for Computational Intelligence, Learning, and Discovery for funding this project. A USDA MGET 2001-52100-11506 Fellowship to O.C. is gratefully acknowledged.

Table 1. Various sources of annotation assigned to assembled sequences.

Annotation	Total Number	Total Consensus	Distinct Number	Distinct Consensus
BLASTN to RefSeq RNA	7,643,037	4,114,965	191,602	105,514
BLASTX to RefSeq Protein	7,624,127	3,372,389	71,332	31,195
BLASTX to Pfam	6,716,663	3,364,108	76,385	40,813
Exonerate	71,116	33,016	34,573	22,083
Associated GO Terms	20,436,544	10,326,109	166,119	87,371
Associated KEGG Pathways	6,695,077	3,438,014	92,263	48,763
Putative SNP	2,025,897	2,025,897	45,099	45,099
ORF	1,200,483	723,047	227,954	127,978

Figure 1. The coverage by Iowa Porcine Assembly of human and mouse RefSeq databases is similar to coverage of human by mouse, and of mouse by human.

Black shows the coverage of subject database by the query database, as measured by BLAST matches, while gray shows the number of sequences unique to the subject database.


Number of Sequences
Pig:Human
Mouse:Human
Pig:Mouse
Human:Mouse
Query:Subject Database