MammoSapiens: eResearch of the lactation program.
Building online facilities for collaborative molecular and evolutionary analysis of lactation and other biological systems from gene sequences and gene expression data.

Kevin R Nicholas[1,3], Amit Kumar, [3] Yvan Strahm[2], David Powell[2], Torsten Seemann[2], Kerry A. Daly[4], Amelia Brennan[1], Karensa Menzies[1], Julie Sharp[1,3], Matthew Digby[1], Christophe Lefèvre[1,2,3].

1. CRC for Innovative Dairy Products, Department of Zoology, University of Melbourne, VIC 3010, Australia.
2. Victorian Bioinformatics Consortium, Monash University, Clayton VIC 3080, Australia.
3. Institute for technology Research and Innovation, Deakin University, Waurn Ponds VIC 3217, Australia.
4. Centre for Advanced Technologies in Animal Genetics and Reproduction, Faculty of Veterinary Science, University of Sydney, NSW 2006, Australia; Cooperative Research Centre for Innovative Dairy Products, Australia.

Abstract:

Delivering bioinformatics power to life science researchers inevitably runs into problems of limited computing resources in the context of exponentially increasing data sources, access time, costs, lack of skills and, rapidly evolving technology and software tools with poorly defined standards. In this context the development of online facilities to best enable collaborative research often needs to be customized to specific project applications in close cooperation with the experimentalist users and, to be concerned with the storage and management of results to allow more consistency and traceability of results on a broad access data mining platform. Here we showcase an Internet based research platform using the PHP/MySQL paradigm for the collaborative, integrative and comparative analysis of lactation related gene sequences and gene expression experiments to support lactation research. We also illustrate how these resources are used, how they enable research by allowing meta-analysis of data and results and, how the bottom-up development of customized eResearch components can lead to the production of more generic functional software tools and eResearch environments for deployment to a larger number of biological researchers working on other bio-systems.

Introduction

Mammals are characterized by the total dependency of the new born on milk produced by the maternal mammary gland. Mammalian species have evolved a variety of lactation strategies providing a rich biodiversity resource to investigate lactation by comparative analysis. For example, marsupials give birth to a relatively immature embryo after a short gestation period. By contrast, in eutherians, including all extant mammals outside Australasia and America, most of the development occurs *in utero*. Thus, the marsupial young depends on milk for a significant period of time of its development. As a result, the study of lactation in marsupials is not only interesting for the exploration of the evolution of the lactation system in mammals, but also provides a unique model to explore the role of marsupial milk factors on the control of mammalian development and their relationship with eutherian intrauterine factors potentially having similar functions *in utero* [1]. Monotremes, including platypus and echidna, are extant representatives of the most ancient mammals [2]. Monotremes lay leathery-shelled eggs from which emerges an altricial young that is also totally dependant on milk provision by the mother during an extended lactation period. Thus, the analysis of monotreme milk may provide key insight into the origins, function and evolution of lactation. Other animals present peculiar adaptations relevant to study particular aspects of lactation physiology. For example the fur seal is able to maintain lactation after extended periods spent foraging at sea while in the majority of mammals lactation is irreversibly turned off after a few days of weaning. Thus, the fur seal presents a unique differential model to analyse the control of lactation during involution [3].

We have deployed a high throughput technology platform, including genomics, transcriptomics, proteomics, metabolomics and bioactivity screens, for the study of lactation in the tammar wallaby (*Macropus eugenii*) and other mammals with extreme lactation strategies. A bioinformatics resource is being developed to support storage, retrieval and analysis of the data generated. An eResearch resource was built to integrate lactation related and other data from a variety of mammalian species using a plethora of technologies and data sources either available in the public domain or generated in house (sequence, expressions, pathways, literature). These efforts should result in high quality annotation for the mammary function and the lactation system. The benefit from integration into an easily accessible and robust annotation framework with a rapidly customisable data-mining interface is illustrated by the way it is enabling research. This provides an interesting test case for *in silico* biology (computer based biology) where the nature of molecular information transmitted between mother and child during lactation is explored.

We have developed a database system for the annotation and analysis of genomics data; short sequence fragments found to be expressed in particular tissues or cells called Expresssed

Sequence Tags (EST) used for gene identification, and gene expression data. An annotation pipeline based on a MySQL database, open source programs and scripts developed in house is presented. A customized web interface allows users to query and retrieve the data. Usage of open source programs such as Phred and Phrap [4], BLAST [5], HMMER [6], ESTScan [7, 8] and, a series of PHP scripts allow the handling of thousands of sequences in tasks like sequence assembly, searching for sequence homology and coding regions, prediction of biological function, integration of experimental expression data from different platforms for meta analysis (EST sequencing, microarray, MPSS) as well as general database management. We have annotated EST libraries derived from mammary glands and milk from a number of species [2, 9-12]. This allowed us, for example, to identify several ESTs containing a putative signal peptide sequence, with the short-term goal of identifying new milk proteins and, the long-term aim to trace their evolution and identify their properties. Similar methods and interfaces have been implemented to allow the representation of expression data obtained by a combination of methodologies across the lactation cycle in order to provide a comprehensive view of genes and the regulation of their expression during the lactation program. Here, we describe our implementation, and discuss how the e-Resource is applied to enable research.

Sequence data

Genomics has enabled the study of biological systems through the global analysis of gene expression. The first step in the collection of information about gene expression in a new animal model usually consists of obtaining information about the genes by complementary DNA (cDNA) or genome sequencing. With the high cost of genome sequencing, cDNA sequencing is the most common approach and usually provides information that is directly related to the gene expression in the biological system under study. One draw back for the use of sequencing to measure gene expression is that a large number of sequences need to be derived as thousands of genes may be expressed in any tissues at variable levels. Interestingly, recent progress in high throughput sequencing technologies is renewing interest in cDNA sequencing for the estimation of gene expression [13]. In any case, before gene expression can be measured and analysed, basic sequence information needs to be obtained in the form of short partial sequences, which need to be assembled into larger genes sequences. These anonymous gene sequences encoding protein products have to be identified during the annotation process by comparison with known genes or gene products from other, better known, evolutionarily related species. Typically, computer programs are used to search biological sequence databases for evolutionary significant similarities. As a diversity of programs and databases may need to be used and reused to improve annotation when new data is updated, the management and retrieval of annotation data may become cumbersome over time. This task can be greatly facilitated by the implementation of automatic procedures to drive the computation, and the database storage of results, as well as the development of online interfaces to manage the annotation process and mine the results.

In order to address these issues, we have developed EST-PAC a web oriented multi-platform software package for expressed sequences tag (EST) annotation [14]. Originally developed to address our annotation issues, the software was made into a more generic platform and released into the open source for general use. EST-PAC provides a solution for the administration of EST and protein sequence annotations accessible through a web interface. Three aspects of EST annotation are automated: 1) searching either local or remote biological databases for sequence similarities using BLAST services, 2) predicting protein coding sequence from EST data and, 3) annotating predicted protein sequences with functional domain predictions. In practice, EST-PAC integrates the BLASTALL suite, EST-Scan2 and HMMER in a relational database system accessible through a simple web interface. EST-PAC also takes advantage of the relational database to allow consistent storage, powerful queries of results and, management of the annotation process. The system allows users to customize annotation strategies and provides an open-source data-management environment for bioinformatics.

Efforts in the open source and in the academic community have been made to provide the scientific community with on line services, examples of which are PipeOnline [15], EST-PAGE [16], or complete packages such as ESTannotator [17], ESTAP [18]PartiGene [19], and Prot4EST [20]. However, these packages often have restrictive system dependencies, do not always allow extensive data mining and, may not always be available for download and customization. Furthermore, few packages allow real sequence management where users can decide to build queries through a web interface and link them to job submission through a web interface allowing the storage and use of complex filters for sequence similarity searches with criteria based on previous results. Our facility offers more flexibility for the annotation process, updating and the optimal use of often-limited computational resources.

EST-PAC also provides a workbench to cluster ESTs onto reference genome sequence data sets when available. Finally, usage of EST-PAC is not restricted to EST sequences and any type of nucleotide or protein sequences can be loaded for the management of sequence analysis results. This allows the compilation, storage and management of a diversity of customized sequence databases for the analysis of ESTs or the cross referencing of other sequence libraries. EST-PAC provides an open framework for rapid prototyping of data mining and on-line visualization of sequence data, presenting an expandable data-management environment for research and education in bioinformatics.

We have used EST-PAC to annotate sequences from a range of species, including wallaby, seal, platypus, echidna and birds to identify sequences and infer biological function based on similarity with other model species such as humans and mice [2, 11, 12, 21]. One empowering advantage of eResources to manage results is that the environment readily allows for more global analysis of the underlying data. For example, rather than the simple task of generating annotation lists, it is possible to conduct a simple meta-analysis of the distribution of similarity throughout the genome or, a particular subset of gene expression, biological properties or functional pathways. This is illustrated in figure 1 showing the distribution of top hit similarities between the seal mammary gene clusters and other genomes. The results show that seals are closely related to dogs. We have used this information to construct experiments with seal sample on a dog genomic platform [21]. Similar approaches can be used to analyse the variability of sequence divergence. As it is becoming accepted that different genes may be diverging at different rates [22], we are now mining this data to analyse the evolutionary properties of genomes.

EST-PAC also allows the mapping of the genome of one species into another. Usually, bi-directional reciprocal mappings need to be conducted for validation and identification of problematic gene family member attributions. Online access to sequences and maps greatly facilitates analysis of such gene family relationships. Furthermore inter-genome maps also provide a useful link between genomes to query expression data across species for the comparative analysis of gene expression as discussed below.

Finally, another application of sequence libraries is the estimation of gene expression in the biological sample from which the library was derived. This requires the clustering of sequence data using dedicated software to analyse the overlap between sequences and reconstruct larger gene connoting sequences. The clustering of sequence similarity onto the full genome sequence of the organism when available or onto other related genome databases can be also used as an alternative or a complement for sequence library clustering. EST-PAC has been valuable for the validation of gene catalogues. Once contiguous sequences are assembled, they can be mapped back onto ESTs to count the number of occurrences of gene ESTs in the library and estimate gene expression in the original sample. This provides a digital estimation of gene expression that is potentially more accurate than analogue measurements obtained by sequence hybridisation platform like DNA microarrays. We have employed these approaches to analyse the gene expression profile of the main milk protein genes in the tammar wallaby, identifying interesting lactation phase specific and regulatory candidates [11]. The main drawback of the sequencing approach is the need to sample a large gene space with several thousand genes. Until recently this has been prohibitive due to the high cost of high throughput sequencing. This is however rapidly changing and technological advances in the area of DNA sequencing might revive this approach in the near future.

Massively parallel signature sequencing (MPSS):

Massively parallel signature sequencing is an emerging technology for the mass sequencing of typically short sequence tags to measure gene expression [23]. Sequence similarity searches are also used for the identification of sequence tags by comparison and clustering against reference genome databases. Figure 2 shows a summary of mammary gland gene expression of the main milk proteins from lactating and non-lactating wallabies estimated by coda library analysis and MPSS. We have presented elsewhere an analysis of such data obtained from tammar wallaby model and discussed the advantage and limitation of the MPSS approach [11]. The availability of a comprehensive gene sequence catalogue was instrumental in the analysis. In the near future, new technology that overcomes some of the current limitations will be used and, a large flow of data might soon become available for the analysis of an increasing number of biological systems. This will potentially open a new biological insight integrating the characterisation and quantification of gene expression. It is possible that flexible e-Bioinformatics workbenches will quickly be in high demand for in depth annotation and analysis of this data.

Microarray data:

The microarray or DNA chip technology allows the measurement of gene expression of a large number of genes when gene sequence information is already available. The technology exploits the property of DNA to hybridise specifically to a complementary sequence. Typically, short sequences, which may be either cloned cDNA or synthetic fragments, are chosen and deposited in individual cells of an array on a small surface where each cell in the array contains many copies of a specific sequence. cDNA is synthetised from the population of messenger RNAs in the biological sample, labelled with a fluorescent dye and, hybridised to the microarray. Specific sequences will bind to their complement sequences at the corresponding location on the array. The array can be scanned to measure the dye intensity at the location of each cell to provide an estimation of the relative quantity of cDNA molecules containing particular sequences representing an estimator of the relative gene expression of the corresponding genes in the sample. One difference between microarray and sequence based expression data is that microarrays provide a more error prone analogue signal source than digital signal obtained from sequence sampling. Statistical analysis is typically used to characterise error rate, normalise data, identify bias, explore and compare the results from multiple samples or experiments. A large number of software packages and methods can be employed, each with their advantages, constraints and limitations. Often, statisticians may be involved in data analysis. The diversity of approaches that may be used for particular experimental designs and the rapidity with which new methods are developed is such that it is quite difficult to standardise and automate the analysis in a research context. However, what is more important is to have ways to retrieve the data online, making it available to end-users with overlapping interests at any time, and compare the results produced by different methods or people. This is particularly important when data starts to accumulate quickly and, enables meta-analysis of multiple experiments or comparison of results obtained on multiple platforms or in a diversity of species.

To address this issue, we have developed simple online databases and interfaces to store and query expression data. For each technology platform and experiment, interfaces are often customised in order to provide appropriate graphical representation and query facilities for particular experimental design (e.g. time course or treatment). In our experience, the availability of simplified, easily accessible and targeted online interfaces to explore results after expert data pre-processing has greatly enhanced the usability of microarray data and, over time, stimulated cross project analysis. A diversity of studies employing the e-Resource has been published [21, 24-29]. We are now also trying to rationalise our approach into a more generic package also based on the PHP-MySQL paradigm in order to facilitate the online retrieval and query of processed data. However, funding for translational research in this area of open software development is needed. Nonetheless, the possibility to quickly built interfaces for the analysis of expression data has expended collaborations in other area of research such as the study of reprogramming during germ cell development [30]. The difficulty to interact efficiently with expression data has often already become a bottleneck for biological research.

The combination of sequence cross-species referencing available for EST-PAC and expression data empowers meta-analysis with the straightforward implementation of queries and interfaces to retrieve and compare gene expression in multiple species [3]. Figure 3 shows an overview of our implementation. The integration of our data in an eResource will facilitate this comparative analysis.

Conclusion

The rapid development of high throughput technology in the life science has empowered biologists by allowing access to a more global view of biological systems with the measurement of many variables in parallel, such as gene sequences, gene expressions, proteins, metabolites or functional screens. As a consequence, bioresearch is increasingly dependant on computational technology for the management and analysis of the large amount of data being produced quickly. This generates new problems and eResearch platforms are needed to empower biological researchers with the full benefit of high throughput technology. Bioinformatics is ideally placed at the interface between biology and computer science to tackle these issues. In a collaborative environment it is important to have broad, comprehensive access to data and results. The Internet platform is an ideal media for this. Here, we have reviewed an eResearch platform developed to facilitate lactation research. Based on the PHP/MySQL paradigm this platform allows for rapid customisation and immediate online availability of user and project oriented interfaces to empower biological researchers by allowing them to interact directly with data and results through a web interface. In the context of rapidly changing techniques or format with poor standard, this is an efficient approach to enhance the level of interaction and feedback between specialists leading to a better, faster and more traceable exploitation of results.

Our development cycle is mainly driven by research priorities. Data is stored in databases and online interfaces are quickly implemented to address urgent issues. When possible and based on experience with customised interface design, we are trying to derive more generic tools for release into the public domain. For example, EST-PAC scripts can be downloaded, installed and configured on most systems to rapidly deploy an online sequence annotation platform. Unfortunately, funding for these translation activities is typically not budgeted in research projects and this not a priority. However, with the rapid adoption of high throughput technologies by a larger number of researchers, the value of eResearch platforms for sequence or gene expression data management is gradually becoming apparent and large initiatives are being undertaken in this area. For example caArrray (http://caarraydb.nci.nih.gov) is a resource for cancer data. This typically requires large computational facilities, which are expensive to maintain. Our approach is to develop relatively simple, easy to install and customised software geared toward individual end users with limited computational resources and computer skills. This allows the rapid development of user-friendly Web interfaces directly addressing the requirements of researchers for data access and analysis. However, the deployment of our tools onto large open access computational facilities should be considered to make adoption of such tools even more straightforward, providing readily available and functional eResearch platform readily available in the area of gene bioinformatics. As high throughput sequencing technology starts to reach the laboratory and, with the emergence of system biology and 'fishing science' where large numbers of experiments are conducted for extensive data mining, eResources will become critical in many areas of biology.

Figure legends:

Figure 1: **BLASTN similarity distribution of the seal gene catalogue.** Distribution of similarity (percent identity) between assembled seal EST sequences and dog (purple), cow (blue), human (green) or chicken (yellow) representative sequences from the Unigene database. Only BLAST alignments with a high score (> 700) were selected. Gene sequence similarity peak at 96% between the seal and the dog, 91 % between seal and man and 92% between seal and cow.

Figure 2**: Comparison of tammar wallaby milk protein gene expression estimated by cDNA or MPSS sequencing**. Expression data where obtained by cDNA sequence or MPSS data analysis. Expression levels (normalised to parts per 10 000 total gene expression) are shown for the main milk proteins in non-lactating and lactating samples. Form left to right, cDNA libraries form the mammary gland of pregnant (23p) animal at 23 day pregnancy, cDNA libraries form 4 day involuting animal (4i), cDNA libraries from lactating animal at day 130 (130L) or day 260 (260L), MPSS data from lactating mammary gland RNA at day 151 (151n_mpss) or day 241 (241n_mpss), results from a normalised cDNA library. Numbers in parenthesis indicate the total sequence counts available for each cDNA library set. The figure shows comparable milk protein gene proportion in lactating animals estimated either by cDNA sequencing or MPSS.

Figure 3: **Overview of the bioinformatics e-platform**. EST-PAC provides a web interface for the management, storage and querying of sequences and results from annotation tools such as BLASTALL, EST-Scan2 and HMMER. Expression data are stored in an expression database. Online query interfaces combined expression and annotation data.

1. Sharp, J.A., et al., *The comparative genomics of tammar wallaby and fur seal lactation; models to examine function of milk proteins.* Milk Proteins: From Expression to Food (Editors: Abby Thompson, Mike Boland, Harjinder Singh) Elsevier. (in press), 2008.
2. Warren, W.C., et al., *Genome analysis of the platypus reveals unique signatures of evolution.* Nature, 2008. **453**(7192): p. 175-183.
3. Brennan, A.J., et al., *The tammar wallaby and fur seal: models to examine local control of lactation.* J Dairy Sci, 2007. **90 Suppl 1**: p. E66-75.
4. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
5. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **17**: p. 3389-402.
6. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**: p. 755 - 763.
7. Iseli, C., C.V. Jongeneel, and P. Bucher, *ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 138 - 148.
8. Lottah, C., et al., *Modeling sequencing errors by combining Hidden Markov models.* Bioinformatics, 2003. **19**(Suppl 2): p. 103 - 112.
9. De Leo, A.A., et al., *Characterization of two whey protein genes in the Australian dasyurid marsupial, the stripe-faced dunnart (Sminthopsis macroura).* Cytogenet Genome Res, 2006. **115**(1): p. 62-9.
10. Sharp, J.A., et al., *Fur seal adaptations to lactation: insights into mammary gland function.* Curr Top Dev Biol, 2006. **72**: p. 275-308.
11. Lefevre, C.M., et al., *Lactation transcriptomics in the Australian marsupial, Macropus eugenii: transcript sequencing and quantification.* BMC Genomics, 2007. **8**: p. 417.
12. Sharp, J.A., C. Lefevre, and K.R. Nicholas, *Molecular evolution of monotreme and*

marsupial whey acidic protein genes. Evol Dev, 2007. **9**(4): p. 378-92.

13. Torres, T.T., et al., *Gene expression profiling by massively parallel sequencing.* Genome Res, 2008. **18**(1): p. 172-7.
14. Strahm, Y., D. Powell, and C. Lefevre, *EST-PAC a web package for EST annotation and protein sequence prediction.* Source Code for Biology and Medicine, 2006. **1**(1): p. 2.
15. Ayoubi, P., et al., *PipeOnline 2.0: automated EST processing and functional data sorting.* Nucleic Acids Res, 2002. **30**: p. 4761 - 4769.
16. Matukumalli, L.K., et al., *EST-PAGE. Managing and analyzing EST data.* Bioinformatics, 2004. **20**: p. 286 - 288.
17. Hotz-Wagenblatt, A., et al., *ESTAnnotator: A tool for high throughput EST annotation.* Nucleic Acids Res, 2003. **31**: p. 3716 - 3719.
18. Mao, C., et al., *ESTAP. An automated system for the analysis of EST data.* Bioinformatics, 2003. **19**: p. 1720 - 1722.
19. Parkinson, J., et al., *PartiGene. Constructing partial genomes.* Bioinformatics, 2004. **20**: p. 1398 - 1404.
20. Wasmuth, J.D. and M.L. Blaxter, *prot4EST: translating expressed sequence tags from neglected genomes.* BMC Bioinformatics, 2004. **5**: p. 187.
21. Sharp, J.A., et al., *The fur seal-a model lactation phenotype to explore molecular factors involved in the initiation of apoptosis at involution.* J Mammary Gland Biol Neoplasia, 2007. **12**(1): p. 47-58.
22. Zhang, P., Z. Gu, and W.-H. Li, *Different evolutionary patterns between young duplicate genes in the human genome.* Genome Biology, 2003. **4**(9): p. R56.
23. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630 - 634.
24. Daly, K.A., et al., *Analysis of the expression of immunoglobulins throughout lactation suggests two periods of immune transfer in the tammar wallaby (Macropus eugenii).* Vet Immunol Immunopathol, 2007. **120**(3-4): p. 187-200.
25. Brennan, A.J., et al., *A population of mammary epithelial cells do not require hormones or growth factors to survive.* J Endocrinol, 2008. **196**(3): p. 483-96.
26. Daly, K.A., et al., *Identification, characterization and expression of cathelicidin in the pouch young of tammar wallaby (Macropus eugenii).* Comp Biochem Physiol B Biochem Mol Biol, 2008. **149**(3): p. 524-33.
27. Daly, K.A., et al., *Characterization and expression of Peroxiredoxin 1 in the neonatal tammar wallaby (Macropus eugenii).* Comp Biochem Physiol B Biochem Mol Biol, 2008. **149**(1): p. 108-19.
28. Daly, K.A., et al., *CD14 and TLR4 are expressed early in tammar (Macropus eugenii) neonate development.* J Exp Biol, 2008. **211**(Pt 8): p. 1344-51.
29. Sharp, J.A., et al., *Identification and transcript analysis of a novel wallaby (Macropus eugenii) basal-like breast cancer cell line.* Mol Cancer, 2008. **7**: p. 1.
30. Lefevre, C. and J.R. Mann, *RNA expression microarray analysis in mouse prospermatogonia: Identification of candidate epigenetic modifiers.* Dev Dyn, 2008. **237**(4): p. 1082-9.

Figure 1: BLASTN similarity distribution of the seal gene catalogue

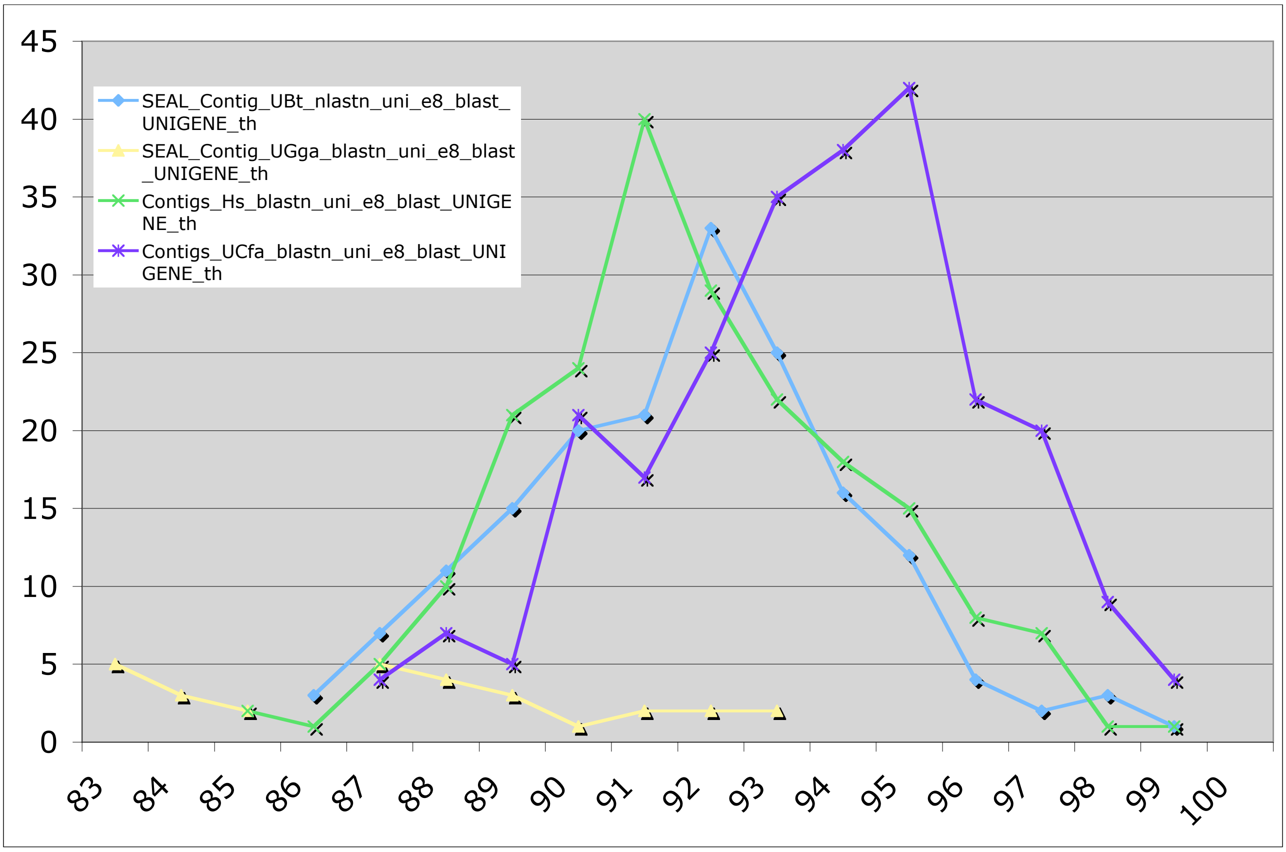


Figure 2:
Comparison of tammar wallaby milk protein gene expression estimated by cDNA or MPSS sequencing.

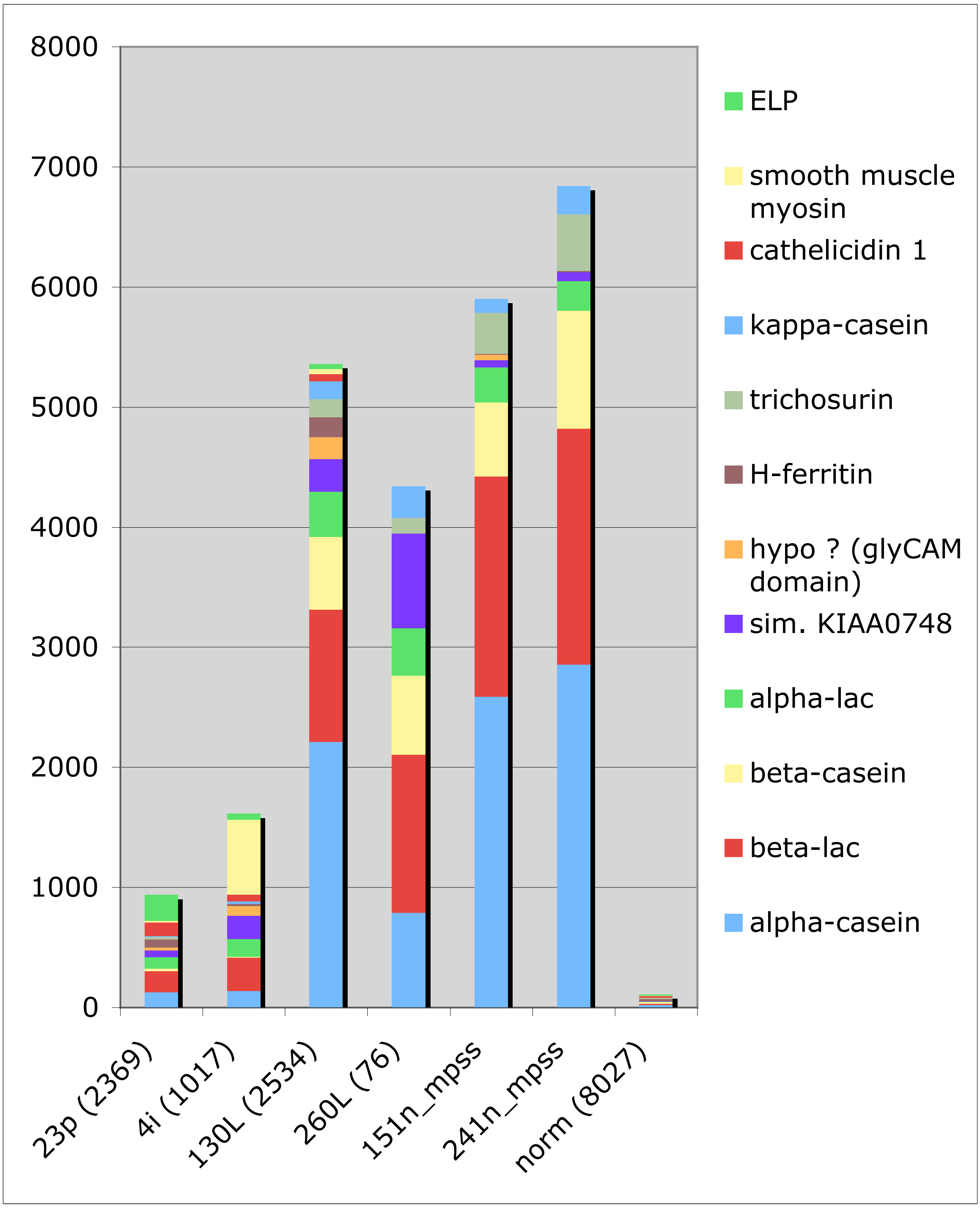
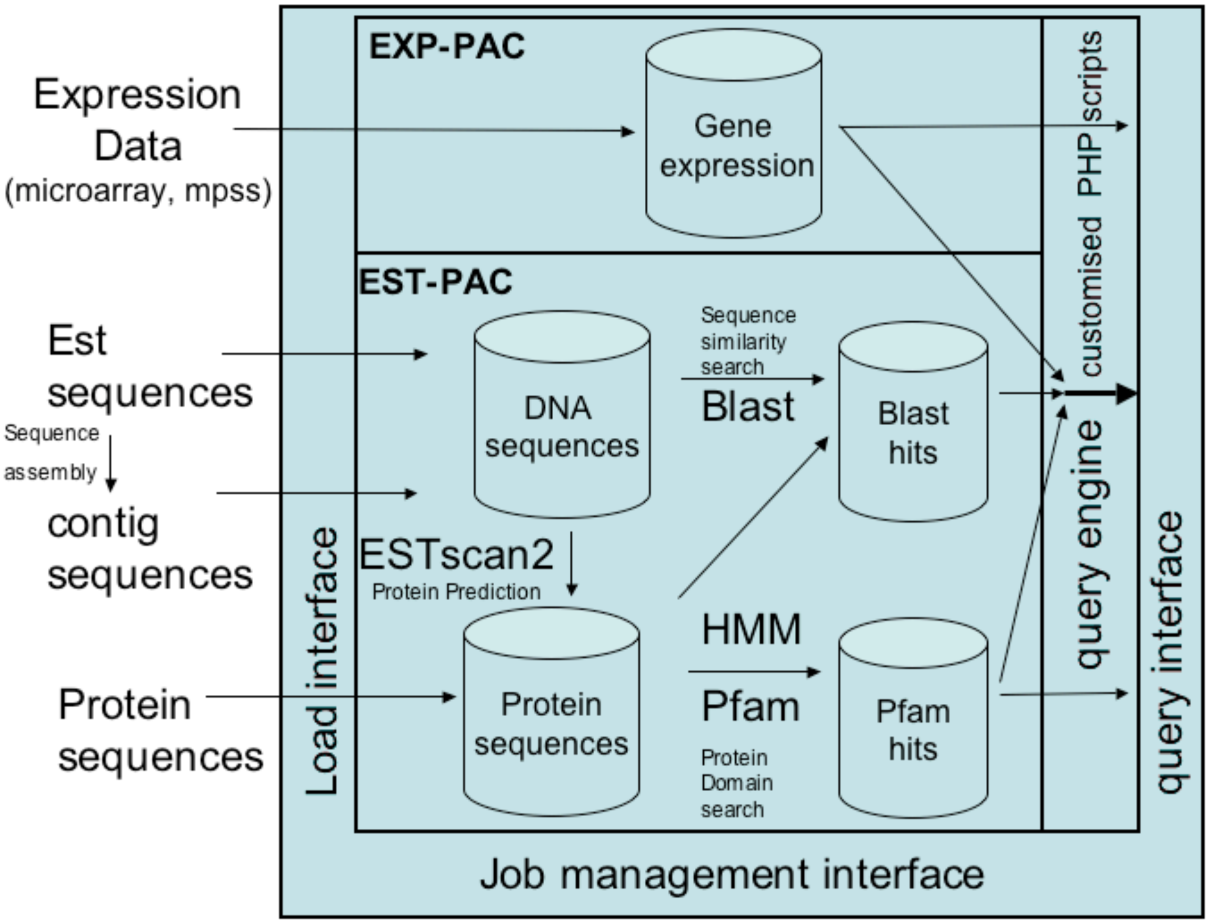


Figure 3: Overview of the bioinformatics platform


EXP-PAC
Expression Data (microarray, mpss)
Gene expression
customised PHP scripts
EST-PAC
Est sequences
Sequence assembly
contig sequences
DNA sequences
Sequence similarity search
Blast
Blast hits
ESTscan2
Protein Prediction
Protein sequences
Protein sequences
HMM
Pfam
Protein Domain search
Pfam hits
Load interface
query engine
query interface
Job management interface