ANNOTATION OF PORCINE EXPRESSED SEQUENCE TAGS AND CREATION OF PORCINE-HUMANE ORTHOLOGOUS GENE RESOURCES

Y. D. Zhang[1,2], C. K. Tuggle[2], M. F. Rothschild[2], K. Garwood[3] and W. Beavis[3]
[1]Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia
[2]Department of Animal Science, Iowa State University, Ames, Iowa, USA
[3]National Center for Genomic Resources, Santa Fe, New Mexico, USA

SUMMARY
A collection of 113,497 public Expressed Sequence Tags (ESTs) from Genbank dbEST were utilized. Nucleotide similarities between pig ESTs and human genes were determined by cDNA alignment against 97,100 human UniGene sequences. Available data for pig-on-human comparative maps and homology between pig and human chromosomes were then used to predict pig EST assignments on pig chromosomes. The similarities among pig ESTs and 3,331 pig genes retrieved from GenBank were also examined. Results showed that 28,113 pig ESTs hits 10,258 human UniGenes, and 6,254 pig ESTs hits 1,362 pig gene sequences (score >= 200). As a results of this comparative analysis, a porcine expressed sequence tagged (EST) database, a set of tools for EST analyses and a web query tool for public access to this database were developed. This database is comprised of a number of tables covering the EST profile, alignment similarity, human ortholog information on cytogenetic, and RH locations. Interactive web query interfaces are developed for public access to the EST database.**Keywords**: Porcine, EST, Orthologous gene, Human Genome, cDNA alignment

INTRODUCTION
Expressed Sequence Tag or EST is a partial sequence of a randomly chosen cDNA, obtained from the results of a single DNA sequencing reaction. ESTs are used both to identify transcribed regions in genomic sequence and to characterize patterns of gene expression in the tissue that was the source of the cDNA. This fast approach to cDNA characterization will facilitate the tagging of most new genes at a fraction of the cost of complete genomic sequencing, provide new genetic markers, and serve as a resource in diverse biological research fields (Adams *et al.* 1991). Although there are 113,497 pig ESTs available in public domain (www.ncbi.nlm.nih.gov/dbEST, accessed on 10 February 2003), a small proportion of them have been annotated (e.g. Tuggle *et al.* 2001, Smith *et al.* 2001). The recently published draft sequence of the human genome is a significant milestone in modern biology, providing a very valuable reference for comparative mapping among mammals (Venter *et al.* 2001). Analysis of this draft sequence estimated that the human genome contains approximately 34,000 genes, among them, 31,000 are protein coding genes (Baltimore, 2001). To date, the Nomenclature and Chromosome Committees of the Human Genome Organization have approved 14,750 genes with functions and cytogenetic locations (www.gdb.org, as at 10 Febuary 2003). Annotation of the ESTs using human genome recourses will facilitate further physical mapping and functional study of the ESTs. Mapped ESTs can be used as candidate genes in quantitative trait loci (QTL) regions, for fine mapping QTL and detecting association between phenotypic performance and haplotypes. Annotated ESTs can also be used as microarray probes. Because sequencing genomes of livestock species seems infeasible at present, EST sequences provide an alternative approach for exploring the genomes of livestock species (Adams *et al.* 1991). This study is to annotate pig ESTs using human genomic

resources to identify porcine-human orthologous gene resources, to develop database to manage data and pipeline tools to implement this process.

MATERIALS AND METHOLDS

The study was carried out in two steps. The first step is to retrieve resources, including 113,497 public pig ESTs, human Unigene sequences (http://www.ncbi.nlm.nih.gov/UniGene/), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), Unists (http://www.ncbi.nlm.nih.gov/genome/sts/) and radiation hybrid (RH) map location (GeneMap'99, http://www.ncbi.nlm.nih.gov/genemap99/). The second step is to analyze ESTs or cDNA sequences and integrate the analytical results with available human and swine genomic resources. Three modules were developed for this step. The first module searches similarity between pig EST sequences and human or pig gene sequences by implementing the BLAST® tool (NCBI, Altschul *et al.* 1990). A shell program was written to wrap up the BLAST implementation for data input and store the analysis output. The second module parses the BLAST results and retrieves significant alignments including relevant items (gene identifier, blast score, expectation value) for the subsequent steps. This module was constructed using PERL and BIOPERL modules (http://bioperl.org). The third module integrates the parsed BLAST results with NCBI databases LocusLink, UniSTS, radiation hybrid map (GeneMap'99), and comparative mapping information between pig and human. Consequently, a number of resource tables were generated for populating the database. To manage this annotation data, a pig EST relational database was created using the Ingres® II Database Management System (Computer Associates®). This relation schema of the database was designed in third normal form to eliminate redundancy and anomaly (Ullman 1982). Database administration and web query tools were developed using the DBI PERL module. In addition, a Java based *in-silico* mapping tool (pigest.genome.iastate.edu/java/) was also developed to predict pig EST map location based on pig-human comparative cytogenetic map information (http://www.toulouse.inra.fr/lgc/pig/cyto/cyto.htm).

RESULTS AND DISCUSSION

Similarities between pig ESTs and human genes or between pig ESTs and pig genes. The BLAST alignment showed that approximately 60% pig ESTs are not homologous to human genes and 40% of pig ESTs are homologous to human genes. Particularly, around 16% of the pig ESTs showed high homology to 10,258 human genes at a BLAST score equal to or greater than 200, suggesting that they are candidate procine-human orhtologous genes. The similarities among pig ESTs and 3,331 pig genes showed that 6,254 pig ESTs are homologous to 1,362 pig gene sequences (score >= 200). This implies that pig ESTs are valuable sources to create pig-human orthgologous gene resources. These resources will facilitate physically mapping of the ESTs on porcine chromosomes.

Pig EST database. This database hosts the porcine-human orthologous gene resources. As shown in Figure 1, the database consists of tables that accommodate ESTs, their annotations (similarities between pig ESTs and human gene sequences or between pig ESTs and pig gene sequences) and prediction of EST mapping location on pig chromosomes. Tables are integrated on the internal relational links.

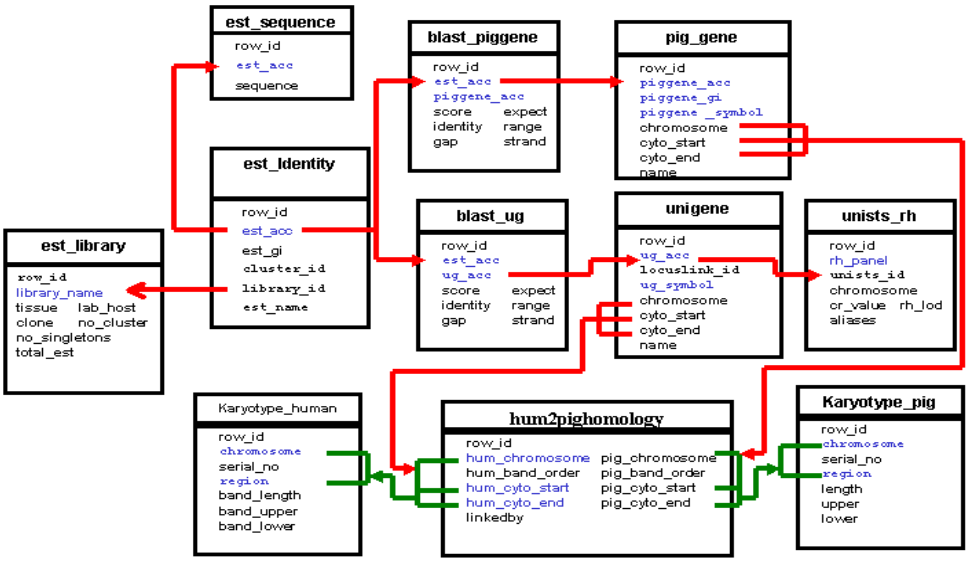

Figure 1. Pig EST database schema

Web query tools. As shown in Figure 2, the DBI PERL web query tools (http://pigest.genomeiastate.edu/query.html) handle the web query to the EST database in two ways to search for: 1) homology between pig ESTs and human UniGenes. This search can be made by either Single Query (fetching ESTs on EST id, Unigene identifier and/or Unigene symbol) or by Multiple Query (fetching ESTs allocated in a library and/or aligned on one human chromosome); 2) homology between pig ESTs and pig genes. The search can be made either by pig ESTs or by pig genes. The items in the query output are linked to related sources for further information.

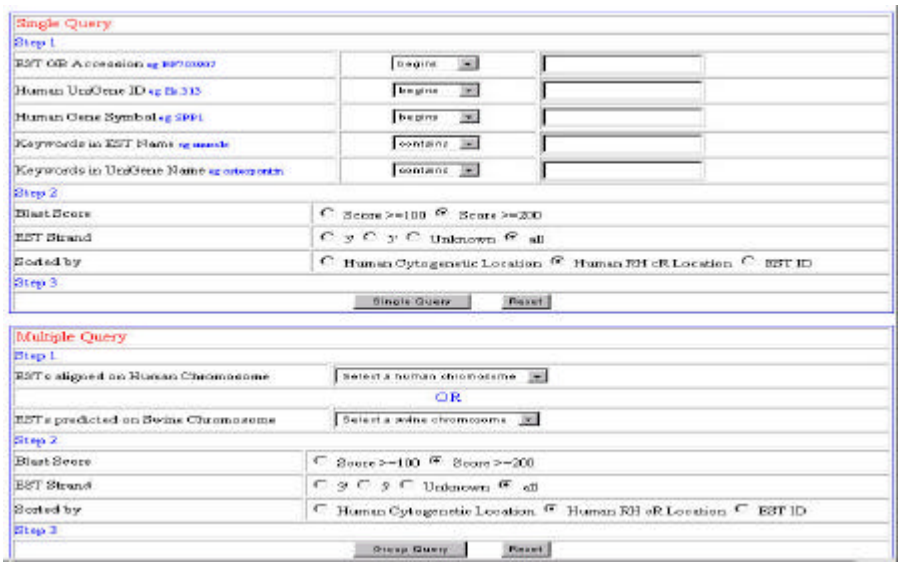

Figure 2. Web form for querying the pig EST database (single and multiple query)

An in-silico mapping tool. A Java based tool (http://pigest.genome.iastate.edu/java/index.html) was also developed for *in-silico* mapping of pig ESTs to human chromosomes and project the mapping locations of pig ESTs. For example, Figure 3 shows ESTs which have homology of human genes on a chromosomal region (HSA10pter-p11). This tool predicts that these ESTs can be mapped on a pig chromosomal region (SSC10q12-qter).

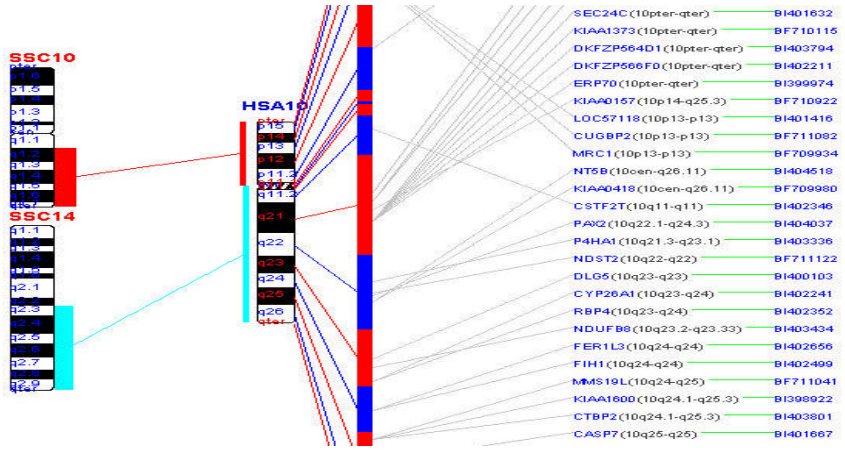

Figure 3. An example of *in-silico* mapping of ESTs to pig chromosomes 10 and 14

SUMMARY

In Summary, pig ESTs are annotated and integrated with human genome information. As a result, the pig-human orthgologous gene resources are created. Using the pig-human orthologous gene resources, a EST database has been developed to host porcine EST annotation information. The analysis revealed a large number of ESTs showing high similarities to human genes. The web query and the *in-silico* mapping interfaces can be a useful approach to facilitate porcine gene mapping. The porcine-human orthgologous genes are valuable resource for 1) identifying new genetic markers, 2) physically mapping them to pig chromosomes, and 3) creating denser maps to assist identifying QTL. This approach can also be applied to other species to create useful genomic resources from EST source, for example in cattle.

REFERENCES

Adams M.D., Kelley J.M., Gocayne J.D., Dubnick M., Polymeropoulos M.H., Xiao H.,Merril C.R., Wu A., Olde B., Moreno R.F., et al. (1991) *Science.* **252**:1651-6.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) J Mol Biol**215,** 403-410.

Smith T.P., Fahrenkrug S.C., Rohrer G.A., Simmen F.A., Rexroad C.E., Keele J.W. (2001).*Anim Genet.* **32**:66-72.

Tuggle C.K., Green J. A., Fitzsimmons C., Woods R., Prather R., Malchenko S., Soares M. B., Roberts C. A, Casavant T., Harger C., Zhang Y., and Rothschild M.F. (2001) *Proceedings of The International Conference On The Status Of Plant And Animal Genome Research IX*, San Diego, CA, USA

Ullman, J (1982).. Computer Science Press, Rockville, Maryland.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., et al. (2001). *Science.* **291**:1304-51.