AN INTEGRATED SUITE OF INFORMATICS TOOLS AND RESOURCES TO SUPPORT POST-GENOMICS INVESTIGATION

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Weizhong Li

January 2008

ABSTRACT

The genome sequencing projects have brought about a massive increase in the scale of bioinformatic analysis. To engage in post-genomic analysis requires the development of techniques for processing these huge datasets automatically, efficiently and effectively, and this requires the discovery of new approaches, the development of new efficient bioinformatics tools and the establishment of high-quality, accessible information resources. This thesis describes the development of bioinformatic tools and resources, and analytical methods for a major post-genomic project directed at an open transcriptomic screen of mechanisms involved in the environmental stress adaptation of an important environmental model species, the common carp *Cyprinus carpio* L..

The project required the identification and characterisation of cDNA resources through expressed sequence tag (EST) analysis, for which a new user-configurable package, EST-ferret, was developed. The package integrates a suite of open source algorithms connected by PERL scripts that includes options for EST sequence cleaning-up, assembly, BLAST homology search, protein domain searches, and Gene Ontology (GO) annotation. ~13,500 ESTs were processed through EST-ferret and the results have been incorporated into a comprehensively annotated and searchable database, carpBASE 2.1. Thus 9202 high-quality EST sequences were assembled into 6033 non-redundant sequences. Extending the alignment search methods to include protein domains, UTRs and repeat elements annotated an additional 12.6% of ESTs. Finally, a 'GOprofiler' programme was developed and embedded in EST-ferret to assign GO annotations to ESTs. Collectively these tools maximised the identification and functional annotations for cDNA clones.

Analysing gene expression profiles from microarrays is fundamental for post-genomic approaches. ExprAlign was developed to cluster and visualise gene expression data. This included CORR, a programme which determines the similarity of gene expressions between genes by computing millions of Pearson correlation coefficients. ExprAlign also implemented the VxInsight package to align ESTs into different expression clusters and ordinate and visualise the resulting clusters as a 3D landscape. ExprAlign was used to suggest identities for unidentified ESTs by relating 522 unclassifiable ESTs in carpBASE 2.1 to other BLAST-identified genes, and separating some unique gene and some gene isoforms. GOmatrix, using Fisher's exact test, was developed to determine which non-redundant gene expression clusters were statistically over- or under-represented in GO categories of interest. This has greatly assisted the understanding of biological roles and molecular functions of different gene groups identified from the transcript profile.

Comparative, cross-species analysis of sequence data and gene expression data is important to functional genomic investigation. Orthology analysis was processed across carp, zebrafish and human and a tool called FindOrthologs was developed for this purpose. ExprAlign was implemented in the orthology analysis for discovering how conserved the correlated gene expressions of orthologous genes were across carp and human. GOmatrix also indicated the conserved biological processes for the orthologous gene groups.

CONTENTS

LIST OF TABLES

DEDICATION

I dedicate this thesis to my family.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Andrew Cossins at University of Liverpool for providing opportunities, offering facilities, offering direction and advice, and giving encouragement. I am very grateful to my co-supervisor Andrew Brass at University of Manchester for his direction and assistance during the period of study. I would especially like to thank Andrew Gracey at University of Southern California for his advice and suggestions.

I acknowledge and appreciate the valuable contribution made by Dr Lu Mello and Mr Chris Duckett at School of Biological Sciences, Dr Anthony Morton at Physics Department, and Dr Cliff Addison at Computer Sciences Department, all of whom helped me at different stages.

Thanks also to Dr. Bela Tiwari at CEH Oxford in providing me with useful idea on improving my programming capability. Thanks also to Prof. Peter Diggle at the University of Lancaster in providing me ideas on statistic approaches. I also want to thanks Dr Margaret Hughes and Dr Daryl Williams for instructing me on technologies for EST sequencing and DNA microarrays. Last, I thank all my colleagues at the Lab for Environmental Gene Regulation for helping me in my study.

ABBREVIATIONS

ANOVA	ANalysis Of VAriance
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
CDD	Conserved Domain Database
cDNA	Complementary DNA
CDs	Coding Sequences
COG	Clusters of Orthologous Groups
DBMS	database management system
DDBJ	DNA Data Bank of Japan
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EGO	Eukaryotic Gene Orthologs
EMBL	European Molecular Biology Laboratory
ESTs	Expression Sequence Tags
ExprAlign	Expression Alignment
dbEST	Database for Expression Sequence Tags
E-value	Expect Value
FET	Fisher Exact Test
GO	Gene Ontology
GEO	Gene Expression Omnibus
GPL	GNU General Public License
HGP	Human Genome Project
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
INSDC	International Nucleotide Sequence Database Collaboration
IUBMB	International Union of Biochemistry and Molecular Biology
JVM	JAVA Virtual Machine
J2EE	JAVA 2 Platform Enterprise Edition
KEGG	Kyoto Encyclopedia of Genes and Genomes
LEGR	Laboratory of Environmental Gene Regulation

MAGE-ML	Microarray Gene Expression Markup Language
MatLab	MATrix LABoratory
MCL	Markov Cluster
MF	Molecular Function
MGED	Microarray Gene Expression Data
MGSC	Mouse Genome Sequencing Consortium
MIAME	Minimum Information About a Microarray Experiment
NCBI	National Center for Biotechnology Information
OMG	Object Management Group
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PHP	Hypertext Preprocessor
PSSMs	Position Specific Score Matrices
RefSeq	Reference Sequences
ROC	Relative Operating Characteristic
SAGE	Serial Analysis of Gene Expression
SCF	Standard Chromatogram Format
SDK	Software Development Kit
SMD	Stanford Microarray Database
SQL	Structured Query Language
SSH	Suppression Subtractive Hybridization
TC	Tentative Consensus
TIGR	The Institute of Genomic Research
TOGA	TIGR Ortholog Gene Alignment
TOGs	Tentative ortholog groups
UniProt	Universal Protein Database
UTRs	Untranslated Regions
UTRdb	Untranslated Region Database
WGS	Whole Genome Shotgun
XML	Extensible Markup Language
YAC	Yeast Artificial Chromosome

CHAPTER 1: GENERAL INTRODUCTION

1.1 Genomics

1.1.1. The Challenge in Biology

As we move into the 21st century, how we view and practice biology continues to change and evolve. Biology had entered a new era with the official publication of the initial sequence and analysis of the human genome comprising 3 billion nucleotide base pairs (International Human Genome Sequencing Consortium 2001). Now biology can be treated as an informational science, since it contains the fundamental of biological information, such as genomes, transcriptomes, proteomes (Lederberg and McCray 2001) and environmental signals. Indeed, to deliver the new high throughput technologies and the associated data processing capability, biology has become increasingly cross-disciplinary as biologists join forces with chemists, computer scientists, engineers, statisticians, mathematicians and physicists.

Since the 1970s, recombinant DNA technology has matured as a suite of technical developments and innovations and is now applied widely in locating, isolating, synthesizing, amplifying and purifying specific DNA molecules (Cohen *et al.* 1973). A key development in the evolution of this technology was the discovery of restriction enzymes (Smith and Nathans 1973; Arber 1974; Nathans and Smith 1975), which allows DNA fragments to be recognized and cut at precise locations. Gel electrophoresis (Poulik and Smithies 1958) provided the standard technique for separating the resulting DNA fragments on the basis of their size and electrical charge. The Southern Blot technique (Southern 1975) transfers DNA fragments from a gel to a membrane for analysis using gene-specific probes, and the DNA library technique allows researchers to clone gene collections in self-replicating libraries which can be used as sources for gene discovery. The Polymerase Chain Reaction (PCR) (Rabinow 1996) allows small amounts of DNA fragments, even single molecules, to be amplified a billion-fold in a test tube within just a few hours. Genetic engineering, the application of recombinant DNA technology to specific biological, medical, or agricultural problems, is

now a well-established scientific discipline that combines all of those techniques described so far. It supplies new information about the structure and the function of genes.

In recent years, following the progress in establishing the genomic DNA sequences of several organisms, a much broader range of approaches have been developed to understand the organization within genomes of large-scale collections of genes and proteins, and to characterize the functional relationships of all expressed transcripts or polypeptides. The genome includes the complete hereditary information of an organism and is encoded in the DNA or RNA. Genomic science is the study of the genomes, which has required substantial improvements particularly in high throughput techniques (DNA sequencing, DNA arrays, mass spectrometry, *etc.*). It provides fundamental information about genome content, organization, function, growth, development, evolution, and the control of gene expression. Genomics encompasses several subfields (Campbell *et al.* 1999). Structural genomics includes the construction of genomic sequence data, gene discovery, and localization and the construction of gene maps. Functional genomics studies the biological function of genes through the analysis of their products. And comparative genomics compares gene or protein sequences from different genomes to elucidate functional and evolutionary relationships. The goals of genomics include compiling the genomic sequences of organisms, establishing the location of all genes in a genome, annotating the gene set in a genome, establishing the functions of all genes in a genome, generating gene expression profiles for cells under differing conditions, and comparing genes across different organisms.

In the pre-genomic era, scientists worked on individual genes or a small set of genes. However, in the post-genomic era, researchers are able to investigate the whole genome or at least a substantial fraction of it. The key to post-genomic science therefore is the generation, analysis, integration and presentation of the large-scale genomic data. This leads to the need for a range of new informatic protocols and approaches to the analysis of sequence and microarray data, the development of efficient tools to deal with these large-

scale data, and the establishment high-quality resources to store data and present results.

One of the biggest challenges in the post-genomic science is to bring an awareness and understanding of how mathematics, computer science, engineering, and statistics play a central role in deciphering the complexities of the genomic science. How should one build up, retrieve and deposit the genomic data? How should one automatically and efficiently process the large-scale dataset? How should one explore, decipher and visualize the content of the data? What are the biological meanings of the data? How should one interpret biological results? How should one compare huge datasets across species? How should one establish static network maps into dynamic mathematical models? Bioinformatics tools for acquiring, storing and analysing biological data are developed to help scientists to find out the answers of these questions.

1.1.2. Bioinformatics

DNA and protein sequences are being collected and deposited in computer data banks that are freely accessible *via* the Internet to researchers all over the world. Bioinformatics is an emerging field linking of biology and computer science to focus on the development of biological databases, computer-search algorithms, gene-prediction software, and other analytical tools to make sense of DNA, RNA, and protein sequences data (Pierce 2002). Major research areas in bioinformatics include sequence alignment, sequence assembly, gene identification, gene expression prediction, protein structure prediction, protein structure alignment, protein-protein interactions, and the modeling of evolution. With the explosion of DNA and protein data available to researchers, bioinformatics has become the key technique both for handling and processing large datasets and complex analytical procedures, and for generating meaningful biological interpretation of the data. A significant amount of effort is being directed at how to warehouse effectively and efficiently and access these data, as well as on new approaches aimed at

analysing, interpreting and displaying these warehouse data in order to make novel biological discoveries.

1.2 Information from sequences

1.2.1 DNA sequencing

A cloned DNA molecule or any DNA, from a clone to a genome, is completely characterized only when its nucleotide sequence is accurately determined. The study of genomics is based on DNA sequencing, which is the process of determining the nucleotide order of a given DNA fragment or sequence. The ability to sequence DNAs has greatly enhanced our understanding of genome organization and our knowledge of genes including structure, function of genes, and the mechanisms of gene regulation.

Modern DNA sequencing began in 1977 with advent of technologies to rapidly decoding DNAs. Walter Gilbert and Allan Maxam's method (Maxam and Gilbert 1977), which was based on the chemical degradation of DNA, involved multiplying, dividing, and carefully fragmenting DNAs. Frederick Sanger's dideoxy sequencing method (Sanger 1977), which was based on the elongation of DNA, used "chain-terminating" or "poison" molecules to reveal precise positions of the bases. Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity, and the Sanger method quickly became the standard procedure for rapid and accurate sequencing of any purified fragment of DNA.

1.2.2 Sequencing entire genomes

A goal of genomics is to determine the ordered nucleotide sequences of entire genomes of organisms, and the industrial scale of DNA sequencing thus lies at the heart of the recently completed genome projects. DNA sequencing in large-scale genome sequencing projects is automated and fast; each instrument being able to generate reads totalling several hundred thousand nucleotides per day. Using a combination of recombinant DNA techniques and DNA sequencing, the genomes of more than 100 prokaryotic species (Klug *et al.* 2005) and several eukaryotic model species have been sequenced, with

hundreds more projects underway. The activity has resulted in very large-scale DNA sequence data that is increasing at a logarithmic rate.

In 1976, Walter Fiers at the University of Ghent was the first to establish the complete nucleotide sequence of a viral RNA-genome (*bacteriophage* MS2) (Fiers *et al.* 1976). The first DNA-genome project to be completed was the bacteriophage phi X174 DNA, with only 5368 base pairs, which was sequenced by Fred Sanger in 1977 (Sanger *et al.* 1977). In 1995, the first completed bacterial genome was that of *Haemophilus influenza,* sequenced by Craig Venter and Claire Fraser of the Institute for Genomic Research (TIGR) and Hamilton Smith of Johns Hopkins University (Fleischmann *et al.* 1995). Yeast (*Saccharomyces cerevisiae*) has been the lead eukaryotic organism in genomics and was the first eukaryote to have its genome fully determined (Mewe *et al.* 1997), followed by the genome of *Eschericia coli* (Selinger *et al.* 2000). The first genome sequence of the multicellular eukaryotes, the nematode *Caenorhabditis elegans* was published in 1998 (C. elegans Sequencing Consortium 1998).

The Human Genome Project (HGP), officially launched in October 1990, aimed to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of all genes. The scale of this project requires the development of novel and automated methods for cloning and sequencing DNA. In 1998, Celera Genomics, a company created by Craig Venter, initiated a private sector effort to sequence the human genome. Both public and private sequencing projects announced the completion of a rough draft of the human genome in June 2000 (Yamey 2000) followed several months later by the completed draft analyses of the human genome (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001). This draft has provided with insight into global characteristics for the human genome. The finished euchromatic sequence of the human genome was reported in 2004 (International Human Genome Sequencing Consortium 2004).

The public and private sequencing projects used different technical approaches: the map-based approach was used by the Human Genome Consortium and the whole genome shotgun (WGS) sequencing approach was

adopted by the Celera Genomics team (Adams *et al.* 2000). Both require breaking the genomic DNA into small overlapping fragments whose DNA sequences can be randomly determined in a sequencing reaction. In the map-based method, completed genetic maps and physical maps of the genome were firstly produced. Genetic mapping, also called linkage mapping, is a way to determine approximate locations of markers relative to the locations of other known markers (normally polymorphic makers such as microsatellites). Physical mapping, based on the direct analysis of the DNA sequence, places genes in relation to each other along physical distances measured in number of base pairs or multiples, kilobases (Kb), or megabases (Mb). Genetic maps and physical maps provide known locations of genetic markers at regularly spaced intervals along each chromosome, then a subset of clones from the physical map is fully sequenced, and finally the overlaps between sequences of the individual clones are used to assemble the entire sequence maps of the genome according to the known order of these clones on the physical maps (Pierce 2002). In the whole-genome shotgun sequencing approach, large-insert clones are not mapped, but small-insert clones are prepared directly from genomic DNA and sequenced without any information on where these clones map in the genome. Powerful computer programmes then assemble the overlapping clones into consensus sequences covering the whole genome. The requirement for overlap in this method means that the genome will be sequenced multiple (often from 10 to 15) times (Pierce 2002).

In 2000, the genome of the fruit fly *Drosophila melanogaster* became the first metazoan genome to have been sequenced by the whole-genome shotgun (WGS) method (Adams *et al*, 2000). In December of 2002, a high-quality draft sequence of the mouse genome was published and the analyses of the genome were reported by the Mouse Genome Sequencing Consortium (MGSC) (Mouse Genome Sequencing Consortium 2002). The WGS was also implemented in genome sequencing of bacteria of *Haemophilus influenzae* Rd. (Fleischmann *et al.* 1995), *Staphylococcus aureus* (Kuroda *et al.* 2001) and pig (Wernersson *et al.* 2005).

1.2.3 cDNAs and EST

In addition to the DNA sequence of an entire genome, other types of sequence data are also useful for genomic projects and have been the focus of sequencing efforts. One consists of the complementary DNA (cDNA), which is essentially a synthetic double-stranded DNA transcribed from a messenger RNA (mRNA) through the action of the enzyme reverse transcriptase. mRNA is the form of ribonucleic acid that directs the production of cellular proteins through the process of translation (McLachlan *et al.* 2004). The molecule of mRNA is relatively fragile and can easily be broken down by the action of enzymes that are prevalent in biological solutions, so researchers commonly manipulate the cDNA that possesses the complementary bases of the mRNA and exists in a more stable state. The full-length cDNA evidence is taken as a gold-standard proof for identification of the sequence of a transcriptional unit, for determination of how it is processed, and for localization of the open reading frame (ORF) it encodes (Griffiths *et al,* 2002). In addition, the techniques for routinely amplifying and purifying individual RNA molecules do not exist, so cDNA sequences are extremely valuable in understanding transcript and polypeptide structure, discovering novel protein-coding genes, and searching for motifs.

Genome sequencing and cDNA sequencing are two approaches for producing sequence data and for improving the annotation of genes (Castelli *et al.* 2004). Genomic DNA sequencing is necessary for cloning entire genes or an entire genome. cDNA sequencing is a suitable approach for seeking specific genes that are active in a specific type of tissue in an organism. In cDNA sequencing, fragments from actively transcribed genes are enriched in the output sequences and introns do not interrupt the cloned sequences. 'Non-model' species generally do not have sufficient funding to allow sequencing of the entire genome. Therefore, cDNA sequencing is widely used to generate sequences of interesting genes without the full cost of a genome sequence. It provides an efficient and effective way to acquire data relevant to expression analysis for non-model species, particularly as unattended informatic

techniques for the automated identification of protein-coding sequences in genome sequences are not completely effective

Expressed Sequences Tags (ESTs) (Adams *et al.* 1991) are short cDNA sequence reads (either the 5' or the 3' ends, or both). The cDNAs are created by isolating RNA from a cell, subjecting it to reverse transcription, producing a set of cDNA fragments that correspond to the expressed RNA molecules (Pierce 2002). ESTs represent the products of active genes in a particular tissue and at the time of sampling the distribution of ESTs in a population of cDNAs indicates the relative abundance of the different transcripts. The identification and analysis of ESTs has formed the basis for gene discovery (Edwards 2007; Peng *et al.* 2007; Xia *et al.* 2007), gene prediction (Wei and Brent 2006; Lu *et al.* 2007) and gene expression studies (Neiman *et al.* 2006; Koutaniemi *et al.* 2007; Tang *et al.* 2007; Wang *et al.* 2007). ESTs can also be aligned with genomic DNA sequence, and thereby used to determine the exon boundaries of the gene and to predict mRNA structure (Griffiths *et al.* 2002). Sequencing the full-length cDNA is more time-consuming than single pass EST sequencing. EST sequences are generated in a single pass, they have a higher error rates than sequences that are verified by multiple sequencing runs (Boguski *et al.* 1993). Although ESTs are not of as high quality as sequences determined by conventional means, they are an excellent source of sequence data. Since the original description of ESTs in 1991 (Adams *et al.* 1991), the growth of ESTs in the public databases has been dramatic. In September 2006, a total of 38.9 million ESTs representing over 1200 different organisms were deposited in the dbEST (Boguski *et al.* 1993). The top seven organisms represented in the EST division (Figure 1.1) were *Homo sapiens* (7.89 million records), *Mus musculus* (4.72 million records), *Oryza sativa* (1.19 million records), *Zea mays* (1.14 million records), *Bos taurus* (1.14 million records), *Danio rerio* (1.13 million records), *Xenopus tropicalis* (1.04 million records). Criticism of EST approaches to gene identification has been their redundancy, and the infrequent representation of ESTs for genes that are rarely expressed.

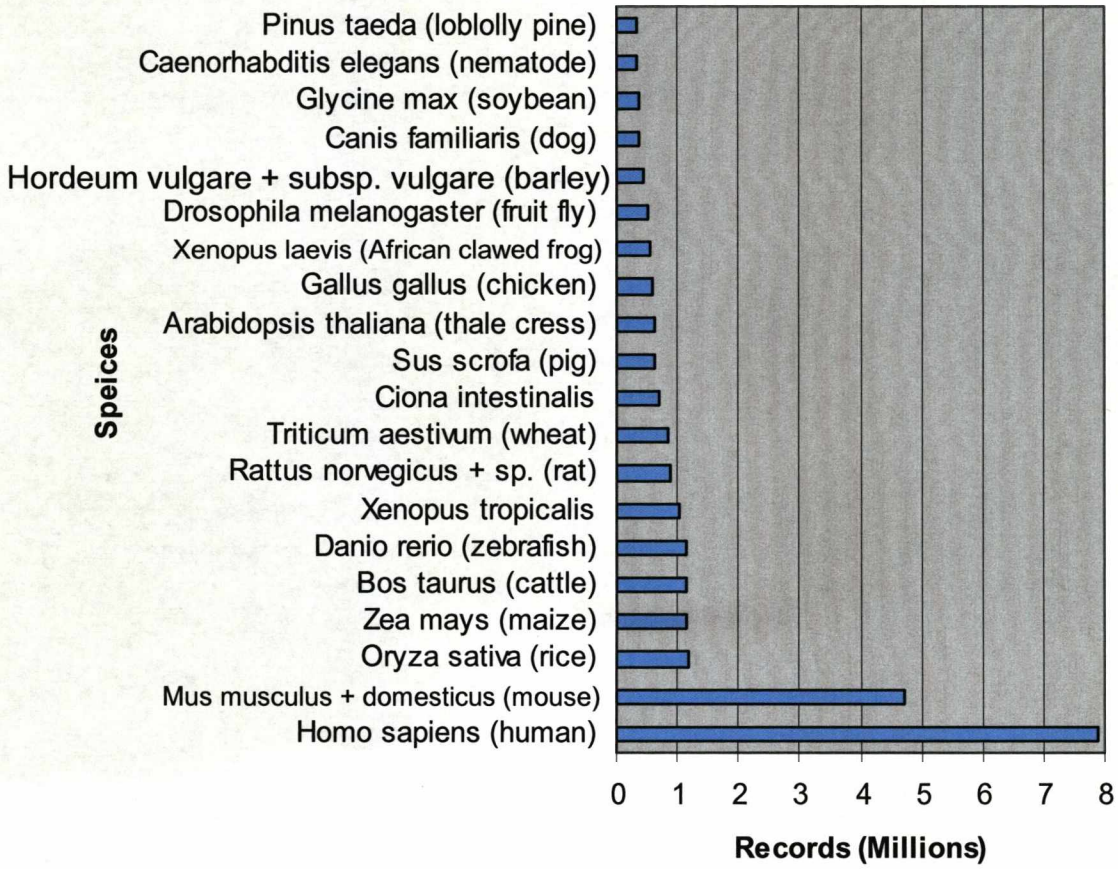


Figure 1.1: ESTs contribution for top 20 species in dbEST. The left side lists the 20 species. The histograms on the right side show the number of records. The data was obtained in September 2006.

1.2.4 Sequence databases

DNA sequencing for genomes, cDNAs, ESTs, *etc*, has established resources of large amounts of sequences for a range of different organisms. To facilitate analysis, this sequence data must be deposited and curated in a way that allows these data to be searched and analysed easily. For this, scientists have put much effort into the design, construction and maintenance of biological sequence databases, such as GenBank (Benson *et al.* 2006), EMBL (Kulikova *et al.* 2007) and dbEST (Boguski *et al.* 1993), dbSTS (Olson *et al.* 1989). These databases have established the standard formats for sequence data storage, and protocols to allow scientists around the world to submit or download sequence data *via* the Internet. GenBank (Benson *et al.* 2006), built and distributed by the National Center for Biotechnology Information (NCBI, USA), is a part of the International Nucleotide Sequence Database Collaboration (INSDC), along with its two partners, the DNA Data Bank of Japan (DDBJ, Mkshima, Japan) (Miyazaki *et al.* 2004) and the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) nucleotide database (Kulikova *et al.* 2007) from the European Bioinformatics Institute (EBI, Hinxton, UK). GenBank incorporated publicly available DNA sequences with supporting bibliographic and biological annotation of more than 205,000 named organisms (Benson *et al.* 2006), obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. GenBank data is accessible through the official website of the NCBI, the EMBL or the DDBJ. For example, Entrez (Wheeler *et al.* 2003) is an integrated retrieval system for sequence data in the NCBI. GenBank is a primary database but not a curated review.

Scientists have analysed these sequence resources and generated databases containing more annotations, such as RefSeq (Pruitt and Maglott 2001), UniProt (Apweiler *et al.* 2004), Ensembl (Hubbard *et al.* 2007), the TIGR gene indices (Quackenbush *et al.* 2001), TIGRFAMs (Haft *et al.* 2003), ProSite (Sigrist *et al.* 2002; Hulo *et al.* 2006), MSD, (Tagari *et al.* 2006) InterPro (Mulder *et al.* 2003; Mulder *et al.* 2007), PDB (Berman *et al.* 2000) and IncAct (Kerrien *et al.* 2007). These curated databases enrich the sequence

data first by removing redundancy and second by providing information validated by expert biologists, ensuring that the data found in these curated collections are highly reliable. In addition, the sequencing projects initiated the sequence analyses of different organisms and sequence databases were also developed to deposit the annotated biological data for those projects, *e.g.* WormBase (Harris *et al.* 2004), FlyBase (FlyBase Consortium 2003), NEMBASE (Parkinson *et al.* 2004), the honeybee EST database (Whitfield *et al.* 2002), the cattle EST database project (Rebeiz and Lewin 2000), the frog EST project (Gilchrist *et al.* 2004), and chicken EST databases (Boardman *et al.* 2002; Hubbard *et al.* 2005; Carre *et al.* 2006). These projects provided resources for the investigations relying on the comparative genomics.

Researchers in different parts of the world can easily use these open accessible genomic resources and tools *via* the Internet. The sharing of the databases accelerates accumulation of data and its analysis. However, different databases possess different levels of data quality. Queries in different databases might generate rather different matches. Therefore, researchers should be careful on selecting sequence databases as references. Databases can be set in a priory order in order to retrieve more confident data from multiple databases.

1.2.5 Sequence analysis

Sequence data are available for downloading from open resources through the Internet and can be produced by sequence machines in laboratories. However, the raw data of sequences are characters representing the nucleotide bases, which in an unprocessed form do not make sense to biologists. To make biological discoveries, computational methods were developed to analyse sequence data.

1.2.5.1 Sequence cleaning and clustering

Sequence raw data contain artificial fragments (e.g. vectors, plasmids and poly-A tails, bad quality fragments and sequences, mitochondrial sequences, *etc*). These fragments or sequences can negatively affect the quality of datasets. Therefore filters are required to remove them. Popular programmes

for sequence cleaning are Cross_match (http://www.phrap.org), LUCY(Chou and Holmes 2001).

Another challenge of sequence analysis is to reduce the sequence redundancy. This requires the clustering and the assembly of the sequence data. Sequence clustering attempts to group sequences based on similarity in sequence alignments. A sequence alignment is a way of arranging sequence to emphasize their regions of similarity, which may indicate functional or evolutionary relationships between sequences. Sequence assembly refers to aligning and merging fragments of a sequence to reconstruct the original sequence. Well-known sequence assemblers are Phrap (http://www.phrap.org), the TIGR Assembler (Sutton *et al.* 1995), and CAP3 (Huang and Madan 1999), GAP4 (Haas 1998), CLOBB (Parkinson *et al.* 2002), the Celera Assembler (Huson 2001), Arachne (Batzoglou 2002; Jaffe 2003) and AMOS (http://amos.sourceforge.net).

1.2.5.2 Sequence annotation

Annotation, a key element of sequence analysis, is a process that identifies genes, their protein products, their regulatory sequences, their structures, and their function(s) (Klug *et al.* 2005). Annotation also identifies non-protein coding genes and finds and characterises mobile genetic elements and repetitive-sequence families present in genomes. Several complete genome projects and other sequence projects have provided well-annotated information of different species. For example, the human (*Homo sapiens*) genome has a size of about 3.4 billion base pairs with about 32,000 genes, and about 1262 annotated protein domain families (International Human Genome Sequencing Consortium 2001); The fly (*Drosophila melanogaster*) is ~180 Mb in size and has ~13,600 genes (Adams *et al.* 2000); The total length of mouse genome was estimated to be about 2.5 Gb, and the number of genes was only in the range of 30,000-40,000 (Mouse Genome Sequencing Consortium 2002). The manual annotation for 60,770 full-length mouse complementary DNA sequences was also reported in 2002 (Okazaki *et al.* 2002). Moreover, secondary databases, such as RefSeq and Swiss-Prot, also provide good quality annotations for their sequences.

One of the specific purposes of sequence annotation is to predict genes and identify their products. Usually, there are two ways to achieve this, (i) signal searching and (ii) homology-based searching. Signal searching is the analysis of sequence signals and sequence motifs, such as exons, which are involved in gene specification. Homology-based searching relies on comparing sequences of interest again known coding sequences to deduce whether the sequences are actually related to one another. Genomic sequences are usually analysed by both signal searching and homology-based searching. cDNAs and ESTs are usually analysed by homology-based searching.

The term homology is different from the term similarity. Similarity is a quantitative measure of how related two sequences are to one another (Baxevanis and Ouellette 2005). Similarity is always based on an observable, usually pairwise alignment of two sequences to one another. High degrees of sequence similarity may imply a common evolutionary history or a possible commonality in biological function. Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character (Fitch 2000). High-level similarity between two sequences could indicate high chance of that they are homolog. Genes either are or are not homologous; homology is not measured in degrees.

Gene homologs can be separated into two classes, orthologs and paralogs (Fitch 1970; Eisen 1998; Gogarten and Olendzenski 1999; Fitch 2000). Orthologs are homologous genes performing the same biological function in different species but derived by historical descent from a single ancestral gene in the last common ancestor of the compared species; in contrast, paralogs are homologous genes within a species, evolving by duplication of an ancestral gene within the lineage (Koonin 2001; Lee *et al.* 2002). In brief, orthology and paralogy differ in that one proceeds from speciation and the other proceeds from gene duplication. There could be only one ortholog in an organism, but this is frequently not the case. There could be more than one ortholog and all can be true (Fitch 2000).

The methods used to assess sequence similarity and homology can be divided into two major types. The global sequence alignment method compares

two sequences along their entire length and provides the best alignment of two sequences across their entire length. The local sequence alignment method compares two sequences and intends to find out the most similar regions in two sequences being aligned. In general, the global alignment method is most applicable to highly similar sequences of approximately the same length. The local alignment method is most capable of finding subsequences within the sequences being compared that may have biological relationship. The local alignment method is best for sequences that share some degree of similarity or for sequences of different lengths. The most popular software using the local sequence alignment method is BLAST (Altschul *et al.* 1997). Others are FASTA (Pearson 2000), BLAT (Kent 2002), *etc.*

Another purpose of sequence annotation is to understand biological roles and molecular functions which genes and their products play in the cell. Popular databases of terms used to annotate genes can be found in the Gene Ontology (GO) project (Gene Ontology Consortium 2004), the Enzyme Commission (EC) database (Bairoch 2000), the KEGG pathways (Kanehisa *et al.* 2002), *etc.* GO offers a controlled vocabulary (an ontology) of terms in 3 separate domains: biological process, cellular component and molecular function. The Enzyme Commission database and the KEGG pathways provide information on enzymes and their metabolism pathways.

1.3 Information from gene expression

Functional genomics is the study of the expression and interaction of gene products of the genome (Griffiths *et al.* 2002). The goals of functional genomics include identifying all mRNA molecules transcribed from a genome, as a means of identifying expressed proteins (Pierce 2002). Therefore, 'transcriptomics' and 'proteomics' are the two outgrowths of genomics. Proteomics, not detailed in this PhD thesis, is the study of the proteome, a set of all proteins encoded by a genome and present at a given time under a given set of conditions; Transcriptomics is the study of the transcriptome, a set of all RNA molecules transcribed from a genome and produced at any given time (Klug *et al.* 2005). Unlike the genome, which is fixed for a given cell line, the

transcriptome can vary with external environmental conditions. Many important clues about gene functions come from knowing when and where the genes are expressed. The transcriptome reflects the genes that are expressed at any given time, as it includes all mRNA transcripts in the cell. Transcriptomics examines the expression level of mRNAs in a given cell population, often using high-throughput techniques based on DNA microarray technology.

mRNA is the form of ribonucleic acid that directs the production of cellular proteins. The process by which the genetic information carried in DNAs is coded into mRNAs and proteins is called gene expression. Since mRNA acts as the intermediary between DNA and protein, measuring mRNA levels in the cell can be used to indirectly infer the amount and the kind of proteins the cell is producing. mRNA measurement is used to estimate cellular changes in response to external signals, specific stimuli or environmental changes. Biologists are interested in testing expression patterns of mRNAs and want to observe what cellular proteins are produced and what functions those proteins play in particular types of tissues or in response to specific external stimuli. However, it is not mRNA which is measured, but a DNA copy of the mRNA known as cDNA that is actually measured. As already described, cDNAs are more stable and easier to be measured than mRNAs.

1.3.1 DNA Microarrays

Common methods used for high throughput measurements of gene expression levels are divided into two general categories: digital and analog. The digital methods, such as the Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.* 1995), were based on the generation of sequence tags (Audic and Claveris 1997). SAGE involves isolations of short unique sequence tags from a specific location within each transcript. These sequence tags are concatenated, cloned, and sequenced (Baxevanis and Ouellette 2005). The relative abundance of sequenced tags in a sample is then analysed to represent the level of gene transcript expressions in the sample. The SAGE technique runs a high risk of error when two or more genes share the same tag and when a gene has more than one tag. Also SAGE can only be effective in a sequenced

organism when gene identity can be established by comparing to tags with the know gene sequences.

The analog methods, such as DNA microarray, are based on sample hybridization to cDNA clones or oligonucleotides on arrays. The use of microarrays for gene expression profiling was first published in 1995 (Schena *et al.* 1995). A microarray is a very small, two-dimensional array, typically on a glass, filter, or a silicon wafer, upon which thousands of gene-specific probes are immobilized on a matrix (Knudsen 2004; McLachlan *et al.* 2004; Baxevanis and Ouellette 2005). The matrix acts as a parallel set of probes to detect the abundance of transcription mixture with labelled nucleic acid synthesized from a tissue type, developmental stage, or other condition of interest. The expression profiles of thousands of genes under that condition can thus be assayed simultaneously. In principle, all the genes in an organism's genome can be represented on a microarray, so the expressions of all genes can be assessed at one time. The DNA microarray technology is much more appropriate than SAGE in the analysis of gene expressions for large numbers of samples.

Microarrays can be used to determine which RNA and DNA sequences are present in a mixture of nucleic acids, and to examine changes of mRNA expression in contrastive tissues, conditions or states. By virtue of their scale, these technologies benefit biological research greatly and further our understanding of biological processes, gene regulation and molecule interactions.

There are two major kinds for high-density DNA microarray in common use: two-color cDNA or oligo arrays, also known as spotted arrays, and high-density oligonucleotide arrays (oligo arrays), also known as gene chips. cDNA arrays utilizes robotic deposition or "spotting" of DNA molecules, while oligonucleotide arrays involves oligonucleotides made by a photolithographic process similar to manufacture of computer chip (Hughes *et al.* 2001).

1.3.1.1 cDNA microarrays

Two-color arrays are produced by placing cDNA sequences in spots, each representing a different gene on a surface such as a microscope slide. RNA transcripts are obtained from two samples, such as a healthy individual and a diseased individual, and reversely transcribed into cDNA and separately labeled with two different dyes (green and red). And the two labelled populations are then combined and hybridized to the array. The cDNAs will hybridize to their complementary spots on the slide and the red/green fluorescence on each spot is determined with a laser scanning microscope. Spots which are either green or red indicate significantly higher expression levels in one sample relative to other, whilst yellow spots indicate similar level of expression in both samples.

1.3.1.2 Oligonucleotide arrays

Oligonucleotides are short fragments, ranging from 15 to 70 bases in length, taken from the hundreds of nucleotides in a DNA segment that function as a gene (Aitman 2001; Jordan 2002). For oligonucleotide arrays (or gene chips), oligos are synthesized and then spotted onto the chip, or they can be synthesized directly on the chip (*in situ* or *in silico*) through a process of photolithography (McLachlan *et al.* 2004). A single RNA sample is reverse transcribed into cDNA, hybridized to the array, and imaged with a scanner. It is common to include some probes which contain a mismatched base in the oligo, and the mismatched values are subtracted from the perfect match reading, and these differences are summarised from probes at intervals along entire probe for a given gene to produce a single measurement (Nielsen *et al.* 2003).

Researchers consider issues such as specificity and efficiency of hybridization, and accuracy and reproducibility of resulting gene transcript expression levels when assessing the advantage and disadvantage of using oligonucleotide arrays versus cDNA arrays. Probe oligos may be more accessible for hybridization than the probe cDNA strands, due to their much shorter chains with single terminal points for attachment to the slide or chip. Oligonucleotide arrays are also used to detect a sub-region of a gene, which is

a valuable tool when there is a family of gene with a high similarity of sequence. Oligoarrays offer greater specificity than cDNAs or PCR products, having the capacity to distinguish single nucleotide polymorphisms and discern splice variants. Additionally, having uniform lengths for the oligo probes enhances the chance of finding optimal hybridization, and it is easier to engineer.

1.3.1.3 Genomic tiling microarray

Recent genomic tiling array experiments have shown evidence of large amount of transcription outside the boundaries of known genes (Selinger *et al.* 2000). The tiling arrays assay transcription at regular intervals throughout the genome without bias towards the location of known and predicted genes. The design of tiling arrays is not dependent on current genome annotations and thus enables rare transcripts to be detected (Johnson *et al.* 2005). Genomic tiling using microarrays becomes an important complement to other efforts to determine the transcriptome. However, the technique requires a completely sequenced genome and thus is only used in the studies of model species currently.

1.3.2 Analysing microarray data

1.3.2.1 Image processing

The results obtained from DNA microarray experiments are usually stored as image files, such as TIFF images. The first step in extracting information from a microarray experiment is image processing. This is carried out to identify the relative fluorescence intensity of each of the features on the array. Images are converted to numerical values and this process is known as quantification. Each spot on the microarray contains two numerical components known as signal and background. Signal values correspond with true intensity data while background values correspond to intensity values unrelated to the binding of target cDNA. Separating signal from background is an important step in the quantification process. Quantification can be absolute (signal intensity) or relative (ratio of absolute signals in two samples). Thus

researchers distinguish DNA microarray data analysis into primarily two types: (1) one-channel DNA data that reflect absolute intensities; versus (2) two-channel DNA data that represent relative intensities or ratio data (McLachlan *et al.* 2004).

1.3.2.2 Normalising expression measurements

Technological problems and biological variation make it difficult to distinguish signal from noise in the analysis of gene expression. Microarrays are usually applied to the comparison of gene expression profiles under different conditions. Normalisation of the measured expression level or ratio values adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparisons can be made. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescence dyes used, or systematic biases in the measured expression levels (Baxevanis and Ouellette 2005).

1.3.2.3 Expression clustering

Data analysis for one experiment and a control will limit itself to a list of regulated genes ranked by the magnitude of up- and down-regulation, or ranked by the significance of regulation determined. Data analysis for more experiments to measure the same genes under different conditions, in different mutants, or at different time points makes sense to group the significantly changed genes into clusters that behave similarly over the different conditions.

One of the most widely used clustering approaches is hierarchical clustering, which determine the relationships based on the Euclidean distances between the respective data points. Hierarchical clustering is an agglomerative approach in which single expression profiles are joined to nodes, which are further joined until the process has been carried to completion, forming a single hierarchical tree.

K-means clustering (McLachlan *et al.* 2004) is another widely used approach in gene expression analysis. The user has to define the number (*K*) of

output clusters before the clustering procedure. The algorithm initiates the clustering by randomly selecting K seeds from the observation data points as the cluster centres. Then it takes each observation data point and associate it to the nearest seed based on its distances to the seeds. All observations are in turn assigned to each of the K clusters and the first loop of clustering is completed. Then the centre for each clustering is defined as a new seed from the first loop of clustering. The new seeds could be different to the old seeds in the previous loop. After that, the algorithm runs the next loop of clustering by re-associate each observation data point to the nearest new seed. The loop of the clustering continues until the seeds and the cluster memberships do not change anymore. Once the loop is terminated, the K-means clustering map is built.

To cluster the microarray data, distances or correlations between gene expressions are usually required. There are a number of other ways to calculate distance between two genes, such as the Euclidean distance, the Pearson correlation coefficient and the Spearman correlation coefficient. Euclidean distance is able to measure the absolute level of gene regulation. Two genes whose expression levels were perfectly parallel to one another across the data points could still be far apart in the Euclidean space if the absolute levels in each experiment were different. However, the Pearson correlation coefficient across data points of genes is able to measure the relative shape of the gene regulations rather than the absolute levels (Kim *et al.* 2001). The Spearman correlation uses ranks rather than raw expression levels which makes it less sensitive to extreme values in gene expression data. Therefore, the Pearson correlation is a natural choice to measure gene correlations.

1.4. Orthology analysis for comparative genomics

Comparative genomics requires analysis across species. Cross-referencing the available genomic data has several important applications, including the identification of homologous genes in eukaryotes. The difference between for ortholog, paralog and homolog were described in the Section 1.2.5.2. The identification of orthologous groups is particularly important because it is assumed that such genes play similar biological, developmental or

physiological roles and consequently, should share conserved functional and regulatory domains. Orthologous groups are also helpful for transferring functional information between genes in different organisms with a high degree of reliability. The analyses of orthologs are useful for sequence identifications, genome annotation, gene/protein evolution studies, and comparative genomics.

The notions of orthology and paralogy are intimately linked because, if duplication occurred after the speciation event that separated the compared species, orthology becomes a relationship between sets of paralogs (co-orthologs), rather than individual genes. Orthology analysis between species is often complex because of large numbers of paralogs within some protein families. Orthologs typically occupy the same functional niche in different species, whereas paralogs tend to evolve toward functional diversification. The common carp (*Cyprinus carpio*) has been considered tetraploid because of its chromosome number (2n = 100) and its high DNA content (Ohno *et al.* 1967). The analyses of microsatellite loci (David *et al.* 2003) also suggests that the common carp is tetraploid and that polyploidy occurred by hybridization. The ortholog analysis between common carp and other species is complex but will provide opportunities to explore the insight of common carp gene duplications and evolutions of its large gene families.

Analysis of gene expressions in model organisms, particularly human and mouse, has become a fundamental reference for the studies of gene expressions in non-model species, such as common carp. Co-expression of orthologous genes between model species and non-model species help to discover conserved gene modules in the evolution and illustrate differences and similarities of expressions for conserved gene modules between different organisms (Stuart *et al.* 2003). This leads to the combined use of data from genes with orthologous relationships and their expression.

1.5 Aims of investigation

The post-genomic screening of DNA and protein expression profiles depends critically upon the ready availability of basic DNA sequence information for the species of choice. In the case of the 'genomic' model

species this is provided by the genome sequencing projects and unattended identification of genes within the genomic DNA, combined with the production and analysis of hundreds of thousands of EST sequences. Altogether these allow prediction of a large fraction if not all of the expressed transcripts and proteins with a high degree of confidence.

The model species attract strong research communities with substantial and highly focused research funding. However, non-model species generally have insufficient data in sequences and annotation to allow genome-wide investigation. It is necessary to generate sufficient resource to enable a genomic approach. Recent work suggests that this is achievable. For example, the hypoxia-induced gene expression profiling in fish *Gillichthys mirabilis* was studied by Andrew Y. Gracey (Gracey *et al.* 2001), based on production of 1700 cDNA clones. The NERC-funded carp genomics project used the same approach to generate over 13,500 cDNA clones, all of which needed to be EST characterized and functionally annotated.

The ultimate goal of this thesis is to produce non-model species genomic resources, to develop novel analysis approaches, and create new informatics tools for understanding the mechanisms underlying environmental responses and adaptation in common carp responding to environmental challenge. Given that in the year 2001 the gene data banks contained a few hundred cDNA sequences it was first necessary to construct common carp cDNA libraries, to clone large numbers of genes and to characterise these by means of EST sequencing. These resources would be used to characterise the expression profile of each gene in animals subjected to different stressors, thereby to provide a meaningful annotation of gene functions.

This study started by identifying a collection of end-sequence characterized cDNAs by BLAST homology searches and organising them into a searchable database. A number of other smaller-scale EST projects were also underway providing additional collections ranging from a just few hundred to a few thousand EST sequences and given this small number there is a particular need to maximise proportion of clones that are identified and annotated. Identification of these ESTs through sequence alignment algorithms relies

heavily upon the more complete identification and functional annotation achieved for the genomic model species. The reliability of these identifications is heavily dependent on the methods used and the sequence relationships between the species of choice and the most appropriate genomic model.

The first aim of this study was the provision of an integrated suite of informatics tools to annotate the common carp ESTs, the main outcome being an EST annotation package, EST-ferret, described in Chapter 2. EST-ferret is a configurable software package that can automatically analyse and annotate ESTs of the non-model species by integrating suitable analysis tools and data resources. Chapter 2 also describes the construction and the applications of the BioCluster, which is a parallel computer grid to speed up bioinformatics programmes in processing biological data of increasing size.

The result from EST data analysis requires a user-friendly interface as a resource for other researchers to gain access to the gene data. Chapter 3 describes a new database, carpBASE, for the results of common carp ESTs analysis. It was built on a Linux server with technologies of the MySQL database (http://www.mysql.com), the Apache HTTPD server (http://httpd.apache.org/) and the PHP Hypertext preprocessor (http://www.php.net) (Petersen 2002).

Combining expression data and sequence data can help biologists to make more biological discoveries. Another aim of the PhD study was the creation of new approaches to understand biological meanings by combining these two kinds of data. Chapter 4 details two new approaches, ExprAlign and GOmatrix. ExprAlign computes the Pearson correlation coefficients of gene expressions, then build up a landscape to visualize relationships of gene expression. This approach was used to suggest identities for microarray probes lacking any meaningful BLAST identity. It was also used to identify groups of co-regulated genes in carp adjusting to stressful situations. The GOmatrix identifies which gene expression groups are over- or under-represented in particular Gene Ontology categories. An important feature of contemporary genomic research is the ability to define conserved and evolving relationships in gene expression properties of different species. There are significant

problems in establishing the appropriate orthologous relationship for specific genes, and these need resolving before the comparative, evolutionary approach can be properly implemented. Chapter 5 describes orthologous genes and their gene co-expressions across common carp and human. Zebrafish was taken as a bridge species connecting common carp and human. This investigation is also helpful for the further analysis on genome duplication of common carp.

The last aim of the PhD was the initiation of constructing rainbow trout oligonucleotide arrays. Chapter 6 details the optimization of the sequence dataset for the trout oligoarray design. The protocol to optimize sequence datasets for oligoarrays was defined in the project.

The final chapter, Chapter 7, draws the conclusions on the PhD project.

CHAPTER 2: DEVELOPING TOOLS FOR ESTS ANALYSIS

2.1. Introduction

2.1.1 Background of ESTs

Expressed Sequences Tags (ESTs) are short cDNA sequence reads, obtained by isolating mRNA from cells and tissues and subjecting it to reverse transcription, producing a set of cDNA fragments that correspond to expressed RNA molecules (Pierce et al, 2002). ESTs can be obtained from the 5' or the 3' ends of directionally cloned cDNAs, or both ends from randomly cloned cDNAs.

EST sequences are generated by a single pass sequencing run, so they have a higher error rate than full length mRNA or genomic DNA sequences that are verified by multiple sequencing runs. Despite their fragmentary and inaccurate nature, ESTs are an excellent source of sequence data. They were originally intended as a way to identify gene transcripts (Adams *et al.* 1991), but have been instrumental in gene discovery and sequence determination (Sutton *et al.* 1995). ESTs are also proved to be useful resources for the annotation of genomes (Haas *et al.* 2002; Haas *et al.* 2003), for designing probes for DNA microarrays (Antipova 2002), for determining the boundaries of the transcript, and for predicting mRNA structure (Griffiths *et al.* 2002). ESTs analyses are now widely applied through genomics and molecular biology communities.

The benefits arising from the rapid generation of large numbers of cDNA sequences were not universally recognized when the cDNA concept was originally proposed (Iyer and Szybalski 1963). Since the initial demonstration of the utility and the cost effectiveness of the original EST approach described by M. Adams in 1991 (Adams *et al.* 1991), many sequencing centres have automated the processing of EST generation and the growth of ESTs in the public databases has been dramatic. Large-scale EST projects have been launched for several organisms of experimental interest, resulting in an ever-increasing number of ESTs (Okubo *et al.* 1992; Adams *et al.* 1993; Hillier *et al.* 1996; Krizman *et al.* 1999; Boardman *et al.* 2002; Whitfield *et al.* 2002;

Clark *et al.* 2003; Palmer *et al.* 2003; Cogburn *et al.* 2004; Kimura *et al.* 2004; Rise *et al.* 2004; Hubbard *et al.* 2005; Carre *et al.* 2006). In 1992, a database called dbEST (http://www.ncbi.nlm.nih.gov/dbEST) (Boguski *et al.* 1993) was established in the NCBI (http://www.ncbi.nlm.nih.gov/) to serve as a collection point for ESTs, which are then distributed to the scientific community as the EST division of GenBank (Benson *et al.* 2006). ESTs are submitted to dbEST firstly, and then the three standard international sequence databases (GenBank, EMBL and DDBJ) (Miyazaki *et al.* 2004; Kulikova *et al.* 2007) will include the entries of the ESTs under the data-sharing agreement. Therefore, all ESTs can be accessed through GenBank, EMBL or DDBJ, regardless of where the sequence was originally submitted. The number of the ESTs in dbEST increased dramatically in the past decade. In July 2007, a total of 44.2 million ESTs representing over 1355 different organisms were deposited in dbEST.

2.1.2 EST analysis of UniGene and the TIGR Gene Indices

The major challenges for ESTs investigations are to correct for the presence of redundant EST data, to make putative gene assignments for the data, and to discover new biological insights arising from the data. The criticism of ESTs in gene libraries for most organisms has been due primarily to redundancies, and an absence of genes that are rarely expressed. Computationally, this can be thought of as a clustering or assembly problem in which the sequences are vertices that may be coalesced into clusters by establishing connections among them. Sequences can be clustered on the basis of overlapping bases and then assembled into a consensus sequence and much effort has expended at reducing the number of ESTs by grouping together records that likely derive from the same gene.

The NCBI also identifies through BLAST alignment searches all homologies for new EST sequences and incorporates that information into the dbEST. The data in dbEST is further processed in the NCBI to produce the UniGene database (Wheeler *et al.* 2003) of gene-oriented sequence clusters. UniGene is a system for automatically partitioning ESTs and other mRNA sequences, along with coding sequences (CDS) annotated on genomic DNA,

into a non-redundant set of gene-oriented clusters. UniGene has clustered ESTs from 10 animal and 7 plant species by 2003 and expanded them to over 70 species by July 2006. UniGene starts with entries in the appropriate organism division of GenBank, combines these with ESTs of that organism and creates clusters of sequences that share virtually identical 3' untranslated regions (3' UTRs). Each UniGene cluster contains sequences that represent a unique gene, and is linked to related information, such as the tissue types in which the gene is expressed, model organism protein similarities. In the human UniGene database, over 3.6 million human ESTs in GenBank have been reduced 35-fold in number to just about 104,000 sequence clusters (Wheeler *et al.* 2003). The UniGene collection has been used as a source of unique sequence for the fabrication of microarrays for the large-scale study of gene expression of model species.

The Gene Indices (http://compbio.dfci.harvard.edu/tgi/), original called TIGR Gene Indices, present another automated system to cluster ESTs and other annotated gene sequences (Quackenbush *et al.* 2001). They are a collection of species-specific databases that use a highly refined protocol to analyse ESTs in an attempt to identify the genes represented by that data and to provide additional annotations regarding those genes. They are constructed by first clustering, then assembling EST and annotated gene sequences from GenBank for the targeted species. This process produces a set of unique, high-fidelity virtual transcripts or Tentative Consensus (TC) sequences. The TC sequences can be used to provide putative genes with functional annotation and to link the transcripts to mapping and genomic sequence data.

UniGene and the TIGR Gene Indices have different protocols to clear, cluster and annotate the sequences. One major difference between these two is that UniGene does not produce the assembled contig for each group, while the TIGR Gene Indices output the contig to represent the sequences in each group. A contig is a continuous sequence of DNA that has been assembled by the alignment of overlapping cloned DNA fragments.

2.1.3 Aims of the EST-ferret project

UniGene and the TIGR Gene Indices provide clustering information for the different species, which are mostly model species and have large-scale public EST resources available. However, we were interested in genes of non-model species, such as common carp, as an essential resource for a microarray approach to understanding environmental/stress responses, but the public sequence resources in 2001 when this project was initiated were insufficient. Both UniGene and the TIGR Gene Indices do not contain information for common carp, and only about 2000 common carp sequences were available in GenBank. So there was a need generate several thousand EST reads from cloned cDNAs. Colleagues thus produced thousands of EST sequences of common carp for our investigation.

UniGene and the TIGR Gene Indices have their own protocols and systems to analyse and establish their data. The routines used in their protocols were suitable for ESTs analysis on model-species which had large-scale EST data. A number of other bioinformatics software solutions have also been developed to automatically clean, cluster, assemble and annotate raw ESTs such as PipeOnline 2.0 (Ayoubi *et al.* 2002), ESTAP (Mao *et al.* 2003), ESTWeb (Paquola 2003), ESTAnnotator (Hotz-Wagenblatt *et al.* 2003), ESTprep (Scheetz 2003), EST Pipeline System (Xu 2003), PartiGene (Parkinson *et al.* 2004) and parpEST (D'Agostino 2005). These software packages are pipeline systems, where EST sequences are passed through different third party bioinformatics programs and searched against different sequence databases. Further details of these packages are discussed in the Discussion section in this chapter. Different criteria, cluster algorithm and resources for cleaning, clustering, annotating and storing sequences data were required in our projects on non-model species. Our project was initiated in 2001 when most of other packages were less matured or unavailable. Therefore, we developed a suite of bioinformatics tools to analyse the carp EST data.

2.1.3.1 Building EST-ferret

A key aspect of this project was the provision of an integrated suite of informatics tools and resources for the convenient analysis of EST data. The first tool is the production of an EST annotating package, called --- "EST-ferret". EST-ferret was designed to have the following major features:

- Integrating suitable bioinformatics analysis tools and biological data resources
- Convenient to ESTs analysis for non-model species
- Containing components of sequence processing, sequence clustering, sequence annotating and result storing
- Portable to other laboratories
- Easy to install with high-level computer knowledge
- User-configurable
- Running automatically

2.1.3.2 Developing GOprofiler

Gene Ontology (Gene Ontology Consortium 2004) annotation system is a fundamental and increasingly powerful tool for describing the biological role of genes. A second tool developed here is a Gene Ontology annotation programme, called GOprofiler. This was designed to assign the Gene Ontology terms and its IDs to each gene in a gene list, and to associate genes to different Gene Ontology sub-categories. GOprofiler was designed to be integrated into EST-ferret.

2.1.3.3 Building BioCluster

Several important bioinformatics programmes, such as BLAST, CDD (Marchler-Bauer *et al.* 2003; Marchler-Bauer *et al.* 2005), can be run on the NCBI BLAST website, on a local standalone computer or on a computer cluster with parallel capabilities. With the dramatic growth of sequence data, running BLAST searches in a standalone computer for thousands of genes is time-consuming since it could take days or weeks to finish the large jobs. By contrast, running BLAST searches in parallel using multiple computers for the

same job can save much time. Thus the third element of the carp EST project was to implement a powerful parallel computer system --- 'BioCluster' --- for using BLAST, CDD, etc., for speeding up the biological data analysis. The BioCluster is a computer grid that has powerful parallel capability for running programmes for analysis and deposition of EST data. The BioCluster should have the following major features:

- Powerful computing capability
- Friendly user-interface for local or external access
- Running bioinformatics tools automatically
- Accessing result easily

2.2. Materials and Methods

2.2.1 Selecting external software and data resources

To develop the EST-ferret package for EST analysis, external bioinformatics tools and data resources were selected for integration into the package.

2.2.1.1 Base-calling and sequence cleaning

Trace data from sequencing machines are stored in chromatogram files, in ABI format or Standard Chromatogram Format (SCF) (Dear and Staden 1992), and are usually displayed in the form of chromatograms consisting of four curves of different colors, each curve representing the signal for one of the four bases (A, T, C, G) and drawn left to right in the direction of increasing time to detection. To analyse the sequences, base-calling must be performed to convert the traces into sequences of bases together with a quality score for each base. The Phred base-caller (Ewing and Green 1998; Ewing *et al.* 1998) was adopted to automate this process. It uses a four-phase procedure to determine a sequence of base-calls from the processed and the entire procedure is rapid on typical computer workstations. The Phred appears to be the first base-calling program to achieve a lower error rate than the ABI software, averaging 40%–50% fewer errors in the data sets examined independent of position in read, machine running conditions, or sequencing chemistry. The output of the base-calling are sequence files and quality files in the standard FASTA format, which consists of a header line followed by the sequence bases or quality scores. The header line begins with a '>' and contains a name and/or a unique identifier for the sequence, and often lots of other information too.

To clone and manipulate DNA/cDNAs from the biological source, researchers usually insert a cloning vector, such as plasmid, phage, cosmid, BAC, YAC, into the DNA/cDNA (Amemiya *et al.* 1999). It is important to eliminate low-quality or apparently artificial sequences (e.g. poly-A tails) before clustering because even a small level of noise can have a large corrupting effect on a result. Failure to identify and remove all of the vector sequence results in a finished sequence that is contaminated. Thus, procedures

are introduced to eliminate sequences of foreign origin and identify regions that are derived from the cloning vector or artificial primers or linkers. LUCY (Chou and Holmes 2001) is a programme for DNA sequence quality trimming and vector removal which provides flexible parameters for cleaning the sequences. But LUCY does not trim quality files properly, which are very important for sequence clustering. Also it has no ability to select the best representative sequence from a set of duplicated sequences for the same clone. Another programme, called Cross_match (http://www.phrap.org/), is an efficient implementation of the Smith-Waterman-Gotoh algorithm. It is able to mask vector sequence segment in EST data. EST-ferret integrates Cross_match to mask vector sequences. PERL scripts were developed to trim vector, poly-A tails and low quality regions from the sequence, to eliminate low quality sequences and mitochondrial sequences, and also to revise quality files for presentation of final clean sequences.

2.2.1.2 Sequence clustering and assembly

To identify unique genes from ESTs, clustering places EST reads into different unique gene groups, and EST assembly generates consensus sequence for each group. ESTs clustering is important in grouping sequences that originate from the same gene before ESTs are assembled to reconstruct the original mRNA. The quality and utility of the assembled sequences relies on the capability of the sequence clustering or assembly programmes to generate high fidelity consensus sequences from the ESTs. Many programmes to cluster and assemble EST data have been generated. The three major programmes for this are CAP3 (Huang and Madan 1999), Phrap (http://www.phrap.org) and the TIGR Assembler (Sutton *et al.* 1995). F. Liang compared these three assembly programmes and found that none of them performed perfectly (Liang *et al.* 2000). TIGR Assembler proved slightly more sensitive to subtle yet consistent differences in sequence, such as those present in closely related members of a gene family. However, this sensitivity, combined with the naturally occurring errors inherent in ESTs, causes both to split transcripts, generating an over-representation of clusters for some genes. One the other hand, Phrap is

insufficiently sensitive to sequence differences, causing it to over-assemble ESTs and sacrifice the fidelity of the consensus sequences in produces by generating a significantly higher number of insertions and incorrect base assignments. Liang's study indicated that CAP3 incorporated the best features of these other programs and was able to produce high fidelity consensus sequences and maintain a high level of sensitivity to gene family members while effectively handling sequencing errors (Liang *et al.* 2000). Based on these, CAP3 was selected to cluster our ESTs into sequence groups and assemble into sequence consensuses.

2.2.1.3 Sequence annotating

BLAST

To discover biological meanings of ESTs, their gene names or their protein products need to be identified. The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1997), one of the most widely used computational methods to conduct searching for sequence similarities for identifying the sequences, was implemented in EST-ferret and the BioCluster project. BLAST is capable of detecting not only the best region of the local alignment between a query sequence and its target, but also whether there are other plausible alignments between the query and the target. The original, standard family of BLAST programs contains BLASTN (nucleotide query sequences against nucleotide subject sequences), BLASTP (protein query sequences against protein subject sequences), BLASTX (nucleotide query sequences translated in all reading frames against protein subject sequences), TBLASTN (protein query sequences against nucleotide subject sequences translated in all reading frames) and TBLASTX (six-frame translations of nucleotide query sequences against six-frame translations of nucleotide subject sequences).

BLAST search can be conducted on the BLAST web servers *via* the Internet or in a local BLAST server. The most widely used portal for these searches is the BLAST home page (http://www.ncbi.nlm.nih.gov/BLAST) at the NCBI. The BLAST programme and the BLAST databases can be

downloaded and then installed in local computers. Running a local BLAST programme provides flexibility on searching against any sequence databases. It is also not affected by the speed limits of the Internet and the web servers. BLAST searching is a parallel algorithm, Therefore it can be implemented in a computer grid to speed up the running in our BioCluster project.

Public, well-annotated sequence databases, such as Swiss-Prot (Boeckmann *et al.* 2003), RefSeq (Pruitt and Maglott 2001) databases for different vertebrate species, were integrated in EST-ferret. The RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq), developed by the NCBI, is a curated secondary database providing a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcripts, and protein products, for over 4700 other organisms (November 2007). RefSeq presents an important effort to curate all of the sequences, but it is currently limited in scope and scale because of the labour-intensive activity it represents.

Swiss-Prot, a major part of the UniProt database (Apweiler *et al.* 2004), contains manually annotated records, based on information from the literature-based curation, together with a curator-evaluated computational analysis. The annotated records describe the properties of the protein, such as its function, any known post-translational modifications, domains, catalytic or other sites, secondary and quaternary structures, similarities to other proteins, diseases caused by mutations in the protein, pathways in which the protein is involved, sequence conflicts, and variants. It is clearly a highly labour-intensive process for producing a fully curated Swiss-Prot entry. Swiss-Prot is obviously a valuable protein sequence database for biological research.

Gene Ontology

A BLAST identity usually does not provide the details on gene functions. However, the Gene Ontology (GO) project (Gene Ontology Consortium 2004) offers a structured, controlled vocabulary and classifications that cover several domains of molecular and cellular biology and is freely available for community use in the annotation of genes, gene products and sequences (http://www.geneontology.org/). It describes the attributes of gene

products in three non-overlapping domains of molecular biology; Molecular Function describes functional activities, such as catalytic or binding activities at the molecular level; Biological Process describes biological goals accomplished by one or more ordered assemblies of molecular functions; and Cellular Component describes locations, at the levels of sub-cellular structures and macromolecular complexes. The GO information gives scientists more opportunities to make biological discoveries and is an important and useful recourse for annotating sequence data. Therefore, GO was also included in EST-ferret.

Enzyme analysis

Enzyme information is helpful to understand the roles of the genes in different metabolic pathways. Thus the ENZYME database (Bairoch 2000) was also implemented into the EST-ferret. The ENZYME database (http://www.expasy.ch/enzyme/) is a repository of information related to the nomenclature of enzymes, primarily based on the recommendations of the Nomenclature Committee or the International Union of Biochemistry and Molecular Biology (IUBMB). It contains data, such as EC (Enzyme Commission) numbers, for each type of characterized enzyme. The EC numbers are helpful in the development of computer software involved in the manipulation of metabolic pathways.

Protein Domain searching

Sequences might represent protein domains, which are distinct units of protein three-dimensional structure, carrying functions. Proteins domains may be treated as building blocks of structure and function, dividing the primary and tertiary structure of a chain into distinct units (Marchler-Bauer *et al.* 2002). Domains are also mobile genetic units, rearranging in various combinations throughout the molecular evolution of proteins. Proteins can be composed of single or multiple domains. The annotation for locating and identifying conserved protein domains has become an indispensable tool in the analysis of genes and genomes and provides valuable insights into the molecular evolution

of single- and multiple-domain proteins, as well as help to validate other annotation. Secondary databases were used for further analysis in order to annotate a larger number of EST sequences, not previously annotated by BLAST.

Two sets of secondary protein databases, InterPro (Mulder *et al.* 2007), and the Conserved Domain Database (CDD) (Marchler-Bauer *et al.* 2005) were integrated into EST-ferret. InterPro is an integrated documentation resource of protein families, domains and functional sites. CDD is the protein classification component, starting out as essentially a mirror of publicly available domain alignment collections, such as SMART (Letunic *et al.* 2004), Pfam (Bateman *et al.* 2004) and COG (Tatusov *et al.* 2003). CDD converts these alignment models into searchable databases of Position Specific Score Matrices (PSSMs) derived from CDD alignments (Marchler-Bauer *et al.* 2002). The search results are calculated using the RPS-BLAST algorithm (Altschul *et al.* 1997) and the PSSMs. The databases in common between CDD and InterPro, Pfam (Bateman *et al.* 2004) and Smart (Letunic *et al.* 2004), are searched using Hidden Markov Models (HMMs) as part of the InterProScan (Zdobnov and Apweiler 2001) analysis, while CDD employs RPS_BLAST (Altschul *et al.* 1997). We therefore analysed which method, RPS-BLAST or HMMs, produced best results in searching Pfam and SMART. The comparison between them determined that CDD is the best way to automate the analysis, since more reliable protein domains were found by RPS-BLAST. The comparison for InterPro and CDD was taken by my colleague Dr. Luciane V. Mello.

Searching for UTRs & repeat elements

The 5'- and 3'-untranslated regions (5' and 3'-UTRs) of eukaryotic mRNAs play a crucial role in post-transcriptional regulation of gene expression (Klausner *et al.* 1993; McCarthy and Kollmus 1995), modulating nucleo-cytoplasmic mRNA transport (Wilhelm and Vale 1993; Bashirullah *et al.* 1998), translation efficiency (Curtis *et al.* 1995) and translation stability (Decker and Parker 1994; Beelman and Parker 1995). UTRdb (Pesole *et al.* 2002), UTRsite (Pesole *et al.* 2002) and PatSearch (Grillo *et al.* 2003) were

integrated in EST-ferret for UTRs searching. UTRdb is a specialized database of non-redundant 5' and 3'-UTRs of eukaryotic mRNAs. UTRdb entries are enriched with specialized information not present in the primary sequence databases. UTRs patterns in UTRdb have been collected in the UTRsite database in order to make it possible to search any sequence for the presence of annotated functional motifs. The PatSearch programme is a flexible and fast pattern matcher able to search for specific combinations of oligonucleotide consensus sequences, such as UTRs. In EST-ferret, PatSearch was implemented to search against UTRsite for locating UTR patterns of query sequences.

RepeatMasker (http://repeatmasker.genome.washington.edu) is a programme adopted in EST-ferret to scan the RepBase database (Jurka 2000), a DNA database and an electronic journal of repetitive elements, for interspersed repeats and low complexity DNA sequences. Sequence comparisons in RepeatMasker are performed by the program Cross_match, which is mentioned in the section of 2.2.1.1.

2.2.2 Computing environments for software developing

2.2.2.1 The Red Hat Linux operating system

Software developing for EST-ferret package were done in the Red Hat Linux (Petersen 2002) 7.2 operating system. The Red Hat Linux operating system has become one of the major Linux distributions and one of standard Linux versions, bringing to the PC all the power and flexibility of a UNIX workstation as well as complete set of Internet applications and fully functional desktop interface. It maintains a strong commitment to open source Linux applications and most of the bioinformatics software and biological data can be implemented in it. These advantages provide environment to allow us to develop our own applications for analysing biological data.

2.2.2.2 PERL, Java and MySQL

The programmes developed were written in JAVA (http://java.sun.com/) (Naughton and Schildt 1999), PERL (Wall *et al.* 1996)

(http://www.perl.org/) and Unix shell scripts. EST-ferret was written in the PERL programming language, which excels at slicing, dicing, and integrating data files and is the language of choice for the many bioinformatics researchers (Baxevanis and Ouellette 2005). It is available in Red Hat Linux platform and easy to learn and use. The open source programmes written in PERL can be easy to integrate with other bioinformatics software and make it available to users.

JAVA (Naughton and Schildt 1999) is an object-oriented programming language developed by the Sun Microsystems (http://java.sun.com/). JAVA programmes are interpreted by the Java Virtual Machine (JVM). Therefore JAVA applications can be run on any computer operating system with the JVM installed regardless of computer architecture. JAVA 2 Platform Enterprise Edition (J2EE) defines the standard for developing component-based multitier enterprise applications. Its features include web services support and the Software Development Kit (SDK).

SQL (Structured Query Language) is a tool for organizing, managing, and retrieving data stored by a computer database (Groff and Weinberg 1999). The computer programme that controls the database is called a database management system (DBMS). MySQL is one of the world's most popular open source database management systems. With the MySQL, databases for ESTs annotation resources can be constructed for local and remote users to access *via* the Internet.

2.3. Results

2.3.1 EST-ferret package

A user-configurable, automated pipeline, EST-ferret, was developed for the convenient analysis of EST sequence data. It includes all of the necessary steps for cleanup and trimming of sequences, submitting to external sequence repositories, clustering, identifying by BLAST homology searches and by searches of protein domain databases, annotation with computer-addressable terms and production of outputs for direct entry into microarray analysis packages. It is an open resource for academic users and available at my project website http://legr.liv.ac.uk.

2.3.1.1 The pipeline in EST-ferret

This package has four major components (Figure 2.1): (1) ESTs coding system (Figure 2.1a); (2) sequence processing (Figure 2.1b); (3) sequence clustering (Figure 2.1c) and (4) sequence annotating (Figure 2.1d). It is composed of several widely used, open-source algorithms, including PHRED, CAP3, BLAST, PatSearch, RepeatMasker, and can interrogate a range of sequence and annotation databases, including GO, CDD, ENZYME database, to run either step-by-step to track the outputs, or as a single batch process for delivery of putative identities and detailed annotations. This makes it particularly useful in supporting microarray analysis of transcriptome responses where it is necessary to profile annotations across gene lists generated in genome-scale expression profiling experiments. User can easily edit the configuration file to define parameter settings and other configuration information for analyses. It outputs result in flat files which can be also imported into the MySQL database and accessed *via* the Internet. The custom-integrated open-source bioinformatics programmes and databases in the EST-ferret are listed in Table 2.1.

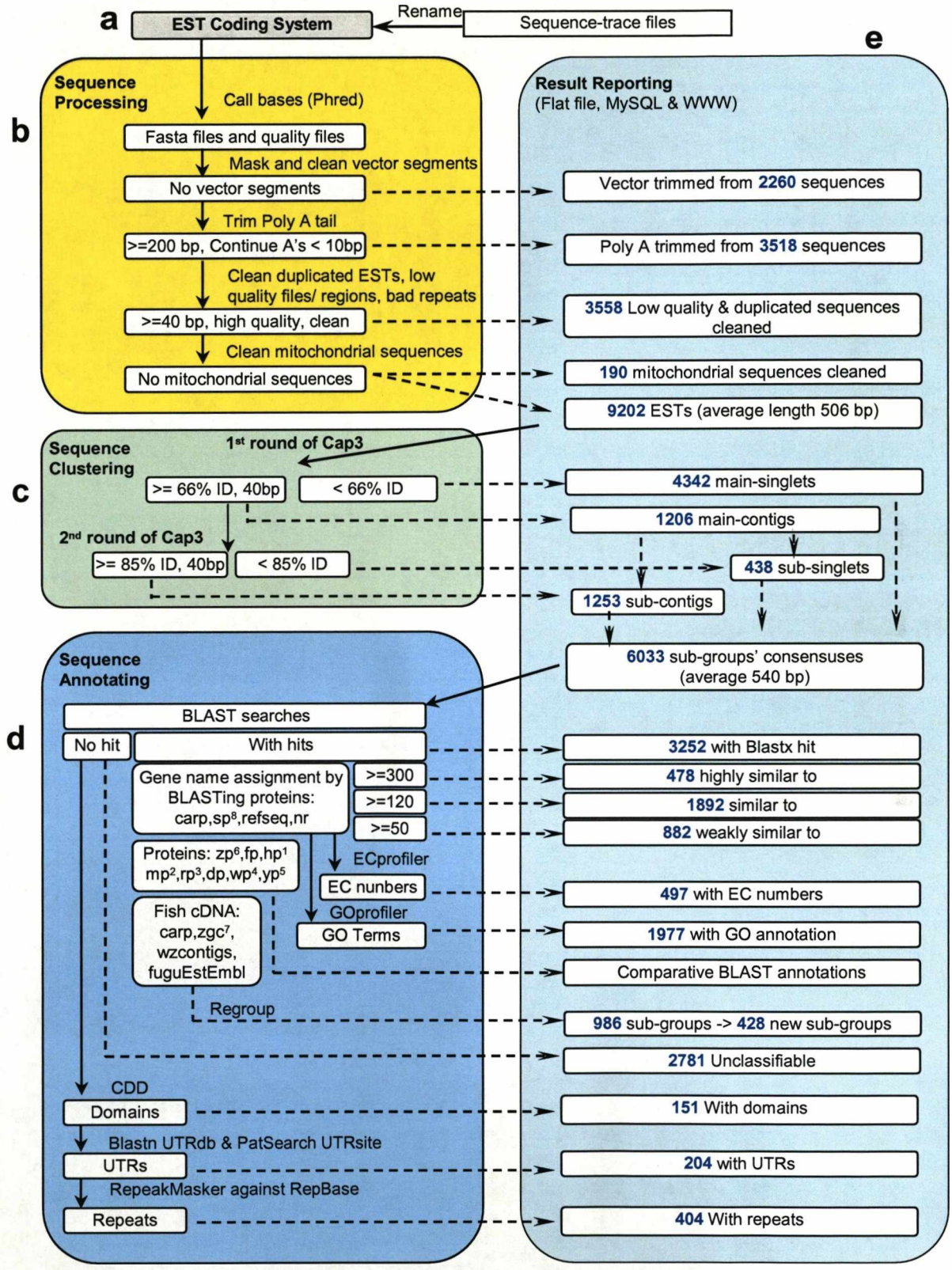


Figure 2.1: Pipeline for in EST-ferret for carpBASE. (a) EST coding system (b) Sequence processing (c) Sequence clustering (d) Sequence annotating (e) Result reporting.

hp[1] --- human protein sequences; mp[2] --- mouse protein sequences; rp[3] --- rat protein sequences; wp[4] --- worm protein sequences; yp[5] --- yeast protein sequences; zp[6] --- zebrafish protein sequences; zgc[7] --- Zebrafish full-length cDNA Collection; sp[8] --- swiss-prot protein sequences;

Table 2.1: Major programmes and databases used in EST-ferret 2.0

a: External programmes

Programme	Resources from	Description
Phred	http://www.phrap.org bge@u.washington.edu	Reads DNA sequencer trace data, calls bases, and assigns quality values to the bases
Cross_match	phg@u.washington.edu	Masks sequences
Cap3	xqhuang@cs.iastate.edu	Assembles sequences and generates consensus sequences
NCBI BLAST	ftp://ftp.ncbi.nlm.nih.gov	Basic Local Alignment Search Tool that includes s a set of similarity search programs
GOprofiler	http://legr.liv.ac.uk/	Assigns GO annotations for sequences according to BLAST best hits against Swiss-Prot
ECprofiler	http://legr.liv.ac.uk/	Scans enzyme database by using BLAST best hits against Swiss-Prot
PatSearch	ftp://www.pesolelab.it/	Finds UTR patterns from UTRsite
RepeatMasker	nilah@geospiza.com or http://www.geospiza.com/	Masks repeats from RepBase
PERL	http://www.perl.org/	A computer language
MySQL	http://www.mysql.com	A SQL Database Management System
Apache HTTP Server	http://www.apache.org	A web server software
PHP	http://www.php.net	The Hypertext Pre-processor

b: Databases

Database	Resource from	Description & reference
Common carp mitochondrial completed genome	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=5835023	GenBank accession number NC_001606
1027 known carp protein sequences	Retrieved *via* NCBI Entrez	Known common carp sequences available in GenBank
UniProtKB/Swiss-Prot protein	ftp://us.expasy.org/databases/uniprot/knowledgebase/uniprot_sprot.fasta.gz	a well curated protein sequence database
RefSeq Vertebrate: for mammalian and other species	ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_mammalian & ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_other/	Through heavily manual curation on know proteins and NCBI's Genome Annotation Projects
Nr	ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.tar.gz	Non-redundant protein collections in NCBI
Zebrafish protein	ftp://ftp.ncbi.nih.gov/refseq/D_rerio/mRNA_Prot/zebrafish.protein.faa.gz	Zebrafish protein sequences generated through NCBI RefSeq and NCBI Genome Annotation projects

Table 2.1b continued

Database	Resource from	Description & reference
1902 fugu known protein sequences	Retrieved *via* NCBI Entrez	
Human protein	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/protein.fa.gz	Human protein sequences generated through NCBI RefSeq and NCBI Genome Annotation projects
Mouse protein	ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/protein/protein.fa.gz	Mouse protein sequences generated through NCBI RefSeq and NCBI Genome Annotation projects
Rat protein	ftp://ftp.ncbi.nih.gov/genomes/R_norvegicus/protein/protein.fa.gz	Rat protein sequences generated through NCBI RefSeq and NCBI Genome Annotation projects
FlyBase:	ftp://flybase.net/genomes/Drosophila_melanogaster/current/fasta/dmel-all-translation-r4.2.1.fasta	Fly proteins sequences from FlyBase project
Worm protein	ftp://ftp.wormbase.org/pub/wormbase/acedb/WS147/wormpep147.tar.gz	Worm protein translations of all predicted and confirmed genes
Yeast protein	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/orf_trans_all.fasta.gz	Yeast translations of all systematically named ORFs.
Zebrafish gene collection (ZGC)	ftp://ftp1.nci.nih.gov/pub/MGC/fasta/dr_mgc_cds_aa.fasta.gz	Zebrafish full length cDNA collection
Zebrafish WZ contigs	http://www.genetics.wustl.edu/fish_lab/assemblies/wzcontigs.gz	Washington University zebrafish EST assembly
Fugu EST assembly	http://fugu.biology.qmul.ac.uk/Download/	HGMP Fugu EST assembly
UTRdb	ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR	Contains UTR sequences
UTRsite	ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/UTRSite	Contains UTR patterns
RepBase	http://www.girinst.org	Updated repeat elements
CDD (Pfam, Smart, Kog and Cog)	ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/	A Conserved Domain Database and Search Service
Gene Ontology data	http://www.geneontology.org/GO.current.annotations.shtml http://www.godatabase.org/dev/database/	GO annotations include molecular function, cellular component & biological process
Enzyme database	ftp://ca.expasy.org/databases/enzyme/release/enzyme.dat	Includes descriptions for all known enzymes
KEGG enzyme search	http://www.genome.ad.jp/kegg-bin/mk_point_html & http://www.genome.ad.jp/kegg/kegg2.html	A biochemical pathway database

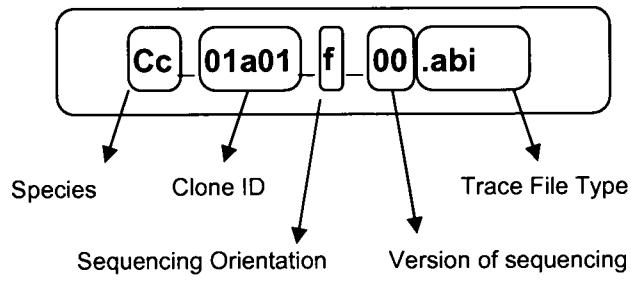
2.3.1.1.1 EST Coding System

EST-ferret requires that ESTs are named using a coding scheme based on that suggested by RIKEN (Konno *et al.* 2001) (Figure 2.1a & Figure 2.2). The EST Coding System includes the Simple Coding System, implemented in the entire pipeline of EST-ferret, and the Full Coding System, implemented in dbEST submission files. In the simple coding system, the sequence is assigned a species or library name, a clone ID that is the coordinate in the plate containing the cDNA, the direction of the sequence read ('f' signifies forward 5' read, 'r' reverse), the number of the sequencing attempt, and the file format (ABI), all with underscored separators. For example, an EST coded Cc_01a01_f_00.abi is *Cyprinus carpio*, microtiter plate 01, row a, column 1, forward read and was the first sequencing attempt (00) of this clone. EST-ferret is also compatible with a full more descriptive coding scheme, which is used in building up dbEST submission files. A *dbEST* (Boguski *et al.* 1993) submission from EST-ferret requires that a unique EST ID is defined for each sequence, and EST-ferret automates this process.

2.3.1.1.2 Sequence processing

Sequencing base-calling was performed using Phred and vector sequences were masked by Cross_match (http://www.phrap.org/). The default criteria of vector masking were set as the default setting, except the minimum length of matching word was 7 and the minimum alignment score was 20. Sequences were judged not to have poly-A tails if the longest continuous sequence of A's was less than 10 bases in length. Reads containing less than 40 high-quality bases after trimming vector segment were considered as low quality reads and discarded. The 5' and 3' regions with bases with a quality score less than 15 were determined as low quality regions of the reads. ESTs were defined as bad repeat ESTs (Appendix 2.1) if the content of any base was greater than 80% of the sequence. Mitochondrial sequences were identified by BLASTN search (E-value < $1e^{-15}$) against available mitochondrial genomes and removed. Vector segments, poly-A tails and low-quality regions of reads were trimmed from the ESTs. If a clone is represented by multiple reads then

a. Simple Coding Scheme



b. Full Coding Scheme

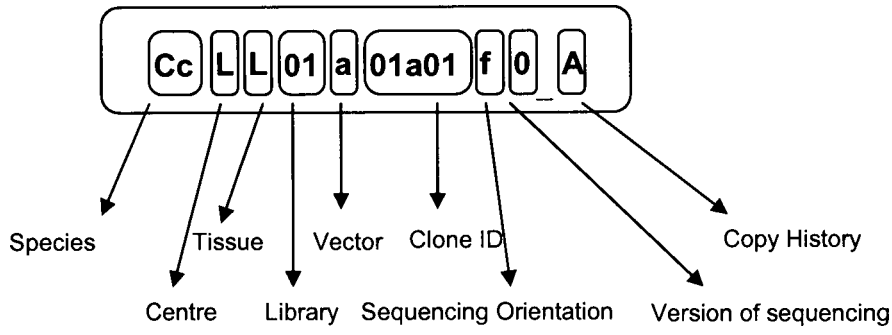


Figure 2.2: Coding system. (a) The Simple Coding System was implemented through the EST-ferret pipeline. (b) The Full Coding System was used in making submission files to dbEST

the highest quality read was selected to represent the clone. This procedure is shown on Figure 2.1b. In addition, EST-ferret generates a user-configurable flat file of the high quality ESTs that is correctly formatted for submission to dbEST.

2.3.1.1.3 Two-round of sequence clustering

The consensus sequence is derived from the overlapping sequences. Large-scale sequences have a good chance of producing strong overlaps which share high similarity, while a small number of sequences have good chance of producing weak overlaps which share low similarities. Sequences with a weak overlap might be the sequences from a same gene or different genes within the same gene family. My colleagues produced ~13,500 common carp ESTs, which might be insufficient enough to build up strong overlaps. Therefore two rounds of CAP3 clustering and assembly (Huang and Madan 1999) were implemented in EST-ferret to analyse the data. The first round with low stringency classified main-groups in which sequences were from the same gene family; and the second round with higher stringency classified sub-groups which represent putative unique genes. This provided a more straightforward understanding of the relationship between genes and members of possibly large gene families (Osato *et al.* 2002). The benefits of this over a single stage assembly can be understood by monitoring how main-groups containing members of possibly extended gene families, and which, because of a recent genome duplication event (David *et al.* 2003), were over-represented within the carp genome, break down to different sub-groups that approximate to unique genes. ESTs not in the same main-group have no overlaps in sequence alignments; ESTs in a same main-group have overlaps in sequence alignments and probably come from a same gene family; ESTs in a same sub-group are more similar to each other than those in a same main-group but not in a same sub-group.

In the default setting, the first round clustered ESTs with > 40 bps overlap and 66% identity together as main contigs, and ungrouped ESTs as main singlets. Then in the second round, the member sequences of each main

contig were subjected to more stringent clustering, default being > 40 bps overlap and 85% identity, to yield sub-groups which were either sub-contigs that contain more than one EST, or sub-singletons that appear to be unique. Sub-contigs or sub-singletons were given annotations by the Sub-groups Coding System (Table 2.2) that indicates that they were products from two-rounds clustering. For sub-groups from main-contigs, the sub-group ID was contributed by the ID of main-contig and the ID of sub-contig or sub-singlet; For sub-groups from main-singlet, the sub-group ID was contributed by the ID of main-singlet and additional part "-1". The addition part "-1" allows the form of sub-group IDs of main-singlets to be the same as those of sub-contigs and sub-singlets. For example, the sub-group 10-2 is sub-contig or sub-singlet number 2 derived main-group 10.

Table 2.2: Sub-group Coding System

a: Sub-groups from main-contigs

Round of clustering	**First round**		**Second round**
Common name	ID of main-contig	Dash	IDs of sub-contig & sub-singlet
Sub-group ID sample	10	-	2
Description	Sub-group 10-2 is the second sub-contig/sub-singlet of main-contig 10		

b: Sub-groups from main-singlets

Round of clustering	**First round**		
Common name	ID of main-singlet	Dash	Additional part
Sub-group ID sample	1207	-	1
Description	Sub-group 1207-1 is main-singlet 1207		

2.3.1.1.4 Sequence annotation

EST-ferret offers a user-configurable annotation pipeline that assigns putative function to each sequence using the results of BLAST homology searches of selected databases, profiles of Gene Ontology, and identifications of enzymes, UTRs, protein domains and repeat regions (Figure 2.1d).

Putative gene names were assigned to ESTs through identification of probable homologs in the public sequence databases. Sequences in well-annotated databases were used as target sequences in similarity searches for unknown query sequences. The default databases for BLAST analysis were chosen to best suit the characteristics of each non-model species. The Swiss-Prot is valued for its high quality annotation with minimal redundancy, the usage of standardised nomenclature and direct links to GO annotations for most of entries (Camon *et al.* 2003). The RefSeq protein records are produced through heavily manual curation on known proteins and the NCBI's Genome Annotation Projects (Pruitt and Maglott 2001). At the same time, users can implement other standard BLAST databases, such as nr, UniGene, TIGR gene indices, etc., or build up their own BLAST databases from their own sequences, for the BLAST searches in EST-ferret.

EST-ferret provided the "Parallel" BLAST searching and the "Priority" BLAST searching. The "Parallel" BLAST searching allows all query sequences to be searched against all selected subject sequence databases. BLAST hits of these searches from different databases species can be compared to each other. Thus, the "Parallel" BLAST searching can help to find the conserved sequences cross-species and the evidence that these sequences have conserved functions through many hundreds of millions of years of evolution. However, the "Parallel" BLAST searching is time-consuming on repeating searches for a same query sequence. To avoid this, the "Priority" BLAST searching can be chosen in EST-ferret. The "Priority" BLAST searches query sequence against the subject databases one by one, which are pre-ordered by user. If the query has hits against the top priority database, the search for this query is stopped; if the query does not, the search will go on for the second priority database; and so on. If user does want to have all BLAST information against different species and also wishes to save time, the BioCluster, mentioned below, is a good choice to speed up the "Parallel" BLAST protocol.

Classifiable ESTs can be further annotated using the Gene Ontology and enzymes nomenclature. GO annotations and Enzyme Commission (EC)

(Bairoch 2000) numbers are assigned to ESTs by using GOprofiler and ECprofiler programs that are embedded in EST-ferret. The GOprofiler programme, written in PERL and Java, interrogates formatted GOA association (Camon *et al.* 2003) to extract Gene Ontology (Gene Ontology Consortium 2004) annotation for each assembled EST which shared homology with an entry in the Swiss-Prot database. It outputs information relating to biological process, molecular function and cellular component. It also provides how many sequences and which sequences are associated with GO sub-categories. Its output can also be used as the input of GOmatrix which will be described in Chapter 4. The ECprofiler programme, written in PERL, interrogates Enzyme database (Bairoch 2000) to retrieve EC (Enzyme Commission) numbers for each assembled EST which shared homology with an entry in the Swiss-Prot database.

For ESTs without a gene name assignment (i.e. unclassifiable ESTs), EST-ferret attempts to annotate them as "with UTR", "with protein domain" or "with repeat". The CDD search for protein domains was introduced by Dr. Luciane V. Mello and integrated in EST-ferret pipeline by myself. PatSearch and BLASTN search can also be performed against UTRsite and UTRdb *via* EST-ferret to identify UTRs. RepeatMasker was embedded in EST-ferret for masking the repeats and low complexity DNA sequences of the ESTs. The output indicates which sequences contain putative protein domains, UTRs and repeat elements and their locations. Taken together these procedures maximize sequence identification and annotation in order to overcome the constraints of relatively small EST collections.

2.3.1.2 Program design

2.3.1.2.1 Source codes and usage of the package

EST-ferret was designed to provide a unified and configurable package for research groups lacking dedicated informatics support or high-level experience. The source codes and other files of EST-ferret are portable onto any Linux machine and are stored in the folder called "EST-ferret". In the package, there are several UNIX shell scripts which are command files for

running EST-ferret package. PERL and Java scripts in the folder of "script" can be explored and re-edited by user for their particular purposes. External programmes and databases need to be installed before running the package. Configurations in the file of "config.txt" are required for editing before data analysis. Users can easily edit the configuration file to define parameter settings and other configuration information for analyses. The key parameters include the E-value threshold for the BLAST search, the overlap length and the minimum identity for the CAP3 clustering, *etc.* Information for dbEST submission to dbEST can be located in the folder "user_dbEST_info". In addition the program can be run step-by-step in order to track the outputs, or as a single batch process. The manual for EST-ferret 2.0 is available at http://legr.liv.ac.uk/EST-ferret/.

2.3.1.2.2 Output reporting and storing

Finally, EST-ferret produces analysis reports in a variety of flat file formats, some of which can be serve as inputs for some gene annotation and gene expression profiling tools, and also as a MySQL (http://www.mysql.com/) database that can be interrogated using a web-based search tool (Figure 2.1e). EST-ferret will create a results folder named by user to store results in flat files. There are 4 subfolders inside the result folder: processing, clustering, annotating and reporting. These subfolders also contain subfolders in which different results files are located. With this folder structure, the user can easily explore the results for different stages without any confusion. EST-ferret can automatically produce EST submission file for dbEST as well.

2.3.1.3 Performance

EST-ferret was tested using several sets of ESTs, including common carp ESTs generated during the construction of a microarray, and was used to create several EST databases, such as carpBASE, squirrelBASE, roachBASE, etc. The annotation by EST-ferret has become a key resource in understanding the genes and supporting the microarray analysis in the Laboratory of

Environmental Gene Regulation (LEGR). The annotated databases analysed by EST-ferret are detailed in Chapter 3.

2.3.2 GOprofiler

Gene Ontology provides the functional annotation of genes from within the biological process, cellular component and molecular function domains. This makes it easier for biologists to interpret their large scale data in high-throughput experiments using computer support. Tools for the automatic annotation of genes with GO terms are available as open resources. Some are web-based programme and can not be integrated into EST-ferret, such as Gene Ontology Statistics (http://www.bioinf.ebc.ee/gost/), GeneTools (Beisvag *et al.* 2006), Goanna (http://agbase.msstate.edu/GOAnna.html), GOannotator (Couto FM 2006), Gotcha (Martin *et al.* 2004) and Manatee (http://manatee.sourceforge.net/). Some only accept individual gene/protein searches, such as GOannotator. Others are standalone packages but difficult to export data into a flat-file format for further interpretation, such as GoMiner (Zeeberg *et al.* 2003). GOprofiler was developed in my project as a standalone package which can be run in a user's local machine and integrated within the EST-ferret package. The output of GOprofiler can be edited and served as input for further analysis by GOmatrix, which is detailed in Chapter 4.

GOprofiler was developed to annotate ESTs with GO information. It is a standalone programme for LINUX system and was also integrated into EST-ferret. Query sequences can be assigned UniProt (Swiss-Prot or TrEMBL) IDs using EST-ferret or other BLAST search tools. UniProt matches of query genes serve as a bridge to connect genes to GO annotations. For this, GOprofiler uses the GOA (Gene Ontology Association) database (Camon *et al.* 2003) to connect the query to the GO terms and IDs. Output from GOprofiler characterises each query gene with GO annotation by indicating the GO matches for each query sequence in a tab-delimited text file. It also characterises a list of query genes with GO annotations by listing query gene matches for each of over 500 GO sub-categories in another tab-delimited text file. This second text file can be used as input for further analysis by GOmatrix.

GOprofiler is available at the LEGR data centre website (http://legr.liv.ac.uk/). It requires the LINUX operation system with the pre-installations of PERL and JAVA.

2.3.3 BioCluster

A Linux machine cluster with 40 Linux machines was constructed by the Physics Department of Liverpool University. My colleague, Chris Duckett set up a Portable Batch System (PBS) (http://www.openpbs.org) in this 40-Linux cluster. The PBS is a batch job and computer system resource management package. It accepts batch jobs, a shell script and control attributes, preserves and protects the job until it is completed, runs the job and delivers output back to the submitter. The PBS in the 40-Linux cluster manages and schedules the jobs into 40 Linux machines. With the help of Dr Cliff Addison (Computing Department), Dr. Anthony Moreton (Physics Department.) and Dr. Michael George (Physics Department), BLAST and CDD searches were implemented in the cluster, and we named it BioCluster. The web interfaces for accessing BLAST and CDD searches in BioCluster are available at http://bioserv2.sbs.liv.ac.uk/~fishomics/ and http://bioserv2.sbs.liv.ac.uk/~fishomics/cdd.html. The Figure 2.3 indicates the webpage of BLAST search in BioCluster. To submit the job to BioCluster, firstly, query sequences need to be zipped into an input file (.zip format), and then BLAST programmes and parameters can be set by chosen by user. BioCluster will automatically send an email to user on completion of the job. The user can then download the results from the BioCluster website.

Table 2.3 indicates the performances of the BLAST searches in BioCluster. It took less than 3 hours to complete a BLASTX search for 9610 sequences against the nr database. By the contrast, a single-CPU standalone computer (P4 CPU 1.3 GH, 512M RAM, Red Hat Linux 7.2 system) spent 4 days to complete the same job. The BioCluster is thus about 30 times faster than the single computer.

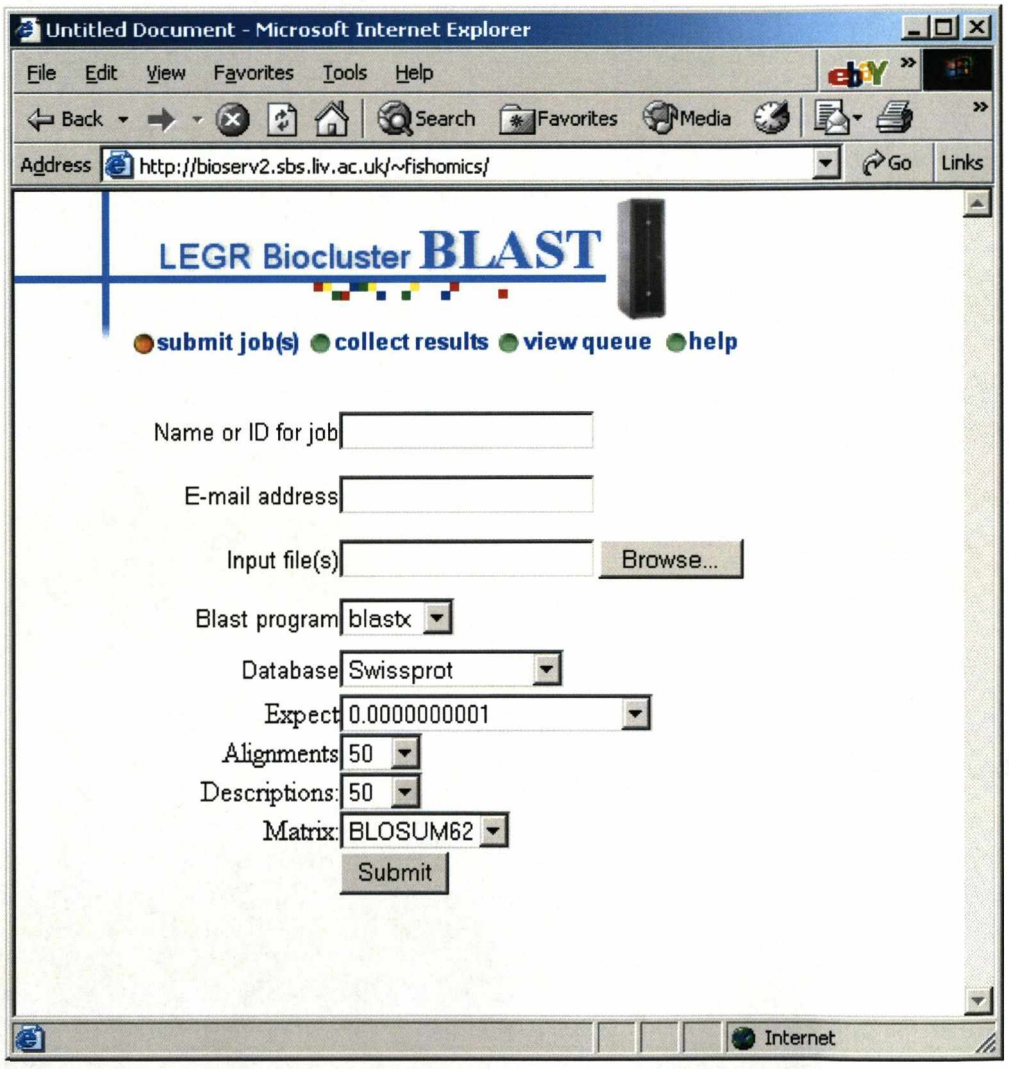


Figure 2.3: A screen shot of the BioCluster web interface

Table 2.3: Performances of BioCluster in BLAST

Tests	BLASTall program	Databases	# of trout sequences	Total run time (Second)	Sequences / Second	Seconds / Sequence
1	BLASTx	Swiss-Prot	7188	729	9.86	0.10
2	BLASTx	Swiss-Prot	6406	706	9.07	0.11
3	BLASTx	RefSeq	4113	453	9.08	0.11
4	BLASTx	RefSeq	4757	581	8.19	0.12
5	BLASTx	nr	4696	9610	0.49	2.05
6	BLASTx	nr	7866	11,352	0.69	1.44

2.4. Discussion

2.4.1 BLAST against FASTA and BLAT

For sequence annotation, there are several other published bioinformatics tools developed to identify sequences. FASTA (Pearson 2000) was the first widely used programme designed for the database similarity searching. The original program FASTP (Lipman and Pearson 1985) was designed in 1985 for protein sequence similarity searching. FASTA, described in 1988 (Pearson 2000), added the ability to undertake searches on DNAs against DNAs and DNAs against proteins. Like BLAST, FASTA enables the user to compare a query sequence against large databases. Both FASTA and BLAST use rigorous algorithms to find sequences that are statistically relevant and both bring significant strengths. Because FASTA uses a version of the more rigorous Smith-Waterman alignment method (Smith 1981), it generally produces better final alignments and is more suitable for identifying more distantly related sequences than BLAST. However, BLAST uses a heuristic approach that approximates the Smith-Waterman algorithm and is slightly less accurate than Smith-Waterman but over 50 times faster (Baxevanis and Ouellette 2005). For highly similar sequences, their performance is fairly similar. FASTA is more computationally intensive, but BLAST runs much faster than FASTA. BLAT (BLAST-Like Alignment Tool) (Kent 2002), similar to BLAST, is designed to rapidly align longer nucleotide sequences having more than 95% similarity. However, it uses a slightly different strategy to BLAST in order to achieve faster speeds. It is commonly used to find the position of a sequence of interest in a genome or to perform cross-species analyses. FASTA is time-consuming in running and BLAT is good for sequence alignment to genomes, therefore, BLAST is the best choice to perform similarity searching for unknown cDNAs/ESTs to provide gene identities or protein product descriptions. The speed and relatively good accuracy of BLAST are the key technical innovation of the BLAST and the reason of why BLAST is currently the most popular bioinformatics search tool.

2.4.2 Cut-off in BLAST

The output from BLAST searching reports alignment information for a list of hits over the selected cut-off criteria. The resulting information for each hit includes the description and the database accession, alignment bit score, alignment length, alignment identities, expected value, etc. The listing of a hit does not inevitably mean that the hit is biologically significant. For gapped alignments, the significance of a given alignment with score S is represented by the expected value (E or E-value), the expected number of chance alignments with a score of S or better (Altschul *et al.* 1997). It describes the number of hits one can "expect" by chance when searching a database of a particular size. It decreases exponentially with the score S that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. This means that the lower the E-value, or the closer it is to "0" the more "significant" the match is. The E-value can also be used as a convenient way to create a significance threshold for reporting results. But it is hard to define the appropriate threshold that should be set in BLAST searches. Andreas Baxevanis (Baxevanis and Ouellette 2005) suggested that one should look for hits with E-values of E-6 or less and sequence identity of 70% or more for nucleotide-based searches; one should look for hits with E-value of E-3 or less and sequence identity of 25% or more for protein sequence searches. The E value also reflects the size of the database used and the scoring system in use. Thus, these cut-offs of any other suggested cut-offs should not be used blindly.

2.4.3 EST-ferret against other EST pipelines

There is also a number of bioinformatics software solutions developed to automatically clean, cluster, assemble and annotate ESTs. These include PipeOnline 2.0 (Ayoubi *et al.* 2002), ESTAP (Mao *et al.* 2003), ESTWeb (Paquola 2003), ESTAnnotator (Hotz-Wagenblatt *et al.* 2003), ESTprep (Scheetz 2003), EST Pipeline System (Xu 2003), LUCY (Chou and Holmes 2001), PartiGene (Parkinson *et al.* 2004), annot8r and parpEST (D'Agostino 2005). Joanne Moran has reviewed advantages and the disadvantages of most of

these packages in respect of computational issues (Moran 2004). Here I will discuss the differences of these packages in respect of performance and analytical capabilities (Table 2.4).

Table 2.4: Analytical capabilities of other ESTs analysis packages

Package	Call base	Clean sequence	Submit to public db	Cluster / assemble sequence	Annotate sequence	Search protein domain	Match GO	Match enzyme/ Pathway
ESTWeb	Yes	Yes	No	No	No	No	No	No
ESTprep	No	Yes	Yes	Yes	No	No	No	No
ESTAP	Yes	Yes	Yes	Yes	Yes	No	No	No
PipeOnline	Yes	Yes	Yes	Yes	Yes	No	No	No
EST Pipeline System	Yes	Yes [a]	No	Yes	Yes	No	yes	No
parpEST	No	Yes	No	Yes	Yes	No	Yes	Yes
EST-Annotator	Yes	Yes	No	Yes	Yes [f]	No	No	No
LUCY	Yes	Yes [b]	No	No	No	No	No	No
PartiGene & annot8r	Yes	Yes	Yes	Yes [d]	Yes	No	Yes	Yes
EST-ferret	Yes	Yes [c]	Yes	Yes [e]	Yes [g]	Yes	Yes [h]	Yes.

[a] No trimming for low quality regions
[b] No trimming for quality files
[c] Cleaning vectors segment, low quality regions, low quality sequences; Cleaning quality files
[d] Using CLOBB
[e] 2 stage of CAP3 clustering
[f] Prior to clustering
[g] Priority BLAST or Parallel BLAST
[h] Using GOprofiler. Output can serve as input of GOmatrix, GoMiner GeneSpring.

2.4.3.1 Features of pipelines

ESTWeb is only a pipeline for.preprocessing and storing cDNA sequence reads. It has no capability on further EST analysis, such as submitting data to public database, clustering, annotating, domain searching.

ESTprep is similar to ESTweb, but has additional functions on sequence clustering and sequence submitting to public databases. It does not have capability for sequence clustering and sequence annotating. It does not have capability of base-calling, sequence annotating, domain searching.

ESTAP and PipeOnline include different sets of analytical procedures that verify, clean, cluster, annotate and transform ESTs data into relational databases. They are similar to EST-ferret but have not included functions on searching secondary database (e.g. InterPro and CDD) and for retrieving GO annotation and KEGG enzyme information.

EST Pipeline System also offers a friendly web interface, but it lacks the ability to submit to public databases and to search on secondary database. It is also unable to analyse enzymes and their metabolic pathways. Moreover, it does not allow trimming of low quality regions of the ESTs.

ParpEST includes functions of sequence cleaning, clustering and annotating. However, it does not support raw trace files and provide function of submitting data to public databases.

ESTAnnotator is similar to EST-ferret. The major difference between them is that BLAST searches for the cleaned ESTs are implemented prior to clustering in ESTAnnotator. This allows ESTAnnotator to obtain a preliminary annotation for the sequences before further analysis. In the analysis by EST-ferret, ESTs sharing overlaps contribute to a contig which is thus longer in length than each of the member ESTs. This contributes to a higher probability to achieve significance by greater proportion of BLAST hits. The ability within EST-ferret for assembly prior to BLAST searches also saves time for the following BLAST searches simply by reducing the number of BLAST searches required. In addition, ESTAnnotator is not able to perform searches on CDD and GO. Also it does not provide the EST submission file for public databases.

LUCY is a tool only for sequence cleaning. Quality files for sequence files are good for sequence clustering. The main disadvantage of LUCY on sequence cleaning is that it does not trim quality files for the low quality regions. It will cause problems when passing the cleaned FASTA formatted sequences and the unclean quality files to the clustering procedure.

PartiGene is a powerful EST analysis package developed by Prof. Mark Blaxter's bioinformatics group at the University of Edinburgh. PartiGene is divided into three segments that process the raw sequence traces (trace2dbest), generate the partial genomes (PartiGene) and derive peptide predictions

(prot4est). It implements a BLAST clustering algorithm, CLOBB (Parkinson *et al.* 2002), to cluster ESTs and then use the program Phrap for assembly of sequences. A major advantage of using CLOBB is that the program performs incremental updates of datasets maintaining previous cluster identities. One disadvantage for CLOBB is that it does not accept quality files for sequence clustering. The CLOBB programme output 7535 clusters for the common carp 9202 cleaned ESTs by using the same stringency level as used in the EST-ferret analysis. Comparing to 6033 clusters (described in Chapter 3) produced by EST-ferret, the clusters from CLOBB output was 1502 (~25%) more. The resulting clusters were also found more redundant based on the BLAST hit interpretations. For example, according to the analysis by BLAST and ClustalW (Thompson *et al.* 1997), the ESTs for main-group 288 in carpBASE 2.1 came from the same gene family for fast skeletal myosin light chain. The main-group 288 contained two unique genes, light chain 1a (sub-group 288-1) and light chain 3 (sub-group 288-2). But CLOBB clustered these 25 ESTs into 8 groups. More discussion for main-group 288 is also mentioned in Section 3.4.3 of Chapter 3. The result from two rounds of CAP3 clustering in EST-ferret was more reliable than that from CLOBB in analysing the common carp ESTs data. CAP3 does not preserve the original clusters through subsequent builds, but it has the capability of clustering sequences and producing assembled contigs and singlets in a single step. This makes it convenient to be implemented in two-round of clustering, which is very important for exploring relationships between genes in a same gene family.

annot8r (http://www.nematodes.org/bioinformatics/annot8r/), also developed by Prof. Mark Blaxter's group, is another tool for annotating ESTs with GO terms, EC number and KEGG pathways. It has similar functions to GOprofiler and ECprofiler which were developed in our lab. In GO annotation, Annot8r provides a resulting table which indicates the GO matches to each query sequence and is useful for analysis of GO-term distributions. However, GOprofiler not only provides the resulting table indicating the GO matches for each query sequence, but also produces another table ready for building the GO-term distribution. And the second resulting table from GOprofiler can be

served as input of GOmatrix (described in Chapter 4). This makes it useful for the analysis of different gene groups in gene expression experiments.

2.4.3.2 Exclusive Features of EST-ferret

EST-ferret has five major components: ESTs coding system, sequence processing, sequence clustering, sequence annotating and result reporting. The user can adopt the ESTs coding system for their ESTs in order to keep the EST datasets in good order and provide correct cDNA library information for the submission to dbEST. The user can also ignore the ESTs coding system for flexibility of analysis.

Sequence processing in EST-ferret accepts FASTA-format sequences and traces files. For input of traces, the cleaning procedure cleans not only FASTA-format sequences but also quality files. The trimming of quality files is important for the subsequent clustering of sequences. Most of other EST pipelines only trim FASTA-format sequences, but do not trim quality files.

Most of other EST pipelines only provide one-stage clustering, in which the user can only group ESTs into unique genes by setting a high stringency for the clustering. But if user also wants to know about the relationship between different unique genes, particularly unique genes within the same gene family, one-round clustering is not able illustrate these relationships. The two-stage of CAP3 clustering in EST-ferret groups genes into gene families (represented as main-groups) in the first-round clustering with a relatively low stringency and then into unique genes (represented as sub-groups) in second-round clustering with a much high stringency. It offers opportunities to explore relationships not only between unique genes, but also between genes and their gene family. For example, 5 sequences, named as A, B, C, D and E, are clustered using two stage system. A and B belong to sub-group 10-1; C and D belongs to sub-group 10-2; and E belongs to sub-group 11-1. This indicates that A and B belong to the same gene, C and D come from the same gene, the gene for A and B and the gene for C and D come from the same gene family but are different genes. E belongs to another gene family. The two-stage clustering is particular suitable to the common carp, which is

believed to have experienced a whole genome duplication event. The species contains twice the content complement of chromosomes and their genes, and as a consequence contains many large gene families.

EST-ferret implements the "Priority" BLAST and the "Parallel" BLAST to process BLAST searches automatically to identify large scale sequences. The "Priority" BLAST allows searching BLAST databases and storing BLAST hits in a pre-defined database order. It saves time in BLAST searching for best quality hits. The "Parallel" BLAST allows searching parallel BLAST databases and saving information for all hits from different BLAST database. This is time-consuming but can indicate matches across multiple databases for different species. The agreement and the difference of hits across different databases allow validation of the BLAST hits. The user can select one or both modes according to their analytical needs.

The additional searching of protein domain or signatures against CDD and InterPro provides more information for suggesting identifications for those un-classifiable ESTs in the BLAST searching. This is practically important when analysing sequences for non-model species. Most of other EST pipelines do not have this function.

The integration of GOprofiler makes EST-ferret more powerful on understanding biological meaning of genes by using GO terms. The compatibility of EST-ferret, GOprofiler and GOmatrix (described in Chapter 4) allows easy identification of their sequences and extends their functional annotation. EST-ferret output can serve as input for other gene expression analysis tools, including GeneSpring (http://www.silicongenetics.com) and GoMiner (Zeeberg *et al.* 2003). This allows users conveniently to explore expression or annotation profiles, respectively, of gene lists arising from their particular experiments. The functional annotations of "Parallel"/"Priority" BLAST, GO, EC, protein domain/signature in EST-ferret make it useful for 'non-model' species ESTs analysis.

EST-ferret played an important role in the project of the common carp transcription in the LEGR (Laboratory of Environmental Gene Regulation). It constructed the carpBASE, described in Chapter 3, to provide EST resource

with good-quality annotation and other relative biological data for the non-model species common carp. This accelerated the analysis of the large-scale cDNA microarray data for the common carp.

CHAPTER 3: carpBASE

3.1. Introduction

3.1.1 Common carp

The common carp, *Cyprinus carpio* L., belongs to the Cyprinid family, which is the largest family of freshwater fishes and the second largest family (after the gobies) of all fishes (Helfman *et al.* 1997). Common carp is the third most cultivated species worldwide (David *et al.* 2003; Peteri 2007) and is economically important in the freshwater aquaculture.

The common carp represents a classic case of a 'non-model' species with attractive attributes for scientific study but in 2001, when this thesis project started, possessed almost no genomic resources. It originates in a continental climate with extremes of both winter and summer. It is also tolerant of a wide range of temperatures for which it exhibits an extremely plastic thermal phenotype (Johnston and Temple 2002; Gracey *et al.* 2004). Responses occur in just a few days or weeks after a change in temperature and include acquired tolerance of both extreme cold and heat (Cossins *et al.* 1987) and physiological adjustments in a wide range of tissues. Thus, Lee and Cossins (Lee and Cossins 1988) have demonstrated adaption in intestinal morphology and functions; Cunningham and Hyde described substantial changes in performance of isolated retinas (Hyde and Cunningham 1995); Langfeld and Johnston have described substantial changes in the myosin composition of slow muscle fibres (Langfeld *et al.* 1991). It is also tolerant of extreme environmental hypoxia for which it displays a series of cardiorespiratory (Stecyk and Farrell 2002; Stecyk and Farrell 2006) and metabolic adaptations (Nilsson and Renshaw 2004). As a consequence the carp has become a well-used model for investigating fundamental mechanisms of the tolerance to environmental stresses. It is also increasingly used as a subject for physiological analysis in situations where the small size of zebrafish is limiting and the relationship between carp, as an environmental/aquaculture model, and zebrafish, as the genomics model within the same taxonomic group, is likely to become of increasing importance.

3.1.2 Reasons for constructing carpBASE

EST-ferret and BioCluster, described in Chapter 2, provide efficient approaches to analyse ESTs data, output results in flat files and a MySQL database, and submit sequences to dbEST. Full utilisation of the data by the wider research community requires that the sequences and their annotations be presented in a user-friendly addressable format. Since flat-files and relational databases can be difficult to handle for inexperienced users, a searchable web-interface was designed and created allowing researchers to browse collections of EST sequences and their annotations.

Much effort has been devoted to common carp microarray experiments in the LEGR, such as the temperature-response experiment (Gracey *et al.* 2004), the hypoxia-stress experiment (Fraser *et al.* 2006), and unpublished starvation experiments. Analyses by EST-ferret produced extensive functional annotations for identified genes in these experiments and these projects have benefited enormously from the construction of this detailed EST database.

Actually, several other communities have constructed web-addressable EST databases, such as FlyBase (FlyBase Consortium 2003), FunnyBase (Paschall *et al.* 2004), NEMBASE (Parkinson *et al.* 2004) and WormBase (Harris *et al.* 2004). FlyBase (FlyBase Consortium 2003) provides an integrated view of the fundamental genomic and genetic data for the major genetic model *Drosophila melanogaster* and related species. FunnyBase (http://genomics.rsmas.miami.edu/funnybase/super_craw4/) has been used to store, annotate, and analyze 40,363 expressed sequence tags (ESTs) from the heart and liver of the fish, *Fundulus heteroclitus*. NEMBASE is a publicly available online database (http://www.nematodes.org/nematodeESTs/nembase.html) providing access to the sequence and associated meta-data currently being generated as part of the Edinburgh-Wellcome Trust Sanger Institute parasitic nematode EST project. NEMBASE 2 currently holds sequences from 37 different species of nematode. WormBase (http://www.wormbase.org) is the central data repository for information about *Caenorhabditis elegans* and related nematodes. These

databases provide annotation information for their sequences in different web-format interfaces. Researchers can explore, search, or download interesting information on/from their websites.

3.1.3 Aims of the project

With the in-house bioinformatics tools of EST-ferret and BioCluster, I aimed to build up biological interpretations for the ESTs data produced by the LEGR, deposit the resulting data into databases, and also create a user-friendly web-interface for lab researchers to view, search and download the data. While only one database, carpBASE will be described in detail here, I have established databases for other organisms using similar a procedure, including squirrelBASE, troutBASE and roachBASE for ESTs of ground squirrel (*Spermophilus lateralis*), rainbow trout (*Oncorhynchus mykiss*) and roach (*Rutilus rutilus*), respectively. The needs of carpBASE were used in doing the development of EST-ferret and also as model for the subsequent EST-DB projects. In addition, I made the EST-ferret analysis package freely available to the wider scientific community and this and other resources are hosted on the LEGR Data Centre website for open access.

3.2. Materials and Methods

3.2.1 ESTs materials

The common carp ESTs were obtained from a collection of 13 normalized and serially subtracted cDNA libraries (Table 3.1) prepared from different tissues (liver, heart, brain, skeletal muscle, kidney, intestinal mucosa and gill). These libraries were constructed by Dr. Andrew Gracey (Gracey *et al.* 2004). A medium-scale collection of 15,000 cDNA clones from multiple tissues was produced by Dr. Margaret Hughes. To maximise representation and reduce redundancy within the clone set, and to reduce the overall sequencing costs, my colleagues have used an extended process of serial subtractive hybridisation, in which groups of already isolated clones were physically subtracted from new cDNA libraries as they were constructed (Carninci *et al.* 2000). The clones were directionally cloned and were selected for sequencing both randomly and for some collections on the basis that the corresponding mRNA exhibited an interesting expression profile. cDNA clones were subjected to single pass sequencing from the 5' end (5'LD primer, 5' CTCGGGAAGCGCGCCATTGTGTTGGT) on ABI3730 96-channel capillary sequencer at either the Wellcome Trust Sanger Institute (Hinxton, UK) or the School of Biosciences of Birmingham University.

3.2.2 Computing environment

The project was implemented on a Linux operating system (Red Hat 7.2). Analysis and annotation of the EST collection was performed using the EST-ferret package and the BioCluster server was used to accelerate the BLAST homology searching as described in Chapter 2. The platform was built using Apache, MySQL and PHP open-source applications, which together are an established and popular platform for the creation of web-addressable databases. In this platform, a MySQL database server (http://www.mysql.com) and an Apache HTTP server (http://httpd.apache.org) were used to store and display results, while PHP scripts (http://www.php.net) were created to allow researchers to browse and search the databases. MySQL is a multithreaded, multi-user, SQL Database Management System (DBMS) available as free

software under the GNU General Public License (GPL). SQL (Structured Query Language) is the most popular computer language used to create, modify, retrieve and manipulate data from relational database management systems. Apache HTTPD Server is an open-source HTTPD server for modern operating systems including UNIX and Windows NT. It provides a secure, efficient and extensible server that provides HTTP services with the current HTTP standards. HTTP (Hyper Text Transfer Protocol) is a communications protocol enabling the web-browsers used by the World Wide Web. With this HTTP server, web pages can be created and maintained to display analysis results on the Internet. PHP is a widely-used general-purpose scripting language that is especially suited for web development and can be embedded into HTML. It is compatible with MySQL and Apache HTTP sever.

Table 3.1: Summary table of the cDNA libraries

Library ID	Tissues	Plates picked	Subtracting Driver (plates)	Number of sequences	Notes
Liver 1	L	01-04	-	887	
Liver 2	L	05-07	01-04	902	
Liver 3	L	08-12	01-07	2490	
Liver 4	L	33&35	01-30	379+416=795	
Liver 5	L	31	-	7	SSH/capture, cold enriched
Mixed 1	M,H,K,B,I,G	13-14	-	631	
Mixed 2	M,H,K,B,I,G	15-20	01-12,12&14	2156	
Mixed 3	M,H,K,B,I,G	21-25	01-20	1928	
Mixed 4	M,H,K,B,I,G	26-30	01-25	1523	
Mixed 5	L,M,H,K,B,I,G	36&37	01-30	28	
Mixed 6	M,H,K,B,I,G	39&40	01-20	789	SSH/capture, cold enriched, hypoxia enriched
Muscle 1	M	32&34	01-30	333+379 = 712	
Muscle 2	M	38	01-30	102	SSH/capture, cold enriched

Tissues: **L**-liver, **M**-skeletal muscle, **H**-heart, **K**-kidney, **B**-brain, **I**-intestinal mucosa, **G**-gill. The libraries were used for the isolation of clones printed on the carp microarray. Information was provided by Dr. Andrew Gracey.

3.2.3 Analysis by EST-ferret

5' end sequences from 12,951 clone traces of most interest were passed through all procedures within the EST-ferret package for clone identification through structured homology searches of a range of DNA databanks. EST-ferret was detailed in Chapter 2. Procedures in sequence processing and sequence clustering used the default configuration settings described there.

For sequence annotation, "Parallel" BLAST searches were performed to maximise the amount of information that could be derived from comparisons between multiple species. EST consensus sequences were subjected to BLASTX (Altschul *et al.* 1997) homology searches against the known common carp protein sequences, the Swiss-Prot (Boeckmann *et al.* 2003), RefSeq vertebrate (Pruitt and Maglott 2001) databases and the nr protein database, and against the predicted peptides found in the zebrafish, human, mouse, rat, worm, and yeast genomes. The known common carp protein sequences were retrieved in GenBank using the Entrez system. Swiss-Prot contains high quality sequences with manually curated annotations. The RefSeq provides non-redundant set of sequences including genomic DNA, transcript (RNA), and protein products, for major research organisms. The nr database mainly contains sequences of non-redundant GenBank CDS translations. Each EST was assigned a putative identification based on the gene with which it had the greatest homology using the BLAST-derived bit score. When available, ESTs were assigned names based on homology with entries in the known common carp protein sequences, otherwise the database priority order was "Swiss-Prot → zebrafish proteins → zebrafish cDNAs (Strausberg *et al.* 1999; Rasooly *et al.* 2003)→ human proteins → mouse proteins → rat Proteins → worm proteins (Harris *et al.* 2004) → yeast proteins (Christie *et al.* 2004)→ RefSeq → nr proteins". If a sequence had BLAST hit(s) from the known common carp protein sequences, the top hit from the carp proteins was assigned as the gene name to the sequence; if not, the top hit from Swiss-Prot was assigned as the gene name to the sequence; and if it still had no hits, the top hit of zebrafish proteins was assigned, and so on. Other databases like

TIGR (Quackenbush *et al.* 2001) and TrEMBL (Boeckmann *et al.* 2003) are available but not manually curated and this limits the reliability of annotations. Thus, the TIGR Indices and the TrEMBL were excluded from the database orders for high quality annotations for the ESTs. BLASTN searches against human and mouse full-length cDNAs were not included for the gene name assignments because of the differences in codon usage of the general evolutionary distance between fish and mammals.

A gene assignment decision tree was developed to explain how similar and confident the gene name assignments are. The gene name assignment tree terms consisted of "Homolog of", "Similar to", and "Weakly similar to" based on their BLAST-derived bit score. This is a sensible measure (Okazaki *et al.* 2002) for similarity searches in order to find target sequences with same functions or evolutionary origin. ESTs that could not be assigned a name due to an absence of a homologous sequence in the public databases were described as "Unclassifiable ESTs".

3.2.4 Chi-square statistics test

A chi-square test (Mendenhall and Sincich 1988; Dawson-Saunders and Trapp 1994; Levine *et al.* 2001; Montgomery and Runger 2003) is a statistical hypothesis test in which the test statistic has a chi-square distribution when the null hypothesis is true. The chi-square test is a rough estimate of confidence. It accepts weaker, less accurate data as input than parametric tests (e.g. t-tests and analysis of variance). Chi-square tests are often used to examine association between two categorical variables.

In this thesis, the chi-square test was used to test whether the number of enzymes found in carpBASE 2.1 was significantly higher in each KEGG pathway category than would be expected by random selection from the entire gene list. In the following bivariate table (Table 3.2): a is the number of enzymes found in carpBASE 2.1 and in KEGG pathway A; b is the number of enzymes not in carpBASE 2.1 but in KEGG pathway A; c is in carpBASE 2.1 but not in pathway A; and d is not in carpBASE 2.1 and not in pathway A. We

wanted to test whether the number of enzymes found in carpBASE 2.1 is significant over-represented in KEGG pathway A.

Table 3.2: A bivariate table

Number of enzymes	In carpBASE 2.1	Not in carpBASE 2.1
In KEGG pathway A	a	b
Not in KEGG pathway A	c	d

Firstly, the chi-square test in this study was defined for the hypothesis:

H_0: There is no association between occurrence of enzyme genes in the KEGG pathway A and occurrence of enzyme genes in carpBASE 2.1.

H_a: The occurrence of enzyme genes of carpBASE 2.1 in the KEGG pathway A does not follow the chi-square distribution.

Secondly, the expected frequencies were calculated under the null hypothesis of no association between the categories of the two variables. For example, the expected number of enzymes found in carpBASE 2.1 and KEGG pathway A was calculated as

$$E_1 = (a + c) * (a + b) / (a + b + c + d) \qquad \text{(Equation 3.1)}$$

The data were divided into k (=4) cells. So the other three expected values (E_2, E_3 *and* E_4) for the other three cells were generated in the same way.

Thirdly, the chi-square statistic value were computed as

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \qquad \text{(Equation 3.2)}$$

where i stands for the cell number 1 to 4, O_i is the observed number for cell i and E_i is the expected number for cell i.

The probability of getting a chi-square value of a minimum given size was used to indicate whether or not the value was significant. The probability depends in part on the degrees of freedom of the table from which our chi-square value is derived. Therefore, we had to known the degree of freedom to determine the significance level. The degrees of freedom (df) for a table can be calculated by the following formula:

$$df = (r - 1) * (c - 1) \quad \text{(Equation 3.3)}$$

where r is the number of rows in the table and c is the number of columns in the table. In this case, the table was a 2 X 2 table and the degree of freedom was 1 ($df = (2-1)*(2-1) = 1$).

The test statistic follows a chi-square distribution with 1 degree of freedom. With this distribution, the p value of the chi-square value can be determined. The significance level α was set as 0.05. If the p value of the chi-square is less than α, H_0 is rejected and the enzymes of carpBASE 2.1 in a KEGG pathway are significantly over-represented or under-represented.

The chi-square test could not address whether the enzymes were over- or under-represented. The expected values were thus compared to the observed values. If the observed number is greater than the expected number in a KEGG pathway, the enzymes in carpBASE 2.1 determined to be over-represented in that KEGG pathway.

3.3 Results

The EST-ferret extends the annotation of cDNA sequences by combining BLASTX searching with searches of protein domains. To explore environmental stress responses in the common carp *Cyprinus carpio* L., 12,951 directionally cloned were generated predominantly full-length, normalized cDNAs. Of these, 9202 high-quality 5' end sequences have been assembled into 6033 non-redundant sequences, 53.9% of which were identified by BLASTX alignment against multiple sequence databases. Non-overlapping clones were also co-located on a zebrafish full-length cDNA collection and were aligned against protein domains, UTRs and repeat element databases, yielding useful annotation data for an additional 2.5, 3.4, and 6.7% clones, respectively. Together these procedures have increased the proportion of EST clusters with annotation from 53.9% obtained by BLAST alone, to 66.5%. The resulting data for carpBASE 2.1 were stored in a web-addressable MySQL database carpBASE 2.1, available at http://legr.liv.ac.uk/carpBASE (Figure 3.1). The interface contains tools for searching and browsing all the annotations and includes information for clustering, BLAST searching, GO, Enzyme, orthology analysis, ExprAlign, etc. The ExprAlign algorithm and orthology analysis are described in Chapters 4 and 5.

3.3.1 Analysis from processing and clustering

The sequence trace files of 12,950 5' carp ESTs were processed through EST-ferret. Of the 12950 ESTs from base-calling, 2260 contained vector segments which were trimmed; 3669 had poly-A tail; 3518 of these were trimmed of poly-A tails, whilst the remaining 151 short ESTs (< 200 bp if ploy-A tail trimmed) were retained untrimmed; 3543 low quality and duplicated ESTs were discarded; 15 bad repeat ESTs were also cleaned; 190 were found to be of mitochondrial origin. Finally, 9202 high quality sequences, with an average length of 506 bases (range 41 up to 984 bp, Figure 3.2) remained after sequence cleaning and were submitted to dbEST. These 9202 ESTs constitute the dataset carpBASE 2.1 (Figure 2.1 b and e, Table 3.3) and were submitted to dbEST (GenBank accession numbers CA963982-

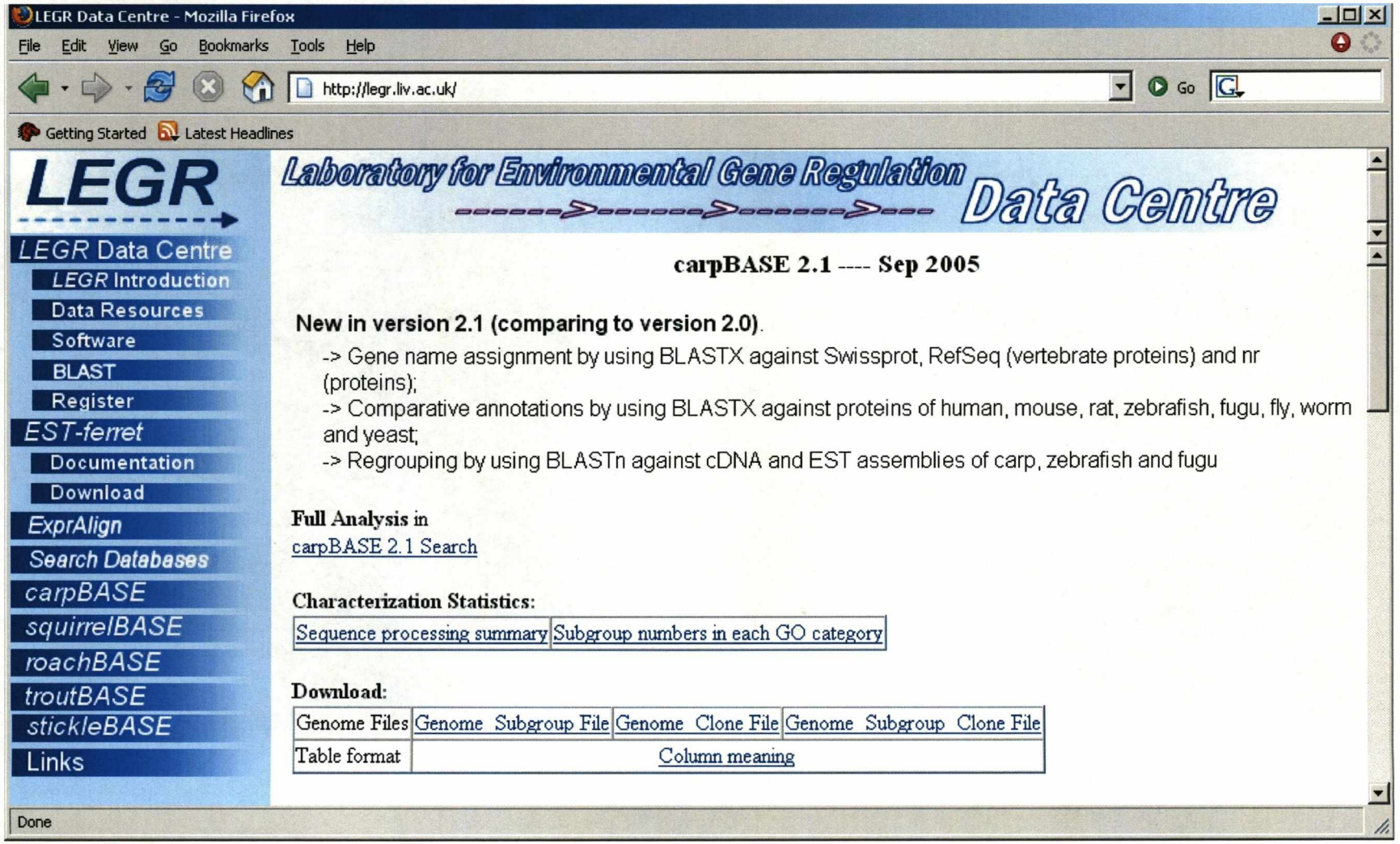


Figure 3.1: A screen shot of carpBASE in the LEGR website (July 2007)

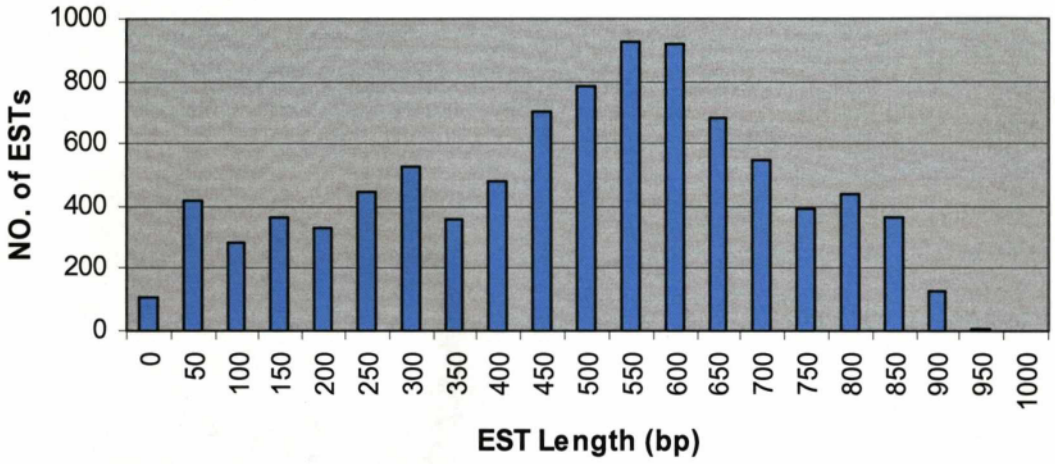


Figure 3.2: Distribution of EST length for ESTs in carpBASE 2.1

CA970467, CF660356-CF663121 & CO729406-CO729448). From the Phred score distribution over 85% bases were scored between 20 and 60 (Table 3.4).

Table 3.3: ESTs summary of sequence processing

	Number of ESTs	
Total traces	12,950	
Total ESTs	12,950	Base called
ESTs with vector segments	2260	Trimmed
ESTs with Poly-A tails	3518	Trimmed
Low quality & duplicated ESTs	3543	Discarded
Bad ESTs	15	Discarded
Mitochondrial ESTs	190	Discarded
Clean ESTs with high quality	9202	Retained
Average length of the clean ESTs	506bp	
Max. length of the clean ESTs	984bp	
Min. length of the clean ESTs	41bp	

Table 3.4: Phred score distribution of bases for 9202 high-quality ESTs in carpBASE 2.1

Phred Score	The number and percentage of bases	
0-10	103,869	2.20%
10-20	537,966	11.60%
20-30	840,957	18.10%
30-40	972,303	20.90%
40-50	1,286,603	27.70%
50-60	910,843	19.60%
60-70	0	0
70-80	0	0
80-90	0	0
90-100	0	0
All	4,652,541	100%

Because the carp genome is likely to have undergone a whole genome duplication (David *et al.* 2003) and is expected to possess larger number of

paralogous and more extended gene families, two rounds of clustering and assembly (Figure 2.1 c and e and Table 3.5) were adopted. The first round of CAP3 clustering assembled the 9202 ESTs into 1206 main-contigs (containing 4860 ESTs) and 4342 main-singlets. The second more stringent round of CAP3 re-clustered the main-contigs to yield 1253 sub-contigs (containing 4422 ESTs) and 438 new sub-singlets. The length distribution of the sub-contigs is shown on Figure 3.3. Thus, two rounds of CAP3 clustering yielded 6033 distinguishable sequences from 6033 sub-groups, comprising 1253 contigs and 4780 singletons, with an average length of 540 bp (range 41-2290 bp). The 6033 sequences were also regarded as transcript units. 29 sub-contigs contain ≥ 15 ESTs each (Table 3.6) and the largest contained 69 ESTs. The redundancy of the carp EST collection was 1.53 (i.e. 9202/6033), which compares favourably with that reported for other EST projects: 2.15 in Atlantic salmon (Rise *et al.* 2004), 2.41 in fugu (Clark *et al.* 2003), 10.6 in medaka (Kimura *et al.* 2004) and 28.4 in mouse (Carninci *et al.* 2003).

Table 3.5: Summary of sequence clustering

	Number of ESTs
ESTs for clustering	9202
Main-groups from 1st clustering	5548
Main-contigs from 1st clustering	1206
Main-singlets from 1st clustering	4342
Sub-contigs from 2nd clustering	1253
Sub-singlets from 2nd clustering	438
Sub-groups from 1st and 2nd clustering	6033
Average size of sub-groups (no. of ESTs)	1.53
Max. size of sub-groups (no. of ESTs)	69
Min. size of sub-groups (no. of ESTs)	1
Average length of sub-groups (bp)	540
Max. length of sub-groups (bp)	2290
Min. length of sub-groups (bp)	41
No. of sub-groups containing	
1 ESTs	4780
2 ESTs	764
3 ESTs	214
4-5 ESTs	136
6-10 ESTs	88
11-20 ESTs	36
21-30 ESTs	6
31-50 ESTs	7
>50 ESTs	2

Note: This table summarizes the information of sequence processing and sequence clustering including numbers of ESTs cleaned and remained in each steps, the size and length of assembled ESTs.

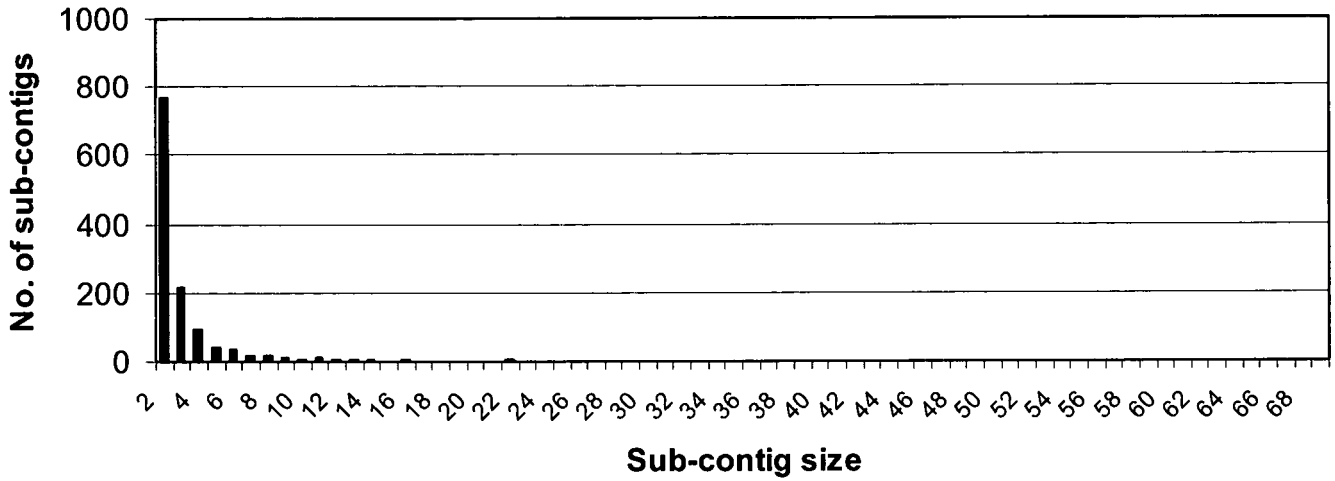


Figure 3.3: Distribution of sub-contig size in carpBASE 2.1

Table 3.6: Gene name assignment for the largest sub-groups

Sub-group ID	No. of ESTs	Length (bp)	DB	Bit score	Identity (%)	Hit Acc. No.	Gene description
296-1	69	1325	cp [a]	687	100	BAA08755.1	Skeletal alpha-actin
43-2	61	798	rp [b]	218	73	XP_698979.1	14 kDa apolipoprotein
971-1	42	1549	cp	776	99	AAC96094.1	Creatine kinase M3-CK
279-2	35	810	cp	214	100	CAC83659.1	Parvalbumin
622-2	35	873	cp	471	94	CAI35911.1	Glyceraldehyde-3-phosphate dehydrogenase
1127-2	34	405	nr				Unclassifiable EST
397-1	33	1215	cp	306	93	BAA89704.1	Myosin regulatory light chain
279-3	32	695	cp	211	100	P02618	Parvalbumin beta
622-1	32	1142	cp	587	93	CAI35911.1	Glyceraldehyde-3-phosphate dehydrogenase
1026-1	30	838	nr				Unclassifiable EST
971-2	25	1300	cp	751	96	AAC96093.1	Creatine kinase M2-CK
142-1	22	902	sp [c]	378	79	P84335	Tropomyosin 1 alpha chain
279-4	22	661	cp	213	100	CAC83658.1	Parvalbumin
490-4	22	880	sp	460	87	P32007	ADP/ATP translocase 3
1160-1	22	474	sp	115	95	P58372	60S ribosomal protein L30
38-1	20	1638	cp	736	83	AAO74862.1	Fetuin long form
1062-1	20	893	cp	360	61		Vimentin
424-2	19	574	rp	111	72	XP_699235.1	Apolipoprotein C-I precursor
488-1	19	1212	sp	596	83	P53447	Fructose-bisphosphate aldolase B
747-1	18	2290	cp	1202	86	AAL57604.1	Transferrin variant A
454-1	17	889	sp	214	60	Q9D7X8	Protein C7orf24 homolog
679-1	17	684	nr				Unclassifiable EST
63-1	16	487	sp	223	85	P80856	Fatty acid-binding protein, liver
69-4	16	255	nr				Unclassifiable EST
216-1	16	639	cp	176	82	CAC83658.1	Parvalbumin
435-3	16	668	rp	94.7	78	XP_525925.1	rRNA intron-encoded homing endonuclease
556-1	16	889	cp	87.8	70	AAB26498.1	Granulin-3 growth modulatory factor
69-2	15	1153	cp	334	98	CAC34942.1	Apolipoprotein A-I
511-2	15	2173	cp	426	98	CAB57858.1	Carp Desaturase 2

[a] cp --- Known carp protein sequences in GenBank

[b] rp --- RefSeq protein database

[c] sp --- Swiss-Prot protein database

3.3.2 Functional inferences from BLAST homology

In order to annotate the sequences generated in this work, a series of different analyses were applied (Figure 2.1 d and e and Figure 3.4). In the first, using BLAST searching, EST-ferret assigned putative gene names to 53.9% of the unique ESTs (3252 of 6033 sub-groups) regarded as having significant similarity (bit score > 50) to sequences present in the public databases. The remaining 2781 (46.1%) remain unclassified. Hundreds of classified sequences have gene names were unclear in meaning like "hypothetical ...", "similar to ... ", "unknown ...", "Probable ...", and so on. Manual curations were performed to obtain another gene name with clear meaning from other BLAST search results. We defined the criteria to classify the degree of similarity of each sequence alignment as 'highly similar to', 'similar to' or 'weakly similar to' as bit score >300, >120 and >50, respectively. In total 6033 sub-groups, 3252 were assigned with hits by Blast, 478 (7.9%) were assigned as "Highly similar to ...", 1892 (31.4%) were assigned as "Similar to ...", 882 (14.2%) were assigned as "Weakly similar to ...", and 2,781 (46.1%) were "unclassifiable" by the gene name assignment. The most abundant genes in carpBASE 2.1 were skeletal alpha-actin, 14 kDa apolipoprotein, creatine kinase M3-CK and parvalbumin (Table 3.6).

ESTs arising from the same gene may fail to cluster together because they share little or no overlapping sequence. To address this problem, we explored whether we could use zebrafish full length cDNA (Zebrafish Gene Collection, ZGC) (Strausberg *et al.* 1999; Rasooly *et al.* 2003) or EST contigs as scaffolds to collapse non-overlapping ESTs for the same carp gene into single contigs. This revealed that 986 of these ESTs exhibited non-overlapping alignments with 428 cDNAs or EST contigs and thus could be collapsed into 428 new sub-groups. The gene name assignments of the majority of the 986 sub-groups also provided evidence to support the regrouping.

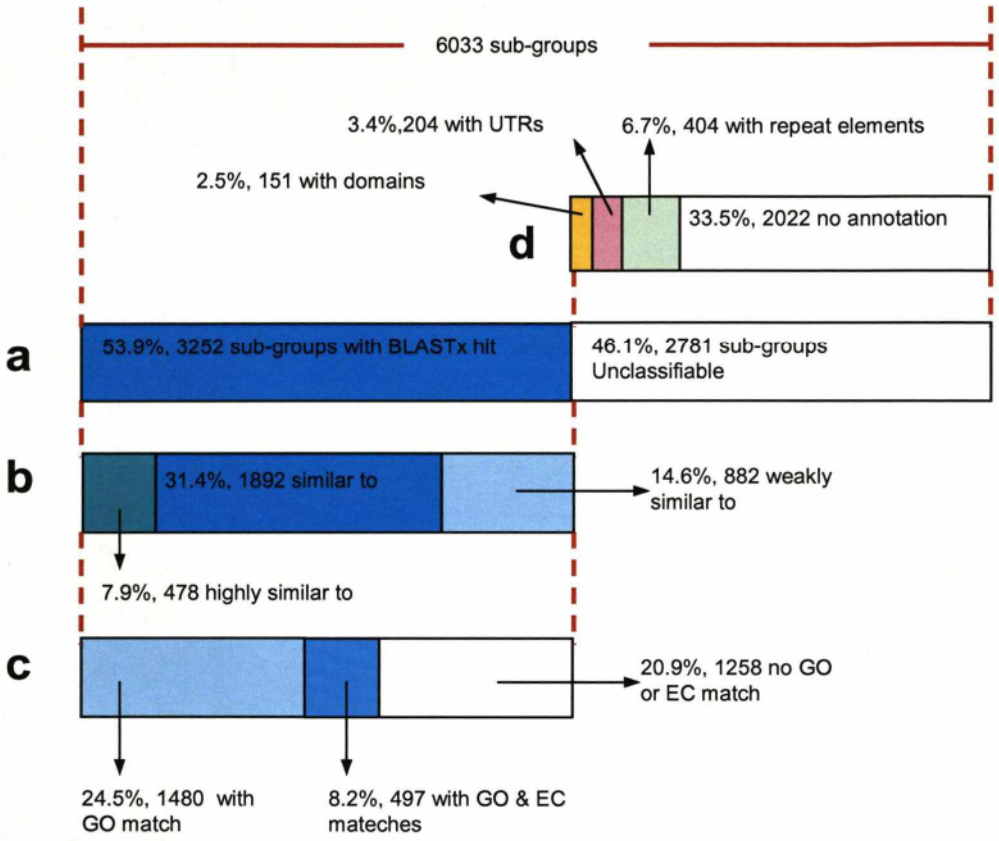


Figure 3.4: Annotations for 6033 sub-groups in carpBASE 2.1. (a) BLAST identities for the sub-groups. The whole bar represents consensuses of 6033 sub-groups. The blue bar represents the identical sub-groups by BLAST and the white represents the unclassifiable sub-groups by BLAST. (b) Different degrees of homology in sequence alignments of BLAST for the identical sup-groups. (c) Proportion of identical sub-groups with GO and enzyme annotations. (d) Additional annotations with protein domains, UTRs and repeats for unclassifiable sub-groups.

3.3.3 Functional annotation with GO and enzyme

GOprofiler assigned Gene Ontology annotations to 1977 of the 3252 BLAST identified ESTs, with the carp EST collection covering a broad range of biological processes (Figure 3.5 & Appendix 3.1). carpBASE 2.1 contains particularly large numbers of sub-groups in "transport", "biosynthesis", "nucleic acid metabolism" and "catabolism" sub-categories of biological process, in "intracellular" sub-category of cellular component, and in "ligand binding or carrier" and "enzyme" sub-categories of molecular function. carpBASE 2.1 also includes ~70 sub-groups implicated in "stress response" and ~120 in "response to external stimulus". 1756 of the classified sub-groups were not annotated with GO annotations and we expect that the ongoing improvements in GO annotation for Swiss-Prot will lead to significant improvements in carp gene annotation in the future.

ECprofiler assigned 256 EC numbers to 497 different sub-groups consistent with some enzymatic reactions being catalyzed by more than one gene. The enzyme database (Bairoch 2000) contained 4005 entries in Jan 2007. The enzymes in carpBASE 2.1 comprised ~6.4% (=256/4005) of them. EC identities were submitted to the KEGG pathway database (Kanehisa *et al.* 2002) web-search tool and Table 3.7 shows that the enzymes in carpBASE 2.1 covered broad categories of KEGG pathways. It also compared the coverage of the KEGG pathways with that of the far more extensive mouse EST database. For many pathways the % coverage of carp was similar to that of mouse, particularly including pathways associated with glycolysis/gluconeogenesis, carbon fixation, and valine/leucine/isoleucine degradation. The chi-square statistics tests (Levine *et al.* 2001) also indicated the enzymes were significant over-represented in these pathways. Moreover, carp enzymes in fatty acid metabolism (Figure 3.6) and pentose phosphate pathway were significantly enriched but mouse enzymes are not. For other pathways the mouse database provide 2-6 times more genes, which reflects the fact that the number of mouse ESTs are over 20 times greater than those in carp ESTs. We used tissues rather than specific cell-types, so we probably have a larger proportion of house-keeping genes than might be achieved.

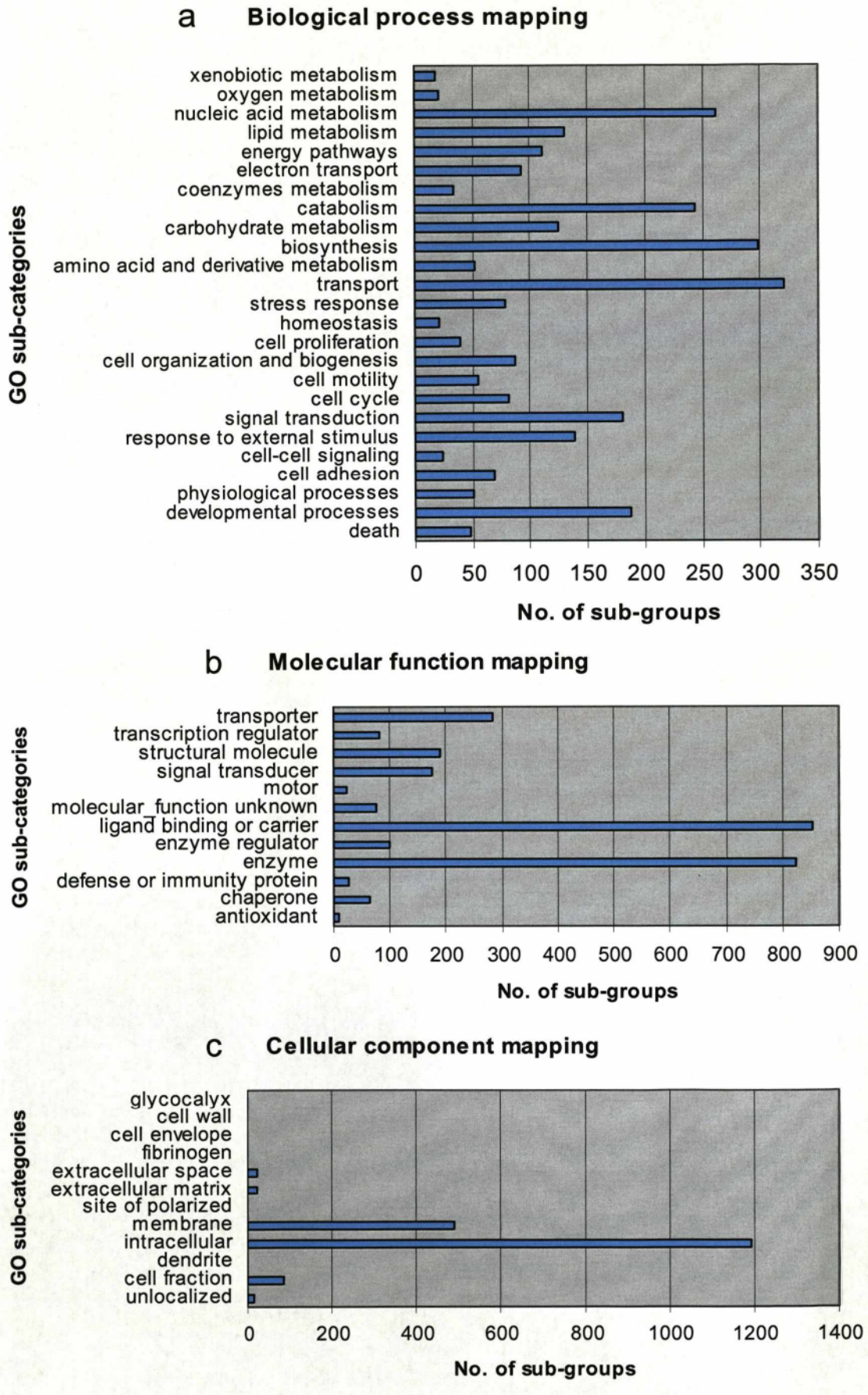


Figure 3.5: carpBASE 2.1 GO annotations from GOprofiler. Figure 3.5: GO annotation by GOprofiler. (a) Distribution of GO matches in biological process; (b) Distribution of GO matches in molecular component. (c) Distribution of GO matches in cellular component.

Table 3.7: Enzymes of carpBASE 2.1 and mouse involved in KEGG pathways

KEGG pathway categories	All EC no. in reference pathway	carpBASE EC no. (%) in reference pathway	Mouse EC no. (%) in reference pathway	Fold of Mouse/ carpBASE
Carbohydrate Metabolism				
Glycolysis / Gluconeogenesis	40	21 (52.5%)	25 (62.5%)	1.19
Citrate cycle (TCA cycle)	23	7 (30.4%)	13 (56.5%)	1.86
Pentose phosphate pathway	34	10 (29.4%)	14 (41.2%)	1.4
Fructose and mannose metabolism	62	10 (16.1%)	15 (24.2%)	1.5
Galactose metabolism	37	5 (13.5%)	14 (37.8%)	2.8
Starch and sucrose metabolism	75	8 (10.7%)	17 (22.7%)	2.13
Aminosugars metabolism	39	4 (10.3%)	17 (43.6%)	4.25
Inositol phosphate metabolism	30	5 (16.7%)	10 (33.3%)	2
Pyruvate metabolism	67	13 (19.4%)	17 (25.4%)	1.31
Glyoxylate and dicarboxylate metabolism	58	4 (6.9%)	5 (8.6%)	1.25
Energy Metabolism				
Oxidative phosphorylation	13	4 (30.8%)	7 (53.8%)	1.75
Methane metabolism	26	2 (7.7%)	5 (19.2%)	2.5
Carbon fixation	23	10 (43.5%)	12 (52.2%)	1.2
Nitrogen metabolism	63	6 (9.5%)	5 (7.9%)	0.83
Lipid Metabolism				
Fatty acid metabolism	28	11 (39.3%)	14 (50%)	1.27
Biosynthesis of steroids	35	9 (25.7%)	14 (40%)	1.56
Bile acid biosynthesis	27	6 (22.2%)	10 (37%)	1.67
Androgen and estrogen metabolism	26	2 (7.7%)	12 (46.2%)	6
Urea cycle and metabolism of amino groups	34	5 (14.7%)	17 (50%)	3.4
Glycerolipid metabolism	80	7 (8.8%)	33 (41.3%)	4.71
Nucleotide Metabolism				
Purine metabolism	99	19 (19.2%)	44 (44.4%)	2.32
Pyrimidine metabolism	61	10 (16.4%)	29 (47.5%)	2.9
Amino Acid Metabolism				
Glutamate metabolism	36	6 (16.7%)	13 (36.1%)	2.17
Alanine and aspartate metabolism	38	7 (18.4%)	14 (36.8%)	2
Glycine, serine and threonine metabolism	57	13 (22.8%)	24 (42.1%)	1.85
Valine, leucine and isoleucine degradation	33	11 (33.3%)	17 (51.5%)	1.55
Arginine and proline metabolism	71	10 (14.1%)	26 (36.6%)	2.6
Histidine metabolism	39	5 (12.8%)	12 (30.8%)	2.4
Tyrosine metabolism	72	5 (6.9%)	22 (30.6%)	4.4
Tryptophan metabolism	61	10 (16.4%)	28 (45.9%)	2.8

Data submitted to KEGG Search tool (http://www.genome.jp/kegg-bin/mk_point_html) to build classifications.
X^2 statistics test and associated *P* values were used to compare sample proportions. Value with positive X^2 statistics have more EC numbers than expected (50% grey = $p < 0.01$; 25% grey = $p < 0.05$).
Fold of mouse/carpBASE = (Mouse EC no. in reference pathway) / (carpBASE EC no. in reference pathway)

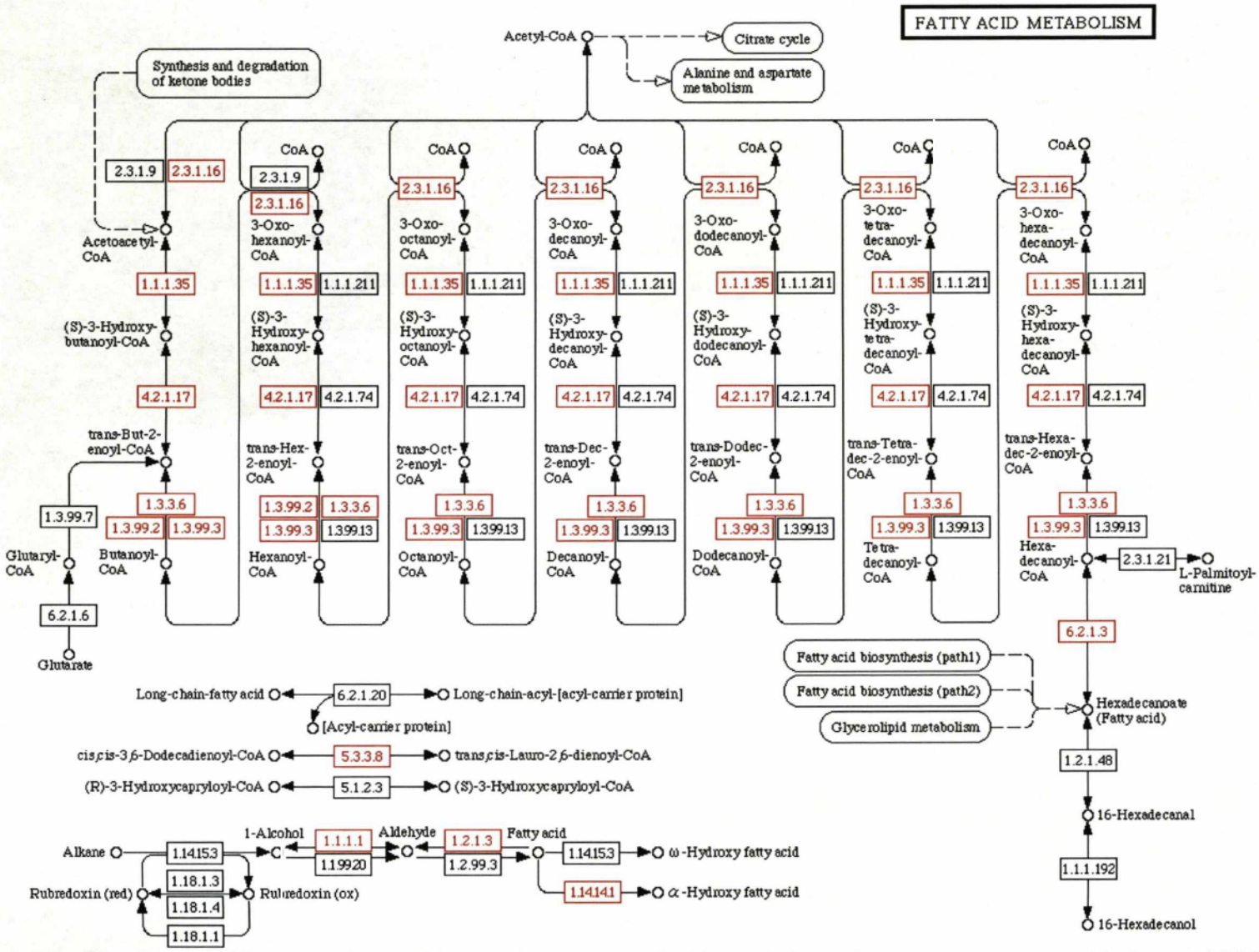


Figure 3 .6: Enzymes of carpBASE 2.1 in pathway of fatty acid metabolism. Red boxes indicate the enzymes represented in the carpbBASE collection

3.3.4 Protein domain analysis

We have also sought to use the available protein domain databases to provide additional functional annotation of ESTs for which no BLAST identification was possible, or which possessed a BLAST alignment of low score. After extensive tests of different search strategies, CDD proved to be the most time-efficient and successful means of searching Pfam (Bateman *et al.* 2004), Smart (Letunic *et al.* 2004), Cog (Tatusov *et al.* 2003) and Kog (Koonin *et al.* 2004) databases (data not shown). The final result of this search of the secondary database yielded an additional 309 annotations, from the four databases in CDD, each of which were carefully checked manually (Figure 3.7a). Of these, 151 (2.5% of 6033) had no gene name assignment (Figure 3.4d).

3.3.5 Analysis of non-coding regions: UTR and repeat elements

The following procedures were used to provide annotation to a further 10.1% of the sequences lacking a BLAST hit. A BLASTN search for all sub-groups against UTRdb (E-value $\leq e^{-5}$) indicated UTR motifs in 947 sub-groups, of which 68.7% were 3' UTRs, 31.3% were 5' UTRs. Table 3.8a indicates the proportions of UTR matches from different species. Using another approach, PatSearch revealed that 16 UTR patterns were found in 1213 sub-groups. The top 3 matches were the K Box (cTGTGATa), the BRD Box (AGCTTTA), and the GY Box (GTCTTCC) which are present in 3'UTRs and mediate negative post-transcriptional regulation (Table 3.8b). Finally, 204 unclassified sub-groups (Figure 3.4d) were revealed containing UTR motifs by using both the BLASTN search and the PatSearch programme. Figure 3.7b indicates the proportions of UTR matches from different species for these 204 sub-groups.

RepeatMasker identified 1.1% of the total length of the 6033 assembled ESTs as known classes of repeat elements found in RepBase. The proportion of repeats in carpBASE 2.1 was substantially smaller than in the salmon EST assemblies (Rise *et al.* 2004), where it comprised 11.9% of the total length of assembled sequences, perhaps a reflection of different cloning strategies. The salmon ESTs project constructed a normalized mixed tissue library, the

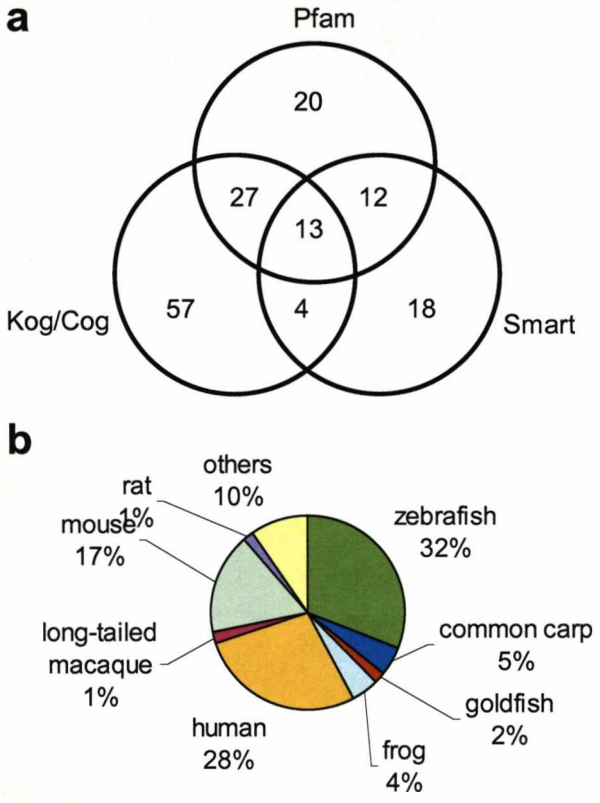


Figure 3.7: Annotation with protein domains and UTRs for unclassifiable sub-groups in carpBASE 2.1. (a) Venn diagram indicates the relationships of protein domain matches in KOG/COG, Pfam and SMART for the unclassifiable sub-groups. (b) Pie chart shows the proportions of UTRs matches in different species for the unclassifiable sub-groups.

suppression subtractive hybridization (SSH) libraries, and the cDNA libraries. The most abundant repeat classes were low complexity (0.58%) or simple repeats (0.46%) (Table 3.9). Other mainly contains repeats of LINEs (long interspersed nucleotide element), SINEs (short interspersed nucleotide element), small RNAs, LTR (long terminal repeats), and DNA elements. Given that the average length of all carp ESTs was 540 bp, the average length of repeats per sequence was only 6 bp. In fact, repeat elements were found in only 699 (11.6% of 6033) sub-groups, of which 404 had no annotations of known sequences, UTR, or protein domain, consistent with them being contaminating genomic intronic DNA (Figure 3.4d).

Table 3.8: UTRs analysis

a. BLASTn matches for sub-groups against UTRdb (E-value <=e-5)

Species	No. of blastn matches	% of sub-groups	Orientation	
			3' UTR	5' UTR
Common Carp (*Cyprinus carpio*)	108	11.40%	99	9
Zebrafish (*Danio rerio*)	455	48.05%	320	135
Goldfish (*Carassius auratus*)	43	4.54%	32	11
Frog (*Xenopus laevis*)	11	1.16%	5	6
Human (*Homo sapiens*)	186	19.64%	103	83
Mouse (*Mus musculus*)	61	6.44%	42	19
Rat (*Rattus norvegicus*)	14	1.48%	8	6
Long-tailed macaque (*Macaca fascicularis*)	16	1.69%	16	0
Red junglefowl (*Gallus gallus*)	19	2.01%	9	10
Others	34	3.59%	17	17

Table 3.8 (continued)

b. Distributions of sub-group matches to UTRsite patterns

UTRsite Acc.	Pattern names	UTRsite name	NO. of match sub-groups
U0002	IRE	Iron Responsive Element	9
U0003	SECIS-1	Selenocysteine Insertion Sequence	4
U0004	SECIS-2	Selenocysteine Insertion Sequence	39
U0006	CPE	Cytoplasmic polyadenylation element	53
U0007	TGE	TGE translational regulation element	6
U0009	15-LOX-DICE	15-Lipoxygenase Differentiation Control Element	1
U0010	ARE2	AU-rich class-2 Element	13
U0011	TOP	Terminal Oligopyrimidine Tract	118
U0012	GLUT1	Glusose transporter type-1 3'UTR cis-acting element	1
U0015	IRES	Internal Ribosome Entry Site	36
U0017	MSL2-3UTR	Male specific lethal 3'UTR cis-acting element	1
U0019	BRE	Bruno 3'UTR responsive element	6
U0020	ADH_DRE	Alcohol dehydrogenase 3'UTR downregulation control element	86
U0023	K-BOX	K-Box	559
U0024	BRD-BOX	Brd-Box	239
U0025	GY-BOX	GY-BOX	216

Table 3.9: Main repeat elements in carpBASE 2.1

	Number of elements	Length occupied (bp)	% of sequences
Low complexity	480	18891	0.58 %
Simple repeats	340	15059	0.46 %
Interspersed repeats		1430	0.04 %
LINEs	9	1218	0.04 %
SINEs	1	71	0.00 %
Small RNA	6	616	0.02 %
LTR elements	1	40	0.00 %
DNA elements	2	101	0.00 %

3.3.6 Properties of ESTs mapping

With EST-ferret, BLASTN searches were performed using carp ESTs to query the WU-contigs (Washington University zebrafish EST assemblies, http://zfish.wustl.edu/) and HGMP fugu ESTs (Clark *et al.* 2003). A tentative homology was ascribed to those sequences with a significance value <= 10^{-10}. This revealed that almost 50% of the carp ESTs (3003 of 6033) shared some homology with the current set of the Washington University zebrafish EST assemblies; whereas only 840 (13.9%) matched fugu ESTs; of which 789 (13.1%) matched both, and just 51 (0.8%) matched Fugu but not zebrafish sequences, and 2979 (49.4%) matched neither species. Thus, carp shares about four times more homologous sequences with zebrafish than Fugu. It was noted that genes which could be assigned a putative identification by homology using EST-ferret, were more likely to possess a homologous zebrafish sequence.

3.3.7 How to access the carpBASE

The EST raw data was processed by using EST-ferret, and results were stored firstly in flat files and then in the MySQL database. The LEGR searchable website was created to allow researchers to access the data *via* the Internet. To access the carpBASE, user can log onto the LEGR Data Centre at http://legr.liv.ac.uk and click on "carpBASE". It provides a tool for searching data, shown on Figure 3.8a. Keywords of clone ID, main-group ID, sub-group ID, BLAST hits, gene names, database accessions, GO terms or IDs and protein domain names are accepted for the searching. For example, for the annotation information for clone 17o10, users can type in 17o10 in the Clone ID field and click button "Search". Figure 3.8b lays out the resulting output page. The first column is for the query clone ID; the second column contains information for clustering; the third shows the top hits of BLAST searches against different databases; the fourth includes information of GO matches, EC numbers, protein domain matches, UTR patterns and repeat elements; the fifth mentions information on gene expression alignment and orthology analysis, which are described in Chapter 4 and Chapter 5. The output also gives links to sequence information and other external database references, like GenBank, Swiss-Prot, AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi),

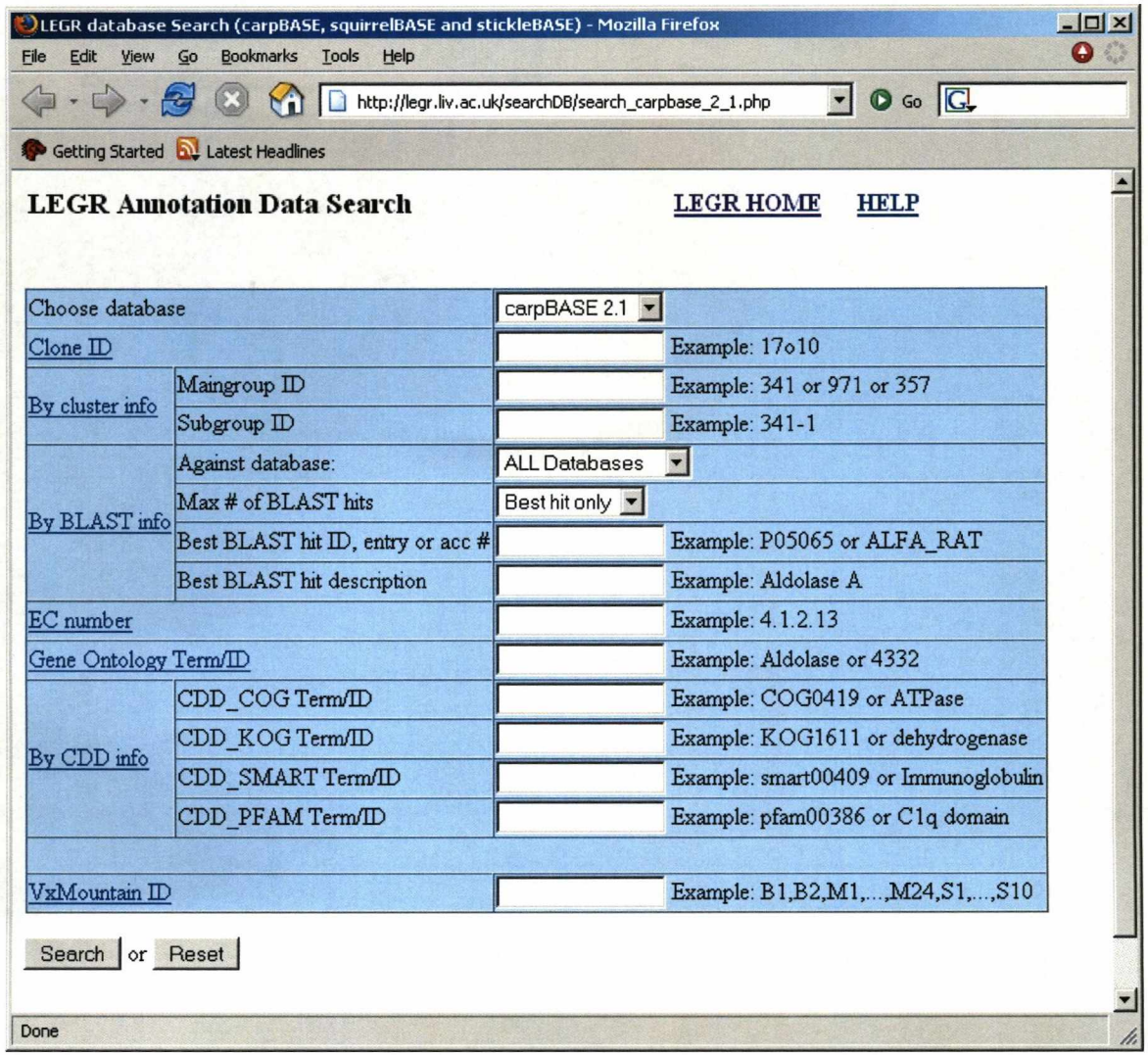


Figure 3.8a: Screen shot of the default web interface for searching in carpBASE 2.1 (July 2007)

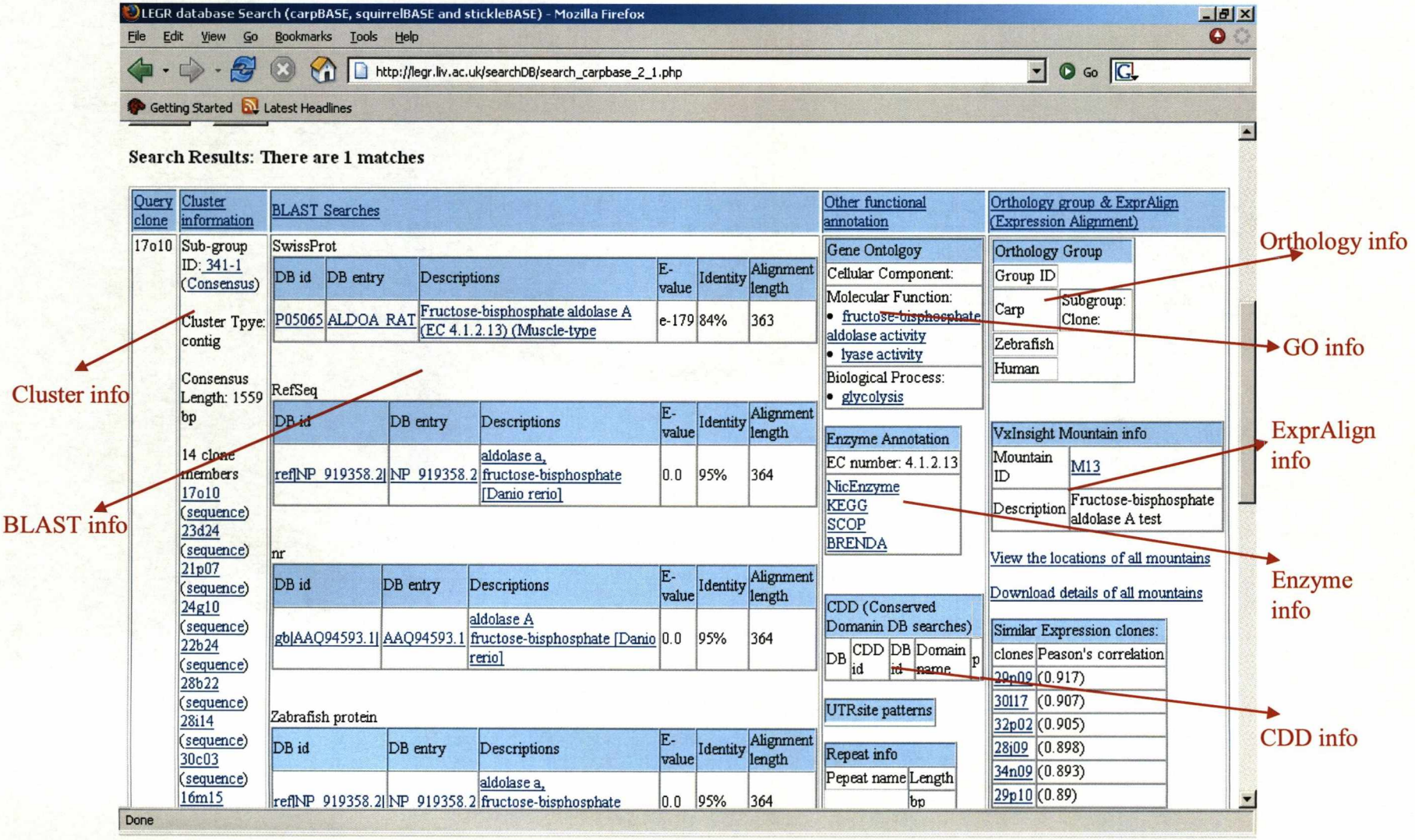


Figure 3.8b: Screen shot of the layout of the results from searching

NicEnzyme (http://www.expasy.ch/), KEGG, SCOP (Andreeva *et al.* 2004) and BRENDA (Barthelmes *et al.* 2007).

3.3.8 Other databases in LEGR

The EST-ferret and the protocols described here for common carp were also applied to establish other EST databases for other species of interest to the LEGR, such as roachBASE, squirrelBASE and troutBASE. Table 3.10 compares the annotation information of these databases. squirrelBASE 3.0 contains ~8800 ESTs and ~5000 EST sub-groups; troutBASE 2.0 included ~11300 ESTs and ~3200 sub-groups; and roachBASE 3.0 contains ~18,500 ESTs and ~9700 sub-groups. The average member size of sub-group for carpBASE 2.1, squirrelBASE 3.0 and roachBASE 3.0 were under 2, while the average member size of sub-group for troutBASE 2.0 was over 3.5. The strategy of constructing cDNA libraries for carp was similar to the strategies for squirrel and roach but different to that for trout. This might cause the significant difference of redundancies in the cDNA libraries for different species.

Table 3.10: Different databases constructed by EST-ferret

Databases	carpBASE 2.1	squirrelBASE 3.0	troutBASE 2.0	RoachBASE 3.0
No. of trace files	12,951	9475	15,035	19,290
No. of high-quality clean ESTs	9202	8803	11,282	18,477
No. of sub-groups	6033	4998	3168	9676
Average size of sub-groups	1.53	1.76	3.56	1.91
Sub-groups with BLAST Hit	53.90%	61.66%	76.77%	51.16%
Sub-groups with GO terms	32.70%	35.97%	38.64%	28.11%
Sub-groups with EC no.	8.20%	8.56%	7.14%	5.75%
Sub-groups with CDD match	2.5% [a]	33.71%	0.54% [a]	/
Sub-groups with UTR pattern	3.4% [a]	23.73%	10.26% [a]	/
Sub-groups with repeats	6.7% [a]	4.83%	2.75% [a]	/

[a]: CDD searching, UTR matching and Repeat searching only applied to unclassified sub-groups.

3.4 Discussions

Prior to the present study, there were few carp sequences in the public databases and a few microsatellite loci analyses (David *et al.* 2003). We have now added 9202 high-quality ESTs to dbEST and have sought to develop a convenient pipeline for the assembly and annotation of these into a fully featured database. We have also sought to map the carp ESTs onto the EST and genome sequence of its nearest genomic model, the zebrafish. These different facilities have all been combined into EST-ferret. The resulting carp database, carpBASE 2.1, presents the assembled contigs and singletons arising from 7 different tissues together with all available collated information in a convenient and searchable form. It also allows preparation of user-specific reports and downloads for incorporation in array analysis packages.

3.4.1 Benefits of techniques used in producing cDNA libraries

The diversity of the common carp clone set was comparatively high, and was efficiently collated; on average just 1.5 EST sequences were required for each sub-group, which compares favourably with equivalent values from fugu (2.41) (Clark *et al.* 2003) and Atlantic salmon (2.15) (Rise *et al.* 2004) EST projects. This arises from the techniques used in the production of the cDNA libraries and clone sets; (i) SMART technology (Zhu *et al.* 2001) to enhance the 5' sequence representation, (ii) serial subtraction to reduce redundancy between successive library extractions and (iii) recapture of full-length clones (Gracey *et al.* 2004). Reduced redundancy goes along with high levels of gene representation and given the relatively small number of sequences generated in this project there is a satisfactory representation of genes from all of the principle pathways of intermediary metabolism, protein catabolism etc. Preparing inserts with full-length cDNA assisted in the assembly process since the different clones were more likely to align than if they were weakly- or non-overlapping. Indeed, we found that only 986 sub-groups could be co-located the homologous full-length cDNA in the zebrafish collection. Also the number of singletons was high and the proportion of putative genes with 2 or more clones was small. Finally, the degree of

representation was indicated by the high proportion of genes comprising pathways of intermediary metabolism, such as glycolysis and lipid metabolism, where again most genes were represented. This broad representation has provided the essential basis on which transcript expression profiles have been interpreted (Gracey *et al.* 2004; Fraser *et al.* 2006).

3.4.2 Benefits of additional searches on CDD, UTRs and repeats

carpBASE 2.1 comprises an assembly of 9202 high quality sequences of which 3252 (53.9%) were identified using conventional BLAST homology searching. A series of additional procedures for the similarity searches against protein domains, UTRs and repeats held in publicly available databanks, were implemented within EST-ferret to explore the remaining 46.1% of sequences. This has suggested UTRs, protein domains and repeat elements for a further 204 (3.4%), 151 (2.5%) and 404 (6.7%) assembled ESTs, respectively, reducing the number of clones lacking any kind of annotation or identity to just 33.5% of the total, all as illustrated in Figure 3.4.

3.4.3 Benefits of the two rounds of clustering

Because of a recent genome duplication event (David *et al.* 2003), common carp genes are over-represented in the carp genome. Two-round of CAP3 clustering breaks ESTs down to different sub-groups that approximate to unique genes. For example, 25 ESTs were clustered into main-group 288 by the first round clustering at a low stringency of 66% identity. This was then split into two sub-groups by the second round of clustering at high stringency of 85% identity. The first (288-1) contained 11 clones while the second (288-2) contained 14 clones (Figure 3.9a) indicating their close relationship with each other was indicated by their clustering in the first round. A single round of CAP3 at 66% would not have generated two groups and a single round at 85% would have generated two groups but with no defined similarity relationship. The two sequential stages generated 2 putative genes within a putative gene family, both being supported by BLAST homology searches and sequence alignments in ClustalX (Thompson *et al.* 1997) (Figure 3.9b) to different but

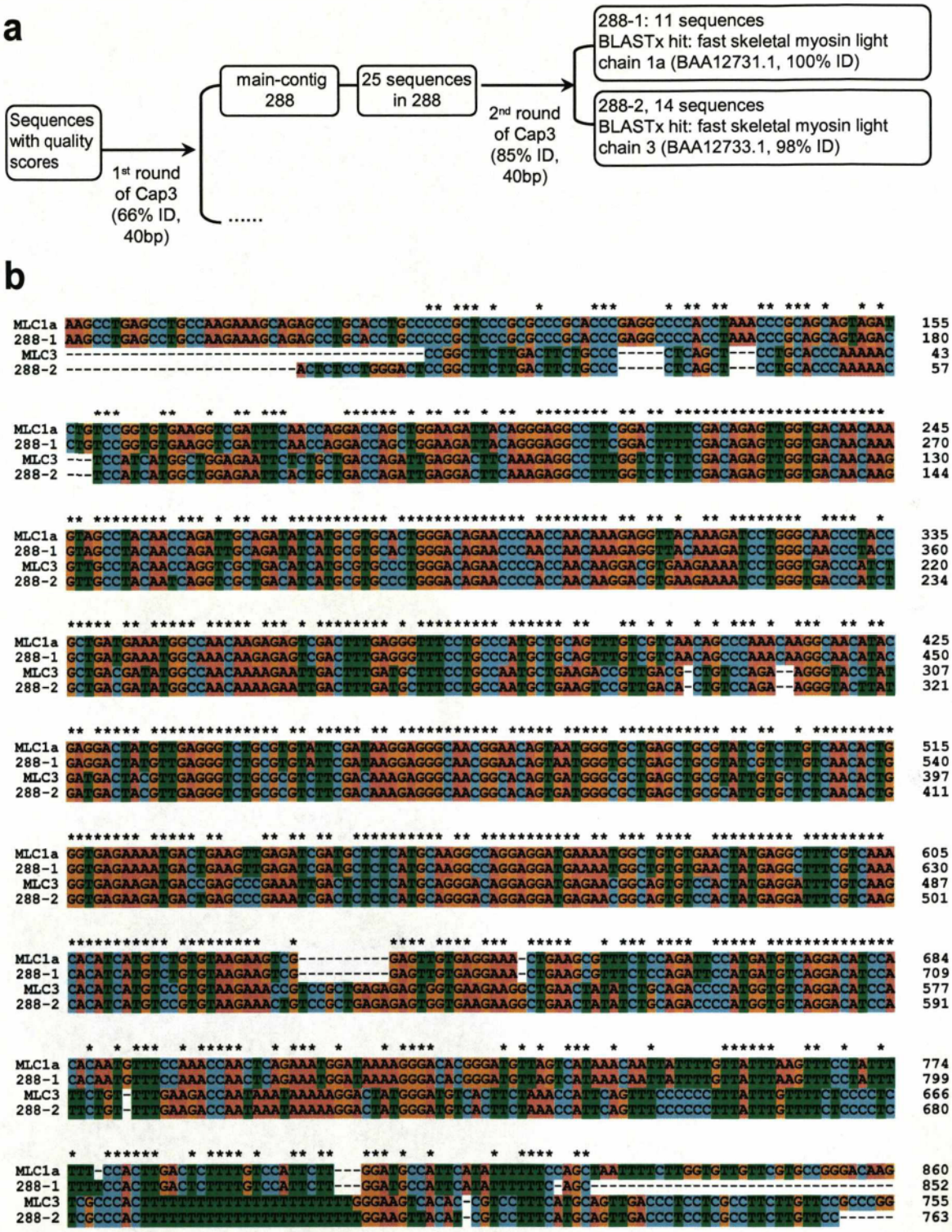


Figure 3.9: The alignment of carpBASE 2.1 main-group 288. (a) Two rounds of CAP3 clustering for main-group 288. (b) The alignments of sub-groups in main-groups 288 by ClustalX.

related myosin light chain genes (Hirayama *et al.* 1997) from skeletal muscle in grey mullet. One round clustering only tells whether sequences were similar or not. However, the two rounds of clustering not only indicate whether the sequences are similar or different, but also show how similar or different they were. For example, the 25 ESTs were clustered into 8 groups by using the PartiGene programme from Edinburgh University. These eight groups shared similar BLAST identities but did not indicate the relationship between each others. Thus, the two rounds of CAP3 clustering have the capability to illustrate the relationships for unique genes within a same gene family and genes in different genes families.

CHAPTER 4: CARP cDNA MICROARRAY DATA ANALYSIS

4.1 Introduction

4.1.1 Microarray data analysis

The microarray-based gene expression profiling technique allows the expression of thousands of genes to be measured simultaneously on a genome-wide scale. This technical advance was based on inverting the paradigm of the Northern blot technique, and on advances in robotic printing technique. The technique offers a massively parallel detection technique which for the first time is sufficient to address the needs of large-scale genome-wide screening experiments. However, the availability of such high dimensional data has necessitated the development of informatic tools and statistical methods to process and visualise the data into a form that is useful for biological inference.

High-density microarrays fall into two major categories: cDNA arrays and oligonucleotide arrays. cDNA arrays are fabricated by robotic deposition or "spotting" of DNA amplified from coding regions predicted in a genome sequence, or from cDNA clone sets. In contrast, oligonucleotide arrays comprise shorter oligonucleotides of 25-70 mer that are either synthesized chemically and then spotted onto the array, or are synthesized *in situ* on the array using a variety of methods such as ink-jet printing (Hughes *et al.* 2001) and photolithography (Barone *et al.* 2001). This chapter describes the data analysis of cDNA microarrays and Chapter 6 details collation of sequence data for rainbow trout oligonucleotide array design.

The raw data obtained from DNA microarray experiments are usually stored as image files, typically as .TIFF format files. The first step in extracting information from an array experiment is the analysis and processing of the image (Yang *et al.* 2001). This firstly identifies the array of features on the array. Secondly, for each feature it then quantifies the intensity of each fluorescence channel, generating numerical values for performing further statistical analysis. Each spot on the microarray contains two numerical components representing the signal and the background. Signal values

correspond with true microarray data while background values correspond to fluorescence unrelated to DNA hybridisation. Separating signal from background is an important step in the quantification process (Jain *et al.* 2002). Quantification can be absolute (signal intensity) or relative (ratio of absolute signals in two samples). According to these two quantifications, researchers separate DNA microarray data analysis into two types: (1) one-channel DNA data that reflect absolute fluorescence intensities; versus (2) two-channel DNA data that represent relative intensities or ratio data (McLachlan *et al.* 2004). The numerical values provide information that can be used to infer the relative concentration of mRNA in the samples.

In gene expression analysis, technical problems and biological variation combine to make it difficult to distinguish signal from noise. Normalisation of the measured expression level or ratio values adjusts the individual hybridization intensities to balance them across a range of arrays used for comparative experiments so that meaningful biological comparisons can be made (Bilban *et al.* 2002; Berger *et al.* 2004). There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labelling or detection efficiencies between the fluorescence dyes used, all giving rise to a systematic bias in the measured expression levels between compared sample (Baxevanis and Ouellette 2005).

An important issue in microarray experiments is the detection of genes that are differentially expressed in tissue samples across a number of specified classes. A single microarray slide that compares only two samples is limited to a list of regulated genes ranked by their magnitude of differential expression between the samples. In order to estimate significance of gene expression changes it becomes necessary to include biological and technical replicates in a statistically-based hybridisation scheme. The *t*-test (Levine *et al.* 2001) can be used to determine whether the expression of a particular gene is significantly different between control and sample. The test statistic t is the estimated difference in means between control and sample populations, divided by the estimated standard deviation of the difference. Then we can calculate the probability of a test statistic at least as extreme as observed, under the null

hypothesis that the population difference is zero (Knudsen 2004). An alternative method is ANalysis Of VAriance (ANOVA) (Mendenhall and Sincich 1988; Churchill 2004), which uses the *F* distribution (Montgomery and Runger 2003) to calculate the probability of finding the observed differences in means between more than two conditions when the null hypothesis is true.

Data analysis of larger experiments that seek to measure the same genes under different conditions, in different mutants, or at different time points frequently seek to group the significantly changed genes into groups, or clusters, that behave similarly over the different conditions. Cluster analysis (Armstrong and van de Wiel 2004) has demonstrated its utility in the interpretation and visualisation of gene expression patterns. One of the most widely used clustering approaches is hierarchical clustering (Jambu and Lebeaux 1983), which determines gene expression relationships based on the Euclidean distances between the respective data points. Hierarchical clustering is an agglomerative approach in which single expression profiles are joined to nodes, which are further joined until the process has been carried through o completion, forming a single hierarchical tree. The *K*-means clustering technique (Hartigan and Wong 1979), where *K* refers to the number of clusters to be imposed on the data, is similar to hierarchical clustering. It skips the calculation of distances between genes and can be implemented by randomly selecting *k* observations to be the initial *k* seeds (cluster centres). The observations are then visited in turn in some pre-specified order with an observation being assigned to the *i*th cluster if it is closest to the current mean of the *i*th cluster. After an observation has been assigned to a cluster, its mean is updated and the next observation is visited. The process is terminated when there is no change in the cluster memberships of the observations (McLachlan *et al.* 2004). Self-Organizing Maps (SOM) (Kohonen 1995) are similar to the *K*-means clustering technique, but clusters in SOM are ordered on a low-dimensional structure, such as a grid. The advantage over the *K*-means clustering is that neighbouring clusters in the grid are more related than clusters that are not neighbours, resulting in a structured ordering of clusters that is absent from the *K*-means clustering. The Principal Component Analysis (PCA)

(Jolliffe 2002) is a clustering method to visualize large-scale gene expression data by reducing dimensionalities. It is a multivariate procedure which rotates the data such that maximum variabilities are projected onto the axes. Distances between expressions of genes are usually required for clustering microarray data.

Most clustering methods rely on measures that calculate the distance between two genes in gene expression space, such as the Euclidean distance, the Pearson correlation coefficient (Cohen 1988) and the Spearman correlation coefficient. The Euclidean distance method measures space distances in absolute levels. The Spearman correlation coefficient uses ranks rather than raw expression levels which makes it less sensitive to extreme values. The Pearson correlation coefficient measures the relative shapes of the gene regulations to describe the similarities and differences between gene expressions. Thus it is a natural choice to compute the correlations between gene expressions (Kim *et al.* 2001).

4.1.2 Bioinformatics databases and tools for microarray data analysis

4.1.2.1 Microarray databases

The volume of microarray data has grown up dramatically over recent years. One early limitation of microarray studies was the lack of the standards for storing, presenting and exchanging the microarray data. So the *Minimum Information About a Microarray Experiment* (MIAME) (Brazma *et al.* 2001), a microarray data annotation standard, was defined to describe the minimum information required to ensure that microarray data could be easily reprocessed and be interpreted, and that results derived from its analysis can be independently verified. It requires reporting of specifications of experiment design, sample treatment, hybridisation conditions, data acquisition, normalisations, etc.

ArrayExpress (Brazma *et al.* 2003) is a public database of microarray gene expression data maintained by the EBI (Brooksbank *et al.* 2005), which is a generic gene expression database designed to hold data from all microarray platforms. It uses the MIAME and the Microarray Gene Expression Markup

Language (MAGE-ML) --- an XML (Extensible Markup Language, http://www.w3.org/XML/) based data exchange format, and it is designed to store well-annotated data in a structured way (Spellman *et al.* 2002). The MIAME and the MAGE-ML were developed by the Microarray Gene Expression Data (MGED) Society (Christian *et al.* 2003) and the Object Management Group (OMG, http://www.omg.org). The ArrayExpress database enables researchers to submit their microarray data online or directly from local databases in the MIAME standards, and allows other researchers to retrieve and understand the data based on these standards. As of January 2007, ArrayExpress contained data from >50,000 hybridizations and >1,500,000 individual expression profiles covering over 200 species (Parkinson *et al.* 2007).

Another database for expression profiles is the NCBI Gene Expression Omnibus (GEO) (Barrett *et al.* 2005), a large public repository for high-throughput molecular abundance data. These data includes not only microarray experiments but also serial analysis of gene expression (SAGE) (Velculescu *et al.* 1995) and mass spectrometry (Hu *et al.* 2005) experiments. The database has a flexible and open design that allows the submission, storage and retrieval of many data types. As of September 2006, GEO holds over 120,000 samples, representing over 3.2 billion measurements, covering over 200 organisms (Barrett *et al.* 2007),

The Stanford Microarray Database (SMD) (Sherlock *et al.* 2001; Ball *et al.* 2005) (http://smd.stanford.edu), one of the major microarray databases, was initially developed in 1999 to serve a small team of researchers using spotted DNA microarrays for human and yeast research at the Stanford University. Since then, it has become a research resource for a much larger scientific community using multiple microarray platforms to study a myriad of biomedical research problems.

4.1.2.2 Tools for analysing and visualizing microarray data

There are many tools for academic or commercial users developed for analysis of gene expression data, such as the maxd package (Hancock *et al.*

2005), GenePublisher server (Knudsen *et al.* 2003), GeneSpring, BioConductor (Gentleman *et al.* 2004; Durinck *et al.* 2005), GSEA (Subramanian *et al.* 2005), VxInsight (Davidson 2001), *etc.* The maxd (Hancock *et al.* 2005) is an academic package for loading and visualising gene expression data. It uses the MGED Ontology, therefore supports the MIAME standard and MAGE-ML format. It contains two sub-programmes: the maxdLoad, a relational database schema, and the maxdBrowse, a web-application for accessing maxdLoad data *via* browser, command line and web service. The maxd package also allows the user to search data, edit records, and generate appreciated-format output for proper submissions to ArrayExpress. The GenePublisher server (Knudsen *et al.* 2003) (http://www.cbs.dtu.dk/services/GenePublisher) is a web-based system for automatic processing data analysis for DNA microarray experiments. The user is allowed to upload the raw data to the server, and then the server will perform normalization and statistical analysis on the data and finally provide the data visualisation. In order to retrieve interested biological property from the data, the server will also search the result of interest against ontology databases, such as signal transduction pathways and metabolic pathways. The GeneSpring software (http://www.agilent.com/chem/genespring) is a commercial package for desktop gene expression data analysis developed by the Agilent Technologies. It is a highly scalable platform designed to meet the needs of the individual researcher, and has a well developed graphic user interface. It integrates data and results from multiple applications, and provides comprehensive statistical analysis, data mining, and visualisation tools for enterprise-level genomic research. The user can also share microarray data and results of analysis by uploading and downloading data to a central server, called SigNet. BioConductor (Durinck *et al.* 2005) is an open source and open development software project that provides a wide range of statistical and graphical tools for the analysis and comprehension of genomic data based on the R programming language (Crawley 2005) (http://www.r-project.org/). These tools are distributed as separate but interoperable packages, each specializing in different sub-areas of analysis such as the 'affy' package to normalize the Affymetrix chip data and the 'graph' package to handle graph

data structures. The BioConductor project is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics (Gentleman *et al.* 2004). Many tools have been developed and implemented in BioConductor for microarray studies, such as BioMart (Durinck *et al.* 2005), Simpleaffy (Wilson and Miller 2005), goCluster (Wrobel *et al.* 2005), MIDAW (Romualdi *et al.* 2005), stam (Lottaz and Spang 2005), maSigPro (Conesa *et al.* 2006), *etc.*

It is important to determine whether a group/set of genes with similar gene expression profiles is enriched for ontology terms. Thus, microarray data analysis is usually focused on gene groups/sets which share common biological processes, such as metabolic pathways, transcriptional programs, and stress responses. GSEA (Gene Set Enrichment Analysis) (Subramanian *et al.* 2005) is a tool for analysing microarray data at the level of gene sets to ease interpretation of a large-scale experiment by identifying pathways and processes. The VxInsight package (Davidson 2001) is a clustering and ordination algorithm used to mine extremely large databases, produce gene groups/sets with similar gene expression profiles, and provide graphical interfaces for visualising data in a 3-D virtual landscape. It can be used for exploring gene relationships in gene expressions using the Pearson correlation coefficients (Zar 1996). It presents a very intuitive visual representation of the data elements in which the geometric placement of the objects conveys significant information. The application of VxInsight to the microarray studies was introduced by Stuart Kim (Kim *et al.* 2001) and M. Martinez (Martinez *et al.* 2004).

4.1.3 Research objectives

In the LEGR, cDNA microarrays for common carp had been generated to investigate gene expressions in terms of environment changes, such as cold stress, hypoxia stress, starvation stress, *etc.* Sequences for the genes were annotated by EST-ferret and stored in carpBASE 2.1, which has been detailed in Chapter 2. But there were difficulties in establishing the identities and biological functions for many genes and gene groups with interesting

expression profiles. These genes or gene groups were deemed unclassifiable in carpBASE 2.1, or were associated with poor annotations that revealed little useful functional information. Examples included clones with poor quality EST sequences, or truncated sequences possessing only 5'- or 3'- UTRs, which had no open reading frames for identifying amino acid homology, or which did not overlap sufficiently with other ESTs for sequence alignment and the creation of contigs. For the purpose of annotating these unclassifiable genes, one option is to re-sequence followed by reanalysis. Another option is to use gene expression profiles to classify the sequences or at least to suggest identities. The first option was excluded in this study due to the extra costs of sequencing. The second option, also an aim of my project, involved the development of an approach, called ExprAlign (Expression Alignment), to relate unclassifiable sequences to annotated sequences by establishing their expression relationships in 3D gene expression space. My colleagues had processed the analyses in the GeneSpring package for the expression data across multiple tissues, but they had not implemented the global analysis across different stresses. In principle, the inclusion of profiles across multiple conditions or treatments increases the opportunities for exploring alignment of expression profiles. ExprAlign also aims to provide approaches to determine these patterns of gene expression across different tissues and across.

Members of co-expressed groupings of genes frequently share related functional properties (Stuart *et al.* 2003). One route to understanding the biological significance of a co-regulated group of genes is to analyse the distribution of Gene Ontology annotations within a grouping relative to their distribution across the expression data as a whole (Gracey *et al.* 2004). Another aim of my microarray studies was to produce a tool, called GOmatrix, to annotate gene groups in gene expression profiles with Gene Ontology annotations, and in turn to identify the statistically significant biological properties of gene groupings.

4.2 Materials and Methods

4.2.1 Common carp microarray data

The gene expression data used in this analysis comprises 707 common carp RNA samples, hybridised to 1414 cDNA microarrays. The carp microarrays were constructed by my colleagues in the LEGR from 13,349 PCR-amplified cDNA clones onto poly-L-lysine coated glass slides by using standard techniques. The cDNA clones were picked from a high-quality collection of common carp cDNA libraries which had been normalized and subtracted for reducing clone redundancy. The cDNA libraries were enriched for environmentally regulated genes and cDNAs were labeled by using amino-allyl adducts coupled to Cy3 and Cy5 fluorescent dyes. The microarrays were hybridized overnight at 65°C, washed and scanned, and finally the images were analyzed (Axon Instruments, Union City, CA). The data which I worked on has been normalised by Dr. Y. Fang (Fang *et al.* 2003) at the University of Manchester. 189 RNA samples were generated from the study of cold stress, 414 for the study of hypoxia stress, and 104 for the study of starvation stress. The data for each of the different stresses contains experimental samples taken at different time points and different tissues. The 'cold' data includes samples for brain, gill, heart, intestine, kidney, liver and muscle for three temperatures (either 23°C, 17°C, or 10°C) ; the 'hypoxia' data contains experiment samples for brain, heart, intestine, liver and muscle for two temperatures (17°C and 30°C); and the 'starvation' data contains samples for liver and muscle.

Individual analyses of gene expressions on the cold stress and the hypoxia stress have been published by Andrew Gracey (Gracey *et al.* 2004) and Jane Fraser (Fraser *et al.* 2006). The 'starvation' data is unpublished work. Here I looked for features of gene expressions across multiple stresses and multiple tissues. The analysis by the Expression Alignment utilised the data from these three stresses; while the study by the GOmatrix focused on the data from the cold stress. Data for cDNA clones which shared a same BLAST identity or had no BLAST identity matches were retained in the analysis by the Expression Alignment, since one of my purpose in this analysis was to related unknown clones to identified clones.

4.2.2 ExprAlign --- Expression Alignment

Expression Alignment, or ExprAlign, was developed to process the gene expression analysis in this study. Figure 4.1 illustrates the main features of the pipeline adopted for ExprAlign.

4.2.2.1 Pearson correlation coefficients for gene expression patterns

To analyse the gene regulations and response to external stresses, three methods were considered to measure and test gene correlations: the Euclidean distance, the Spearman correlation coefficient and the Pearson correlation coefficient (Zar 1996). The most familiar distance measure is the Euclidean distance (Knudsen 2004; Baxevanis and Ouellette 2005), which is the straight-line distance between two points in Euclidean space. In a Euclidean three-dimensional space, the distance between points (x_1, y_1, z_1) and (x_2, y_2, z_2) is given by

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}. \qquad \text{(Equation 4.1)}$$

The generalisation of this to higher-dimensional expression spaces is straightforward. In general, the distance between points x and y in a Euclidean space is given by

$$d = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}. \qquad \text{(Equation 4.2)}$$

Where x_i and y_i are the measured expression values, respectively, for genes x and y in hybridization i, and the summation run over the n hybridizations under analysis. The equations above shows the Euclidean distance measures of the absolute level of gene regulation. But two genes whose expression levels were perfectly parallel to one another across the database could still be far apart in Euclidean space if the absolute levels in each experiment were different. Therefore, the Euclidean distance would not be appropriate for this study.

The Spearman (Rank) correlation coefficient is a non-parametric test for the strength of the relationship between pairs of variables. It uses ranks rather than raw expression levels which makes it less sensitive to extreme values in gene expression data (Kim *et al.* 2001).

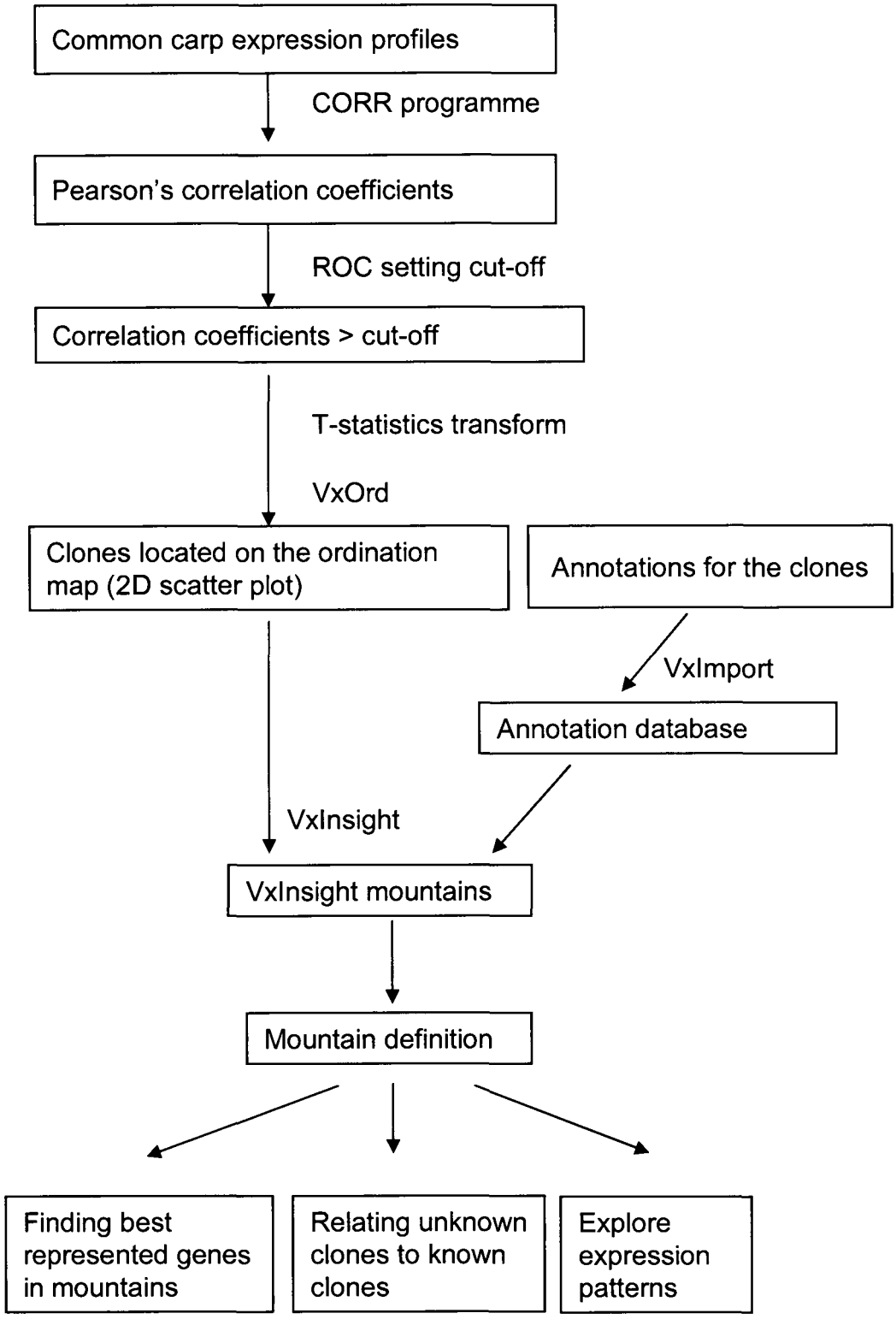


Figure 4.1:ExprAlign Pipeline for carp global expression data

The Pearson correlation coefficient measures the strength of the linear relationship between two variables. It can measure the relative shapes of gene regulations using gene expression data rather than the absolute levels (Kim *et al.* 2001). Therefore, it describes the similarities and differences in patterns between two genes, and is a natural choice to measure gene expression relationship. The Pearson correlation coefficient, used in the ExprAlign to measure the expression similarity, is calculated as following for two variables x and y:

$$r = \frac{\Sigma [(X_i - u_x) (Y_i - u_y)]}{\sqrt{\Sigma (X_i - u_x)^2 \Sigma (Y_i - u_y)^2}} \qquad \text{(Equation 4.3)}$$

The denominator in the equation is always positive, though the numerator may be negative, zero, or positive; also, the numerator can never be larger than the denominator which means r's value must always be between -1 and +1. A positive correlation means that with the increase in the value of one variable, the value of the other variable will increase as well; a negative correlation indicates that an increase in one of the variable's value is accompanied by a decrease in the value of the other variable. A correlation of 0 implies there is no linear association between the two variables. A high positive correlation indicates a strong association between variables while a high negative correlation indicates a strong inverse association (Zar 1996). Genes with no expression similarity will have a value near 0.0, while genes that are strongly similar in expression will have a value near 1.0.

4.2.2.2 Programming to calculate Pearson correlation coefficients

A number of packages for gene expression analysis, including as GeneSpring and maxd, provide multiple tools for processing, filtering, clustering and displaying expression data in user-friendly interfaces. However, analysis of the data obtained from microarray images still poses a number of challenges due to huge data sets. The original data set contained a large number of records (columns) spread over a large number of fields (rows). Given 1000 genes, the calculation of the Pearson correlation coefficients will require 1

million individual calculations. Therefore, the computing and mining application for the purpose of dealing with large data sets must be robust enough to be able to deal with the large dimensionality and be able to complete the calculations in a timely manner for large data sets.

Most of the current available popular packages for microarray data analysis do not provide powerful applications for computing the Pearson correlation coefficients. Our collaborator, Mr. Faraaz Yusufi produced a MatLab application (Yusufi 2004) for the calculations of Pearson correlation coefficients. MatLab (MATrix LABoratory, http://www.mathworks.com/) is a software package for high-performance numerical computation and visualisation (Pratap 2002). It provides an interactive environment with hundreds of built-in functions for technical computations, graphics and animations. It also offers easy extensibility with its own high-level programming language. Faraaz Yusufi output the coefficients for 386 common carp cDNA microarrays (~14,112 probes each) using his MatLab application. It took about 7 days to finish the job in 7 parallel computers.

MatLab is a useful tool for statistical analysis of microarray data (Bevington and Robinson 1992; Drapner and Smith 1998; Branch *et al.* 1999). But the speed is unacceptably slow for undertaking the calculations of Pearson correlation coefficients for large-scale microarray datasets. Therefore I produced a C programme called CORR to speed up the calculations in the Linux system. C codes are compiled into machine assembly language, thus applications written in C programme language (Oualline 1991) tend to be comparatively fast. The C code itself is highly portable for different computer operation systems (Ullman and Liyanage 2005). The CORR programme skips the procedure of depositing coefficient matrix in the RAM, but stores coefficient for each pair of genes directly into a flat text file. These allow it to run much faster than the MatLab application generated by F. Yusufi. On a computer of 1.70GHz CPU and 512MB RAM, the CORR just took about 40 minutes to finish the same job as described above.

4.2.2.3 ROC curves to optimise thresholds for correlation scores

A pair of cDNA clones derived from a same gene should share sequence overlap and should in turn poses similar gene expression profiles. Probes constructed from these clones should be clustered together if they are reverse-transcribed from a same region of the gene. In theory their expression correlation coefficient of these two probes should be close to 1. But in practice, gene expression data is noisy and differences in the exact sequence presented in the probe leads to expression differences. This raises the question of how large a correlation score needs to be in order to assign a pair of cDNA clones as copies of the same gene. The first task was to define a threshold for correlation scores such that if the correlation for two clones was larger than the threshold, then the two clones were likely to be derived the same gene and that their ESTs should be clustered in a same group in carpBASE 2.1.

The EST clustering as implemented in EST-ferret and in the construction of carpBASE 2.1 was used to define the sequence relatedness of each clone in the dataset. The Relative Operating Characteristic (ROC) was implemented to test the usefulness of the search statistics (http://www.anaesthetist.com/mnm/stats/roc/) in order to optimize threshold for correlation scores. The optimized threshold can be obtained by plotting the sensitivity (True Positive, P^+) of the comparison against the selectivity (False Positive, P^-) (Anderson and Brass 1998).

Sensitivity: $P^+ = t^+ / (t^+ + f^-)$ (Equation 4.4)

Selectivity: $P^- = t^- / (t^- + f^+)$ (Equation 4.5)

where t^+ is a true positive: two sequences are in a same sub-group (in carpBASE 2.1 the second round of CAP3 clustering) and have a gene correlation score above threshold. f^- is false negative: two sequences are in a same sub-group but have gene correlation score below threshold; t^- is a true negative: two sequences are in different sub-groups and have gene correlation score below threshold; and f^+ is false positive: two sequences are in different sub-groups but have gene correlation above threshold. If a specific threshold value is defined, it is therefore possible to assign all corrections as true positives, false negatives, true negatives or false positives. The key question is

how to select a best threshold for the correlation scores. The criterion is that the best threshold should be able to minimise the total number of errors. The sensitivity P^+ indicates the probability of the observed true positives at a threshold, so the probability of the missed true positives at a threshold can be given by $(1 - P^+)$. On the other hand, the selectivity P^- shows the probability for the observed true negatives, so the probability of the missed true negatives at a threshold can be given by $(1- P^-)$. Finally, the total probability of the missing of the true positives and the true negatives can be given by $E = (1 - P^+) + (1 - P^-)$. The best threshold should be able to minimise E. A set of thresholds can be tested to calculate the values of P^+, P^- and E. When E is minimised, the optimal threshold is found.

If the transcript expression of two array probes possessed a significant correlation coefficient, we defined these two sequences as a sequence-pair. If sequences in a sequence-pair came from a same sub-group, the sequence-pair was defined as a matched sequence-pair; otherwise, an un-matched sequence-pair. PERL scripts were written to extract the matched sequence-pairs and the un-matched sequence-pairs by examination of their correlation scores. With this information, distributions of t^+, f^-, t^-, and f^+ were established. With the distributions, the minimum of E could be found and the optimal cut-off could be identified. For example, if the threshold was 0.9, the true positive (t^+) would be matched sequence-pairs with correlation scores above 0.9; the false negative (f^-) would be matched sequence-pairs with the correlation scores under 0.9; the false positive (f^+) would be un-matched sequence-pairs with correlation scores above 0.9; and the true negative (t^-) would be un-matched sequence-pairs with the correlation scores under 0.9. The curve for E can be illustrated by a line chart in the Microsoft Excel programme.

4.2.2.4 VxInsight to visualise expression alignments

The Pearson correlation coefficients of the carp gene expressions were be produced by the in-house C programme CORR and the ROC method determined the threshold for the coefficients. These scores stand for the similarities for each pair of genes in expressions. The resulting data were stored in a spreadsheet consisting of a few thousand gene-pairs (the rows), with

coefficient scores, and were visualised using the VxInsight package (Davidson *et al.* 1998; Davidson 2001).

Other visualisation tools for clustering, such as TreeView (http://rana.lbl.gov/EisenSoftware.htm) and GeneSpring (a product of Agilent Technologies), can display the gene expression patterns of clusters on 2 dimensional space. The visualisations of these packages are useful for exploring relationships between two individual genes or among a small set of genes but difficult for interpreting relationships for a large scale of genes.

The VxInsight package uses a terrain metaphor to describe large collections of data, summarizing clusters of similar elements by placing them physically close to each other in the terrain. It consists of three parts: VxOrd, VxInsight and VxImport. VxOrd implements the force-directed ordination algorithm (Fruchtermann and Rheingold 1990) to assign X, Y coordinates in a 2-dimensional surface to each gene based on the coefficients of the gene pairs. Then these coordinates are used to generate the 3-dimensional mountain terrains in which mountains are separated by valleys and open spaces. The heights of the mountains indicate the number of elements clustered together under each mountain. The local groupings and separations between mountains also carry information about the inter-cluster similarities. The data elements in widely separated mountains will have less similarity than those in neighbouring mountains (Davidson 2001). The landscape map can be zoomed in or out in VxInsight to allow user to view data over different scales from the complete overview down to an individual gene. Also it is easy to label the features of the landscape with descriptors of gene identity or gene function. These features make it extremely useful to visualise and interpret complex patterns of gene expression and explore relationships between genes and gene groups in large data set. VxImport loaded the gene annotation into VxInsight for biological interpretations.

Using the raw correlations unduly weights the low similarities and does not adequately represent the information content contained in a strong similarity. The non-linearity of this information is extreme and can change the total range of observed similarity weights by orders of magnitude. Here gene

pair similarities were based on the t-statistic of the correlation coefficient, not on the correlation coefficient itself. The calculation for the t-statistic was as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{(Equation 4.6)}$$

where r = correlation score, n = number of data point. In this study for the data from the multiple tissues across multiple stresses, n was 707.

The expression clusters were created in the VxOrd 1.58 (Davidson 2001) using clone pair similarities based on the t-statistic of the correlation coefficients and mapped on a two-dimensional scatter plot. The VxInsight 2.145 converted the 2-dimensional map into 3-dimensional terrain map, in which the Z axis denoted the density of the clones within an area. The number of hills and the numbers of clones in a hill were not predetermined. The VxImport 0.4.06 generated an annotation database for clones in the visualisation map using annotations in carpBASE 2.1.

4.2.3 GOmatrix

4.2.3.1 Gene expression groups and its GO annotations

Dr. Andrew Gracey provided 23 gene groups or clusters that exhibited differential expression with cold and showed tissue-specific patterns of expression as defined by the *K*-means clustering (Appendix 4.1) (Gasch and Eisen 2002; Gracey *et al.* 2004), as well as a single gene group that was differentially expressed in all tissues with cooling. These 24 gene groups were used as gene lists for input in the GOmatrix analysis. carpBASE 2.1 provided GO annotations for genes in the microarray experiments. GOprofiler, embedded in the EST-ferret package, was used to retrieve GO annotations for each of 24 gene groups. It was able to extract GO annotations for non-redundant gene groups. To identify GO categories in which a gene group was enriched (over-represented) or depleted (under-represented), it was necessary to remove all redundant entries for each gene group. The pipeline (Figure 4.2) shows the steps to make the GOmatrix.

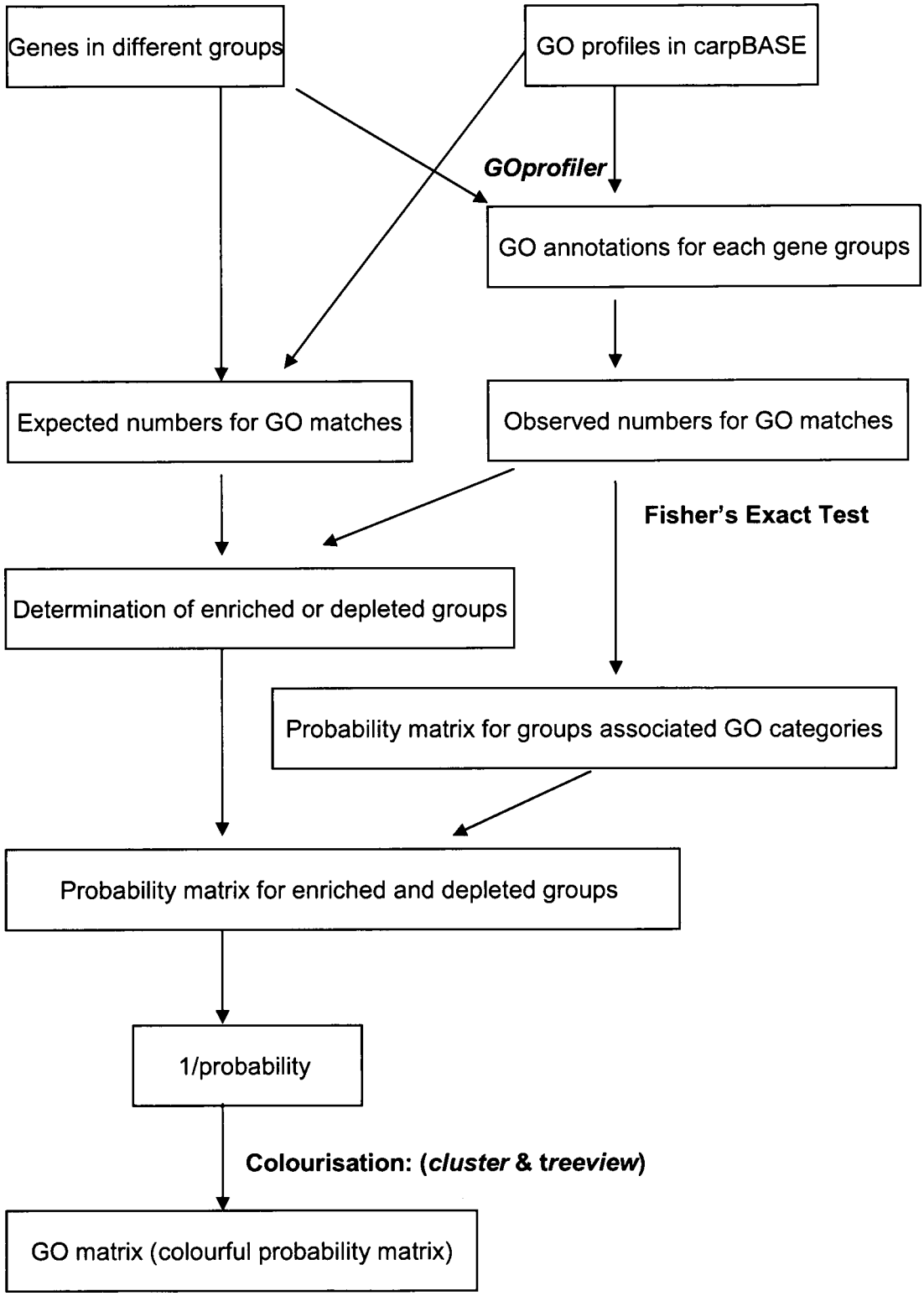


Figure 4.2: Pipeline for GOmatrix

4.2.3.2 Fisher's exact test to build the probability matrix

The Fisher's exact test (Dawson-Saunders and Trapp 1994; Weisstein 2006) is a statistical test used to test association between two categorical variables in a 2X2 contingency table. Unlike the chi-square test (described in Section 3.2.4), which generates estimate probabilities, the Fisher's exact test calculates an exact probability value for the relationship between two variables.

In the GOmatrix, the Fisher's exact test was used to test whether genes within an expression group were significantly over- or under-represented (enriched or depleted) for a GO sub-category in relation to the representation in the gene collection as a whole. The 2x2 contingency table (Table 4.1 a) indicates the observed frequencies according to two categorical variables: the gene group X and the Gene Ontology sub-category Y. *a* is the observed number of genes in the gene group X associated with GO sub-category Y; *b* is the observed number of genes in the gene group X not associated with GO sub-category Y; *c* is the observed number of genes not in the gene group X but associated with GO sub-category Y; and *d* is the observed number of genes not in the gene group X and not associated with GO sub-category Y. The marginal totals are represented by $a+b$, $c+d$, $a+c$, $b+d$, and the grand total is represented by n.

Table 4.1 a: A contingency table

	Genes associated with GO sub-category Y	Genes not associated with GO sub-category Y	Totals
Genes in group X	*a*	*b*	*a+b*
Genes not in group X	*c*	*d*	*c+d*
Totals	*a+c*	*b+d*	*n (=a+b+c+d)*

Table 4.1 b-f: Sample contingency tables

b

2	5	7
6	5	11
8	1	18

c

1	6	7
7	4	11
8	10	18

d

0	7	7
8	3	11
8	10	18

e

6	1	7
4	7	11
8	10	18

f

7	0	7
3	8	11
8	10	18

Given the fixed marginal totals, the hypothesis was defined for the data as:

H_0: There is no association between occurrence of genes associated with GO sub-category Y and occurrence of genes within gene group X.

H_a: Genes within gene group X are significantly over-represented or under-represented for a GO sub-category Y.

Given the fixed marginal totals, if there was no association between the variables X and Y, the probability of observing such an arrangement of the data in the table can be given by:

$$P = \frac{\frac{(a+c)!}{a!c!} \times \frac{(b+d)!}{b!d!}}{\frac{n!}{(a+b)!\,(c+d)!}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!a!b!c!d!} \qquad \text{Equation 4.7}$$

Given the observed marginal totals, Fisher's exact test computes the probability of observing the data as extreme or more extreme. This means it also generates other probabilities by Equation 4.7 for other data tables which are as extreme or more extreme in the same direction (one-tailed) or in both directions (two-tailed). Given the observed marginal totals, a more extreme table has a smaller probability of occurrence in the same direction (one-tailed) or in both directions (two-tailed). For example, if a sample observed table is shown as Table 4.1 b, Table 4.1 c and d are the extreme tables in the same direction and Table 4.1 e and f are the extreme tables in the opposite direction. Probabilities can be calculated by Equation 4.7 for each of these tables. For one-tail Fisher's exact test, the final probability is the sum of the probabilities from Table 4.1 b, c and d; for two-tails Fisher's exact test, the final probability is the sum of the probabilities from Table 4.1 b, c, d, e and f. One tail test can be applied when the association and the association direction between X and Y are already known. Because we did not know whether the association exists and what the association direction is, two-tails Fisher's exact test was used in the calculation of the sum of probabilities. If the final *P* is less than

significance level 0.05, genes within a gene group X are significantly over- or under-represented (enriched or depleted) for a GO sub-category Y.

The Chi-square test (described in Section 3.2.4) and the Fisher exact test work for similar purposes. However, the Chi-square only gives estimated *P*-values and the Fisher exact test returns exact *P*-values. An estimate might be insufficient if the marginal totals are very uneven or the observed value is small (less than 5) in one of the cells. In my study, the number of genes in different gene groups falling into GO sub-categories could be very small, so the Fisher exact test is a better choice than the Chi-square in this case.

Given 10 expression gene groups and 20 GO sub-categories, 200 2X2 contingency tables were constructed and 200 probability values were calculated. A Java programme was written in the Linux system for computing the two-tailed probability values of each 2 x 2 contingency table and laying them onto a 2-dimensional matrix.

4.2.3.3 Determining over-represented and under-represented gene groups

The probability matrix indicates the significance of gene representations in the particular GO sub-categories within each gene groups, but does not describe whether genes are over-represented or under-represented in particular GO sub-categories. Comparisons the observed gene numbers and the expected gene numbers can give the answer. Knowing the number of genes in the whole gene collection and the number of genes associated with the GO sub-categories in the whole gene collection, we were able to generate expected number of genes associated with the GO sub-categories in different expression groups. Comparing each pair of the observed gene number (observed values, *O*) and the expected gene number (expected values, *E*), we can determine over-represented (enriched) and under-represented (depleted) for each gene group associated with particular GO sub-categories. If $O-E > 0$, we can judge that genes in the gene group are enriched in the GO sub-category; and if $O-E < 0$, we can judge that genes in the gene group are depleted in that GO sub-category.

4.2.3.4 GOmatrix coloration

The probability matrix was labeled as enriched or depleted in a Microsoft Excel spreadsheet. Cluster and TreeView (Eisen *et al.* 1998), normally employed for gene expression analysis, was also used to colorize the GOmatrix. These two programmes can be used to cluster and analyse gene expression profiles, but here they were used only for clustering the probabilities and colourising the GOmatrix. In the coloured GOmatrix, red was used to indicate the enriched and blue the depleted GO categories. The density of the colour illustrates the level of significance.

4.3 Results

4.3.1 VxInsight mountains from ExprAlign

4.3.1.1 Optimizing correlation cut-off

Gene expression data combining the experiments of cold stress, hypoxia stress and starvation stress were termed the global data. Theoretically, clones derived from the same gene should share identical expression profiles in the global data and it should be possible to assign an identity to unclassified clones if their gene expression profiles were sufficiently similar to other well annotated genes.

The Pearson correlation coefficients for ~13,500 clones were computed across the global data. The Relative Operating Characteristic (ROC) was used to optimize the threshold. Due to the large size of the data set only those spots that had a correlation higher than 0.8 were recorded. If the threshold was > 0.9, the false negative (f^-) is defined as matched sequence-pairs with the correlation scores between 0.8 and 0.9; similarly, the true negative (t^-) is defined as un-matched sequence-pairs with the correlation scores between 0.8 and 0.9. The optimal cut-off was the one that minimizes the total error ($E = (1- P^+) + (1- P^-)$) in the ROC method. In Figure 4.3, the blue solid line is the curve of E. Actually the ROC curves were not plotted in Figure 4.3. But we found the cut-offs for correlation scores were 0.858 by minimizing the E. 2121 true positive (matched) clone-pairs were found using this threshold.

4.3.1.2 VxInsight mountains

Sequence-pairs with correlation scores over the cut-off of 0.858 served as input for the VxInsight package for gene expression clustering and visualisation. The data after *t*-statistic transformation were imported into the VxOrd program and a 2-dimensional coordination map (Figure 4.4) was established. 3039 clones were positioned on the coordination map. The 2-dimensional map was then transformed into a 3-dimensional landscape using VxInsight and annotations for the 3039 clones were imported onto the map through the VxImport. In the VxInsight 3D map (Figure 4.5), mountains were composed of clusters of clones with high gene expression correlations. The

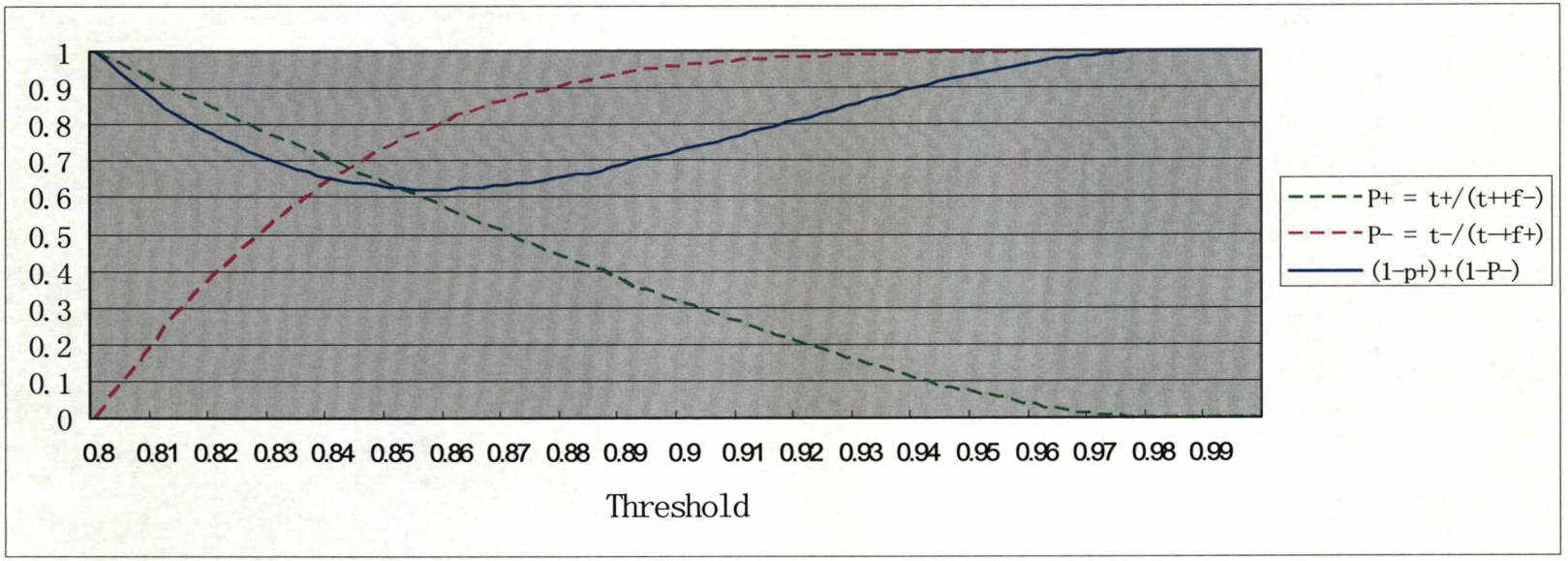


Figure 4.3: Determining the cut-off value of the Pearson correlation coefficient for carp global expression data

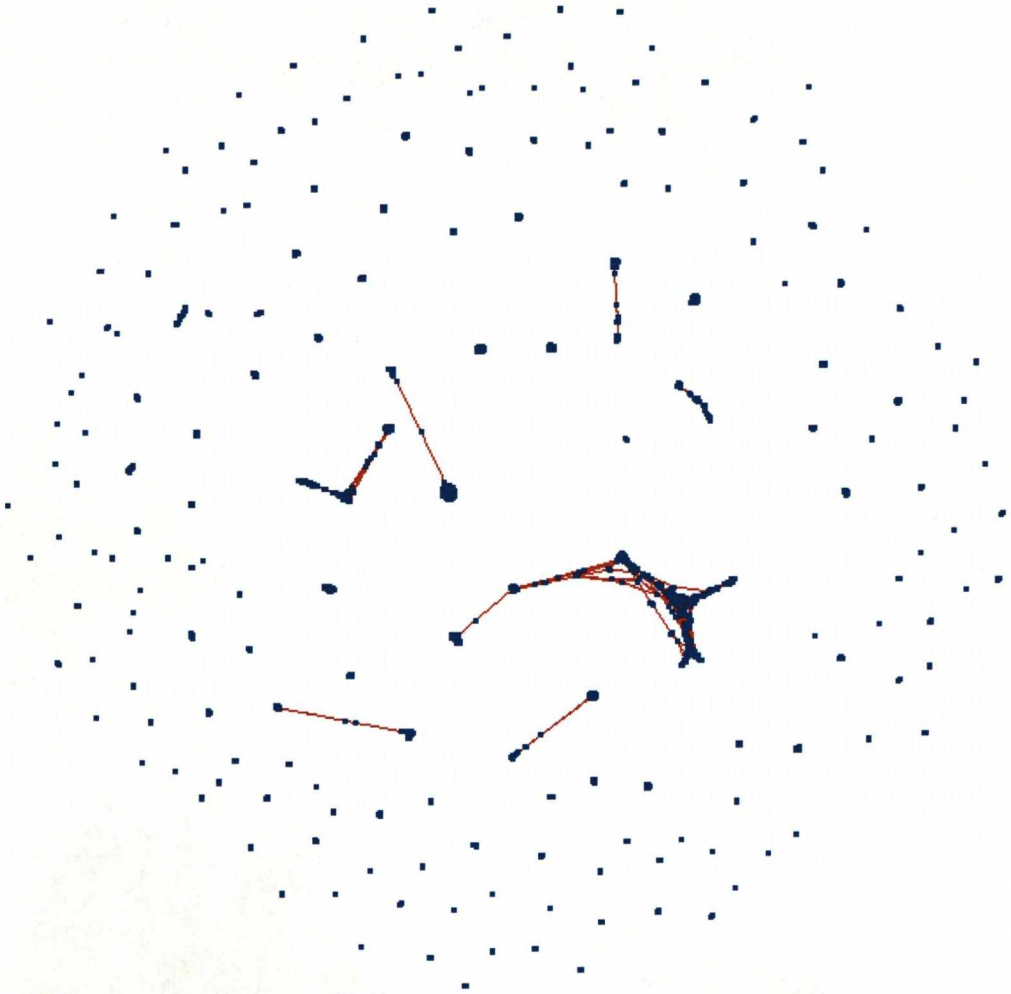


Figure 4.4: 2D map for carp global expression data by VxOrd in ExprAlign

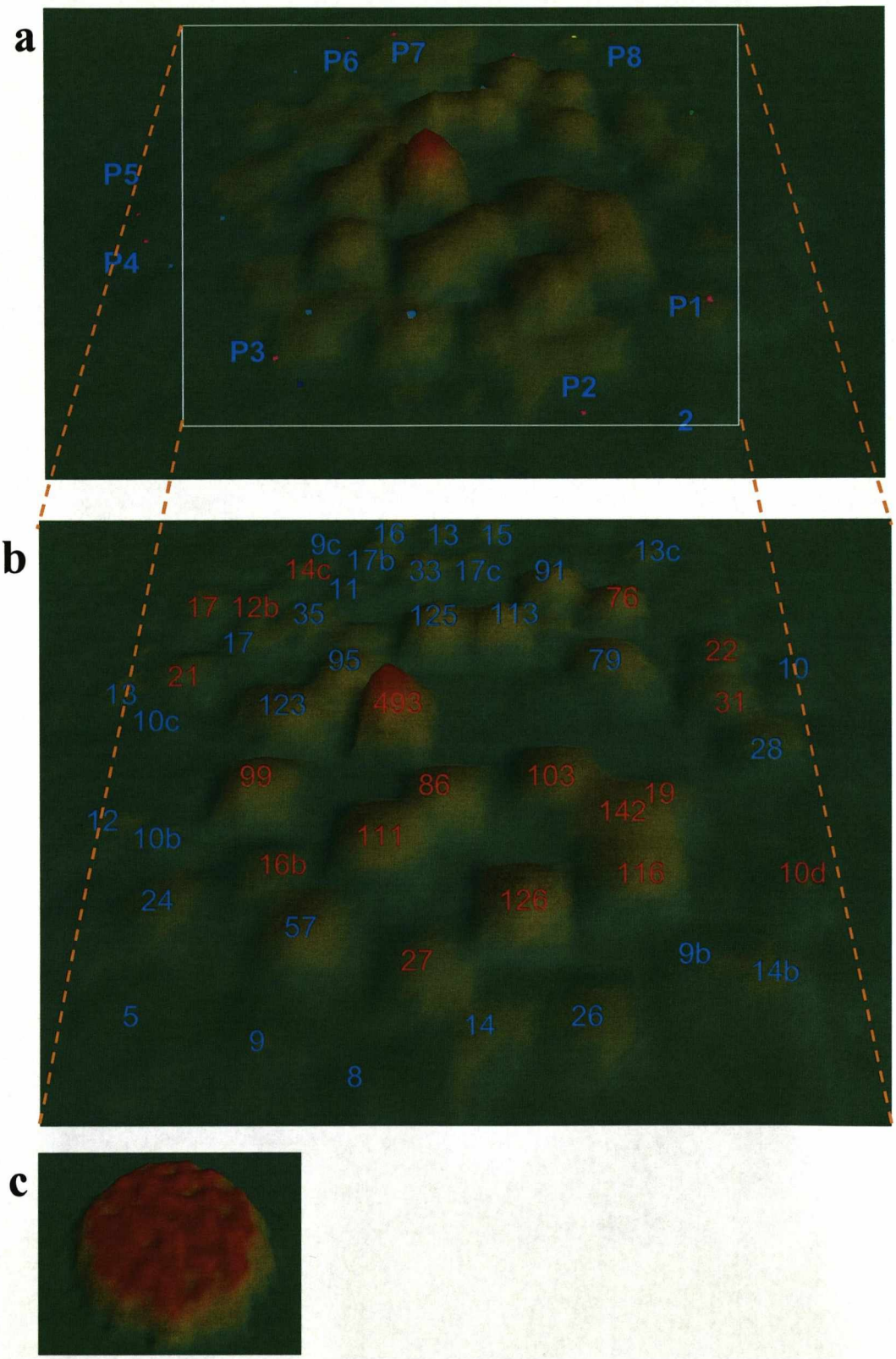


Figure 4.5: GE mountains for carp global expression data in VxInsight. (a) A overview map for the mountains. P1 to P8 indicate the locations of different parvalbumin isoforms. (b) A zoomed map for main mountains. Each GE mountain is labelled with the number of clones with the mountain. Identified mountains are labelled in blue and unknown mountains are labelled in red. (c) Terrain map derived from random-shuffled data.

height of the mountains corresponded to the number of clones. For any clone positioned on the 3D map, the neighbours could be identified or else they would remain unknown. Of the 3039 clones, 1192 were identified and 1847 were unknown in carpBASE 2.1. The annotations of the classified clones offered a route to establishing the identity for adjacent but unclassifiable clones.

Mountains contained sets of highly correlated clones and were named with a prefix "GE", followed by the number of clones within the mountain (Figure 4.5). "GE" here indicates that the groups were "Global ExprAlign" groups generated from the combined gene expression data for cold, hypoxia and starvation. If mountains had the same numbers of clones, a suffix of 'a', 'b', or 'c', and so on, was applied to distinguish them. If a mountain was dominated by a single gene, this single gene was taken as the best representative gene to describe the mountain and summarise the biological identities of the mountain. A best representative gene only represents the largest proportion of clones with a single identity in a mountain and clones represented must be greater than 2 in number and over 60% of identified clones in the mountain. Otherwise, mountains containing non-identified clones or small amount identities or diverged clone identities would be described as unknown. 50 mountains containing 2590 clones were located on the map. Each had 5 clones or more each. 18 mountains were unknown and 32 mountains were identified (Table 4.2). The biggest mountain GE493, unknown, contained 493 diverged or unclassified clones. Identified mountains were labelled with blue numbers in Figure 4.5b. For example, GE35 represented apolipoprotein A-I, GE113 represented glyceraldehyde-3-phosphate dehydrogenase, GE91 represented skeletal alpha-actin, GE17b represented fructose-bisphosphate aldolase A, GE33 represented fructose-bisphosphate aldolase B, GE28 represented ADP/ATP translocases, etc. Other unknown mountains are labelled with red numbers in Figure 4.5b.

To assess the significance of the topographical patterns shown in Figure 4.5b, we randomized the expression table by shuffling the values for all probes across all arrays and then re-clustered the genes in the VxInsight package. We

Table 4.2: Summary for identified GE mountains

GE Mou-ntain	No. of identified clones	Best represent clones						No. of Relatable unknown clones
		Protein description	No. in mountain	% of identified clones in mountain	% of clones in mountain	No. in carpBASE 2.1	*p* to the carpBASE 2.1	
125	48	14 kDa apolipoprotein	34	70.8	27.2	65	3.17E-60	77
123	53	Ribosomal proteins	39	73.6	31.7	339	5.86E-38	70
113	55	Glyceraldehyde-3-phosphate dehydrogenase	50	90.9	44.2	70	6.26E-101	58
95	23	Ribosomal proteins	19	82.6	20	339	1.20E-20	72
91	46	Skeletal alpha-actin	36	78.3	39.6	71	6.49E-65	45
79	39	Apolipoproteins	33	84.6	41.8	113	3.87E-53	40
57	24	Creatine kinases	20	83.3	35.1	74	5.93E-36	33
35	15	Apolipoprotein A-I	13	86.7	37.1	47	8.90E-27	20
33	28	Fructose-bisphosphate aldolase B	26	92.9	78.8	30	4.11E-65	5
28	26	ADP/ATP translocases	26	100	92.9	43	1.67E-60	2
26	16	Fibrinogen	14	87.5	53.8	35	4.01E-31	10
24	6	Creatine kinases	4	66.7	16.7	74	3.37E-07	18
17	9	Transferrin variant A	8	88.9	47.1	44	4.41E-17	8
17b	15	Fructose-bisphosphate aldolase A	13	86.7	76.5	30	7.62E-30	2
17c	13	Fatty acid-binding protein	13	100	76.5	28	2.28E-32	4
16	9	Parvalbumins	8	88.9	50	114	1.36E-13	7
15	6	Vitellogenin	6	100	40	10	3.62E-18	9
14	5	Apolipoprotein Eb precursor	5	100	35.7	19	1.97E-13	9
14b	14	Parvalbumins	14	100	100	114	4.53E-25	0
13	9	Transferrin variant A	7	77.8	53.8	44	2.79E-14	4
13b	7	Acidic mammalian chitinase precursor	7	100	53.8	11	6.75E-21	6
13c	10	Carp Desaturase 2 (CDS2)	10	100	76.9	15	2.17E-28	3
12	9	Troponin T, fast skeletal muscle isoforms	9	100	75	22	2.12E-23	3
11	9	Apolipoprotein C-I precursor	9	100	81.8	24	5.56E-23	2
10	5	Myoglobin	5	100	50	6	1.01E-16	5
10b	8	Warm-temperature-acclimation-related-65 kDa-protein	5	62.5	50	15	2.83E-12	2
10c	9	Uncoupling protein 1	9	100	90	10	4.25E-28	1
9	8	C-type lectin	8	100	88.9	19	2.10E-21	1
9b	9	Invariant chain like protein 2	9	100	100	20	7.14E-24	0
9c	6	Elongation factor 1-alpha; EF-1-alpha	6	100	66.7	13	2.95E-17	3
8	6	Alcohol dehydrogenase	6	100	75	24	2.32E-15	2
5	4	RING finger protein 28	4	100	80	6	2.99E-13	1
Total	**549**							**522**

Number of a best represented gene in a mountain must be > 2, and its percentage of clones in mountain must be over 20%. *P* represents the significance of the gene which is over- or under-represented in the mountain comparing to the whole data set in carpBASE 2.1.

observed no appreciable structure in the randomized terrain map (Figure 4.5c), suggesting that the geography observed in the actual expression map (Figure 4.5a and b) arises from the structure of the data rather than being a property of no biological significance.

4.3.1.3 Data independency

To determine whether the structure of the gene co-expression network was affected by the dataset used in its construction, the GE dataset was recalculated using only the data from the cooling experiment. For the cold data only, the optimal ROC cut-off of correlation scores was determined to be 0.864. 2656 true positive (matched) sequence-pairs were identified using this threshold. 4236 clones were positioned on the VxInsight 3D map (Figure 4.6), of which, 1776 were identified and 2460 were unknown in carpBASE 2.1. Mountains were again numbered according to size. 46 mountains were found and 22 of those were identified. Mountains were named as described above, but with prefix "CE", which indicated they related to Cold ExprAlign groups.

If the resulting gene expression network was independent of the data used in its construction, the component contained within the CE mountains should be identical or similar to those in the GE mountains. Therefore, the components in each GE mountain were compared to those in each CE mountain to judge the differences and the agreements between mountains. Figure 4.7 illustrates the agreement between the global map and the cold map as a matrix of gene identities for each of the identified mountains. For example, 100 clones (~85%) of the 113 clones in GE113 were also found in CE109b. So GE113 was highly similar to CE109b, and both were comprised largely of clones for the glyocolytic enzyme, glyceraldehyde-3-phosphate dehydrogenase. The number 100 was displayed in the cell linking GE113 and CE109b in Figure 4.7. The colour of the box represents the percentage of the agreement; dark-red indicates high level of agreement, while light-red shows a low level of agreement, and an open cell indicates no agreement. The agreement of GE113 and CE109b was calculated as (100/113)*100%. There were 21 highly similar mountain-pairs between GE mountains and CD mountains, listed in Table 4.3.

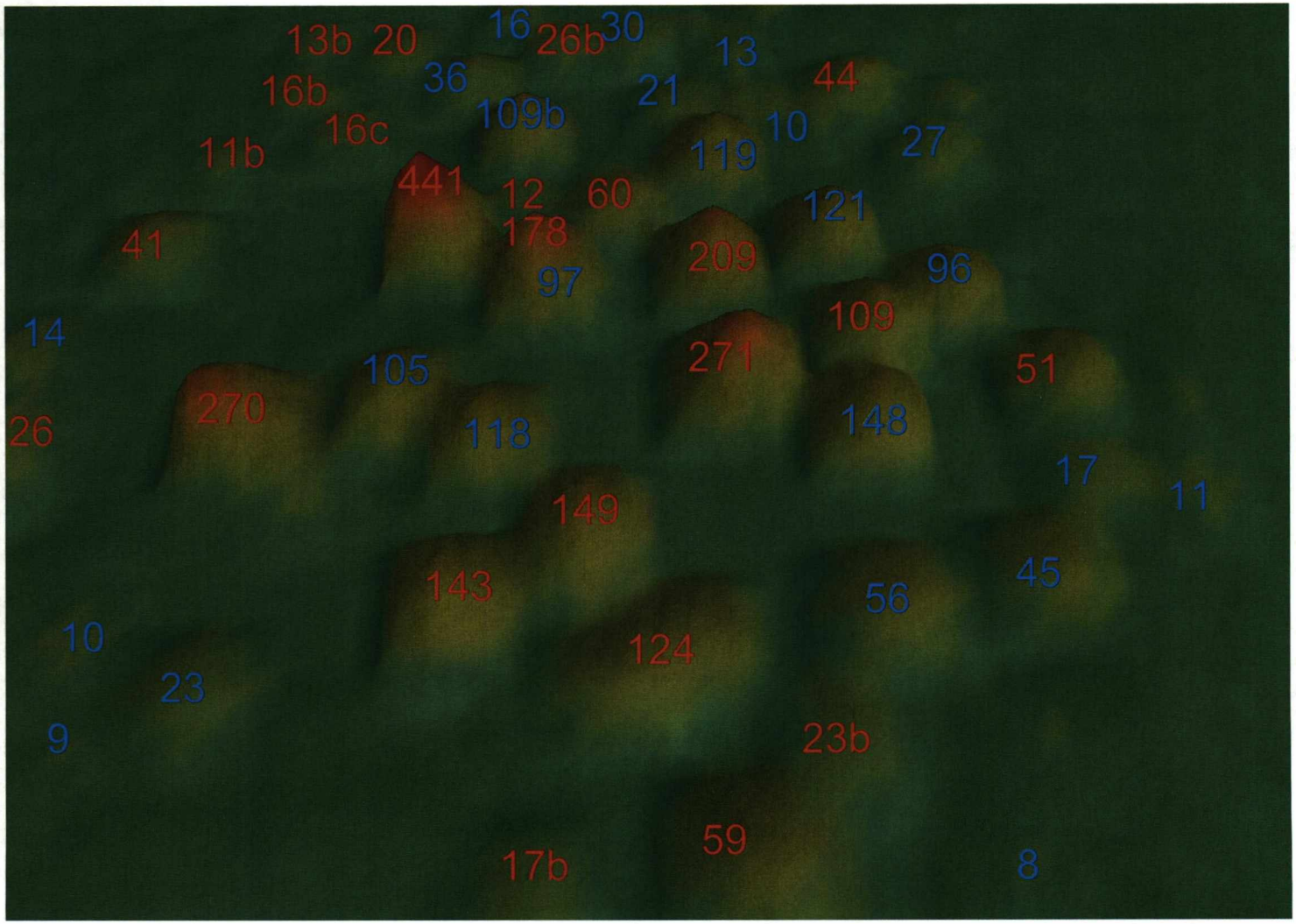


Figure 4.6: VxInsight mountains for carp cold expression data

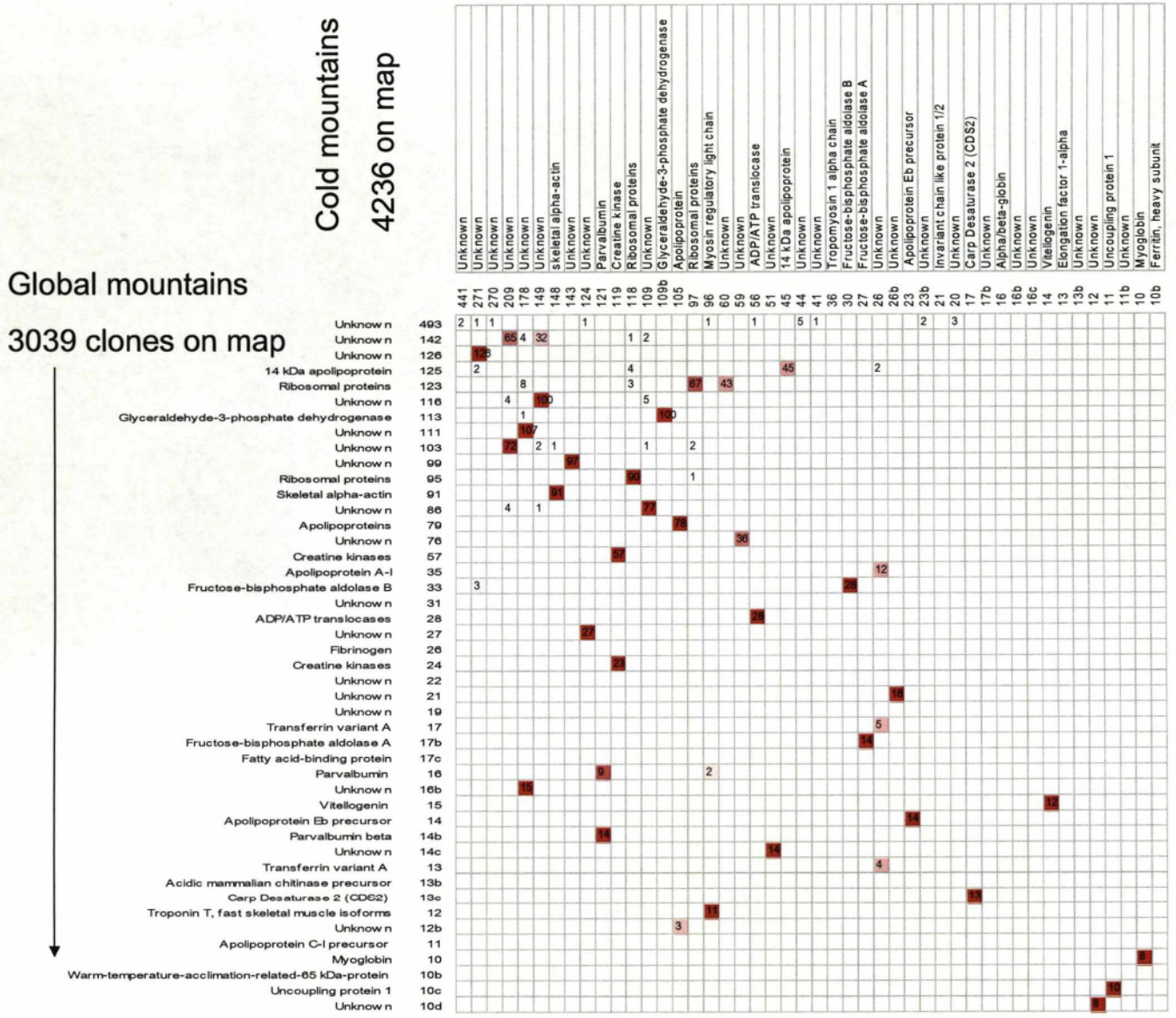


Figure 4.7: Matrix to compare the distribution of gene identities in global and cold mountains

Table 4.3: Comparison of GE mountains and CE mountains indicated in Figure 4.7

Best represent genes	GE mountains	CE mountains	Agreements
Unknown	126	271	126
Unknown	116	149	100
Glyceraldehyde-3-phosphate dehydrogenase	113	109b	100
Unknown	111	178	107
Unknown	99	143	97
Ribosomal proteins	95	118	90
Skeletal alpha-actin	91	148	91
Unknown	86	109	77
Apolipoproteins	79	105	76
Fructose-bisphosphate aldolase B	33	30	28
ADP/ATP translocases	28	56	28
Unknown	27	124	27
Unknown	21	26b	18
Fructose-bisphosphate aldolase A	17b	27	14
Vitellogenin	15	14	12
Apolipoprotein Eb precursor	14	23	14
Unknown	14c	51	14
Carp Desaturase 2 (CDS2)	13c	17	13
Myoglobin	10	10	8
Uncoupling protein 1	10c	11	10
Unknown	10d	12	8

The foregoing comparison indicates that the underlying structure of the gene co-expression network was largely independent of the scale of dataset used in its construction. However, there were also differences between the GE and CE landscapes. As shown on the Figure 4.7, the clones in CE209 were mainly separately located in GE142 and GE103; the clones in CE178 were separately located in GE111 and GE16b; and the clones in CE149 were separately located in GE142 and GE116. Although these mountains contained probes of unknown identity, the global data illustrates its benefits of separating a gene group generated from the cold data alone into two parts with more extensive expression datasets. Additionally, CE26 linked to GE35 (apolipoprotein A-I), GE17 (transferrin variant A) and GE13 (transferrin variant A). Moreover, CE119 (creatine kinases) linked to GE57 (creatine

kinases) and GE24 (creatine kinases); and CE121 (parvalbumins) were broken down to GE16 (parvalbumins) and GE14b (parvalbumins). Creatine kinases and parvalbumins have many isoforms. Genes in GE57 and GE24 could be different isoforms of creatine kinases and genes in GE16 and GE14b could be different isoforms of parvalbumins. This analysis suggests that more expression data can give greater definition to the extent of gene clustering.

4.3.1.4 Data robustness

To access the data robustness of the approach, the clones and the arrays were both randomly reduced 50% and resulted in a random dataset containing 25% of the global data. The random dataset was used to construct another mountain map (Figure 4.8). 2444 clones were located on the RE (Random ExprAlign) mountains. 27 mountains were found and 12 were identified. Mountains were named in a same way mentioned above, but with prefix "RE", which indicated they were Random ExprAlign groups. The Figure 4.9 illustrates the agreement between the global map and the random map as a matrix of gene identities for each of the identified mountains. Each RE mountain possessing an identity was highly similar to a GE mountain also possessing the same identity. 9 of the unknown RE mountains were linked to GE mountains of un known identity. These show high similarities of the components between the GE mountains and the RE mountains and suggest the robustness of the ExprAlign method implemented in the clustering analysis. Of course, there were also differences between the GE map and the RE map. The RE26 (Transferrin variant A) linked to two GE mountains: GE17b (transferrin variant A), GE33 (transferrin variant B) and GE10c (uncoupling protein 1). This also suggested more data can give more definition to the clustering.

4.3.1.5 Relating unclassifiable clones to identified genes

For the identical mountains in the GE map, we computed *p*-values using Fisher's exact test to determine the significance of best represented genes compared to the whole gene set in all mountains. If the *p*-value was less than the critical significance level (0.001), unknown clones were relatable to the

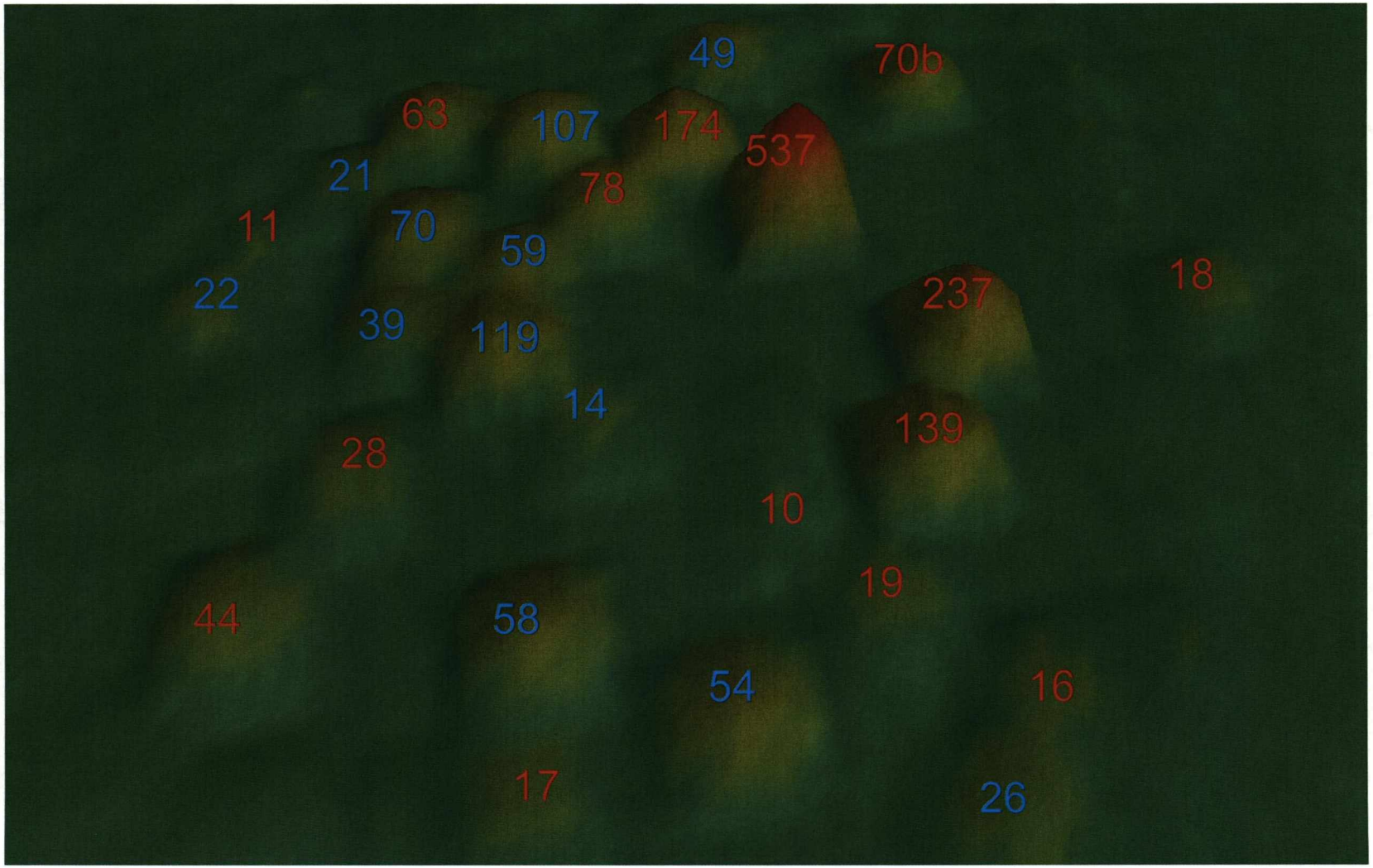


Figure 4.8: RE mountains of carp random expression data in VxInsight. Each RE mountain is labelled with the number of clones with the mountain. Identified mountains are labelled in blue and unknown mountains are labelled in red.

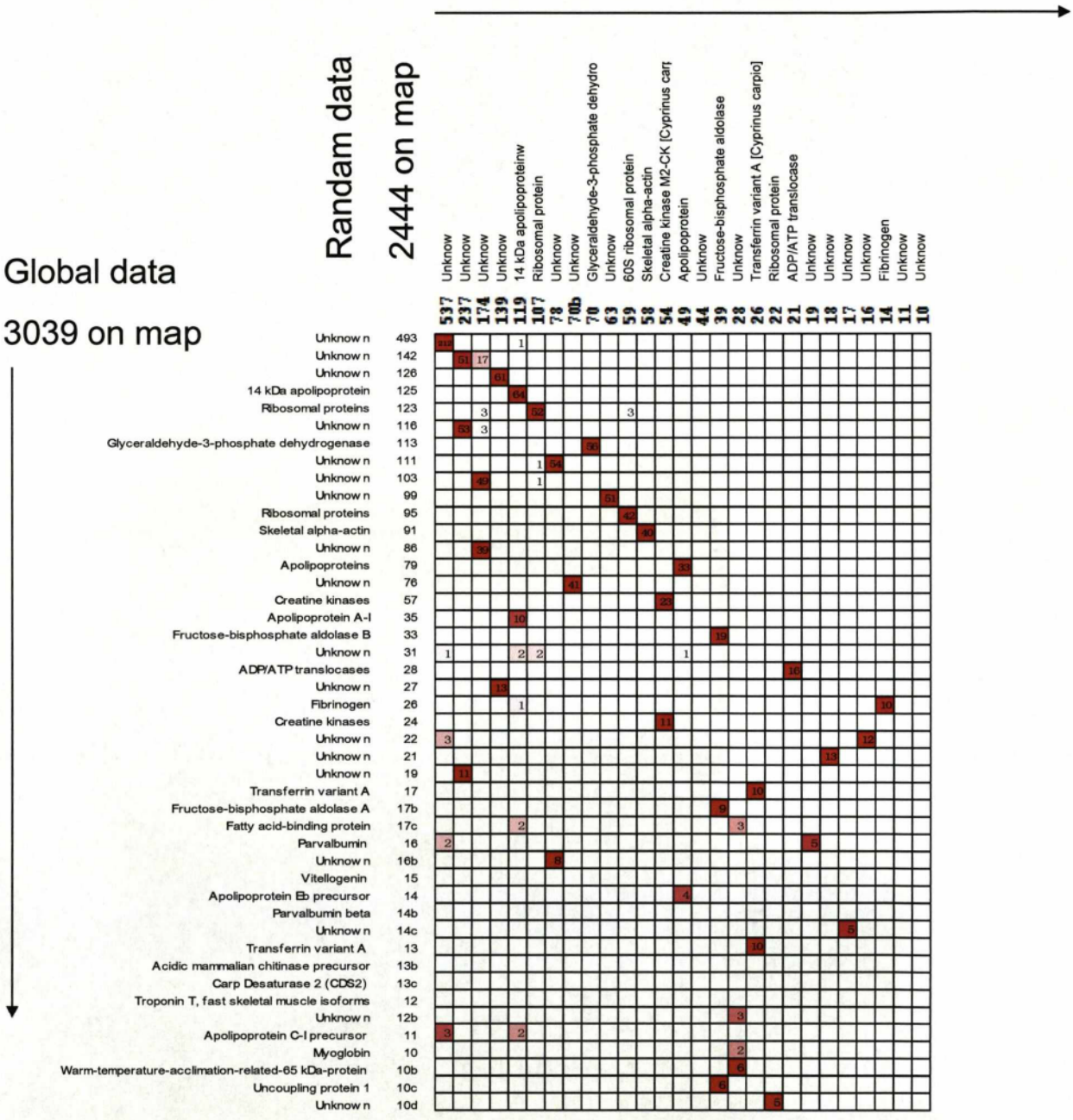


Figure 4.9: Matrix for agreement and difference of global mountains and random (half-half) mountains

best represented gene in the classified mountains (Table 4.2). For example, in GE113, 55 clones were identified as glyceraldehyde 3-phosphate dehydrogenase, which represented ~91% identified clones. The *p*-value (6e-101) was under the critical level. The 58 unknown clones were thus inferred as being related to glyceraldehyde 3-phosphate dehydrogenase. This implies that the unknown clones might be glyceraldehyde 3-phosphate dehydrogenase, or other genes relatable to glyceraldehyde 3-phosphate dehydrogenase in biological functions or processes. In the 31 classified mountains, there were 522 clones, over 17% of the 3039 that were relatable to the BLAST-identified gene in their mountains. They could be the same gene as the represented gene in the mountain or other gene relatable to the represented gene in the biological functions.

4.3.1.6 Expression patterns in GE mountains

Unknown mountains provided insufficient information for the identities of some particular clones. But the gene co-expression relationship implies some biologically meaningful relationship for the probes in each of those mountains. The heatmaps of gene expression for genes in unknown mountains are shown on Figure 4.10. The numbers on the left side of the figure indicates the mountain names and their size. On the header, "Cold" shows the data range for cold experiments, "Hypoxia" for hypoxia experiment data and "Starvation" for starvation experiment. "B" indicates the experiment was taken for the tissue brain, "G" for gill, "H" for heart, "I" for intestine, "K" for kidney, "L" for liver, and "M" for muscle. The heatmaps in the centre show the patterns of gene expression for genes in different mountains. Red colour indicates up-regulated and green colour shows down-regulated. For example, genes in GE21 and GE12b were hypoxia-inducible in tissue liver. The biggest unknown mountain, GE443, contained genes that were substantially up-regulated in the hypoxia liver at 17°C, while the second biggest unknown mountain, GE142, contained genes that were down-regulated in the hypoxia liver at 17°C. Transcription of genes in the former was also up-regulated in hypoxia liver and cold intestine.

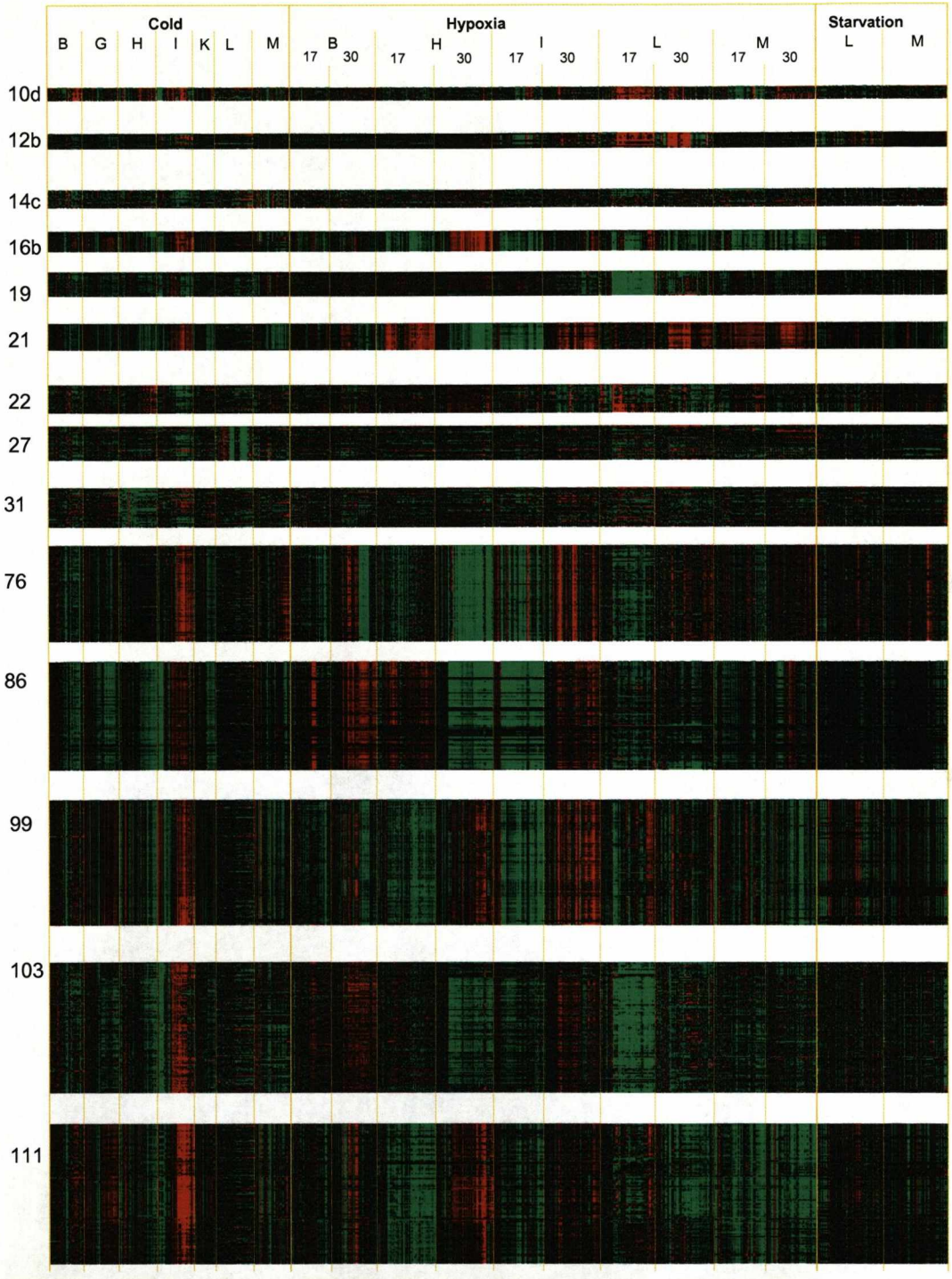


Figure 4.10: Heatmaps for unknown GE mountains. The numbers on the left side indicates the mountain names and their size. On the header, cold shows the data range for cold experiments, Hypoxia for hypoxia experiment data, and Starvation for starvation experiment data. B stands for brain, G for gill, H for heart, I for intestine, K for kidney, L for liver, and M for muscle.

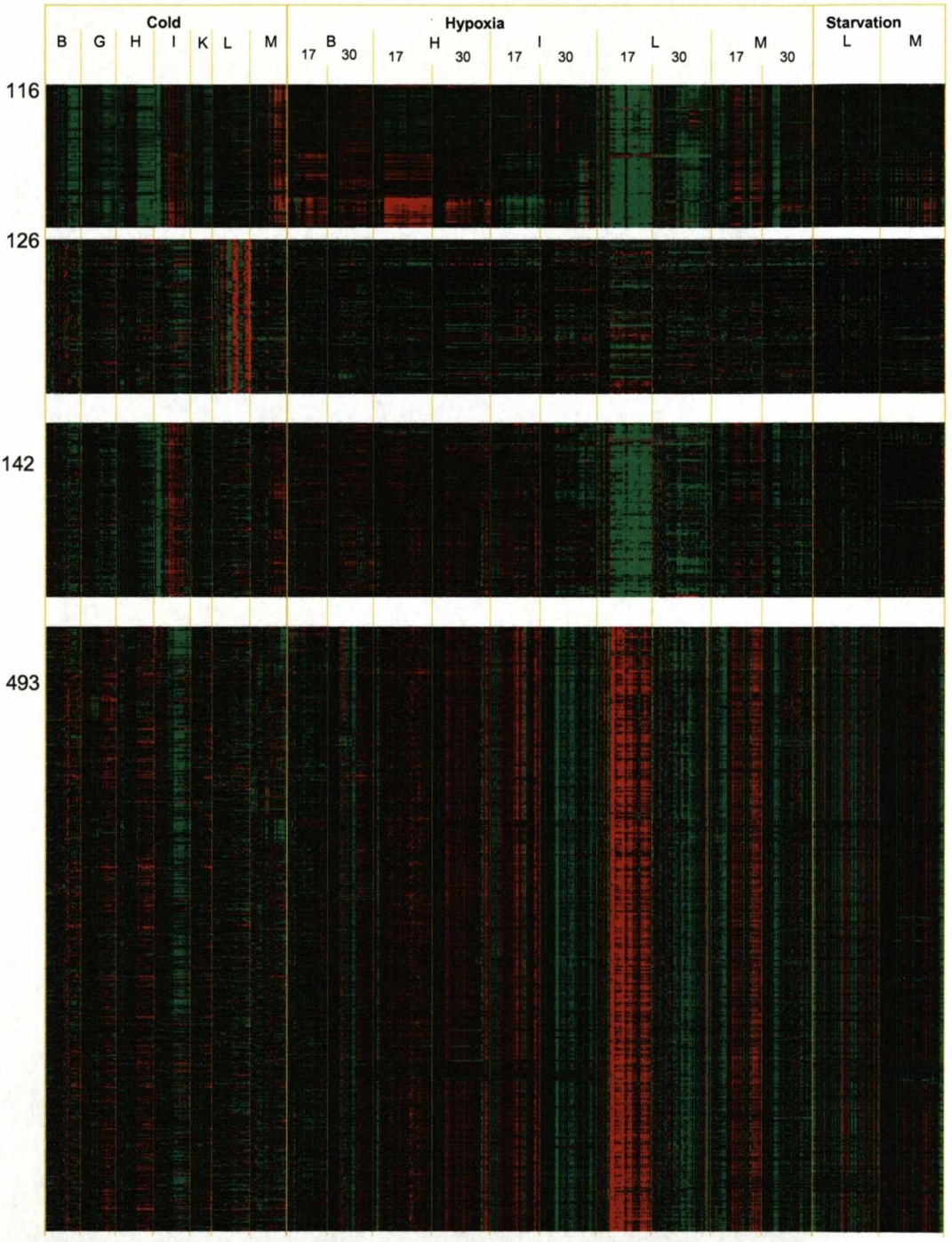


Figure 4.10 (Continued): Heatmaps for unknown GE mountains.

As mentioned above, unknown clones can be functionally related to the identified genes contained within the same mountain. They could be the same gene, or different isoforms of the same gene. Alternatively, they may represent other genes that participate in the same biological process as the most representative gene. What are the expressions patterns of genes within a mountain and why the genes are clustered together? For these, each numbered identified mountain was studied to find expression patterns of the clones and suggest the underlying biological properties for that group of genes. The heatmap was constructed for these identified mountains (Figure 4.11). The numbers on the right-hand side indicates the mountain names and their size. K1, K2 ..., Kn shows the sub-clusters generated by the *K*-means clustering technique in the mountains. The study shown that the ExprAlign was able to separate isoforms for a same gene and genes in a gene family. These outcomes as described in detail for specific genes in the following sections.

4.3.1.6.1 Discriminating genes and gene isoforms using mountains

--- Fructose-bisphosphate aldolases

The fructose-bisphosphate aldolase isoform A and isoform B were separated into two major mountains GE17b and GE33. Fructose-bisphosphate aldolase (Perham 1990; Marsh and Lebherz 1992; Shiokawa *et al.* 2002) is a glycolytic enzyme that catalyses the reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. In vertebrates, three tissue-specific isoforms of aldolase have been defined: aldolase A (muscle and red blood cells), aldolase B (liver, kidney, stomach and intestine) and aldolase C (brain, heart and ovary) (Shaw-Lee *et al.* 1992). carpBASE 2.1 sequence alignments clustered the fructose-bisphosphate aldolases into three main-groups: S341 (aldolase A), S488 (aldolase B) and S698 (aldolase C). Here the main-groups and sub-groups (transcript units) of carpBASE 2.1 were named with a prefix "S" standing for sequence alignments. Mountain GE17b contained clones from main-group S341 and mountain GE33 contains clones from main-group S488. These indicate the expression alignments had a precise agreement with the sequence

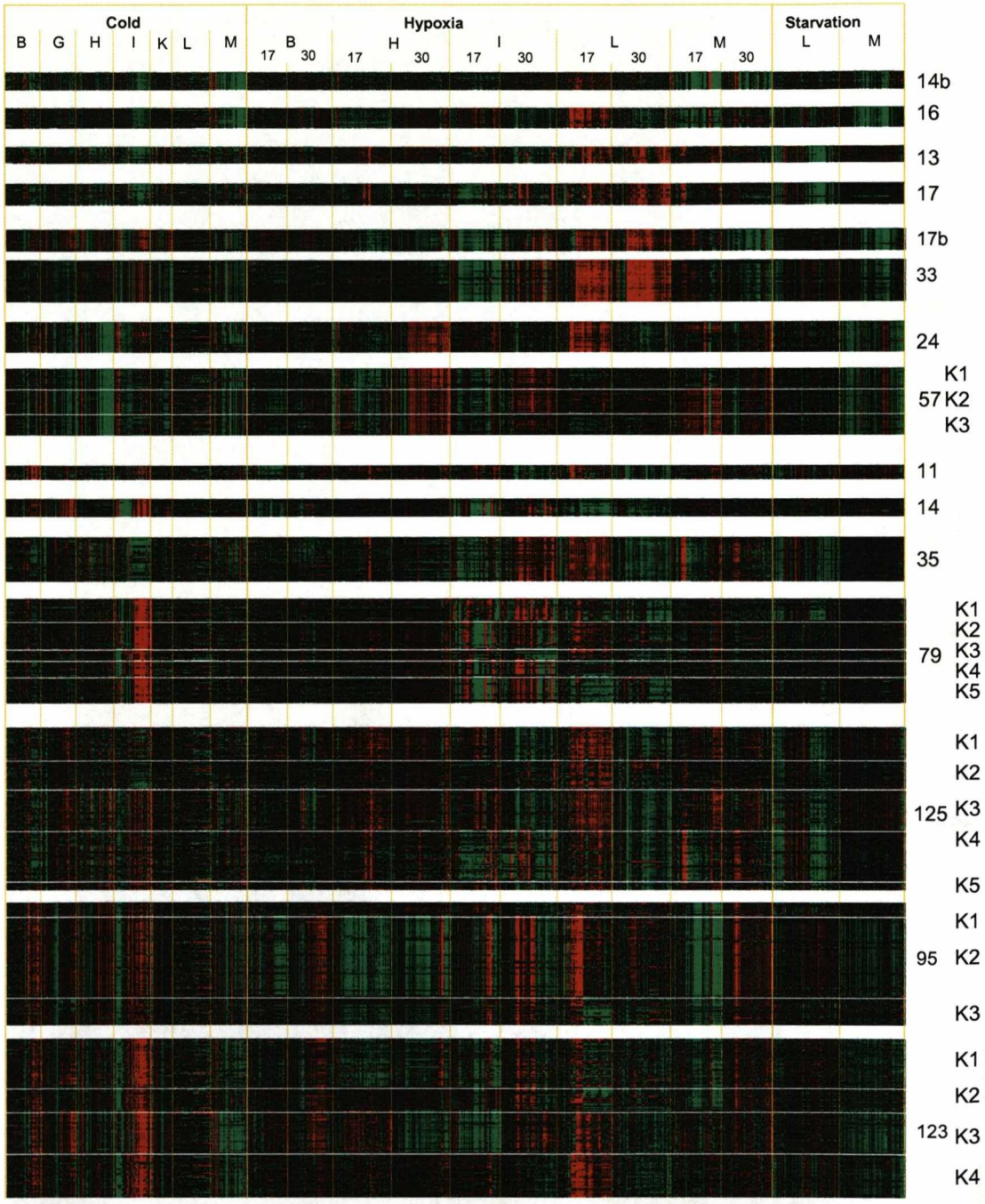


Figure 4.11: Heatmaps for identified GE mountains. The numbers on the right side indicates the mountain names and their size. K1, K2, …,Kn shows the sub-clusters generated by the *K*-means clustering technique in the mountains. On the header, cold shows the data range for cold experiments, Hypoxia for hypoxia experiment data, and Starvation for starvation experiment data. B stands for brain, G for gill, H for heart, I for intestine, K for kidney, L for liver, and M for muscle.

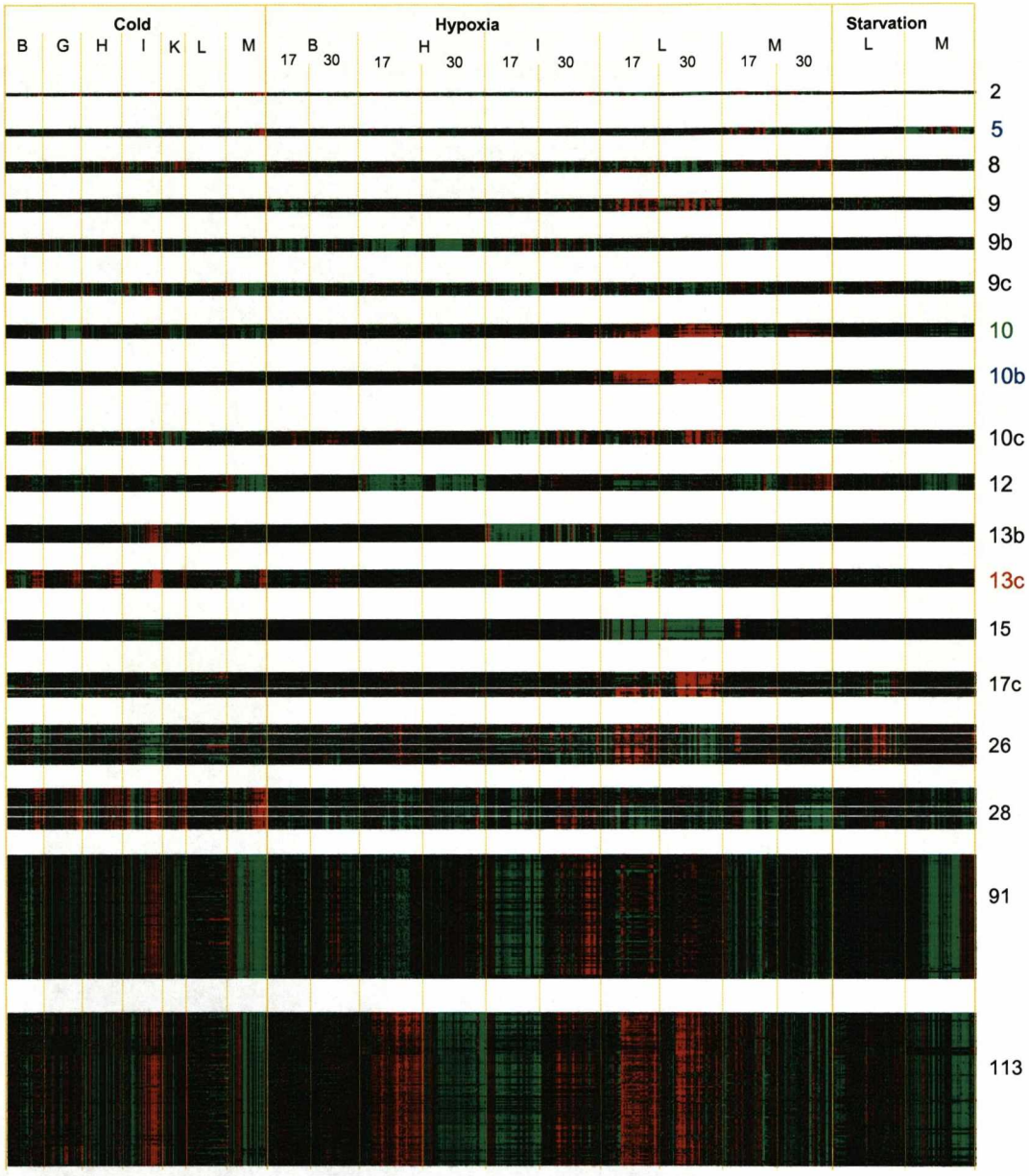


Figure 4.11 (continued): Heatmaps for identified GE mountains.

alignments for this particular gene family. Although aldolase A and B were discrete their location on the landscape was close to each other (Figure 4.5), because of their similar expression patterns (Figure 4.11). On the other hand, clones in GE17b and GE33 were separated by the expression alignments because of small difference in expression mainly in the cold-expressed tissues. This shows that our clustering method was able to separate the aldolase gene into isoforms.

Despite the established tissue-specificity, Figure 4.11 shows that the aldolase A was up-regulated not only in hypoxia muscle (17°C) but also in hypoxia liver, hypoxia brain, hypoxia intestine (30°C), cold brain, cold gill, cold heart, cold intestine and cold kidney. The aldolase B was up-regulated in cold-intestine, cold heart, hypoxia muscle 17°C, hypoxia intestine 30°C and hypoxia liver. The previous study (Shaw-Lee *et al.* 1992) focussed on the absolute level of the gene expression of aldolases, but this study describes the regulated changes of gene expression for aldolases.

--- *Parvalbumins*

Parvalbumin genes exhibited diverse expression patterns and were located in 8 positions on the landscape (Figure 4.5a: P1 to P8). This suggests that parvalbumin might have over 8 isoforms. There are two distinct phylogenetic lineages for parvalbumins: alpha and beta. Most muscle tissue contains parvalbumin of only alpha or beta origin (Elsayed and Bennich 1975; Lindstrom *et al.* 1996). Luciane V. Mello's study (unpublished work) on parvalbumin sequence alignments discovered 9 isoforms for this gene in the carp. They were named beta 1 to beta 8, and alpha 1. By comparing to sequences of the 9 isoforms from L. Mello's study, I found that P7 represented ß6, P6 represented ß7 and p3 represented ß5, P1 contained ß6, ß7 and ß1, and others were unclassified isoforms. P1 was GE14b and P7 was GE16. The heatmap shows the parvalbumin ß6 expression was unregulated in hypoxia liver 17°C.

--- *Transferrin variant A*

Transferrin variants were located in the mountains GE13 and GE17. Previous studies found two transferrin variants (A1 and B1) in crucian carp (*Carassius auratus*) (Yang *et al.* 2004), 5 (A, B, C, D and E) in silver crucian carp (*C. auratus gibelio*) and 3 (A, B and C) in white crucian carp (*C. auratus cuvieri*) (Yang and Gui 2004). The two mountains suggest common carp has at least two transferrin variants. The expression patterns were slightly different betweens the clones of mountain GE13 and GE17 (Figure 4.11). This implies the existence of two transferrin variants in common carp and which await confirmation. Sequence alignments for sequences in GE13 and GE17 indicated GE13 had high possibilities to be transferrin variants A and GE17 contained other transferrin variants.

4.3.1.6.2Distinguishing genes within a gene family using K-means sub-clusters of expression mountain

--- Apolipoproteins

Figure 4.11 shows the apolipoprotein gene family and its precursors were clearly separated onto five different mountains: GE11 (apolipoprotein C-I precursor), GE14 (apolipoprotein Eb precursor), GE35 (apolipoprotein A-I), GE125 (14-kDa apolipoprotein) and GE79 (mixed apolipoprotein & its precursor). Apolipoproteins, synthesized mainly in liver and intestine, are a class of apoproteins and the only protein component of lipoproteins, combining with free cholesterol, phospholipids, cholesterol esters, and some triacylglycerols to form lipoproteins (Eichner *et al.* 2002). In general, the role of apolipoproteins in lipid metabolism includes maintaining the structural integrity of lipoproteins, serving as cofactors in enzymatic reactions, and acting as ligands for lipoprotein receptors. Genes for apolipoproteins (Haddad *et al.* 1986; Hoffer *et al.* 1993; Eichner *et al.* 2002; Kondo *et al.* 2005; Zhou *et al.* 2005) are from a large gene family, which includes apoA-I, apoA-II, apoA-IV, apoB, apoC-I, apoC-II, apoC-III, apoC-IV, apoE and apo-14 kDa. They were extensively investigated for their functions in lipid transport (Luo *et al.* 1986; Kondo *et al.* 2005). ApoA-I, apoA-IV, apoE, apoC-I, apoC-II, apoC-III and apoC-IV, are common in mammals (Haddad *et al.* 1986; Hoffer *et al.* 1993). The 14-kDa apolipoprotein (apo-14 kDa) is a fish-specific apolipoprotein

(Kondo *et al.* 2005; Zhou *et al.* 2005). Hidehiro Kondo's studies (Kondo 2005) showed that the transcripts of pufferfish apoA-I were expressed mainly in liver and the apo-14 kDa gene transcripts were mainly expressed in liver and less abundantly in brain. The similar conclusion can be drawn from the expression heatmaps for mountains GE35. GE11 and GE14 illustrated the transcripts of apo-Eb and apoC-I genes were mainly expressed in brain and intestine. The transcripts of the former gene were also observed in gill.

~85% of identified clones in GE79 were apolipoproteins. However, the expressions of the clones were different in intestine, liver and muscle. *K*-means clustering was implemented to re-cluster the expressions within GE79 to break down the mountain into 5 *K*-means sub-clusters (named as GE79-K1 to GE79-K5) in which clones had highly similar expressions, and helped to explore the expression patterns and relationships between the sub-clusters in GE79. Clones in each sub-cluster had highly similar expressions. It was difficult to find a single gene to represent each of the *K*-means groups because of the limitation of the sequence annotations in carpBASE 2.1. But the *K*-means groups at least provide a direction for separating different genes of apolipoproteins using functional information. For example, based on carpBASE 2.1, GE79-K3 was related to fatty acid-binding protein, GE79-K1 was related to apoA-I, GE79-K4 was related to apoEb precursor, andGE79-K2 and GE79-K5 are related to apoA-IV precursors. By using the same method, mountain GE125 was split into 4 *K*-means sub-clusters: GE125-K3 and GE125-K4 were 14kDa (fish-specified) apolipoproteins, GE125-K2 was related to other apolipoproteins, and GE125-K1 was unknown. If more sequence information was provided, the *K*-means sub-clusters might be identified with confidence. The five mountains and the *K*-means sub-cluster indicate that the methods implemented were able to separate the apolipoprotein gene family into individual genes.

--- *Creatine kinases*

Creatine kinases were mainly located in mountain GE57 and GE24. In vertebrates, the creatine kinase isoenzyme family consists of four types of

isoforms: cytosolic muscle type (M-CK), cytosolic brain type (B-CK), mitochondrial ubiquitous, acidic type (Miu-CK), and mitochondrial sarcomeric, basic type (Mis-CK). H.W Sun (Sun *et al.* 1998) reported at least three M-CK subisoforms (M1-CK, M2-CK, and M3-CK) for common carp, and the deduced amino acid sequences of these three subisoforms of carp M-CK show about 85% identity to mammalian M-CK isoenzyme. *K*-means clustering split GE57 into 3 sub-clusters. Based on the sequence alignment and the BLAST identities of the sequences, GE57-K2 was M2-CK, GE57-K3 was M3-CK, and GE57-K1 was unknown. E57-K1 could not be identified as M1 with the current limited sequence information. The different expression of GE57-K1 to GE57-K2 and GE57-K3 (Figure 4.11) suggested GE57-K1 might be M1 or other sub-isoform. Creatine kinases were also located in GE24, but the gene identities were diverged with other different genes in GE24. Creatine kinase M2 and M3 were up-regulated in hypoxia heart 30°C, hypoxia intestine 30°C, hypoxia muscle 17°C, and down-regulated in cold heart. The former one was also up-regulated in hypoxia liver. These examples shown *K*-means sub-clusters implemented here had the capability of separating genes in a gene family.

--- *Ribosomal proteins*

Ribosomal proteins were positioned on mountain GE95 and GE123. *K*-means clustering separated the two mountains into 3 and 4 subgroups. GE95-K2 was revealed as 60S ribosomal protein L30; and GE95-K3 was designated as 40S/60S Ribosomal proteins. GE123-K2, GE123-K3 and GE123-K4 of GE123 were 40S/60S ribosomal proteins.

4.3.1.6.3 Other genes on GE mountains

Carp desaturase 2 (CDS2) expressions in the mountain GE13c were up-regulated in tissues brain, gill, heart, intestine and muscle of cold experiments. Previous studies using RNase protection assays (Tiku *et al.* 1996) had shown that the transcription of CDS2 was cold-induced in liver.

Myoglobin expression in mountain GE10 was up-regulated in liver of fish exposed to hypoxia. This remarkable observation subsequently led to a

full-scale assessment of protein expression patterns, demonstrating that the myoglobin protein was indeed expressed in a range of non-muscle tissue in contrast to the widely held view that this gene restricted to oxidative muscle only (Fraser *et al.* 2006).

Mountain GE10b represents the warm-temperature-acclimation-related-65 kDa-protein (Wap65). Previous studies suggest that Wap65 was expressed in muscle in response to increased temperature in a different way from heat shock proteins (HSPs) (Kikuchi *et al.* 1995). The mountain in this study implies the discovery of the highly represented Wap65 in liver of hypoxia exposed carp. This suggests that like myoglobin, this gene is expressed in a far wider range of tissues and circumstances that has been appreciated to date.

Fatty acid-binding protein (FABP) was identified as a major constituent of mountain GE17c and its expression was increased in the liver of hypoxia-treated fish. Recent study on the inverse relationship between liver FABP and DCF (dichlorofuorescin) fluorescence intensity suggested that intracellular liver FABP was able to function as an intracellular antioxidant and has the ability to play a major role in the oxidative stress induced by hydrogen peroxide (Wang *et al.* 2005).

GE15 represents the vitellogenin (VG) which was down-regulated in the liver of hypoxia-treated fish. There were different reports in the literature showing that fish various stressful conditions have a negative effect on VG expression (Lethimonier *et al.* 2000).

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in GE113 was up-regulated in the liver of hypoxia experiments and down-regulated in muscle of cold and starvation experiments. Its expression pattern was thus complex with up-regulation in heart 17°C hypoxia but down-regulation in heart under 30°C hypoxia. Also it was down-regulated in intestine under 17°C hypoxia but up-regulated in intestine of 30°C hypoxia intestine. Xie *et al's* study found GAPDH were significantly increased in rat liver and lungs after treatment with bacterial endotoxin or lipopolysaccharide (Xie *et al.* 2006).

4.3.2 GOmatrix for common carp gene expressions

24 interesting *K*-means carp gene expression groups (Appendix 4.1) were provided by Dr. Andrew Gracey as part of the analysis of cold-expressed tissues, comprising ~1700 BLAST identified genes that exhibited statistically significant changes during cold stress. GOprofiler retrieved GO information from carpBASE 2.1 and output the members of genes in each expression groups associated with 25 biological process sub-categories.

The GOmatrix programme, available on the LEGR Data Centre website (http://legr.liv.ac.uk), provides a graphic user interface (GUI) for user to input their data. The GOmatrix (Figure 4.12) result illustrated the patterns for enriched or depleted genes for each biological process category for each of the 24 gene groups. The programme compared the expected values and the observed values, computed two-tailed probability values using the Fisher Exact test, and built up the probability GOmatrix. A colour-coded probability GOmatrix shows not only the up/down representing groups but also the significance. In the colour-coded GOmatrix (Figure 4.12) red indicates over-represented (enriched) and blue shows under-represented (depleted) categories. For example, group 5, which comprised genes up-regulated principally in the intestinal mucosa (Appendix 4.1), and to a lesser extent liver, was enriched for genes involved in transport and oxygen metabolism but depleted for genes involved in nucleotide metabolism and biosynthesis. The GOmatrix indicates the transcriptional regulation of both electron transport (groups 1, 7, and 14) and energy pathways (groups 6, 12, 15, and 18) in the cold response. Group 1, which describes genes that increased in six of the seven tissues, was highly enriched for electron transport genes. The expression of genes involved in carbohydrate metabolism was altered in almost all tissues (groups 2, 5, 12, 15, and 19). The groups came from almost all important tissues indicating a core response to cold stress throughout the body of common carp.

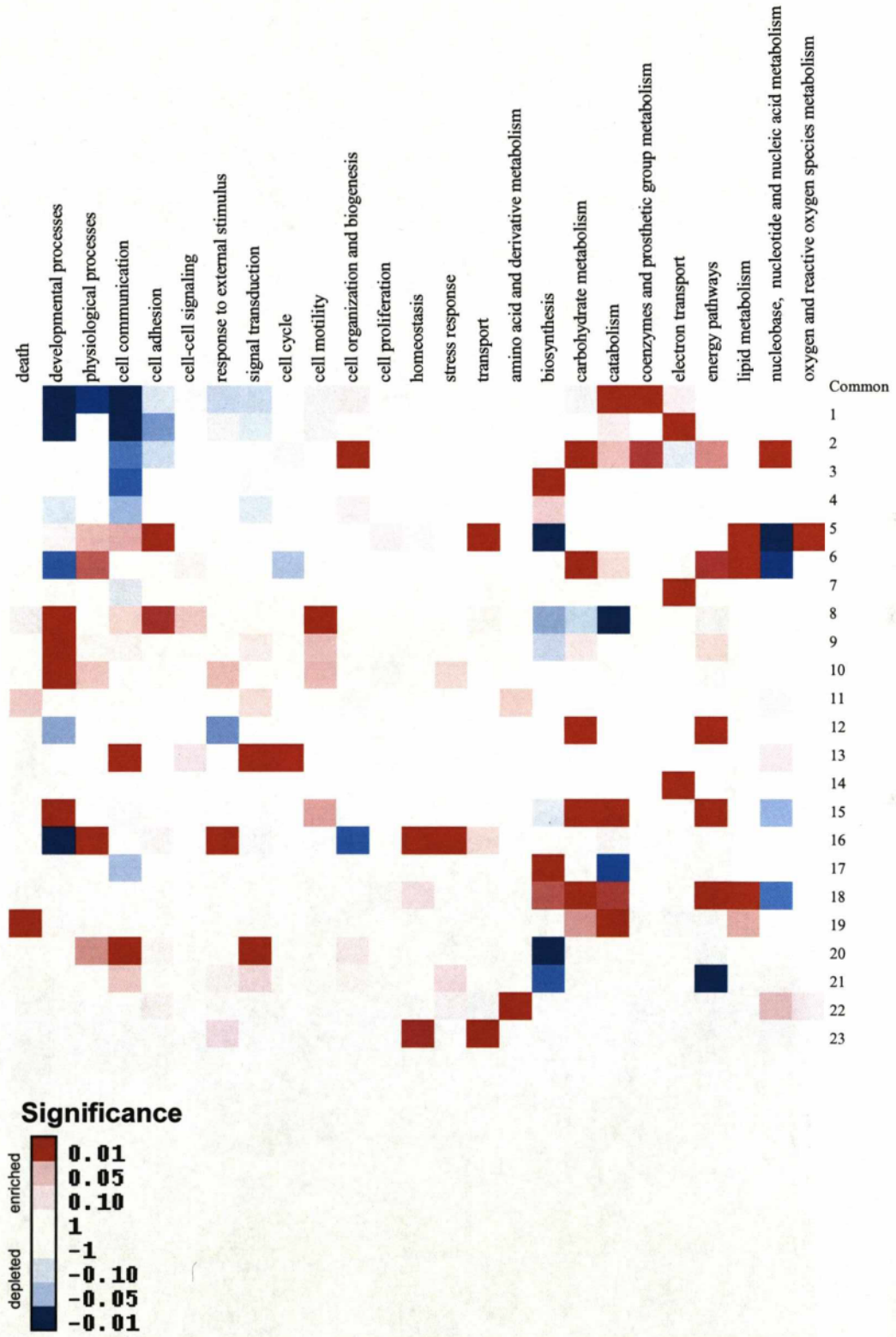


Figure 4.12: GOmatrix for 24 K-means groups in cold data

4.4. Discussions

Genomic investigations have been traditionally led by sequence determination and annotation, followed by the wider investigation of large-scale gene sets using post-genomic techniques. All eukaryotic EST collections contain very large numbers of transcripts which remain unidentified using conventional techniques and the source of these sequences have not been subject to intense system-wide scrutiny. Thus, Chapter 3 described protocols for the identification of the EST sequences obtained for the common carp, and how this was used to direct the interpretation of the gene expression profiles. In carpBASE 2.1, more than 40% of the EST assemblies did not yield BLAST identities and more than 30% had no functional annotations. These non-identified entities might have several different origins. First, they may represent new, undiscovered protein-coding genes. Second, they may be members of the newly recognised non-protein-coding transcripts, including a wide range of different specific RNA forms (Frith *et al.* 2006). Third, they may be spliced segments of RNA which have been cloned into collections, including those which are untranslated regions of spliced transcripts. Finally, they may be concatenated constructs which are generated artificially during the generation of cDNA libraries.

4.4.1 ExprAlign and the profiling of gene expression properties

It is most likely that sequence information and gene expression data are both helpful for understanding properties for each other, and that the two kinds of data validate each other, leading to new biological discoveries. On the other hand expression profiles across a range of experimental treatments for different probes derived from a same gene should be highly correlated.

This Chapter tests the idea that expression profiles from microarray experiments can be used as a technique to aid gene identification as well as gene characterisation. The technique is based on the comparison of correlation coefficients for expression values between pairs of probes on the microarray, the highest values perhaps being used as evidence of a common identity. For this, we first created a fast algorithm for calculating the Pearsons correlation

coefficient, this being necessary due of the very large number of paired comparisons necessary for a complete search of correlated expression properties. These values were then used by VxOrd to create a network linking genes together on the basis of their shared correlations. A force-repulsion mechanism within the VxInsight algorithm transforms the gene networks into discrete clusters. We show that the resulting landscape features, and the associated clusters are entirely robust, firstly because permuting and randomising the expression values leads to a complete loss of landscape features and thus the associated gene clusters, and second, because the form of the clusters are largely retained when using different scales of array data from small to large. We show that larger datasets which include a wider range of experimental treatments can fragment the gene clusters into smaller forms, each with distinctive character. Finally, we show that the major clusters are enriched in particular GO categories, and the mountains or clusters are distinctive from each other. This is consistent with the coherent regulation of specific biological processes or pathways in the phenomenon under investigation.

4.4.2 Gene identification using ExprAlign

The ExprAlign approach aligns gene expression profiles based on correlations between individual genes, just as sequence alignment is based on the comparison of sequences. Both seek to identify the most highly aligned sequence or probe, with the implication that the unidentified sequence or probe can then be ascribed an identity. We show how this technique works in practise. For this analysis we chose to define a high criterion of correlation coefficient based on the ROC optimisation procedure. Many of the resulting landscape features or mountains contained predominantly just one kind of BLAST-identified gene. The unidentified probes that were collocated into that mountain were also tentatively identified with that gene name. Of course, the validity of this technique depends on subsequently demonstrating that the newly annotated genes do indeed have that identity. This was achieved for myoglobin in that mountain GE10 possessed 5 probes that BLASTed as myoglobin using the automated EST-Ferret pipeline, and 5 probes for which

there was no identify. Closer inspection of the corresponding sequences, and manual attempts at alignment, were able to demonstrate that all of the unidentified ESTs were indeed clones of the myoglobin gene. The analysis of the data from the entire treatment experiment identified 32 mountains containing 522 unknown clones, which was ~17% of unknown clones on the map. Based on the procedures described in the Results section based on the inclusion of unidentified genes in landscape features containing only or predominantly a single BLAST identity we were able to suggest 522 identifies.

4.4.3 Separation of isoforms using ExprAlign

A final use for the ExprAlign procedure is to distinguish the properties of different isoforms from within a gene family. The carp was generated from a collection of cDNA clones. Whilst this collection was normalised to moderate the representation of abundant and rare transcripts, there was still some considerable variation in the number of representative clones for each gene. Some abundant genes were represented by as many as 80 clones whilst many more were represented by just one clone. If the repeated clones for the former were all sourced from a single gene then they would be expected to display identical expression patterns across samples and treatment groups. On the other hand if the clones were created from different isoforms which have different properties, despite them generating an identical BLAST alignment, then we would expect this would be reflected in a distinctive expression profile.

We show that these expectations are largely met. Thus in the case of fatty acid-binding protein (mountain GE17c), where we identified 13 clones all possessing the same sequence alignment, we demonstrate that some clones display quite distinctive expression profiles. This was evident in the *K*-means clustering of all of the probes contained within a single ExprAlign mountain. In this particular case the differences were rather subtle and applied to just one tissue and one treatment condition. Nevertheless, this represents a distinctive

feature that is reflected by the clustering by the *K*-means method of the ESTs themselves.

4.4.4 Advantages of the VxInsight package for cluster determination

The ExprAlign approach has a number of practical advantages for the exploration of expression responses, and the collation across all genes of the main clusters of response. It positions gene expression clusters on a three-dimensional landscape in a series of landscape features, which allows users to flexibly explore data by zooming in from large landscape features, to smaller gene clusters and down to an individual gene. Also, because the VxInsight package allows each gene and landscape feature to be labelled with one of a range of annotation terms, the meaning and significance of the genes and features can be better understood. The ExprAlign approach was also implemented in the study for co-expression of conserved genes cross-species (Chapter 5).

It might be that this approach can be as diagnostic in ascribing a meaningful sequence identity as the more conventional sequence alignment techniques, such as BLAST. Both approaches compare the unknown DNA sequence or expression pattern with that of known genes or array probe, respectively, and are thus both limited by the available data. However, the expression approach might be able to identify subtle differences in expression characteristics which may not be evident from sequence data, particularly if the sequence data is limited to the 800 bp provided by a typical single pass read. Thus ESTs possessing the same BLAST identity might well contain different motifs in the microarray probe, and this allows them to discriminate.

4.4.5 Alternative packages for global expression analysis

Most of other analysis packages used for gene expression analysis also depend on data clustering techniques, including both GeneSpring and maxd (Hancock *et al.* 2005). They are complex platforms for analysing large-scale gene expression data and normally require users to be trained before using them. However, these tools only visualise output clusters on two-dimensional

alignments, which are difficult to interpret when dealing with large-scale gene expression datasets. GeneSpring, a commercial software package from the Agilent Technology (http://www.agilent.com), is frequently regarded as a "black-box" for biologist since the fine details of the algorithms inside the package are not revealed. BioConductor (Gentleman *et al.* 2004) provides powerful tools for statistic analysis of gene expression, but does not offer good visualisation tools. GenePublisher (Knudsen *et al.* 2003) is a web server for automated analysis of gene expression profiles. This makes it convenient to be accessed but limited for large-scale input data.

4.4.6 Benefit of using GOmatrix

In gene expression analysis, genes usually are clustered into different gene groups within which genes have similar expression patterns. The GO information is essential for understanding the biological insights of each gene group. In this study, EST-ferret provided GO annotation for individual genes using GOprofiler but did not generate GO annotations for particular gene groups in gene expressions. The GOmatrix algorithm uses Fisher's exact test to compute probabilities to determine the significance of each gene group enriched or depleted within the particular GO sub-categories. It displays results in a coloured matrix, which makes it easy to interpret biological meanings with GO annotations for different gene expression groups and compare the GO annotation patterns between groups.

Other tools for functional annotating gene expression groups includes GoMiner (Zeeberg *et al.* 2003), GoSurfer (Zhong *et al.* 2004), MAPPFinder (Doniger *et al.* 2003), GSEA (Subramanian *et al.* 2005), FuncAssociate (Berriz *et al.* 2003), *ect.* GoMiner, GoSurfer and MAPPFinder generate GO annotation for one or two gene groups each time, but their output is difficult to be transferred to other formats to compare patterns of functional property between multiple gene groups. FuncAssociate is a web-based tool to characterize gene sets with Gene Ontology attributes but only support analysis for 10 species. GSEA is able to analyse multiple gene sets to ease interpretation of a large-scale experiment by identifying pathways and processes. But only GOmatrix is

perfectly working together with GOprofiler to functional annotating multiple gene sets for non-model species and allow comparison of patterns between groups.

CHAPTER 5: ORTHOLOGY ANALYSIS FOR METAGENES

5.1. Introduction

5.1.1 Conservation of gene co-expression patterns

The completed genome sequences for the human and other model organisms offer a near complete listing of genes to undertake the full repertoire of cellular and molecular function for complex life forms to exist. However, just as the functioning of a motor cannot be defined simply by the list of mechanical and electrical components, the functioning of cells, organs and individuals cannot be defined simply from a list of genes and encoded products. Many genes have well described functional roles, notably those involved in intermediary metabolism, but many genes especially those involved in regulation and control, have poorly-defined or unknown functional roles. Moreover, the more recent discovery of transcripts from non protein-coding genes dramatically increases the number of transcriptional products yet to be identified and annotated (Mattick 2007). Clearly, a major task for molecular biology is the complete description of these roles.

Microarrays have become the principal means of addressing the functional responses at the level of the protein-coding gene, and their massively parallel operation yields an unsurpassed breadth of information across large numbers of genes. One of the early organising features of these data sets has been the clustering of many genes into co-regulated groups, and the recognition that many of these clustered genes are functionally related to each other. Despite this the co-regulation of genes is not unequivocal evidence of their functional relatedness. For one thing, some co-regulation might represent a false positive in that the outcome is probabilistic. For another, some genes might be accidentally regulated by the activation of adjacent genes.

Stuart *et al* (2003) have made the important point that the case for the functional significance of gene expression clusters is made more powerful if that cluster was observed in more than one study of a single species, or in studies of other species. They contend that because small differences in fitness confer an advantage for particular expression properties, then the conservation of co-regulated groups of genes, or 'gene modules', constitutes a more

demanding test of phenotypic associations that the loss-of-function mutations of a single gene. Given the intense research activity in array applications over the past decade there is now a large amount of expression data with which to quantitatively to test these ideas.

Stuart *et al* (2003) undertook a comparison of 4 model species, including the human *H. sapiens*, *Drosophila*, *C. elegans* and the yeast, *S. cerevisiae*. They compared genes not so much by exploring expression responses of a single gene, but by comparing the co-expression for two genes between species. This has the advantage of focusing not on treatment-specific responses but on the associations between genes, allowing the comparison to be drawn irrespective of the treatment conditions imposed on either of the compared species. Stuart *et al* (2003) also indicated that the more species that are compared the more powerful does the statistical test of conserved co-expression become.

The identification of cold-response genes and of co-expression properties in the tissues of the common carp, prompted us to suggest that the coordinated cold response system might be expressed in other related species, and may also be expressed in other species in response to other kinds of treatment or stress. Thus the gene modules involved in cold responses might be a reflection of gene regulation more generally rather than temperature specifically, and we can offer a critical test of conserved co-expression responses by comparing carp genes responding to cold with human genes responding to a range of different treatments.

We have used the basic protocol described by Stuart *et al* (2003) which consists of 4 stages. First, it was necessary to identify the known genes in carp that have corresponding homologs in humans, constituting metagenes. A metagene was defined as a set of genes across multiple organisms whose protein sequences are one another's best reciprocal BLAST hit (Kim *et al.* 2001; Stuart *et al.* 2003). Secondly, the co-expression properties of these metagenes need to be determined and ranked for the carp, and separately for the human. Third, the statistical probability of a conserved ranking position for each gene will be calculated by a rank order statistic and the Monte Carlo

simulation. Finally, the P-value for this rank statistic can define the co-expression landscape of compared metagenes using the VxInsight ordination and visualisation package (Davidson 2001). A key stage is the first one, the determination of which carp genes have a corresponding human ortholog.

5.1.2 Orthology

Homology refers to the relationship of two characters that have descended, usually with divergence, from a common ancestral character (Fitch 2000). Homology implies an evolutionary relationship: orthology or paralogy. The difference between for orthology, paralogy and homology were described in the Section 1.2.5.2. Orthology analysis is often complex because of large numbers of paralogs within gene families.

The identification of orthologous gene groups can help the association of functional information between genes in different organisms with a high degree of reliability. It is useful for the functional annotation of sequences, genome annotation, and studies on gene/protein evolution. Studies on orthologs and paralogs can also provide clues for the understanding of genome duplications. Whole genome duplications are likely to have played an important role in generating complexity during the early stages of vertebrate evolution, near the time of divergence of the lamprey linage (Ohno 1970). It may explain the variation in chromosome numbers as well as the multiple gene copies and chromosome segments in species of vertebrates (Postlethwait *et al.* 1998; Wolfe 2001). Phylogenetic analyses of sequences from human, mouse, chicken, frog, zebrafish and pufferfish suggest that ray-finned fishes (Actinopterygii) are likely to have undergone a whole genome duplication event between 200 and 450 MYA (Van de Peer *et al.* 2003). Additional genome duplication, specifically in ray-finned fish (Helfman *et al.* 1997), may have occurred before the divergence of the Teleosts (Taylor *et al.* 2001; David *et al.* 2003; Taylor *et al.* 2003), which is a infraclass in the class *Actinopterygii* (ray-finned fish). The common carp (*Cyprinus carpio* L.), belonging to the same Cyprinid family of fish as zebrafish, has been considered tetraploid because of its chromosome number (2n = 100) and its high DNA content (Ohno *et al.* 1967; David *et al.* 2003; Taylor *et al.* 2003). However, the studies

of orthologs and paralogs between common carp and other species using available limited sequences, such as ESTs, cDNAs, offer opportunities to explore the role of carp gene duplications and the subsequent evolution of its large gene families in generating additional gene diversity.

5.1.3 Objectives for the investigations

The carp stress experiment described in previous Chapters has shown that there are strong patterns of gene expression between different tissues, and between different stressors. In particular we identified patterns of gene co-expression suggesting the involvement of gene co-regulation perhaps through coordinating control processes (Gracey *et al.* 2004). Of course these patterns can be unique to the carp in coping with its seasonal environmental challenges of cold, hypoxia and starvation. On the other hand the expression and co-expression patterns may be shared with other species, particularly if those patterns originated before the divergence of the different vertebrate lineages. In this Chapter we explore the extent to which conserved patterns of gene regulation across the entire vertebrate range can be detected, by comparing the expression profiles from the common carp with that of a distant, model species, namely humans, for which array data is readily available. This experiment also addresses the possibility of undertaking taxon-wide comparisons with less than the full genome sequence coverage. For this we used the methods of Stuart *et al* (2003), in which pairs of genes which are significantly correlated in the two species are identified and subjected to a network ordination using the VxInsight package (Davidson 2001). This analysis was preceded by the identification of orthologous relationships among common carp and human through an all-versus-all reciprocal BLAST search between the two species. To maximise the yield of orthologous relationships we devised a bridge procedure whereby the carp was first BLASTed against the zebrafish for which much greater sequence resources were available and genome coverage was near complete. The metagenes between zebrafish and human were then explored again by a reciprocal BLAST protocol. Human sequence information and gene expression data can be downloaded from the open resources on the Internet.

Common carp ESTs sequence information (described in Chapter 3) and its microarray resource (described in Chapter 4) were also available in our LEGR lab.

5.2 Materials and Methods

5.2.1 Sequences resources

29,267 human protein sequences, downloaded from the NCBI RefSeq website, were used to construct the orthology groups. Sequences in RefSeq were output after processing a series of sequencing cleaning, clustering and annotation steps for the human sequence in GenBank. Each sequence in RefSeq stands for a gene or a gene product and contains high quality descriptive and functional annotations. Moreover, human sequences in RefSeq include all known human genes or gene products in GenBank and this provides greater ability to identify the correct human orthologs of common carp genes.

Common carp ESTs and EST assemblies, described in Chapter 3 and included in carpBASE 2.1, were both used to establish the best available ortholog relationship between human and carp. A problem for the orthology group construction between human and common carp was that for the non-model species we possessed less than the full genome sequence, consisting only of ~13,000 ESTs. These resources were insufficient to precisely identify correct orthologous gene groups between the two species. If another more closely related model-species can link common carp to human as a bridge, we should have a greater chance to define the correct orthologous relationships. Zebrafish, a model species, is closer to common carp in phylogenetic terms than any of other model species for which substantial sequence resources are available. Thus, zebrafish was chosen to be a bridge species to relate common carp to human. 30,583 zebrafish protein sequences were downloaded from the NCBI RefSeq database for this purpose.

5.2.2 Gene expression resources

The common carp gene expression data in the cold experiment described in Chapter 4 was used to calculate the Pearson's correlation coefficients for the co-expression analysis. Coefficient scores between 0.25 and 1 were implemented in the analysis pipeline. The method of calculating was also described in Section 4.2.2 of Chapter 4. Human expression data was that

used in the analysis of metagenes provided by the lab of Stuart Kim, downloaded from http://cmgm.stanford.edu/~kimlab/multispecies/Data/.

5.2.3 Orthology group construction

Firstly, all-versus-all reciprocal BLAST searches were implemented to find out orthologous genes between carp ESTs and zebrafish proteins. Secondly, genes for zebrafish which possessed an orthologous relationship with a carp sequence were used in another reciprocal BLAST search between zebrafish proteins and human proteins (Figure 5.1a). Here BLASTX was used for nucleotides against peptides; TBLASTN was used for peptides against nucleotides; and BLASTP was used for peptides against peptides. This method generated a set of orthologous groups for carp ESTs and human proteins. Another set of orthologous groups for carp ESTs assembly and human proteins were also produced using the same method. Only the orthologous groups which existed in both sets of orthologous groups were defined to be orthologous genes (metagenes) between human and common carp in the co-expressive analysis. If we only use one set of the orthologous groups, the faulty orthologs would not be filtered out and the orthologs were not identified with a high level of confidence. Moreover, the orthologs from common carp ESTs and human proteins provided clone IDs which were consistent in carp microarray data. The clone IDs directly connected the sequence information to the microarray data.

5.2.4 Rank statistics and Monte Carlo simulation

Correlation in expression between pairs of gene within a species reflects the similarity of gene expression patterns for a pair of genes within a species. But what do expression correlations of orthologous genes tell us? Given that a pair of orthologous genes exists in each of two species, a question is that whether the co-expressions of these two orthologous genes are conserved in both species. A Monte Carlo simulation was implemented to identify the orthologous genes which have conserved expression interactions between common carp and human.

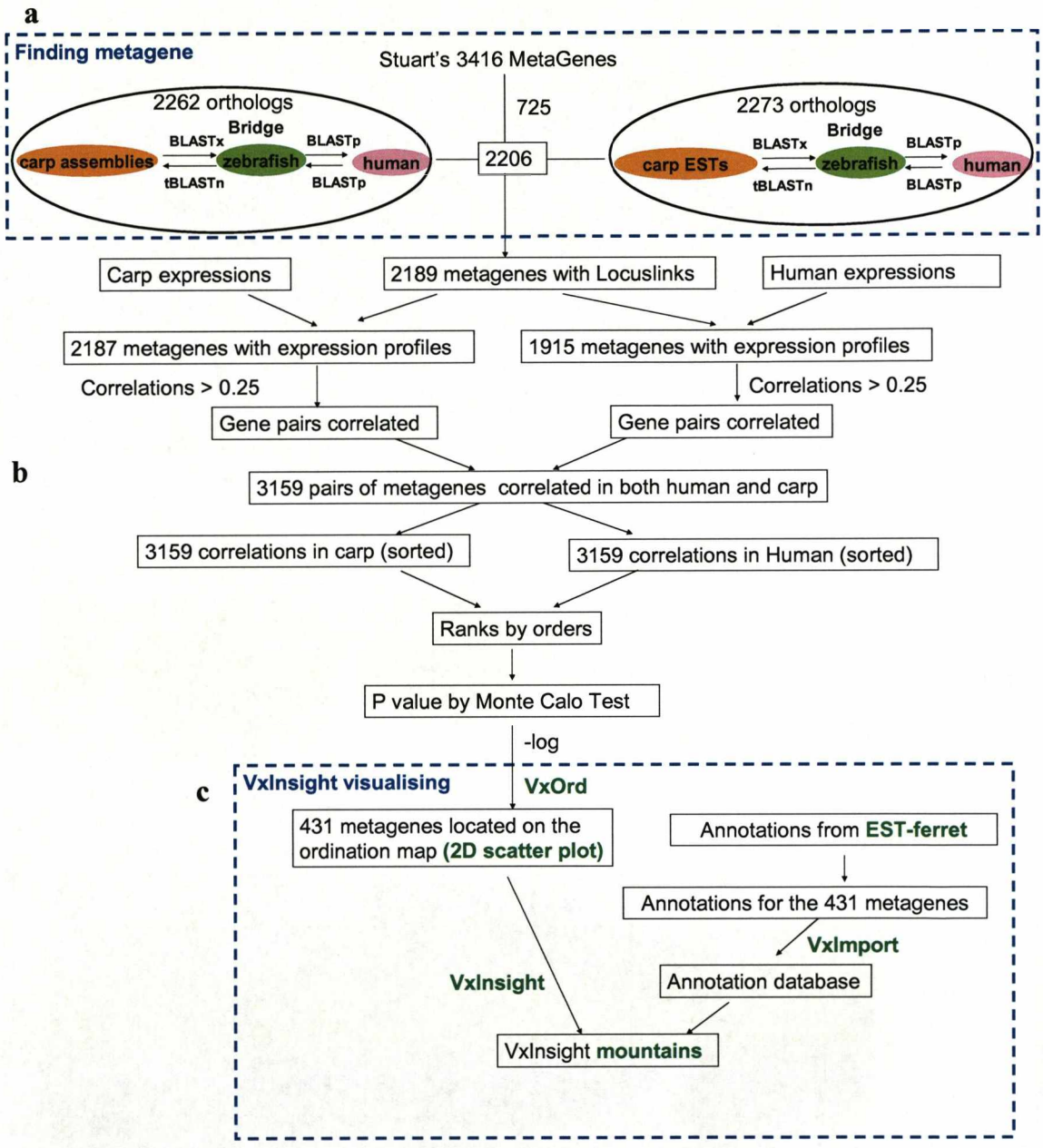


Figure 5.1: Pipeline for constructing orthologous genes and visualising their expressions across human and carp

The Monte Carlo simulation (Hope 1968; Tarantola 2005) is a class of computational algorithms for randomizing a large number of samples and repeating a large amount of computation. Usually, it firstly defines a domain and produces samples randomly within the domain, secondly performs the calculation on the samples, then repeats the sampling and the calculations, and finally aggregates the results. It is suited to calculation by computer and often used to simulate physical, mathematical or biological systems.

The Pearson correlation coefficients for transcript expression between genes in either (common carp or human) were computed using the methods described in Chapter 4. Given a metagene $G_{m,m}$ in species A and B, for example $G_{1,1}$, let $G_{a,1}$ be a gene belonging to $G_{1,1}$ in species A and $G_{b,1}$ be a gene belonging to $G_{1,1}$ in species B. We ranked all of the other genes ($G_{a,2}$, $G_{a,3}$, $G_{a,4}$, ..., $G_{a,n}$) in species A relative to $G_{a,1}$ based on their Pearson correlation coefficients and then divided the rank by the total number of metagenes I, yielding n rank ratios $R_{a,1,2}$, $R_{a,1,3}$, ..., $R_{a,1,n}$ (Shown in Figure 5.2). Using the same approach, $R_{b,1,2}$, $R_{b,1,3}$, ..., $R_{b,1,n}$ were generated for the species B. Given another metagene $G_{2,2}$, another two sets of rank ratios ($R_{a,2,1}$, $R_{a,2,3}$, $R_{a,2,4}$, ..., $R_{a,2,n}$ and $R_{b,2,1}$, $R_{b,2,3}$, $R_{b,2,4}$, ..., $R_{b,2,n}$) were generated. If $G_{a,1}$ and $G_{a,2}$ are correlated in gene expression and $G_{b,1}$ and $G_{b,2}$ are correlated in gene expression as well, the related rank ratios include $R_{a,1,2}$, $R_{a,2,1}$, $R_{b,1,2}$ and $R_{b,2,1}$. $R_{a,1,2}$ is the rank of ratio for correlation of $G_{a,1}$ to $G_{a,2}$ in the species A, $R_{a,2,1}$ is the rank of ratio for correlation of $G_{a,2}$ to $G_{a,1}$ in the species A, $R_{b,1,2}$ is the rank of ratio for correlation of $G_{b,1}$ to $G_{b,2}$ in the species B, and $R_{b,2,1}$ is the rank of ratio for correlation of $G_{b,2}$ to $G_{b,1}$ in the species B. The sum of ratios for $G_{1,1}$ and $G_{2,2}$ was defined as:

$$T_{1,2} = R_{a,1,2} + R_{a,2,1} + R_{b,1,2} + R_{b,2,1} \qquad \text{(Equation 5.1)}$$

T here is the observed rank ratio. Different T ratios were computed for different pairs of metagenes across two species.

The probability of getting the observed rank ratios T can indicate how conserved the correlation of gene expressions for a pair of metagenes across

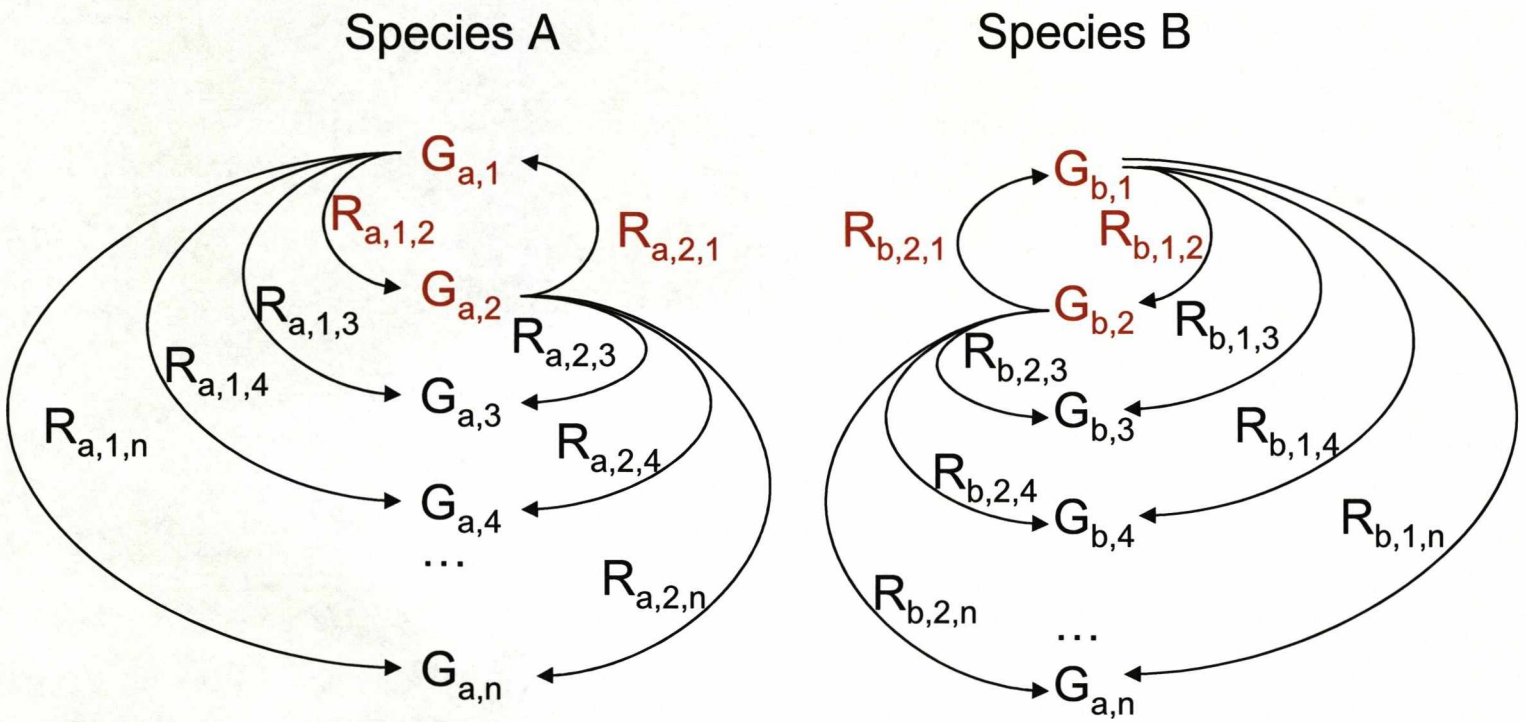


Figure 5.2: Ranking orders for correlations between metagenes

two species are. The Monte Carlo simulation was used to compute this probability (P). The first step of the process generated a new variable $K = 0$. The second step produced 4 random numbers between 0 and I (the total number of metagenes). The third step divided the sum of these 4 random numbers by I and generated the result as T'. If T' is less than T (the sum of ratios), K will be increased by 1. The next step repeated 100,000 times and produced the final value of K (of course, this step can be repeated over 100,000 times if necessary). The final step generated the P value for a pair of metagenes as:

$$P = (K+1)/100{,}000 \qquad \text{(Equation 5.2)}$$

P here is the probability of getting T. The lower the P value, the more conserved are the gene expression levels of the pair of genes across human and carp. This method was similar to the method developed by Kim's lab (Stuart *et al.* 2003) but used different statistic approach.

5.2.5 VxInsight and GOmatrix

The negative logarithms of the P values obtained from the Monte Carlo simulation were used to position the metagenes of common carp and human on a two dimensional ordination in VxOrd followed by the generation of three dimensional representations in VxInsight which resemble landscape mountains (Stuart *et al.* 2003). The method of generating landscape images, described in Chapter 4, displayed the relationships of orthologous genes across carp and human, which possess similar expression profiles.

To have better understanding of their functional profiles, metagene groups classified in the landscape as mountains were separately imported into the GOmatrix analysis. The method of producing GOmatrix is also described in Chapter 4. These two steps aimed to define the significant biological functions of sub-groups of the conserved genes between carp and human.

5.2.6 Programming

5.2.6.1 Programming for constructing ortholog groups

A PERL programme, called FindOrthologs and available at http://legr.liv.ac.uk/orthology/index.htm, was created for processing the all-versus-all reciprocal BLAST searches automatically in a Linux machine. This programme was designed to find orthologs between two or three species. Its computing performance was better than INPARANOID (Remm *et al.* 2001) and OrthoMCL (Li *et al.* 2003), since it outputs orthologous information directly into flat files which do not occupy much machine memory. It can be used for reciprocal BLAST of nucleotides against peptides, nucleotides against nucleotides, and peptides against peptides.

5.2.6.2 Programming for computing Monte Carlo simulation

PERL scripts were also created to process the Monte Carlo simulations. These scripts were compatible with both Linux systems and Windows systems. They were necessary to compute the test with any convenience, since each test requires 100,000 repeats of each calculation.

5.3 Results

5.3.1 Metagenes between human and common carp

Zebrafish is a model vertebrate species for which much more sequence data is available than for the common carp. However common carp is close to zebrafish in phylogenetic terms and as a consequence their protein coding sequences share much homology (Roest and Weissenbach 2005). Sequence data of common carp, zebrafish and human were processed to build up the multi-species orthologous gene groups (Figure 5.1a). Using zebrafish data as a bridge, the all-versus-all reciprocal BLAST searches (p < E-5 and bit score > 50) identified a set of 2262 orthology gene groups between carp EST assemblies and human proteins. It also identified another set of 2273 orthology gene groups between individual carp ESTs and human proteins. Of these two sets of orthology groups, 2206 were found in both sets and were defined as the putative orthologs. The 2206 gene groups were thus defined as the identified metagenes representing homologs that are common to the common carp and human. These metagenes includes ~68% of the BLAST-identified genes in carpBASE 2.1. They were compared with metagenes produced in Stuart Kim's group (Stuart *et al.* 2003) and 725 were found to exist in both sets, which indicates that their approach discovered only one third of human-carp orthologs from the current available data. Obviously, the number of putative carp-human orthologs was small due to the limited sequences data for the common carp. It can be expected that the number of carp-human orthologs would be increased as and when more carp sequences become available.

5.3.2 Expression alignment for the metagenes

Gene expression profiles of both human and common carp were only available for some of the metagenes. The Locuslinks (Pruitt and Maglott 2001) connects the RefSeq human protein sequences (Pruitt and Maglott 2001) to the human gene expression profiles and 2189 of the 2206 metagenes had Locuslinks IDs (Pruitt and Maglott 2001). 2187 had carp gene expression profiles and 1915 had human gene expression profiles (Figure 5.1b). The ExprAlign approach computed the Pearson correlation coefficients (>0.25) of

gene expression for the metagenes in the common carp and for the same metagenes in the human. 3159 pairs of metagenes had correlations greater than 0.25 in both carp and human expression datasets.

The correlation scores for the 3159 pairs of metagenes were sorted in common carp and human respectively, and then ranked by order. *P* values were calculated from the ranked orders by Monte Carlo simulation to indicate how conserved the correlation of gene expressions for the pair of metagenes were. The negative logarithms of these rank order *P* values were used to position the metagenes in the 3D landscape using the VxInsight package (Figure 5.1c). Here the VxOrd algorithm was used to build the 2D ordination map; VxImport was used to create the annotation database with the carpBASE 2.1 data; and VxInsight was used to visualise the data using a landscape metaphor. The landscape located 431 orthologous genes (~20% of all metagenes) into four large mountains B1, B2, B3 and B4 (Figure 5.3a) composed of 102, 82, 60 and 56 metagenes, respectively. Several small mountains containing smaller numbers of metagenes were positioned around these 4 big mountains, making up the remainder.

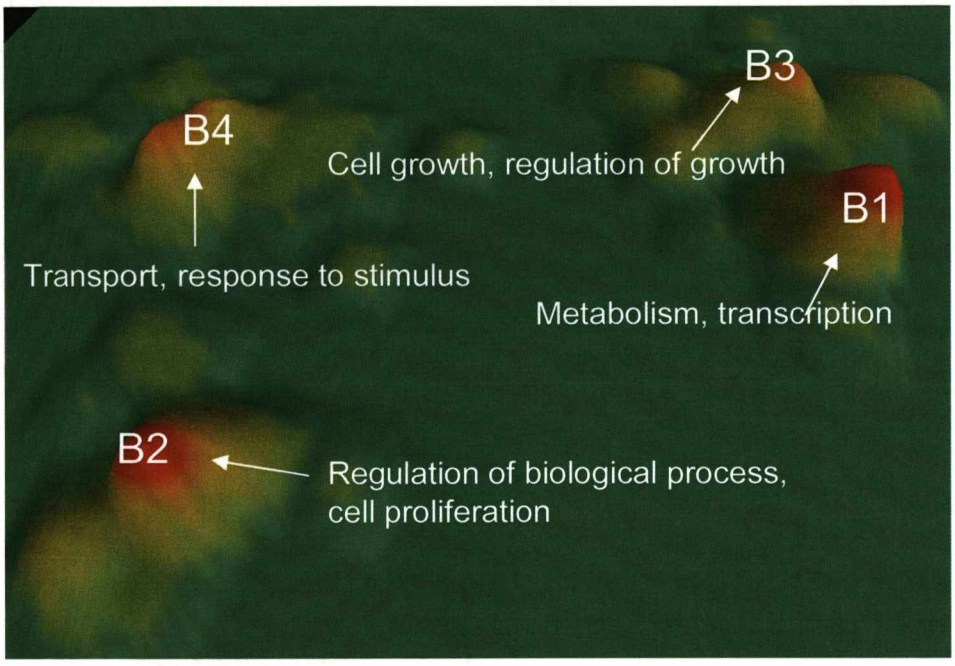
In Chapter 4, the landscape features were used to indicate the correlated expression of genes into groups within a single species. Here the mountains refer to the correlated rank order of metagene correlations within each of the two compared species.

To validate the robustness of the visualisation in this approach, I randomised and shuffled the rank order data for the 431 metagenes, then constructed a random VxInsight map (Figure 5.3b). We did not find any structure to the resulting landscape, which suggested that the method for visualisation was robust in that it produced structures that arose from the structure of the data.

5.3.3 GOmatrix for metagenes

The conservation of metagenes relationships between the two species, as indicated by the 4 mountains, suggests that for each mountain the biological functions or processes associated with these metagenes would also be

a



b

Figure 5.3: (a) Metagene mountains. There are 4 big mountains which are labelled as B1, B2, B3 and B4. (b) Random map for the metagenes does show any structure.

c

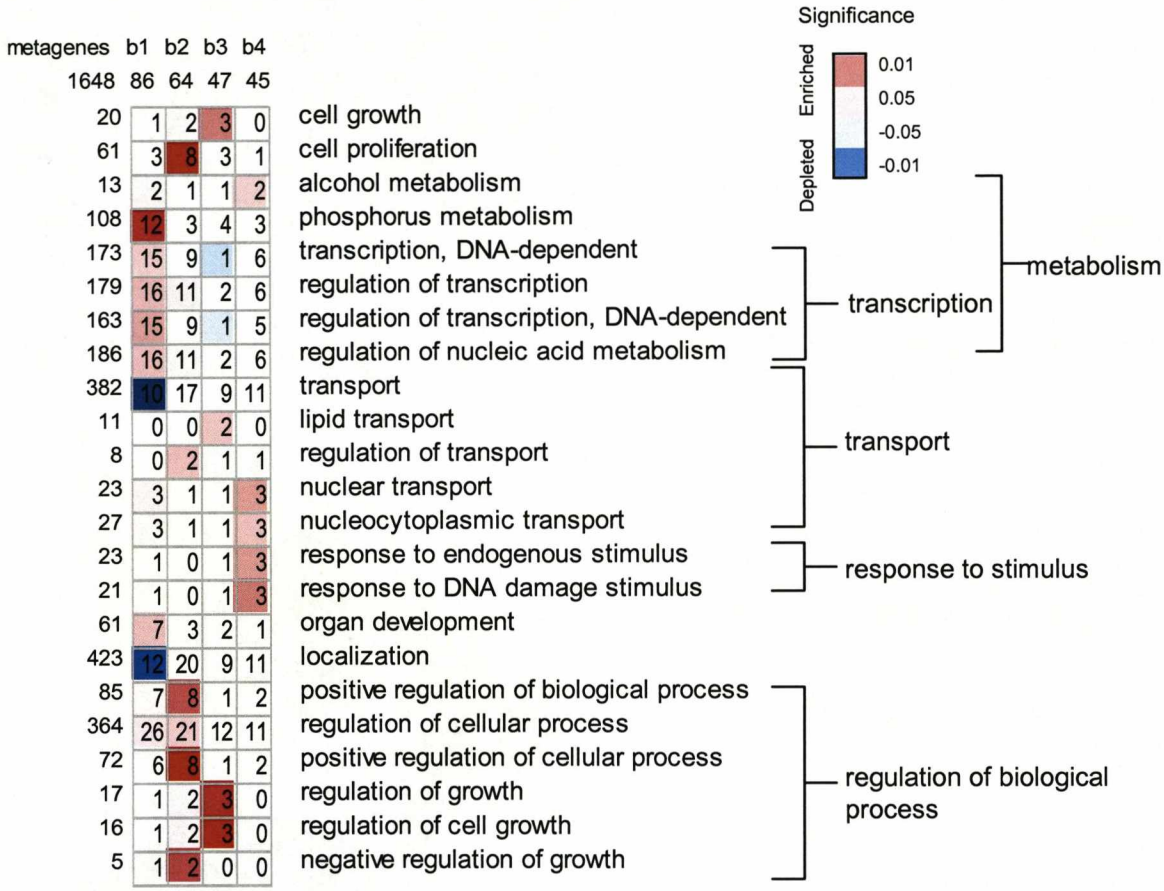


Figure 5.3 (continued): (c) GOmatrix for the 4 big mountains. The number inside each box is the number of matches for the genes in the mountain associated with the particular GO sub-category.

conserved between two species. The GOmatrix technique was used to collate the functional annotations of the metagenes within each mountain and to assess whether any of the 4 mountains possessed any significant over-representation of any GO category. Figure 5.3c indicates that each of the 4 mountains were significantly enriched (over-represented) or depleted (under-represented) in several sub-categories of the GO biological process domain. Thus, B1 was significantly depleted in 'Transport' and 'Localisation' categories but enriched in 'Phosphorus metabolism' and 4 sub-categories of transcription categories ('DNA-dependent transcription', 'regulation of transcription', 'DNA-dependent regulation of transcription', and 'Regulation of nucleic acid metabolism'). B3 was under-represented in these 4 sub-categories of transcription, but was over-represented in 'cell growth', 'regulation of growth' and regulation of cell growth'. By contrast B2 was significantly enriched in 'Cell proliferation', 'Regulation of transport', and 4 sub-categories of regulation of biological process, namely 'Positive regulation of biological process', 'Regulation of cellular process', 'Positive regulation of cellular process', and 'Negative regulation of growth'. B4 was mainly over-represented in 'Nuclear transport', 'Nucleocytoplasmic transport', 'Response to endogenous stimulus', and 'Response to DNA damage stimulus'.

5.3.4 Reactome annotations

Skypainter (http://www.reactome.com/cgi-bin/skypainter2?DB=gk_current) was also used to search the Reactome database (Joshi-Tope *et al.* 2003; Joshi-Tope *et al.* 2005) in order to represent diverse reactions of biological processes for the 431 metagenes in the landscape. The Reactome database is a curated, peer-reviewed resource of biological processes. The basic unit of the Reactome database is a reaction, which is grouped into causal chains to form pathways. The Reactome's primary domain is pathways in human, but it is relevant to other model organism by projecting human pathways onto those species *via* putative orthologs. The Skypainter is a web-interface tool allowing user input query genes to search the Reactome database. It determines which reactions or

pathways are statistically overrepresented in a set of genes as specified by submitted list of identifiers. In other words, given a list of genes, Skypainter can identify common reactions or pathways for these genes. Figure 5.4a indicates that metagenes in B1 were enriched in reactions/pathways of transcription and HIV infection; Figure 5.4b and 5.4d show that metagenes in B2 and B3 were over-represented in reactions/pathways of translation. Figure 5.4e indicates that the 431 metagenes with conserved order between human and carp contain genes that were enriched in reactions/pathways for transcription and translation, and also mainly participate in reactions/pathways for nucleotide metabolism, DNA repair, insulin receptor-mediated signalling and HIV infection. Figure 5.4f shows that the 2206 metagenes were over-represented in reactions/pathway for transcription and translation, and also mainly participate in other current available Reactome reactions/pathways except influenza infection and xenobiotic metabolism.

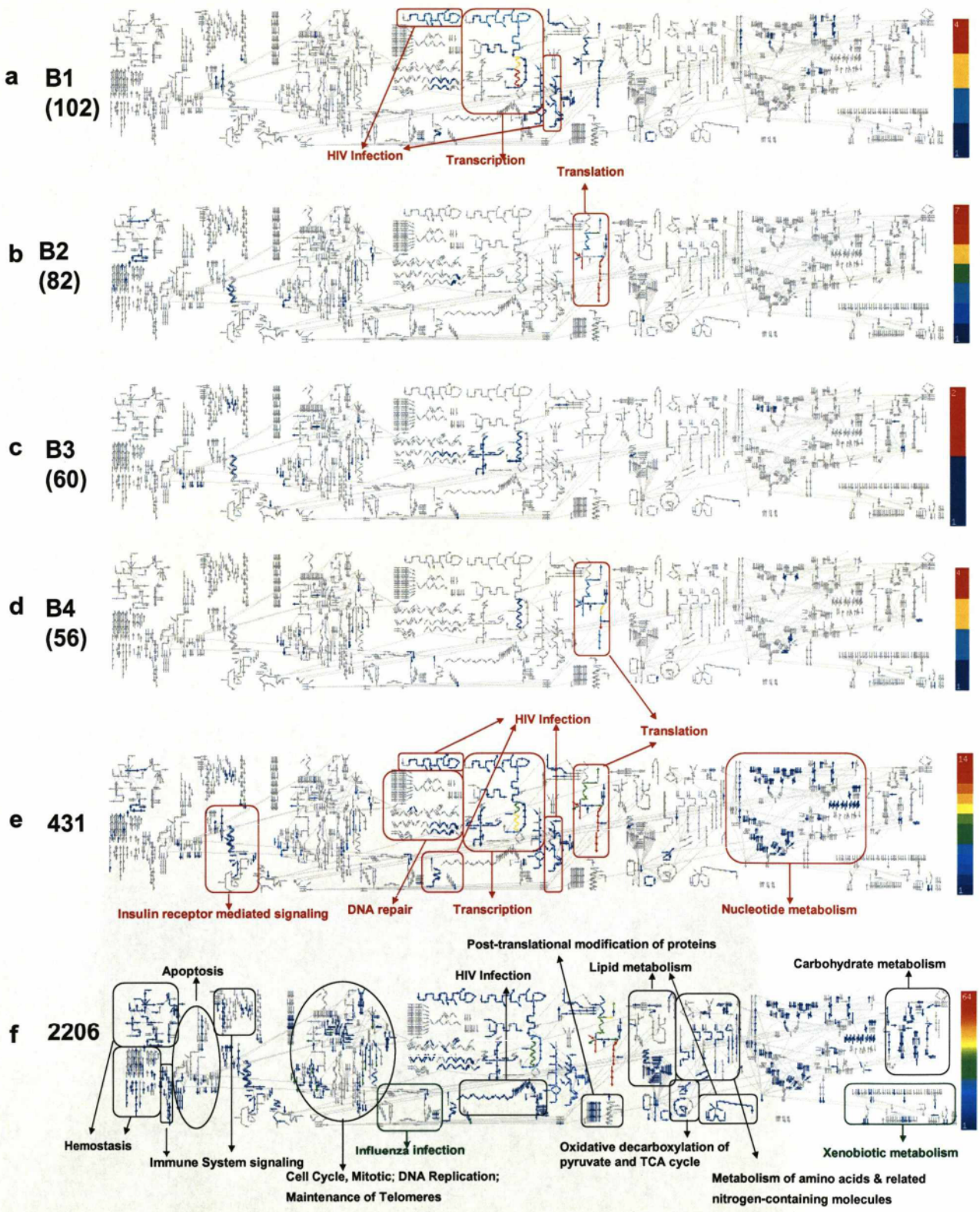


Figure 5.4: Reactome matches for sequences in the mountains. (a) Reactome map for metagenes in B1. (b) Map for metagenes in B2. (c) Map for metagenes in B3. (d) Map for metagenes in B4. (e) Map for the 431 metagenes. (f) Map for 2206 metagenes.

5.4 Discussion

5.4.1 Reciprocal BLAST and the metagene method

As orthologs are related through evolutionary history, phylogenetic trees are the most natural way to detect orthologs. However, on the scale of complete genomes, the analysis of phylogenetic tree is both extremely labour-intensive and error-prone due to the inherent difficulties of phylogenetic tree construction. An alternative method is to use all-versus-all sequence comparison between two genomes to detect orthologs. The underlying premise is that orthologs are more similar to each other in sequence alignment than they are to any other sequence from the respective genomes. It is undertaken by taking a sequence from species A and BLASTing it against the sequence database for species B. The sequence of the highest-scoring gene is taken and BLASTed against the sequence database of species A. If this returns the gene originally used as the highest scorer, then the two genes are considered as putative orthologs.

Many studies have presented algorithms based on reciprocal all-versus-all BLAST for constructing gene orthology relationships (Arvestad *et al.* 2003; Frazer *et al.* 2003; Grigoryev *et al.* 2004; He and Goldwasser 2005). Several resources for orthology analysis are also already available to the public: databases such as COG (Clusters of Orthologous Groups) (Tatusov *et al.* 2003; Marchler-Bauer *et al.* 2005), EGO (Eukaryotic Gene Orthologs) (Lee *et al.* 2002), and OFAM (http://cgg.ebi.ac.uk/services/ortho-fam/), and programmes such as INPARANOID (Remm *et al.* 2001) and OrthoMCL (Li *et al.* 2003). COG recognises relationships among at least three distinct lineages but is highly inconsistent (Stuart *et al.* 2003). EGO, previously named TOGA (TIGR Ortholog Gene Alignment), uses transitive reciprocal best BLAST hits to define their tentative ortholog groups (TOGs) between multiple species. It is thus very restrictive and is easily misled by the functional redundancy of multiple paralogs, and by the absence of true orthologs within incomplete genome data sets (Remm *et al.* 2001). OFAM is another database of protein ortholog families from complete genomes protein database generated by the MCL (Markov Cluster) algorithm (Enright *et al.* 2002) (http://micans.org/mcl/)

for rapid and accurate clustering of protein sequences. It is available from http://cgg.ebi.ac.uk/services/ortho-fam, but it has not been published formally and its description information is not available. The INPARANOID (IN-PARAlog aNd Ortholog IDentification) (Remm *et al.* 2001) algorithm exploits a BLAST-based strategy to identify orthologs as reciprocal best hits between two species, and defines out-paralogs as paralogs that predate the species split, and in-paralogs as paralogs that arose after the species split. It can only process pair-wise comparisons between two species and works for protein sequences. Moreover, it is relatively inefficient in that it requires long computing times when implemented in a single computer. The OrthoMCL method (Li *et al.* 2003) performs similarly to the INPARANOID algorithm but can be extended to cluster orthologs from multiple species.

Several studies have also leveraged microarray analysis by establishing orthologous genes between species. Publicly available packages include Metagenes (Kim *et al.* 2001; Stuart *et al.* 2003), Ensembl (Hubbard *et al.* 2007), RESOURCERER (Tsai *et al.* 2001; Grigoryev *et al.* 2004), *etc.* The Metagene approach was introduced by J. Stuart *et al.* (Stuart *et al.* 2003) to explore the co-expression relationships cross-species used the reciprocal best BLAST hit method. The approach was performed *via* a reciprocal all-against-all BLAST between every pair of protein sequences from each of the compared organisms to define orthologous genes. It is a good method for linking orthologs based on co-expression relationships, but it tends to miss non-orthologous information on duplicated genes. L. Huminiecki and K. Wolfe (Huminiecki and Wolfe 2004) studied the gene expression profiles of orthologous gene sets in human and mouse using the Ensembl orthology groups (Clamp *et al.* 2003), which provides substantial amounts of information relating orthologous gene groups between species and gene expression profiles for the species. However, most of the data for the Ensembl orthologous gene groups was available for model species only. RESOURCERER (Tsai *et al.* 2001; Grigoryev *et al.* 2004) is a microarray-resource annotation and cross-reference database based on the TIGR Eukaryotic Gene Ortholog (EGO) database. It contains information for all commercially available Affymetrix

GeneChips, but only allows comparison of two chips simultaneously in model species.

The approach developed in EGO (Lee *et al.* 2002) is very restrictive, so duplicated genes will not be identified; data from COGS (Tatusov *et al.* 2003) is highly inconsistent; OFAM is poorly developed; INPARANOID is limited in two-species comparison and also requires massive amount of computer memory easily; OrthoMCL successfully completes multiple species comparisons but uses the INPARANOID package as a core, so it also has the same problem of INPARANOID; Ensembl and RESOURCERER provide only information for model species. So we conclude that none of the available databases or tools is ideal for this study of common carp and human through orthology. The Metagene approach missed the duplicated genes, but its ideas on constructing the orthology groups and the cross-species gene expression map were judged most suitable to this study, and a further analysis on paralogs can be introduced after the Metagene method in order to understand the genome duplication for common carp.

5.4.2 The bridge species

Orthology analyses across species are important firstly for identifying conserved genes and secondly to enable comparative evolutionary approaches to the determination of gene function. Thus a definitive indication of gene function in one species can then be used as a model for understanding gene function in related species. Non-model species, such as common carp, tend to have poor sequence resources which impede the construction of orthologous relationship to the well-annotated model species, such as human, for which confidence of gene coverage is high. Common carp is phylogenetically close to zebrafish and their sequences share much homology (Roest and Weissenbach 2005). For this study, zebrafish sequence data acted as a 'bridge' to connect carp sequence data with the corresponding human sequence data. A direct reciprocal BLAST alignment between carp and human identified 2091 carp-human orthologous groups (metagenes), whilst the use of zebrafish sequences as a bridge identified 2206 metagene, an increase of 5.5%. This

indicates that the bridging with a nearer phylogenetic species increased the discovery of putative orthologs with a distantly related species by a small but significant amount. This method might prove useful in exploring the sequence relationships across the vertebrate lineage.

5.4.3 Co-expression between metagenes

We were interested in finding whether the metagenes had conserved expression relationships in gene expression patterns established within a species were conserved across species. Whilst gene and gene group responses may be stress-specific it would be necessary to undertake identical treatments in the compared species. This is practically impossible in comparing the non-model carp with the model human simply because of their different physiological status and approaches to altered temperature; thus, cold exposure of humans elicits an entirely different thermoregulatory response to carp (Cossins and Bowler 1987). However, it is possible to ask whether the associations between genes in expression responses are conserved between species since this can be addressed independently of the experimental treatments imposed using a probabilistic method (Stuart *et al.* 2003).

As in Chapter 5 we have generated a gene co-expression matrix based on the Pearsons correlation coefficient between all pairs of carp-human metagenes for carp. We have created another co-expression profile for all pairs of carp-human metagenes for human. The conservation of the rank position for each metagene pair for both species was tested using the rank order statistic. This generates a negative logarithm of the *P*-values of the conserved rank order between species for each pair of metagenes as the similarity measure, of which 431 of the available 2206 metagenes were judged significantly conserved. These *P*-values were used by VxOrd to direct the force-directed placement of genes onto a landscape placed onto the X-Y plane by VxInsight, the height of the features indicating the density of genes contained within it. The resulting image placed the 431 metagenes onto a landscape map, revealing in 4 large and distinct mountains, labelled B1-4, and some smaller surrounding features. In interpreting this image it is important to be clear as to the meaning of these

landscape features; each mountain contains a network of metagenes with connections being made to other metagenes based on significantly low *P*-values. Thus each mountain contains connected metagenes with highly correlated expression profiles across the arrays surveyed, and with a rank order of correlation coefficient across all metagenes which is conserved between the compared species.

Metagenes contained within the same mountain might participate in the same biological process since they possess correlated expression properties and thus might be subjected to common regulatory pathways. This was tested in two ways on each of the 4 large gene clusters. First, the GOmatrix approach described in Chapter 4 was used to test the significance of GO category representation in each of the 4 gene lists. This GOmatrix outcome indicates that the different mountains were related to distinctive GO biological processes. Thus B1 was broadly related to DNA transcription, B2 was related to regulation of biological process, and B4 was related to response to stimulus.

Second, we used the Reactome system which applies genes contained within a gene list onto a near complete map of all gene-mediated steps in many important biological processes. This map has been created separately from the GO annotation, and thus represents a quite separate test of process/pathway involvement within gene lists, and a more extensive list of processes than included within the KEGG metabolic pathways. One limitation for the current Reactome database is that there are not many reactions and pathways defined. Nevertheless, the outcome corroborated the GOmatrix assignments for mountain B1 as enriched in genes contributing to DNA transcription and responding to HIV infection. B2 and B3 were linked to protein translation. In this study, it suggested the 431 metagenes in the landscape were significantly over-represented in reactions and pathways in transcriptions and translations.

Basing on the ortholog information, further efforts can be emphasised on the investigation of the whole genome duplications for common carp. This requires an introduction of paralog analysis based on the already known orthologs across common carp, zebrafish and human. Two-round of CAP3 clustering was implemented in the carp sequence analyses (Chapter 3) and

provided information for unique genes in different gene main-groups. The genes inside a main-group might contain potential information of paralogs. This offers a clue to introduce further analysis on the common carp whole-genome duplication. The gene expression profiles would also help to understand the functional diversification of paralogs. This study can be extended to multiple species in the future work if more expression data available, such as zebrafish.

CHAPTER 6: BIOINFORMATIC COLLATION OF SEQUENCE DATA AND DESIGN OF AN OPTIMISED OLIGOARRAY FOR A NON-MODEL SPECIES

6.1 Introduction

Two major platforms for high-density microarray manufacture are in common use depending on the type of probe employed. The first uses cDNA probes which are deposited using contact printing devices, as mentioned in Chapter 4. The second uses oligonucleotide probes that are either synthesized chemically and spotted onto the array, or are synthesized on the array *in situ* using a variety of methods such as photolithography (Barone *et al.* 2001) and ink-jet deposition onto the growing oligo chain (Hughes *et al.* 2001). As sequence data accumulates, and as array fabrication technology evolves, the options for design and construction of arrays are changing, particularly for projects that require limited production runs. The result is that on-chip oligoarrays are now the preferred route for production (Li *et al.* In press), and the user needs to define the list of target sequences against which probes are required. Given that the probe capacity of most production platforms is limited, whilst the sequence data and list of potential probes is much greater, then decision have to be made about which target sequences are included and which are discarded. This chapter describes a route to achieving this for the rainbow trout, a non-model species for which at the time of initiating this project (September 2005) there were ~220K EST and ~2K mRNA entries available in GenBank.

6.1.1 Oligonucleotides and oligoarrays

Oligonucleotides are short sequential base-pair segments, ranging from 15 to 150 nucleotides in length, taken from hundreds of nucleotides in a DNA segment that functions as a gene (McLachlan *et al.* 2004). In array experiments the 'probe' is defined as the immobilised DNA placed in a known location on the array surface. In oligonucleotide microarrays, the probes are designed to complement parts of the mRNA sequences. Typically the probe has a known

identity in that it is designed to complement the target. The 'target' is the complex mixture of labelled mRNAs isolated from the tissue of interest, which have been reverse transcribed to first strand cDNA and fluorescence-labelled. Oligoprobes may be more accessible for hybridization than the cDNA probe, due to their much shorter chains and single terminal points for attachment to the array surface. Moreover, oligonucleotides also offers greater specificity than cDNA or PCR products, having the capacity to distinguish single-nucleotide polymorphisms and splice variants (Hughes *et al.* 2001).

Of course, there are also disadvantages to the use of oligoprobes. Firstly, the sensitivity of fluorescence detection declines substantially as probes become shorter (Kreil *et al.* 2007). Thus, cDNA probes allow the detection of low levels of target in tissue extracts, whilst oligoprobes frequently require larger amounts of tissue RNA for adequate detection of target, or, as in the case of the Affymetrix platform, one- or two-cycle amplification of target using a PCR-based approach. This carries significant costs since it is applied to each and every RNA sample, and it also introduces another source of error. In part this difference in hybridisation properties is due to the longer and more flexible cDNA probe offering multiple sites and reduced steric constraints for hybridisation of target to probe. Secondly, the potential for cross-hybridization of multiple targets with short oligoprobes is reduced especially if the probes are defined to be specific to just one the known multiple targets. These two issues of hybridisation performance and specificity represent a trade-off and it is now common to specify 65mer probes as the optimal configuration since they avoid the lack of sensitivity of shorter probes (25-35mer) without the extra costs of building longer 100-150mer probes, or incurring the lack of specificity of cDNA probes. 65mer oligoprobes also match the accuracy of on-chip fabrication on the commercial production platforms.

Several commercially available oligoarray manufactures are now commonplace, notably including Affymetrix (www.affymetrix.com), Agilent Technologies (www.agilent.com), Nimblegene (www.nimblegen.com), OGT (Oxford Gene Technology, www.ogt.co.uk), GE Healthcare (www.gehealthcare.com), Ocimum Biosolutions (www.ocimumbio.com), and

the Illumina (www.illumina.com) platforms. Until very recently, the only platform capable of flexible and rapid production with the ability to specify low (>10) production runs was the Agilent on-chip, ink-jet deposited oligosynthesis array platform originally developed by Rosetta (Slatter *et al.* 2006). This specified 22,000 oligoprobes that could be designed from a FASTA sequence file, and arrays could be supplied within 4 weeks of ordering. Very recent developments have greatly expanded the number of probes on each platform; thus in early 2007 the Agilent platform increased the number of probes from 22K to 240K, with the option of printing multiple arrays on each slide, such as 4 x 44K Agilent project. This company has indicated a further increase to 750K by mid-2008. The Nimblegene platform has similarly projected an increase in the number of features from 450K to 2.1M in Q4 of 2007. By increasing the capacity of arrays by such a large margin it is now possible to interrogate a larger number of potential probes, and the need to restrict choice of probes will not apply from mid-2007 onwards.

6.1.2 Oligoarray design

Taking advantage of the new flexible technologies in microarray fabrication requires only (i) sequence data (Gao *et al.* 2001; Hughes *et al.* 2001) and (ii) an ability to predict and specify the optimal probe design for each target (Charbonnier *et al.* 2005). Sequence data is increasingly available due to the rapid accumulation of EST and mRNA sequences in GenBank. In particular, the genomic model species, such as *C. elegans*, *Drosophila*, mouse and human, have very extensive EST resources and an increasing number of other non-model species are currently experiencing rapid increases in EST coverage (Cossins and Crawford 2005). Even non-model species are experiencing an increase in EST resources; for example, rainbow trout possessed ~220K ESTs in dbEST in mid 2005, whilst the common carp possessed ~25K. These collections provide the raw material for probe design software but because they contain substantial redundancy they need assembly into non-redundant collection, using pipelines such as those described in this Chapter.

The design of oligoprobes is a complex task in that it seeks to specify the optimal probe taking into account several different considerations. First, oligonucleotide probes should be specific to their respective targets with little or no tendency for cross-hybridization to other targets. Second, they should be free of secondary structures that may interfere with its hybridisation of target to immobilised probe (Rouillard *et al.* 2003). Third, the melting temperatures of the defined probes should correspond to the chosen hybridisation temperature, and this means balancing the base content of the chosen design. Fortunately, there are several bioinformatics programmes to design specific oligonucleotides for microarrays, such as the open source OligoArray (Rouillard *et al.* 2003) (http://berry.engin.umich.edu/oligoarray2/) and the commercially available ArrayDesigner (www.premierbiosoft.com/dnamicroarray/). Whilst these probe design algorithms have achieved some sophistication there is no guarantee that the resulting predicted probes capture all of the important features of the best performing probe. It is thus necessary to test the performance of several predicted probes so that the best performing can be selected under the conditions to be used in the following microarray analysis.

6.1.3 Objectives

The substantial costs of generating cDNA libraries, cDNA clones and the associated ESTs for poorly resourced species make it impractical for many laboratories to generate sufficient sequence information to construct a microarray. Even when funding allows there are problems in generating a useful list of sequences. Thus, in conjunction with the European Community funded STRESSGENES project, the Cossins lab has been involved in the production of rainbow trout cDNA collection. The production processes employed in that programme generated huge redundancy in the clone collection, thereby limiting the representation of unique sequences on the resulting array despite the large-scale effort and expenditure. Meanwhile there is a continuing demand worldwide for a high quality arrays for analysis of trout and salmon, Fortunately, during the STRESSGENES project a large number of

ESTs were generated by laboratories worldwide and this provide the necessary scale of sequence data to offer a starting point for array construction. Using this, we developed an informatic pipeline for sequence selection and probe design specifically for the Agilent 22K platform. This Chapter describes this project and the outcome. Another potential option would be to transfer interest to a phylogenetically-related species for which substantial resources are available (i.e. zebrafish). In the case of salmon and trout, which are both species of considerable, aquacultural importance, this would not be acceptable to the stakeholders commissioning this research.

Trout ESTs comprised ~220K at the time of analysis, and it was necessary to reduce this redundant dataset to a non-redundant collection of sequences. To generate a minimally-redundant collection of sequences it is necessary to assemble the ESTs, and for that we sourced an assembly of ~50K contigs from the publicly available Gene Indices project, now placed at Harvard Medical School (http://compbio.dfci.harvard.edu/tgi/). This list was informatically processed in a series of discrete stages to focus down on a core group of sequences which were non-redundant, and which were successfully annotated with functional and GO terms. The resulting list of 22K sequences was then submitted to a commercial probe design algorithm to define several putative probes, each of which was tested experimentally. The collection of best performing probes was then placed onto an Agilent array platform and used in several ongoing experiments.

6.2 Materials and Methods

The method includes procedures to generate a minimally-redundant but maximally representative list of sequences from available sequence data, and to refine that dataset to provide the most informative sequence data consistent with the capacity of the array platform being used.

6.2.1 Sequences resources from RTGI and GenBank

In theory, all of the rainbow trout sequences from GenBank should be included for maximising the informative sequences for the oligoarray design. But in practice, ESTs are usually used to predict oligonucleotides. dbEST, a component part of GenBank, is an excellent resource for this project. As of April 1st 2005 the dbEST contained 227,018 *Oncorhynchus mykiss* (rainbow trout) ESTs.

An assembly of this sequence collection was fortunately available from the TIGR gene indices project (the current indices have since been moved to http://compbio.dfci.harvard.edu/tgi/), and the RTGI (Rainbow Trout Index) Release 5.0 (January 31, 2005) was used. This included 199,167 ESTs and 1197 ETs of rainbow trout and produced 25,427 Tentative Consensus (TC) sequences (TIGR Clusters / contigs), 199 ET sequences and 31,394 singletons. TCs are created by assembling ESTs into virtual transcripts. ETs are a non-redundant set of non-human mature transcript sequences curated by the TIGR's Expressed Gene Anatomy Database (EGAD, http://compbio.dfci.harvard.edu/tgi/definitions.html). The ESTs included in another rainbow trout collection (UniGene Build #14) was 68.9% of the rainbow trout ESTs (release on 1 April 2005) in dbEST, whilst, RTGI 5.0 contained 87.7% of the rainbow trout ESTs (release on 1 April 2005). Because of this, and that only the RTGI build had generated consensuses of the cluster, the RTGI build was selected as the starting point for this analysis and the constituent TCs, the ETs and singletons were downloaded as the principal sequence resource for this project.

Some nucleotide sequences were excluded in the RGTI, including full-length sequences included in GenBank. Ignoring those sequences would

significantly affect the accuracy of the result for oligonucleotides predictions. Therefore, another 1344 rainbow trout mRNAs, excluding ESTs and mitochondrial sequences, were retrieved from the GenBank by using the Entrez tool and were included in the sequence resource for this project. Entrez (Maglott *et al.* 2005; Benson *et al.* 2006) is a cross-database search engine for life science data, such as sequence data, expression profiles, structure data, *etc.* It can be accessed *via* the web-interface at http://www.ncbi.nlm.nih.gov/Entrez.

6.2.2 How to select the consensuses?

A good oligonucleotide set should be informative, non-redundant and specified in the sensed direction. If the oligonucleotides related to unidentified targets, the biological meaning of expression patterns for this target would be unknown and the resulting oligoarray data would be difficult to interpret. Also in the present case where sequence data exceeds the probe capacity of the platform the probes for unidentified targets are arguably the least useful source for probe design. Also an unacceptably high redundancy of the sequence resource would also cause some inefficiency and reduced cost-effectiveness in the production cycle. The sequences need to be specified in the sense direction since this serves as the appropriate hybridisation probe for the anti-sense target. To implement these different needs, a pipeline (Figure 6.1) was developed to optimise the collation of meaningful sequences from open source sequence data, and the specification of a minimally redundant probe set.

6.2.2.1 Identify the sequences

Firstly, the BLAST search was implemented in the pipeline to identify the informative and sensed sequences. For this, the BLASTX tool was used to search the ~57,000 sequences of TCs, ETs and singletons from the RTGI and ~1350 mRNA from GenBank against selected protein sequence databases (Figure 6.1a). The priority order of the selected protein sequence databases in BLSAT searches was Swiss-Prot → RefSeq → nr. Swiss-Prot (Apweiler *et al.* 2004) is a very high quality sequence database for real proteins due to its manually curated annotation. RefSeq (Pruitt and Maglott 2001) is another good quality protein database developed in the NCBI. Sequences in the nr database

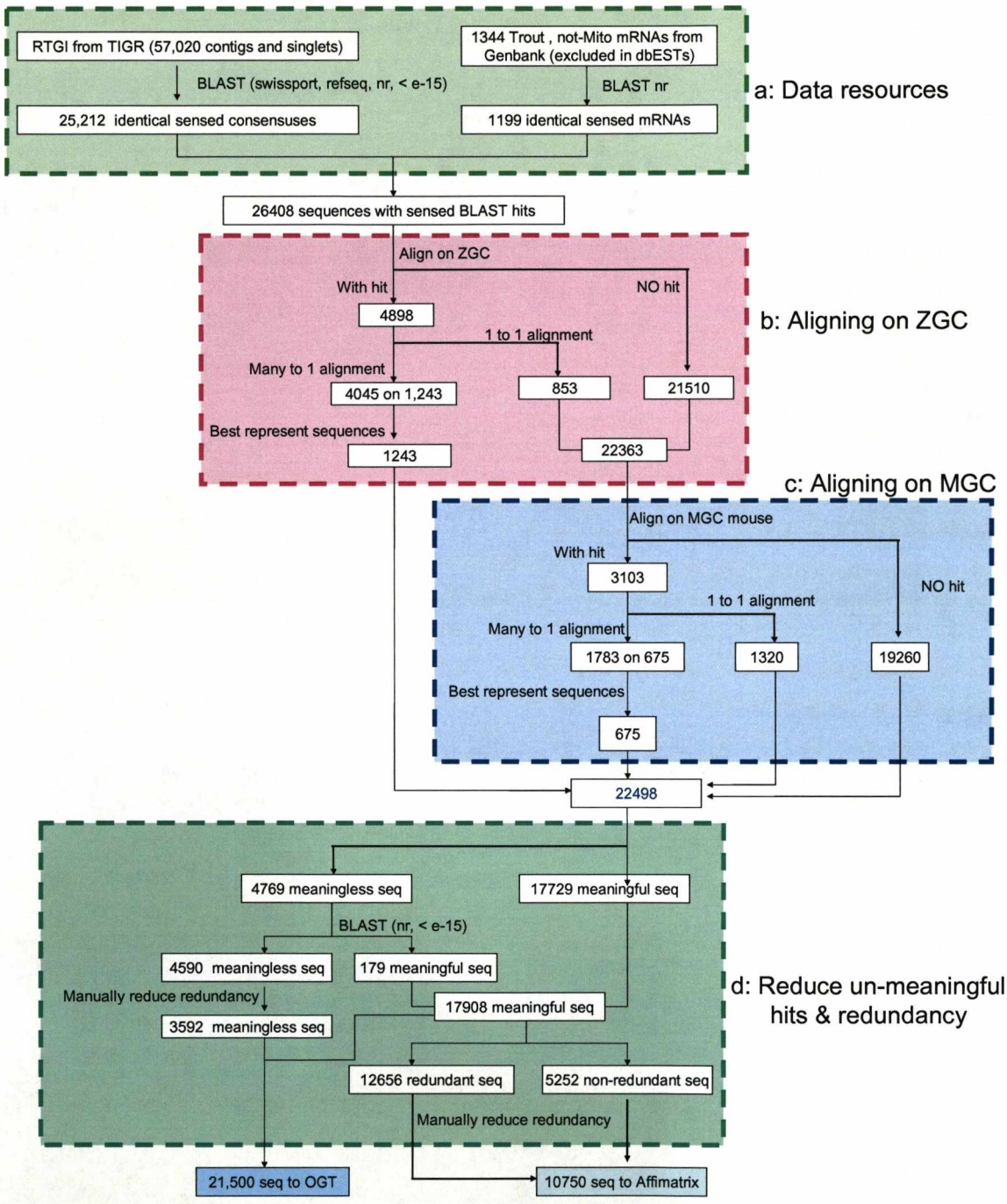


Figure 6.1: Optimising sequence collection for rainbow trout oligoarray design

might be not of good quality, but they can provide additional information for those sequences which have no hits from Swiss-port and RefSeq. The aim for using this order is to assign the best quality identity to each query sequence. More details on the "Priority" BLAST and selecting BLAST databases have been mentioned in Chapter 3. Sequences with sensed BLASTX top hits were processed in the further sequence filters. If the open reading frame for the query sequence in the BLASTX alignment is positive (+1, +2 or +3) frame, the query sequence is classified as a sensed sequence; otherwise, the query sequence is un-sensed.

6.2.2.2 Reduce redundancies by aligning on the ZGC and the Mouse full-length cDNAs

In some cases the BLAST identities established for the sequences generated by Section 6.2.2.1 were identical despite the sequences not overlapping and not comprising a single contig. It is likely that they arise from the same gene but are represented as entirely different TCs, or alternatively they represent different isoforms. To remove the former kind of redundancy, the sequences were firstly aligned onto the zebrafish full-length cDNA collections (ZGC) (Gerhard *et al.* 2004) using BLASTN (Figure 6.1b). Sequences without BLASTN hits to the ZGC were stored for the further filter of aligning onto the mouse full-length cDNAs, described below. Sequences with hits on ZGC were placed into one of two categories: sequences in '1 to 1 Alignment' and sequences in 'Many to 1 Alignment'. '1 to 1 Alignment' means only 1 query sequence aligns on a BLASTN top hit sequence; 'Many to 1 Alignment' means multiple sequences have the same BLASTN top hit sequence. Query sequences in '1 to 1 Alignment' were stored for the further filter of aligning on the mouse full-length cDNAs, described below. Query sequences in 'Many to 1 Alignment' were regarded as the same gene even though they have weak overlaps or no overlaps between each other. The longest sequence candidate of each gene in 'Many to 1 Alignment' was processed in the further filters in Section 6.2.2.3.

Following the same way, sequence candidates having no hit or having '1 to 1 alignments' in the ZGC were aligned again on the mouse full-length cDNA (Figure 6.1c) (MGC Project Team 2004). Sequences having no hits, having '1 to 1 Alignment', or the longest sequence candidates of genes in 'Many to 1 Alignments', were stored for the further filters in Section 6.2.2.3.

6.2.2.3 Recovery of non-informative sequences

After filtering by the BLAST searches and the full-length cDNA alignments, some of the TCs possessed non-informative functional annotations, such as "hypothetical protein ...", "predicted protein ...", "similar to ...", "mRNA ...", "cDNA ..." and "unknown ...", *etc.* These TCs were BLASTXed against the nr database to recover those for which a stronger identity could be provided, and the remainder were discarded (Figure 6.1d).

6.2.3 Tools developed for the project

BLAST alignments can be undertaken in a standalone PC but it will take days for a standalone computer (P4 CPU 1.3 GH, 512M RAM) to process large sequence datasets such as those offered by public EST programme. Using a cluster of PC offers considerable time-saving. We developed a Linux cluster to process the BLAST searches. The cluster was composed of 40 commodity Linux machines and can be accessed *via* the web-interface http://bioserv2.sbs.liv.ac.uk/~fishomics/. The ~57K rainbow trout ESTs aligned against the nr database took just about 3 hours. PERL scripts (http://legr.liv.ac.uk/oligoarray/parser.htm) are available to perform BLAST searches for large-scale sequences.

6.3 Results

6.3.1 Filtered sequences

BLASTX searches identified 26,408 sequences in the sense direction (bit score > 50 and e-value < e-5) from ~57,000 sequences in the RTGI and ~1350 rainbow trout mRNA from GenBank (Figure 6.1a). Of these, 4898 sequences had hits on the ZGC, 3103 had hits on MGC and 19,260 had no hit on both. Finally, 3910 sequences were filtered and 22,498 sequences (Figure 6.1b&c) remained after aligning on the full-length cDNAs of zebrafish and mouse. Of the 22,498, 4769 were non-informative in terms of functional annotation and thus biological meaning. Additional BLASTX searches against the nr database recovered the meanings for 179 sequences of the 4769. In total, 4590 were meaningless and 17,908 were meaningful.

6.3.2 Submission to oligoarray manufacturers

A key issue in array design is to provide the best set of gene probes to take up the full capacity of the array. Thus the array capacity of the chosen platform defines the extent to which the sequence dataset must be trimmed back. This project was undertaken at a time when both Agilent and Oxford Gene Technology offer slides containing up to 22,000 features, whilst the NimbleExpress platform from Affymetrix requires 11,000 sequences from which multiple probes (11 match and 11 mismatch in short 25mers) would be designed. It was thus necessary to prune the sequence lists remaining from Section 6.3.1 by rejecting more sequences, either by manual inspection of BLAST identities and by adoption of more stringent BLAST alignments.

From Section 6.3.1, 22,498 sequences were identified. Approximately 1000 sequences from the 4590 meaningless sequences were discarded by using the criterion of BLAST bit score under 90. This left a total of 21,500 sequences comprising 3592 meaningless and the 17,908 meaningful sequences. This list was submitted to OGT for prediction of oligoprobe design (Figure 6.1d).

Because we were unsure which platform would provide the appropriate support for the fabrication process, we also explored the way in which we would reduce the list of target sequences to match the capacity of the

NimbleExpress platform, which at that time was 485,000 features generated against just 11,000 target sequences. The 17,908 meaningful sequences from Section 6.3.1 were thus manually checked for redundancy again to reduce the number of sequence to 10,750 (Figure 6.1d). Firstly, if two or more sequences have a same BLAST hit with identical accession number, the one having highest bit score was kept and others were discarded. Some sequences derived from the same gene had similar BLAST hit descriptions but were linked to different accession numbers. Thus, secondly, a manual check was performed by carefully reading the BLAST hit descriptions for these sequences one-by-one to identify those linked to the same gene identity. If these sequences were confirmed as arising from the same gene, the ones with the highest bit score were retained and others will be discarded. This step involved the comparisons of BLAST hits from Swiss-Prot, RefSeq and nr for a same sequence. The comparisons gave greater confidence in confirming the correct gene descriptions. Finally, the 10,750 sequences were ready to submit to the Affymetrix oligoarray manufacture. Subsequently, a decision was made to submit only to the OGT/Agilent platform and this gene list was not processed any further.

6.3.3 High-quality production of the oligoarrays from this protocols

The arrays designed from the EST collection produced good quality oligoarray data. Figure 6.2, generated by Dr Lisa Olohan at the Liverpool Microarray Facility (LMF, www.liv.ac.uk/lmf/), illustrates the performance of oligoarrays. Fig 6.2A shows a detail from the array image generated after scanning. This shows adequate hybridisation intensities, and also the quality of the spot morphology and placement, and also that the background fluorescence was low compared with spot intensity. Fig 6.2B shows a 'self-self' plot in which both Cy-labelled channels of a single array sampled the same trout liver cDNA sample. The data points should be identical in each channel and thus should lie on the diagonal line. Dispersion reflects the technical errors inherent in the approach. This illustrates a large range of different intensities for different probes which is consistent with the wide dynamic range of the

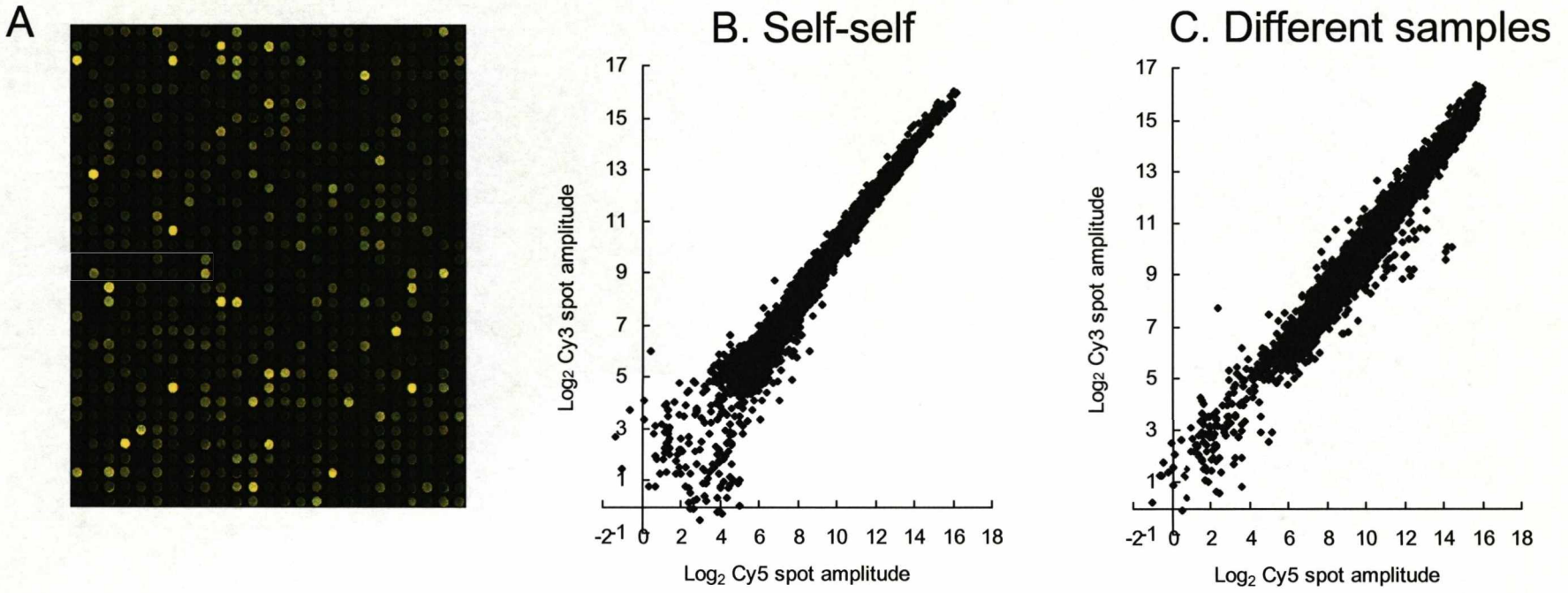


Figure 6.2: Trout oligoarray images. (A) The array image generated after scanning; (B) a 'self-self' plot; (C) a typical experiment in which different Cy-labelled cDNA samples were used in each channel.

transcriptome. Fig 6.2C shows a typical experiment in which different Cy-labelled cDNA samples were used in each channel. Again the data broadly fits the diagonal line, but the spread of data is greater than in A reflecting differences between samples in a subset of genes. Both B and C illustrate LOWESS-normalised data.

6.3.4 Good gene representation of the EST collection from this protocols

The common carp cDNA microarray project cost much money and time in sequencing and assembling ESTs in order to have a good gene representation. However, the protocol in the construction of EST collection for the trout oligoarray saved much money and time but provided good gene representation at the same time. For example, we compared the numbers of enzyme genes in 30 categories of KEGG pathways (http://www.genome.jp/kegg-bin/mk_point_html) (Kanehisa *et al.* 2002) represented in carpBASE 2.1, KEGG mouse genes and our trout EST collection. Figure 6.3 indicates that the number of enzyme genes in the trout EST collection represented in any of the 30 KEGG pathways was larger than that in carpBASE 2.1. This number was also close to that in the KEGG mouse genes except the categories of glycerolipid metabolism. This suggested the gene representation in the trout EST collection covered a large range of trout genes.

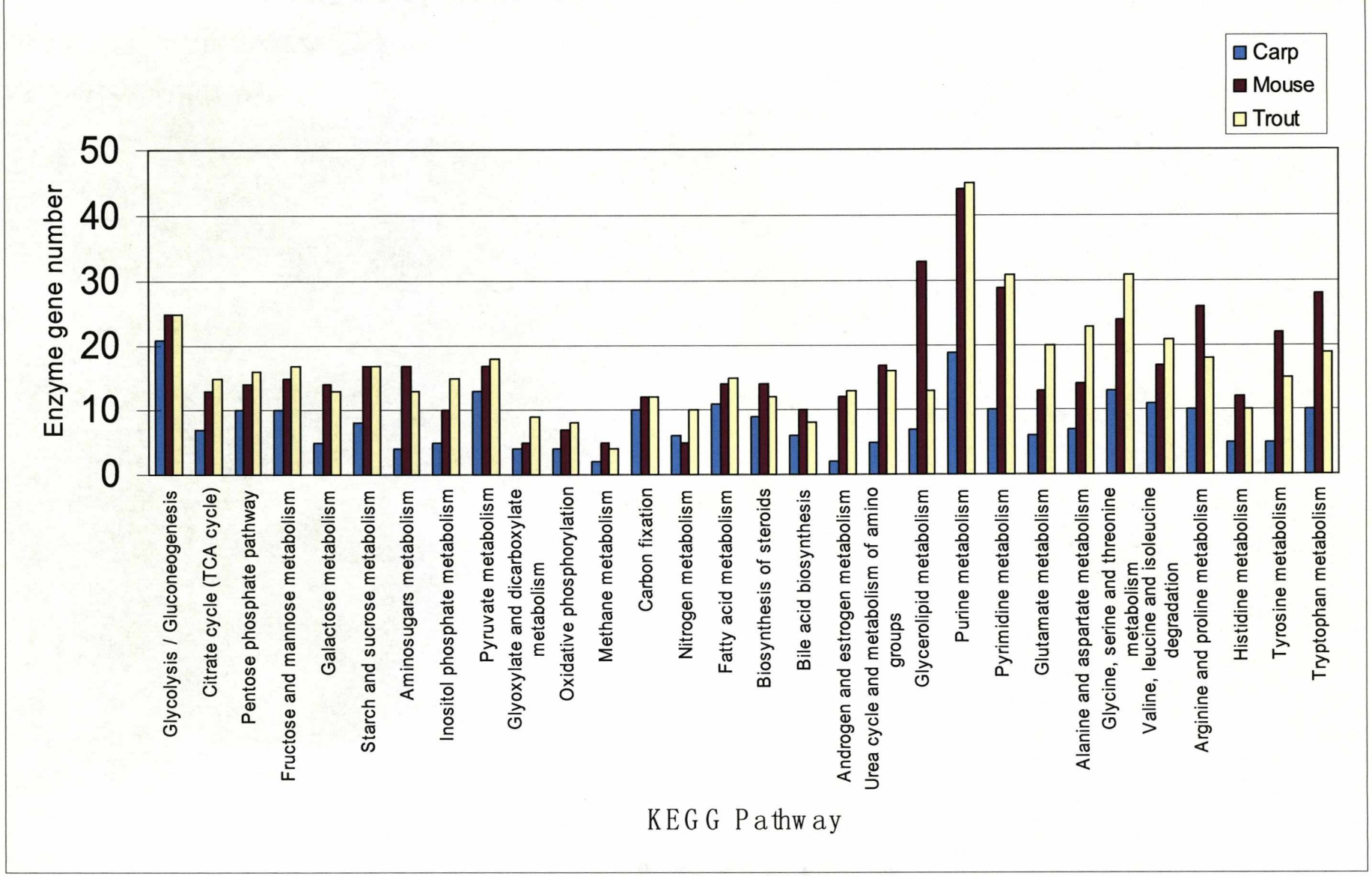


Figure 6.3: Comparison of enzyme gene representation for ESTs in carpBASE, mouse enzymes in KEGG and the EST selection in trout oligoarray

6.4 Discussions

We aimed to devise simple procedures to design and fabricate oligoarrays from the rainbow trout sequence data. In optimising the utility sequence data sets, the principle objective is to maximise the representation of non-redundant and meaningful sequences and minimise the meaningless or the redundant sequences. In this study of oligoarray design for rainbow trout, the TIGR RIGI (Quackenbush *et al.* 2001) and the rainbow trout mRNA from GenBank (Benson *et al.* 2006), containing most of the available rainbow trout sequence information, were used as a convenient and comprehensive resource to design the oligoarray. The source material from dbEST (Boguski *et al.* 1993) was clustered and assembled to generate consensus sequences before further procedures using different tools, such as CAP3 (Huang and Madan 1999), TIGR assembler (Sutton *et al.* 1995), the TGI Clustering tools (TGICL), CD-HIT (Li and Godzik 2006) and BLASTclust (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html). However, the RTGI provided the required listings using credible procedures and it was not necessary for us to undertake assembly.

The BLAST searching in the optimising pipeline discovered the meaningful and sensed sequences. The reasons for the implementations of Swiss-Prot (Boeckmann *et al.* 2003), RefSeq (Pruitt and Maglott 2001) and the nr database have been discussed in Chapter 3. The alignments on full-length cDNAs of zebrafish and mouse decreased the redundancy of the sequences. In fact, more full-length cDNA resources were available in the Mammalian Gene Collection (MGC, http://mgc.nci.nih.gov) (Strausberg *et al.* 1999). The MGC provided full-length open reading frame clones for human, mouse, rat and cow genes. In our study on rainbow trout, full-length cDNAs from zebrafish and mouse were used to identify sequences from a same gene. For studies on other species, the MGC will be a good choice of being subject databases for alignments on full-length cDNAs. The final manual checking was time-consuming, but it did curate the dataset by reducing the redundancy and number of meaningless sequence for limited capacity of the arrays. The output data sets suited our allocated budget for oligoarray fabrication.

Different oligoarray projects might use different approaches to select probes (Rahmann 2002; Sung and Lee 2003; Charbonnier *et al.* 2005; Nordberg 2005; Kreil *et al.* 2006; Verjovski-Almeida *et al.* 2007). For example, Verjovski-Almeida (Verjovski-Almeida *et al.* 2007) implemented a similar protocol to design 44k oligoarray for *Schistosoma mansoni* adult worms. They used CAP3 (Huang and Madan 1999) to assemble transcript information without using the available DFCI *S. mansoni* Gene Index (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=s_mansoni). We used the RTGI as assembly (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=r_trout) so that we did not need to do extra sequence clustering. They used all-versus-all BLASTN to reduce sequence redundancy and could not reduce redundant sequences which were derived from a same gene but had no overlaps. We aligned sequences to the full-length cDNAs of zebrafish and mouse to clean this kind of redundancy. Two protocols both used BLAST to identified sequences but we also implemented automated and manual curation to make the BLAST-identities more meaningful and reliable.

David Kreil has described the principal considerations of probe sequence design, the exploitation of probing multiple target regions and the modelling of probe sequence-specific signals in his study (Kreil *et al.* 2006). Charbonnier used OliCheck to perform selection of probes for the design of whole-genome oligoarrays (Charbonnier *et al.* 2005). These protocols were much more thorough than our simpler procedure. But the very nice distribution of feature intensities found in our experiment (Figure 6.2) suggests that we have generated a high proportion of functioning probes.

Certainly our procedure is much more cost effective than undertaking the further production of cDNAs. For example, 20K cDNAs minimally might take 6 months to generate and sequence, at a cost of 6 x £2400 per month salary, and £4 per EST, giving £14,400 salary and £80,000 sequencing costs. Cost of generating print plates for in-house array fabrication is also not cheap, since we have to PCR-amplify all clones at £1 per clone giving an additional cost of £20,000. Array costs are maybe £20 per array, thus 20 arrays costs £400. So the entire in-house exercise for cDNA production and cDNA array fabrication

costs £114,800. The availability of open source ESTs means that most of these cost is saved. The time taken to do the informatics and design the array is much smaller, say 2 months, comprising £4800 for staff costs, plus oligo and array design costs of £3K, and fabrication costs for 20 slides of 20 x £200 = £4000. Also it might be necessary to use amplification at 20 slides x 2 channels x £30 per amplification = £1200. Total cost is ~£13,100. So the oligoarrays designed by using public ESTs can save much more time and cost than the *de novo* generation of sequence and cDNA probes.

CHAPTER 7: CONCLUSIONS

7.1 Summary of informatic products and their utility

The aim of this Doctoral thesis was to develop a suite of informatics tools and resources and to integrate them into a pipeline in support of a substantial post-genomics investigation of an environmental model species, the common carp. At the time of starting there were few sequence gene resources and DNA sequence data available for this species. Several tools were developed and applied to firstly assemble and analyse EST sequence data, and secondly, to analyse large-scale gene expression profiles in order to interpret the pattern of responses between tissues, between pathways and biological processes, and between individuals exposed to environmental stress or disease-causing organisms. Finally, open source databases were constructed for local and internet access of the resulting sequence and expression data sets.

Table 7.1 lists the bioinformatics tools and resources developed within this project which are schematically connected into a workflow in Figure 7.1 to indicate their working relationships.

1. EST-ferret is the main pipeline that accepts input of FASTA-format and trace-format EST data, including quality scores, and automates the cleaning, assembly, identification and functional annotation of the each sequence file. In doing this it integrates a series of internal and external bioinformatics analysis tools and biological data resources through a scripted package. These tools include ECprofiler and GOprofiler. ECprofiler can produce enzyme information based on the output gene BLAST-identities generated by EST-ferret. Finally, EST-ferret generates the data as flat-files for the construction of carpBASE 2.1 and other EST databases as a MySQL database.
2. GOprofiler was developed to annotate ESTs with GO information (Gene Ontology Consortium 2004) based on the output BLAST-identities (Altschul *et al.* 1997) from EST-ferret. It can also operate as a standalone LINUX programme (Petersen 2002). GOprofiler generates two major tables. One characterises each query gene with GO

annotations whilst the other lists matches of query genes in each of over 500 GO sub-categories. This feature is particularly suitable for characterising gene groups arising from statistically-based microarray profiling experiments; it provides GO annotations for each gene list individually and the output can serve as input for further analysis by GOmatrix.

3. The carpBASE 2.1 and other databases, generated by EST-ferret provides the input data for expression analysis packages such as GeneSpring or the CORR programme in statistical and interpretational analysis of gene expression data. They are also the basic on-line resources for research collaborators worldwide who wish to take access the Liverpool EST dataset.
4. Data from the databases can be submitted to the FindOrthologs PERL programme to construct putative orthologous relationships ('metagenes') between species. Thus, FindOrthologs was used to define 2206 metagenes across carp, zebrafish and human by using data from carpBASE 2.1, together with ESTs for zebrafish and human.
5. Gene expression profiles of carp and human for the metagenes identified above were passed through the ExprAlign pipeline in order to compare the co-expression relationships of pairs of orthologs between species. This has allowed the identification of conserved genetic modules in both carp and humans, each composed of many expression-correlated genes. The ExprAlign approach, originally developed to align the gene expression data for the common carp following exposure to cold, hypoxia and starvation, incorporates an efficient tool, called "CORR", which computes the millions of Pearson correlation coefficients between genes in large scale expression data. The resulting data was modelled using the VxInsight clustering and visualisation package to group, position and visualise gene expression as a 3D landscape.
6. GO information for different non-redundant gene groups identified by array profiling experiments can be extracted by using GOprofiler.

GOmatrix can then be used to discover whether each gene group is significantly enriched (over-represented) or depleted (under-represented) in each of the selected GO sub-categories, by application of Fisher exact test. The properties of each sub-category were thus represented across all statistical comparisons on a coloured 2D matrix.

7. BioCluster, a parallel computer grid containing 40 Linux machines, was developed with the help of my colleagues. This substantially increased the speed of some particularly processor-hungry bioinformatics programmes, such as BLAST, and this saved considerable time in repeated analysis of large-scale data. The "Parallel_BLAST" required by EST-ferret can be processed in the BioCluster. BLAST searching and the CDD protein searching (Marchler-Bauer *et al.* 2005) can be accessed *via* the BioCluster web interfaces http://bioserv2.sbs.liv.ac.uk/~fishomics/ and http://bioserv2.sbs.liv.ac.uk/~fishomics/cdd.html. The user can easily upload query sequences to the BioCluster server and download output from it *via* the Internet.
8. An informatics pipeline was also established to collate the selection of EST data sets for the design of an oligoarray for the rainbow trout. It generated two sets of data: one with ~21,500 sequences submitted to a commercial company for probe design and array fabrication (http://www.ogt.co.uk/), and the other with 11,000 sequences for the submission to the NimbleExpress platform (http://www.affymetrix.com). The pipeline maximised the number of informative sequences and minimised the representation of redundant or un-informative sequences.

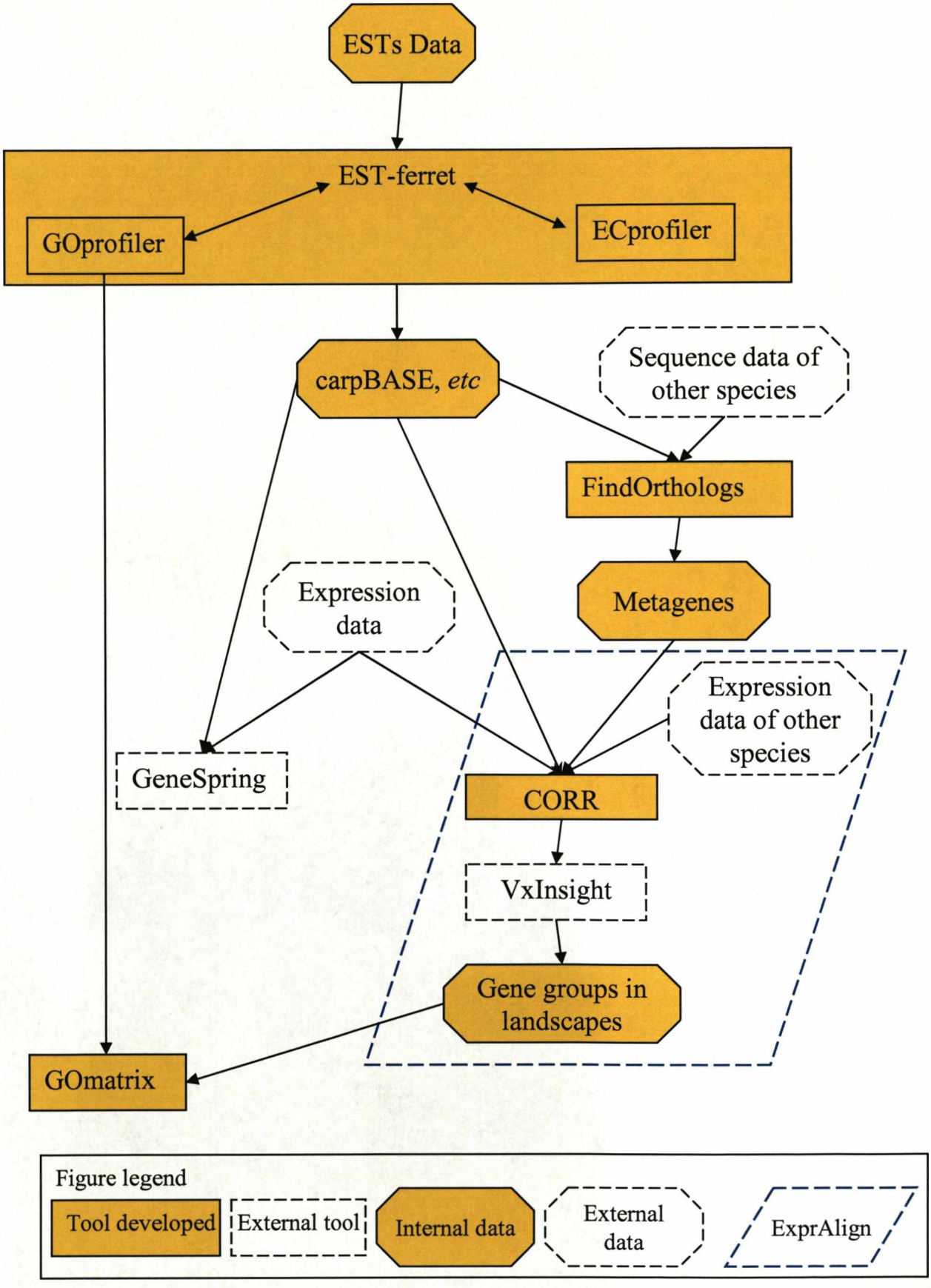


Figure 7.1: Connection of tools and resources developed in this thesis

Table 7.1: Summary for major tools and resources developed in the PhD project

Software or databases	**Use of software or database**	**Platform**
EST-ferret	EST analysis pipeline	PERL & Shell scripts in LINUX
ECprofiler	Assigning EC numbers to sequences	PERL programme in LINUX
GOprofiler	Assigning GO annotations to sequences	PERL & JAVA programme in LINUX
carpBASE and other EST databases	ESTs annotation databases, available at http://legr.liv.ac.uk	Developed in Linux with technologies of PHP, Apache HTTPD server, MySQL
FindOrthologs	Identifying orthologs across 3 species	PERL programme in LINUX
CORR	Used in ExprAlign for computing Pearson Correlation coefficients	C programme in LINUX
GOmatrix	Determining gene groups which are over-/under-represented in particular GO categories.	JAVA programme with a GUI for LINUX and Windows
BioCluster	A parallel computer grid with web interface to access the BLAST programme, *etc.*	Built up under helps of colleagues
Sequence collection of trout oligoarray	A collated sequence collection for rainbow trout oligoarray design.	Established by PERL scripts in LINUX

7.2 Post-genomic analysis for non-model species

7.2.1 ESTs for non-model species

A fundamental requirement for functional genomics research is possession of DNA sequence data. Even though the number of genomic sequences has increased dramatically in the past decade, non-model species still lag well behind the more widely use model species. It is now well established that EST programmes of an appropriate scale can be implemented

even within small laboratories possessing modest budgets (Gracey *et al.* 2001; Gracey *et al.* 2004; Fraser *et al.* 2006). These ESTs can be used to generate both oligo- and cDNA arrays as part of screening experiments to identify genes that respond to experimental treatment (Gracey and Cossins 2003). Efficient bioinformatics tools for ESTs analysis are thus a necessary and critically important requirement for undertaking these post-genomic studies.

The common carp is tolerant of a wide range of temperatures (Johnston and Temple 2002; Gracey *et al.* 2004), environmental hypoxia (Stecyk and Farrell 2002; Stecyk and Farrell 2006) and metabolic adaptations (Nilsson and Renshaw 2004). As a consequence the common carp has become a well-used non-model species for investigating fundamental mechanisms of intrinsic tolerance and acquired resistance to environmental stressors. The work described here is the bioinformatics component of a large study of gene regulatory responses of different carp tissues to a range of environmental stressors. The tools so generated led directly to the identification of a large number of carp genes, some of which, because of the supposed whole genome duplication, displayed unusually large numbers of isoforms, including notably myoglobin (Fraser *et al.* 2006), but also the 9 isoforms of parvalbumin.

7.2.2 EST-ferret and GOprofiler

EST-ferret was constructed at a time when there were few other available assembly pipelines with the appropriate characteristics. It was designed for specific purposes related to the expected problems presented by characterising a duplicated, large eukaryotic genome, and as a result has a number of distinctive features. The EST coding system adopted in EST-ferret helps the user to keep EST datasets in good order and provides the appropriate cDNA library information for submission to dbEST (Boguski *et al.* 1993). It accepts input of FASTA-format sequences and traces files, and not only cleans FASTA-format sequences but also takes regard of the quality files. This feature increases the reliability of subsequent clustering. By contrast, most of other EST pipelines in operation at the time of this work trimmed FASTA-format sequences only within the sequence-cleaning steps.

A second significant feature of EST-ferret in this context is the use of two-stages of CAP3 clustering (Huang and Madan 1999). This groups genes into gene families (represented as main-groups) in the first-round clustering with low stringency and then into unique genes (represented as sub-groups) in second-round clustering with higher stringency. This offers opportunities to explore relationships not only between unique genes, but also between genes and their gene family. This two-stage clustering has proved particularly valuable in the analysis of the common carp genes, since this species is widely believed to have undergone a whole genome duplication to generate unusually large gene families (David *et al.* 2003).

Third, the "Priority" BLAST option in EST-ferret allows searching BLAST (Altschul *et al.* 1997) against a series of defined databases in a pre-defined order. It saves time in BLAST searching and also yields the best quality hits across options. The "Parallel" BLAST feature allows searching of parallel BLAST databases. In this mode, the agreement and the difference of hits across different databases of species allow a better validation of BLAST hit assignments.

GOprofiler, developed to annotate ESTs with GO information (Camon *et al.* 2003; Gene Ontology Consortium 2004), is a standalone LINUX programme but which can be invoked in EST-ferret. It accepts gene lists as input with UniProt (Swiss-Prot or TrEMBL) (Boeckmann *et al.* 2003; Apweiler *et al.* 2004) matches. UniProt matches for query genes serve as a bridge to connect genes to GO annotations. Another benefit of GOprofiler is that it not only furnishes each query gene with GO annotations, but also characterises a list of query genes with GO annotations. Given more lists of genes from gene expression analysis, GOprofiler can provide GO annotations for each list of genes, which can be served as input for further GOmatrix analysis and display. The user can interpret the result across different lists of query genes. The integration of GOprofiler within EST-ferret provides a more powerful means of generating the biological interpretation of gene lists than the straightforward analysis of the GO terms.

Fourth, the additional searching of protein domain or signatures against CDD (Marchler-Bauer *et al.* 2005) and InterPro (Mulder *et al.* 2007) provides additional clues for identifying ESTs which were labelled as un-classifiable during the original BLAST search. This is important for analysing sequences of non-model species, where a sizeable proportion of sequences fail to achieve a significant BLAST hit against sequences. All of the above exclusive functions of EST-ferret maximised the yield of information for the common carp ESTs, and these attributes make it valuable for the analysis of other 'non-model' species.

It is time-consuming to run some bioinformatics programmes, such as BLAST, due to the very large scale of input data running on a single processor. Moreover, these programmes usually need to be run repeatedly using different parameters and criteria in order to optimise the output. In the construction of carpBASE and other databases, BLAST searching was performed repeatedly for query sequences against ~10 different BLAST databases and this took ~2 weeks in a single machine. Therefore to overcome these limitations and to make more convenient the exploration of different projects, it was necessary to develop the BioCluster. The processing of BioCluster was ~30 times faster than a standalone computer with similar hardware for running BLAST.

7.2.3 carpBASE and other databases

EST-ferret efficiently annotated the EST's for a number of non-model species, including common carp, the rainbow trout, the roach, and the ground squirrel. The resulting data was used to construct carpBASE, troutBASE, roachBASE and squirrelBASE, all of which were made available at http://legr.liv.ac.uk. These databases have supported several research projects resulting in a number of landmark publications all based on annotations furnished by this package (Gracey *et al.* 2004; Williams *et al.* 2005; Fraser *et al.* 2006; Gonzalez *et al.* 2007). The databases have also been extensively used by collaborators in Liverpool and elsewhere to search for particular genes of interest, together with the associated sequence and annotation data.

In carpBASE 2.1, 9202 high quality ESTs were processed to generate 6033 gene clusters, of which 53.9% were successfully identified by BLAST searches. The most redundant genes in carpBASE 2.1 were "skeletal alpha-actin", "14 kDa apolipoprotein" and "creatine kinase M3-CK". Results from GO analysis and EC analysis for these ESTs clusters generated more information on biological functions and metabolic pathways of the constituent sequences and indicated that gene diversity was high, given the number of ESTs. carpBASE 2.1 contained particularly large numbers of genes in "transport", "biosynthesis", "nucleic acid metabolism" and "catabolism" sub-categories of biological process, in "intracellular" sub-category of cellular component, and in "ligand binding or carrier" and "enzyme" sub-categories of molecular function. 32.8% of carpBASE 2.1 assembled ESTs were annotated with GO annotations. *S. salar* EST data contained ESTs ~9 times in number greater than ESTs in carpBASE 2.1 but only 26% of them matched sequences in the GO database (Gene Ontology Consortium 2004). The enzymes in carpBASE 2.1 comprised ~6.4% of enzyme entries in the enzyme database (Bairoch 2000) Jan 2007 and covered broad categories of KEGG pathways. The comparison of the coverage of the KEGG pathways for carpBASE 2.1 with that of the far more extensive mouse EST database, we found many pathways the % coverage of carp was similar to that of mouse. For other pathways the mouse database provide 2-6 times more genes, which reflects the fact that there are 20 times more mouse than carp ESTs.

Additional information on the identity and function of cDNA clones came from the analysis of protein domains, UTRs (Pesole *et al.* 2002) and sequence repeats (Jurka 2000) which together provided a further 12.6% annotations of the 6033 clusters. These secondary database annotations significantly increased the biological information for the EST collection as a whole and offered more clues for understanding the identity of the unclassifiable ESTs. carpBASE 2.1 was also adapted to include data generated from the correlation analysis of the expression data correlations giving useful indications of the likely identity of unclassified ESTs.

7.3 Relating Gene Expression Data to Sequence Data

7.3.1 The ExprAlign approach

Whilst EST-ferret maximised the yield of usable sequence information for the common carp ESTs, there were still 33.5% of the gene clusters not having any form of meaningful annotation or identity. ExprAlign (**Expr**ession **Align**ment) was originally developed as a means of suggesting an identity for these ESTs, based on the idea that a microarray probe composed of an unidentified DNA that was a component part of an identified gene should display identical expression properties across all experimental treatments. Conversely, probes that appear to represent identical genes on the basis of BLAST might display divergent properties if they represented different isoforms. Thus sequence alignment could both extend our understanding of sequence identity, and of isoform divergence. This was believed to be particularly true of the common carp which because of the suggested recent whole genome duplication (Larhammar and Risinger 1994; David *et al.* 2003) possesses substantial numbers of duplicated genes for which identification using conventional BLAST homology alignment techniques (Altschul *et al.* 1997) might not be entirely informative.

Using common carp sequence data and an extended gene expression profile, the ExprAlign protocol succeeded in relating 522 unknown ESTs to BLAST-identified ESTs contained within the same landscape feature. These represented 17% of the ESTs on the landscape map. It is worth noting that in order to reduce the presence of false positive probes within a landscape feature or cluster, we used a high cut-off of expression correlation as the criterion of agreement between identified and unidentified probes. Confidence in the fidelity of the gene associations thus generated was increased by two observations, first, by the clear-cut separation of known gene isoforms of fructose-bisphosphate aldolase isoform A and isoform B and discrimination of sub-categories within other gene families (e.g. parvalbumin, apolipoproteins, creatine kinases, *etc*) that BLAST described as having a common identity. The success of the procedure was tested by seeking coherent sequence alignments by manual means, thereby revealing features that were not revealed by

automated BLAST procedures. A good example was the search for additional myoglobin isoforms, with the resulting identification of several previously unidentified ESTs as having that identity, ExprAlign provided critical support for the discovery of a second myoglobin gene which was subsequently confirmed by the discovery of tissue-specific patterns of expression using Northern analysis based on isoform specific probes. At present the common carp is the only vertebrate animal possessing two myoglobin genes, and which expresses both gene in non-muscle tissues, since all previous published descriptions placed myoglobin in oxidative muscle only (Fraser *et al.* 2006). It is very unlikely that the common carp is the only species to display the property of non-muscle expression, and it may be well be that this discovery has widespread general physiological implications of biomedical significance. ExprAlign also suggests the existence of multiple isoforms of apolipoprotein, although resource and time constraints have prevented a detailed follow-up laboratory investigation. It can also provide a useful tool for exploring properties of un-sequenced clones in order to pre-assemble them for subsequent sequencing. This can save time and money, though the cost of sequencing has declined significantly over the period of this project. Finally, in situations where we reinvestigated sequence identity of unclassified clones within a landscape feature we found complete agreement with the sequences of the identified gene, i.e. the myoglobin feature.

It is commonly believed that cDNA microarray probes are unable to differentiate closely related sequences due to the possession of multiple DNA-binding domains over the full length of the cDNA. But this work presented in Chapter 4 clearly demonstrates that this is not true and that different microarray probes from the same sequence assembly cluster may generate quite different patterns of transcript expression. It is worth pointing out that the level of discrimination achieved in searching for divergent expression patterns is greatly increased by including the widest possible range of experimental treatments in the array analysis, each condition presenting a new opportunity for the similarity or divergence to become evident. The carp programme included the exploration of multiple stressors (cold, hypoxia at different

temperatures, starvation and refeeding), and responses of genes were compared across up to 7 different tissues. Breaking up sequence clusters into sub-categories and the discovery of true isoforms indicates that automated sequence alignment procedures alone were entirely unable to estimate the correct number of genes. The outcome of ExprAlign protocol direct the investigator to re-examine specific sequence data more closely, and perhaps to commission a additional sequencing analyses to resolve the issue where the EST lacks sufficient quality or read length.

These results demonstrate that ExprAlign can be a useful alternative approach to sequence identification, and that it complements and extends the more conventional sequence alignments methods. It can separate putative isoforms into differentially responding genes, and it can suggest an identity for unclassified sequences. These benefits of ExprAlign are particularly relevant to studies on non-model species which tend to lack sequence information, at least by comparison with model species.

Finally, the visualisation of these clusters on a 3D landscapes by the ExprAlign package allows the user to zoom in and zoom out over the landscape features (or clusters). This offers a much more intuitive visual interface to understand the relationship between genes and sequences than the non-graphical output of other less tractable statistical packages, where genes or gene clusters are not so conveniently labelled.

7.3.2 Orthologous Genes Relating to Gene Expression

We were interested in exploring the extent to which gene clusters identified in carp responding to environmental stress could be conserved in other vertebrates particularly those with contrasting environmental phenotypes and phylogenetic status. In other words which of the gene relationships discovered in our work are conserved across the vertebrate phylogeny and which are restricted to the common carp and its closest relatives. The cross-species analysis first requires the identification of orthologous gene relationships between species and the construction of so-called metagenes.

The programme FindOrthologs defined 2206 metagenes across carp, zebrafish and human based on a best reciprocal BLAST relationship between pairs of species. 2079 metagenes were identified and covered 63.9% (=2079/3252) identified genes in carpBASE 2.1. Comparing to other available programmes, the programme FindOrthologs performed well on detecting ortholog groups across three species. We chose to compare carp with human because of the rich genomic and array data resources that are freely available. But to maximise the discovery of carp-human orthologs we used a two stage procedure first to map the carp genes onto the nearest model species with well annotated genes lists, and second to compare this list with the human genes. This 'bridging' procedure generated an additional 5.5% of orthologous relationships than when the carp and human were compared directly. The mapping of carp genes onto both zebrafish and human may also prove useful for identifying which carp genes have retained multiple gene copies following from the widely accepted whole genome duplication since.

Orthologous genes are conserved in sequence but are the expression properties of these orthologs conserved between species? Also, are the correlated expression properties of groups of gene in one species conserved with the corresponding gene groups of another distantly related species, given that the correlated profiles may have functional importance in delivering regulated pathways and processes? Specifically, how do these conserved genes and gene groups relate to each other in gene expression properties when comparing a non-model carp and the genomic model species, the human, and can the understanding of gene expression in the former be informed by comparison with the latter? To answer these questions, the ExprAlign approach together with the Monte Carlo simulation was implemented to determine how conserved were the properties of correlated gene pairs in carp and human. The critical calculation of the rank order statistic defines the confidence that the correlated profile of a gene pair in carp is conserved with the corresponding gene pair in human. The probability of getting the observed rank ratios indicates this confidence and the Monte Carlo simulation (Besag and Diggle 1977) was used to compute the *P*. This procedure identified 431 metagenes

whose correlations of gene expressions were conserved between human and carp. These were clustered into 4 groups (B1-4) using the VxInsight 3D landscape package each of which were analysed by GOmatrix and the Reactome (Joshi-Tope *et al.* 2005) to provide an overview of biological processes and reaction pathways that were conserved across the vertebrates. For example, both the GOmatrix and the Reactome indicated the metagenes in the mountain B1 were related predominantly to 'transcription'. In GOmatrix, B2 and B3 were related to 'regulation of biological process' and B4 was related to 'response to stimulus' using mapping within the Reactome package, B2 and B4 were related to 'translation'. Whilst it is not possible to relate these conserved gene association lists to highly specific biological processes and pathways the results do indicate that the combination of the Kim method and ExprAlign is capable of clustering non-redundant genes into groups whose constituent genes undertake functionally related roles. This provides a tractable approach for defining the evolution of conserved and evolving gene modules, and perhaps for relating these to specific features of the environmental phenotypes of species.

7.3.3 GOmatrix

In studies of gene expression, gene expression data is usually clustered to produce different gene lists, each of which represents genes with correlated expression profiles. It is important to understand the underlying biological meanings of the gene lists. Analysis by EST-ferret can provide the identities for genes within each list and GOprofiler places these genes in to the appropriate categories of the GO. But to understand whether the particular GO category is notable within a specific context requires that there is a significant difference between the number of differentially expressed genes in that category and the number expected by chance. GOmatrix, uses Fisher's Exact test for this statistical comparison, providing a *P* value for each GO category being significantly enriched (over-represented) or depleted (under-represented). Undertaking this calculation across all relevant GO categories generates a list of categories which should be included in the interpretation and

drawing of inferences. To aid the identification of patterns in the types of categories the resulting *P* values are presented as a matrix of gene clusters against GO category, with coloured cells according the corresponding P values. This technique offers an objective statistically-based test that directs subsequent hypothesis generation and that replaces the subjective selection of genes for interpretation. The GOmatrix technique has been used in the analysis of carp response to chronic cold (Gracey *et al.* 2004) where it identified a consistent response across all of the major tissues, for example, the transcriptional regulation of both electron transport (in groups 1, 7, and 14) and energy pathways (in groups 6, 12, 15, and 18). The groups came from almost all important tissues indicating a core response to cold stress throughout the body of common carp. GOmatrix has also been used in gene expression analysis for other species being assessed in the LEGR, such as zebrafish, roach, and rainbow trout.

7.4 Concluding Comment

Non-model species lack of genomics data, usually due to lack of resources and the attention of large research communities. As a result getting best value from post-genomic studies of non-model species relies heavily on maximising the yield of information from sequence data by more detailed comparative analysis of sequence resources. This involves the comparison of gene sequence and expression with public data sources for different species (model species and non-model species) using integrated bioinformatics tools. Here we have not only developed tools for analysing sequence and gene expression data separately, but also have connected the two resources to deliver extra biological knowledge.

DNA sequence data is essential to the genomic age of biological research and in the post-genomic stage it is necessary for it to be identified. EST-ferret's exclusive features (Chapter 2) including the ESTs coding system, the trimming of quality files, the two-stage clustering, the "Priority" BLAST and the "Parallel" BLAST, GOprofiler, the protein domain/signature searching make EST-ferret a powerful pipeline for ESTs analysis on non-model species.

carpBASE (Chapter 3) provided useful source of common carp ESTs sequences and information of common carp gene which has supported a series of post-genomics studies on carp.

Gene expression data is another important kind of biological data. Investigations of gene expression profiles by ExprAlign, described in Chapter 4, provided a new approach to understand genes lacking sequence information or with weak sequence alignment to sequence databases. ExprAlign also separated gene isoforms into different clusters demonstrating the power of cDNA array to distinguish sequences. The power of the programme CORR allows researcher to compute large scale of correlations of gene expression in minutes. The combined use of CORR, VxInsight (Davidson 2001), GOprofiler and GOmatrix in ExprAlign provides network visualisations to understand underlying significant biological functions for different gene groups in gene expressions. We have plenty of capability in terms of analysing patterns using clustering, but we are short of tools to help drive to an understanding of the underlying biology. Network visualisation might prove useful for appreciating the complexity of interaction in the cellular system.

In studies of non-model species, it is useful to connect it to model species by establishing orthology relationships. The analysis on orthology and gene expressions between carp and human, described in Chapter 5, combined sequence information and gene expression information to explore the conserved properties of genes which underscores the biological significance of the gene correlations. Using a bridge species is helpful for identifying more orthologs between non-model species and model-species.

Works described in Chapter 6 initiated the oligoarray project for rainbow trout and this project is currently processed by our oligoarray team. This initiative generated two sets of rainbow trout sequences, which maximised the informative sequences and minimised the redundant sequences and the un-informative sequences. The method used in the rainbow trout oligoarray project is also suitable to oligoarray projects for other non-model species and a description of this protocol is now established (Li *et al.* In press; Olohan *et al.* Submission).

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C.A., Gocayne, J.D. and Amanatides, P. G., *et al.* (2000). "The genome sequence of *Drosophila melanogaster*." Science **287**: 2185-2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. and Moreno, R. F., *et al.* (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **21**: 1651-1656.

Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. and Venter, J. C. (1993). "Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library." Nat. Genet. **4**: 373–380.

Aitman, T. J. (2001). "DNA microarrays in medical practice." Br. Med. J. **323**: 611-615.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. **25**: 3389-3402.

Amemiya, C. T., Zhong, T. P., Silverman, G. A., Fishman, M. C. and Zon, L. I. (1999). "Zebrafish YAC, BAC, and PAC genomic libraries." Methods Cell Biol. **60**: 235-58.

Anderson, I. and Brass, A. (1998). "Searching DNA databases for similarities to DNA sequences: when is a match significant." Bioinformatics **14**: 349-356.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." Nucleic Acids Res. **32**(D226-D229).

Antipova, A. A., Tamayo, P. and Golub, T.R. (2002). "A strategy for oligonucleotide microarray probe reduction." Genome Biol. **3**: RESEARCH0073.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Res. **1**: D115-D119.

Arber, W. (1974). "DNA modification and restriction." Prog. Nucleic Acid Res. Mol. Biol. **14**: 1-37.

Armstrong, N. J. and van de Wiel, M. A. (2004). "Microarray data analysis: from hypotheses to conclusions using gene expression data." Cell Oncol. **26**: 279-90.

Arvestad, L., Berglund, A. C., Lagergren, J. and Sennblad, B. (2003). "Bayesian gene/species tree reconciliation and orthology analysis using MCMC." Bioinformatics **19**: i7-i15.

Audic, S. and Claveris, J.-M. (1997). "The significance of digital gene expression profiles." Genome Res. **7**: 986-995.

Ayoubi, P., Jin, X., Leite, S., Liu, X., Martajaja, J., Abduraham, A., Wan, Q., Yan, W., Misawa, E. and Prade, R. A. (2002). "PipeOnline 2.0: automated EST processing and functional data sorting." Nucleic Acids Res. **30**: 4761-4769.

Bairoch, A. (2000). "The ENZYME database in 2000." Nucleic Acids Res. **28**: 304-5.

Ball, C. A., Awad, I. A., Demeter, J., Gollub, J., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Matese, J.C., Nitzberg, M., Wymore, F., Zachariah, Z.K., Brown, P.O. and Sherlock, G. (2005). "The Stanford Microarray Database accommodates additional microarray platforms and data formats." Nucleic Acids Res. **33**: D580-2.

Barone, A. D., Beecher, J. E., Bury, P. A., Chen, C., Doede, T., Fidanza, J. A. and McGall, G. H. (2001). "Photolithographic synthesis of high-density oligonucleotide probe arrays." Nucleosides Nucleotides Nucleic Acids **20**: 525-31.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. and Edgar, R. (2005). "NCBI GEO: mining millions of expression profiles—database and tools." Nucleic Acids Res. **33**: D562–D566.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res. **35**: D760-5.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I. and Schomburg, D. (2007). "BRENDA, AMENDA and FRENDA: the enzyme information system in 2007." Nucleic Acids Res. **35**: D511-4. .

Bashirullah, A., Cooperstock, R. L. and Lipshitz, H. D. (1998). "RNA localization in development." Annu. Rev. Biochem. **67**: 335-94.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J.,

Yeats, C. and Eddy, S. R. (2004). "The Pfam protein families database." Nucleic Acids Res. **32**: D138-141.

Batzoglou, S., Jaffe, D.B. Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. and Lander, E.S. (2002). "ARACHNE: a whole-genome shotgun assembler." Genome Res. **12**: 177-89.

Baxevanis, A. D. and Ouellette, B. F. F. (2005). BIOINFORMATICS A practical guide to the analysis of genes and proteins. New York, A John Wiley & Sons, Inc.

Beelman, C. A. and Parker, R. (1995). "Degradation of mRNA in eukaryotes." Cell **81**: 179-83.

Beisvag, V., Junge, F. K., Bergum, H., Jolsum, L., Lydersen, S., Gunther, C.C., Ramampiaro, H., Langaas, M., Sandvik, A.K. and Laegreid, A. (2006). "GeneTools--application for functional annotation and statistical hypothesis testing." BMC Bioinformatics **7**: 470.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2006). "GenBank." Nucleic Acids Res. **1**: D16-20.

Berger, J. A., Hautaniemi, S., Jarvinen, A. K., Edgren, H, Mitra, S.K. and Astola, J. (2004). "Optimized LOWESS normalization parameter selection for DNA microarray data." BMC Bioinformatics **5**: 194.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P. E. (2000). "The Protein Data Bank." Nucleic Acids Res. **28**: 235-42.

Berriz, G. F., King, O. D., Bryant, B., Sander, C. and Roth, F. P. (2003). "Characterizing gene sets with FuncAssociate." Bioinformatics **19**: 2502-4.

Besag, J. and Diggle, P. (1977). "Simple Monte Carlo tests for spatial pattern." Appl. Statist. **26**: 327-333.

Bevington, P. R. and Robinson, D. K. (1992). Data reduction and error analysis for the physical sciences. Boston, WCB McGraw Hill.

Bilban, M., Buehler, L. K., Head, S., Desoye, G. and Quaranta, V. (2002). "Normalizing DNA microarray data." Curr Issues Mol. Biol. **4**: 57-64.

Boardman, P. E., Sanz-Ezquerro, J., Overton, I. M., Burt, D. W., Bosch, E., Fong, W. T., Tickle, C., Brown, W. R. A., Wilson, S. A. and Hubbard, S. J. (2002). "A Comprehensive Collection of Chicken cDNAs." Current Biology **12**: 1965-1969.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, MC., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout S., and Schneider, M. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res. **31**: 365–370.

Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993). "dbEST--database for "expressed sequence tags"." Nat. Genet. **4**: 332-3.

Branch, M. A., Coleman, T. F. and Li., Y. (1999). "A subspace, interior, and con-jugate gradient method for large-scale bound-constrained minimiza-tion problems." SIAM Journal on Scientific Computing **21**: 1-23.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo. J. and Vingron, M. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat. Genet. **29**: 365-371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P. and Sansone, S. A. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res. **31**: 68-71.

Brooksbank, C., Cameron, G. and Thornton, J. (2005). "The European Bioinformatics Institute's data resources: towards systems biology." Nucleic Acids Res. **33**: D46-53.

C. elegans Sequencing Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**: 2012-2018.

Camon, E., Barrell, D., Brooksbank, C., Magrane, M. and Apweiler, R. (2003). "The Gene Ontology Annotation (GOA) project — application of GO in SWISS-PROT, TrEMBL and InterPro." Comp. Funct. Genomics **4**: 71–74.

Campbell, N. A., Reece, J. B. and Mitchell, L. G. (1999). Biology, Addison-Wesley.

Carninci, P., Shibata, Y. and Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000). "Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes." Genome Res. **10**: 1617-1630.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., Bono, H., Kondo, S., Sugahara, Y., Saito,

R., Osato, N., Fukuda, S., Sato, K., Watahiki, A., Hirozane-Kishikawa, T., Nakamura, M., Shibata, Y., Yasunishi, A., Kikuchi, N., Yoshiki, A., Kusakabe, M., Gustincich, S., Beisel, K., Pavan, W., Aidinis, V., Nakagawara, A., Held, W.A., Iwata, H., Kono, T., Nakauchi, H., Lyons, P., Wells, C., Hume, D.A., Fagiolini, M., Hensch, T.K., Brinkmeier, M., Camper, S., Hirota, J., Mombaerts, P., Muramatsu, M., Okazaki, Y., Kawai, J. and Hayashizaki, Y. (2003). "Targeting a complex transcriptome: The construction of the mouse mull-Length cDNA encyclopedia." Genome Res. **13**: 1273–1289.

Carre, W., Wang, X., Porter, T. E., Nys, Y., Tang, J., Bernberg, E., Morgan, R., Burnside, J., Aggrey, S. E., Simon, J. and Cogburn, L. A. (2006). "Chicken genomics resource: sequencing and annotation of 35,407 ESTs from single and multiple tissue cDNA libraries and CAP3 assembly of a chicken gene index." Physiol Genomics. **25**: 514-24.

Castelli, V., Aury, J. M., Jaillon, O., Wincker, P., Clepet, C., Menard, M., Cruaud, C., Quetier, F., Scarpelli, C., Schachter, V., Temple, G., Caboche, M., Weissenbach, J. and Salanoubat, M. (2004). "Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation." Genome Res. **14**: 406-13.

Charbonnier, Y., Gettler, B., François, P., Bento, M., Renzoni, A., Vaudaux, P., Schlegel, W. and Schrenzel, J. (2005). "A generic approach for the design of whole-genome oligoarrays, validated for genomotyping, deletion mapping and gene expression analysis on *Staphylococcus aureus*." BMC Genomics **6**: 95.

Chou, H. and Holmes, M. H. (2001). "DNA sequence quality trimming and vector removal." Bioinformatics **17**: 1093-1104.

Christian, J., Stoeckert, J. R. and Parkinson, H. (2003). "The MGED ontology: a framework for describing functional genomics experiments." Comp. Funct. Genomics **4**: 127–132.

Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D. and Cherry, J. M. (2004). "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms." Nucleic Acids Res. **32**: D311-4. .

Churchill, G. A. (2004). "Using ANOVA to analyze microarray data." BioTechniques **37**: 173-5.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V.,

Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Birney, E. (2003). "Ensembl 2002: accommodating comparative genomics." Nucleic Acids Res. **31**: 38-42.

Clark, M. S., Edwards, Y. J., Peterson, D., Clifton, S.W., Thompson, A.J., Sasaki, M., Suzuki. Y., Kikuchi. K., Watabe, S., Kawakami, K., Sugano, S., Elgar, G. and Johnson, S. L. (2003). "Fugu ESTs: new resources for transcription analysis and genome annotation." Genome Res. **13**: 2747-53.

Cogburn, L. A., Wang, X., Carre, W., Rejto, L., Aggrey, S. E., Duclos, M. J., Simon, J. and Porter, T. E. (2004). "Functional genomics in chickens: development of integrated-systems microarrays for transcriptional profiling and discovery of regulatory pathways." Comp. Funct. Genomics **5**: 253–261.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ, Lawrence Erlbaum Associates.

Cohen, S. N., Chang, A. C., Boyer, H. W. and Helling, R. B. (1973). "Construction of biologically functional bacterial plasmids *in vitro*." Proc. Natl. Acad. Sci. U.S.A. **70**: 3240-4.

Conesa, A., Nueda, M. J., Ferrer, A. and Talon, M. (2006). "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments." Bioinformatics **22**: 1096-102.

Cossins, A. R., Behan, M., Jones, G. and Bowler, K. (1987). "Lipid-protein interactions in the adaptive regulation of membrane function." Biochem. Soc. Trans. **15**: 77-81.

Cossins, A. R. and Bowler, K. (1987). Temperature Biology of Animals. New York, Chapman and Hall.

Cossins, A. R. and Crawford, D. L. (2005). "Fish as models for environmental genomics." Nat. Rev. Genet. **6**: 324-33.

Couto FM, S. M., Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. (2006). "GOAnnotator: linking protein GO annotations to evidence text." Biomed Discov Collab. **1**: 19.

Crawley, M. J. (2005). Statistics: An introduction using R. New York, John Wiley.

Curtis, D., Lehmann, R. and Zamore, P. D. (1995). "Translational regulation in development." Cell **81**: 171-178.

D'Agostino, N., Aversano, M. and Chiusano, M.L. (2005). "ParPEST: a pipeline for EST data analysis based on parallel computing." BMC Bioinformatics **6**: S9.

David, L., Blum, S., Feldman, M. W., Lavi, U. and Hillel, J. (2003). "Recent duplication of the common carp (*Cyprinus carpio L.*) genome as revealed by analyses of microsatellite loci." Mol. Biol. Evol. **20**: 1425-1434.

Davidson, G. S., Hedrickson, B., Johnson, D. K., Meyers, C. E. and Wyle, B. N. (1998). "Knowledge mining with VxInsight: discovery through interaction." Journal of Intelligent Information Systems **11**: 259–285.

Davidson, G. S., Wylie, B. N., and Boyack, K. W. (2001). "Cluster Stability and the Use of Noise in Interpretation of Clustering." Proceedings of the IEEE Symposium on Information Visualization: 23-30.

Dawson-Saunders, B. and Trapp, R. G. (1994). Basic and clinical biostatistics. Norwalk, Appleton and Lange.

Dear, S. and Staden, R. (1992). "A standard file format for data from DNA sequencing instruments." DNA Seq. **3**: 107–110.

Decker, C. J. and Parker, R. (1994). "Mechanisms of mRNA degradation in eukaryotes." Trends Biochem. Sci. **19**: 336-40.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2003). "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol. **4**: R7.

Drapner, N. R. and Smith, H. (1998). Applied regression analysis. New York, John Wiley and Sons.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." Bioinformatics **21**: 3439-3440.

Edwards, N. J. (2007). "Novel peptide identification from tandem mass spectra using ESTs and sequence database compression." Mol. Syst. Biol. **3**: 102.

Eichner, J. E., Dunn, S. T., Perveen, G., Thompson, D.M., Stewart, K.E. and Stroehla, B. C. (2002). "Apolipoprotein E polymorphism and cardiovascular disease: a HuGE review." Am. J. Epidemiol. **155**: 487-95.

Eisen, J. A. (1998). "Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis." Genome Res. **8**: 163–167.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc. Natl. Acad. Sci. U.S.A. **95**: 14863-14868.

Elsayed, S. and Bennich, H. (1975). "The primary structure of allergen M from cod." Scand. J. Immunol. **4**: 203-208.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res. **30**: 1575-84.

Ewing, B. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res. **8**: 186-194.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res. **8**: 175-185.

Fang, Y., Brass, A., Hoyle, D. C., Hayes, A., Bashein, A., Oliver, S.G., Waddington, D. and Rattray, M. (2003). "A model-based analysis of microarray experimental error and normalisation." Nucleic Acids Res. **31**: e96.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. and Ysebaert, M. (1976). "Complete nucleotide-sequence of bacteriophage MS2-RNA - primary and secondary structure of replicase gene." Nature **260**: 500-507.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst. Zool. **19**: 99-113.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." Trends Genet. **16**: 227-231.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., , Dougherty, B. A. and Merrick, J. M. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." Science **269**: 496-512.

FlyBase Consortium (2003). "The FlyBase database of the Drosophila genome projects and community literature." Nucleic Acids Res. **1**: 172-175.

Fraser, J., Vieira de Mello, L., Ward, D., Rees, H. H., Williams, D. R., Fang, Y., Brass, A., Gracey, A. Y. and Cossins, A. R. (2006). "Hypoxia-inducible myoglobin expression in nonmuscle tissues." Proc. Natl. Acad. Sci. U.S.A. **103**: 2977-81.

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I. and Hardison, R. C. (2003). "Cross-species sequence comparisons: A review of methods and available resources." Genome Res. **13**: 1-12.

Frith, M. C., Bailey, T. L., Kasukawa, T., Mignone, F., Kummerfeld, S. K., Madera, M., Sunkara, S., Furuno, M., Bult, C.J., Quackenbush, J., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Pesole, G. and Mattick, J. S. (2006). "Discrimination of non-protein-coding transcripts from protein-coding mRNA." RNA Biol. **3**: 40-8.

Fruchtermann, T. and Rheingold, E. (1990). Graph drawing by force-directed placement, Technical report UIUCDCS-R-90-1609, Computer Science, Univ. Illinois, Urbana-Champagne, Il.

Gao, X., LeProust, E., Zhang, H., Srivannavit, O., Gulari, E., Yu, P., Nishiguchi, C., Xiang, Q. and Zhou, X. (2001). "A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids." Nucleic Acids Res. **29**: 4744-4750.

Gasch, A. P. and Eisen, M. B. (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol. **3**: RESEARCH0059.

Gene Ontology Consortium (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res. **32**: D258-D261.

Gentleman, R. C., Carey, V., J., , Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol. **5**: R80.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M. and Grouse, L. H., *et al* (2004). "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)." Genome Res. **14**: 2121-7.

Gilchrist, M. J., Zorn, A. M., Voigt, J., Smith, J.C., Papalopulu, N. and Amaya, E. (2004). "Defining a large set of full-length clones from a Xenopus tropicalis EST project." Dev. Biol. **271**: 498-516.

Gogarten, J. P. and Olendzenski, L. (1999). "Orthologs, paralogs and genome comparisons." Curr. Opin. Genet. Dev. **9**: 630–636.

Gonzalez, S. F., Chatziandreou, N., Nielsen, M. E., Li, W., Rogers, J., Taylor, R., Santos, Y. and Cossins, A. (2007). "Cutaneous immune responses in the common carp detected using transcript analysis." Mo.l Immunol. **44**: 1664-79.

Gracey, A. Y. and Cossins, A. R. (2003). "Application of microarray technology in environmental and comparative physiology." Annu. Rev. Physiol. **65**: 231-59.

Gracey, A. Y., Fraser, E. J., Li, W., Fang, Y., Brass, A., Rogers, J. and Cossins, A. R. (2004). "Coping with cold: an integrative, multi-tissue analysis of the transcriptome of a poikilothermic vertebrate." Proc. Natl. Acad. Sci. U.S.A. **101**: 16970–16975.

Gracey, A. Y., Troll, J. V. and Somero, G. N. (2001). "Hypoxia-induced gene expression profiling in the euryoxic fish Gillichthys mirabilis." Proc. Natl. Acad. Sci. U.S.A. **98**: 1993-1998.

Griffiths, A. J. F., Gelbart, W. M., Lewontin, R. C. and Miller, J. H. (2002). Modern genetic analysis integrating genes and genomes, W.H. Freeman and Compary.

Grigoryev, D. N., Ma, S. F., Irizarry, R. A., Ye, S. Q., Quackenbush, J. and Garcia, J. G. (2004). "Orthologous gene-expression profiling in multi-species models: search for candidate genes." Genome Biol. **5**: R34.

Grillo, G., Licciulli, F., Liuni, S., Sbisà, E. and Pesole, G. (2003). "PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences." Nucleic Acids Res. **31**: 3608–3612.

Groff, J. R. and Weinberg, P. N. (1999). SQL: The complete reference, Osborne/McGraw-Hill.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R.K. Jr, Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S,L. and White, O. (2003). "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies." Nucleic Acids Res. **31**: 5654-66.

Haas, B. J., Volfovsky, N., Town, C. D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O. and Salzberg, S. L. (2002). "Full-length messenger RNA sequences greatly improve genome annotation." Genome Biol. **3**: RESEARCH0029.

Haas, S., Vingron, M., Poustka, A. and Wiemann, S. (1998). "Primer design for large scale sequencing." Nucleic Acids Res. **26**: 3006-12.

Haddad, I. A., Ordovas, J. M., Fitzpatrick, T. and Karathanasis, S. K. (1986). "Linkage, evolution, and expression of the rat apolipoprotein A-I, C-III, and A-IV genes." J. Biol. Chem. **261**: 13268-77.

Haft, D. H., Selengut, J. D. and White, O. (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res. **31**: 371-3.

Hancock, D., Wilson, M., Velarde, G., Morrison, N., Hayes, A., Hulme, H., Wood, A. J., Nashar, K., Kell, D. B. and Brass, A. (2005). "maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination." BMC Bioinformatics **6**: 264.

Harris, T. W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., Chen, C.K., Chen, W.J., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E.M., Auken, K.V., Wang, Q., Durbin, R., Spieth, J., Sternberg, P.W. and Stein, L. D. (2004). "WormBase: a multi-species resource for nematode biology and genomics." Nucleic Acids Res. **32**: D411-D417.

Hartigan, J. A. and Wong, M. A. (1979). "A *K*-means clustering algorithm." Applied Statistics **28**: 100-108.

He, X. and Goldwasser, M. H. (2005). "Identifying conserved gene clusters in the presence of homology families." J. Comput. Biol. **12**: 638-656.

Helfman, G. S., Collette, B. B. and Facey, D. E. (1997). The diversity of fishes, Blackwell Sciences, Inc.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A. and Gish, W., *et al.* (1996). "Generation and analysis of 280,000 human expressed sequence tags." Genome Res. **6**: 807–828.

Hirayama, Y., Kanoh, S., Nakaya, M. and Watabe, S. (1997). "The two essential light chains of carp fast skeletal myosin, LC1 and LC3, are encoded by distinct genes and change their molar ratio following temperature acclimation." J. Exp. Biol. **200**: 693–701.

Hoffer, M. J., Hofker, M. H., van Eck, M. M., Havekes, L.M. and Frants, R. R. (1993). "Evolutionary conservation of the mouse apolipoprotein e-c1-c2 gene cluster: structure and genetic variability in inbred mice." Genomics **15**: 62-7.

Hope, A. C. (1968). "A simplified Monte Carlo significance test procedure." J. R. Statist. Soc. B **30**: 582-598.

Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K., Schmidt, E.R. and Suhai, S. (2003). "ESTAnnotator: a tool for high throughput EST annotation." Nucleic Acids Res. **31**: 3716–3719.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R. (2005). "The Orbitrap: a new mass spectrometer." J. Mass Spectrom. **40**: 430-43.

Huang, X. and Madan, A. (1999). "CAP3: A DNA sequence assembly program." Genome Res. **9**: 868-877.

Hubbard, S. J., Grafham, D. V., Beattie, K. J., Overton, I.M., McLaren, S.R., Croning, M.D.R., Boardman, P.E., Bonfield, J.K., Burnside, J., Davies, R.M., Farrell, E.R., Francis, M.D., Griffiths-Jones, S., Humphray, S.J., Hyland, C., Scott, C.E., Tang, H., Taylor, R.G., Tickle, C., Brown ,W.R.A., Birney, E., Rogers, J. and Wilson, S. A. (2005). "Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags." Genome Res. **15**: 174-183.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester B., C. M., Chen Y., Clarke L., Coates G., Cunningham F., Cutts T., Down T., Dyer S.C., Fitzgerald S., Fernandez-Banet J., Graf S., Haider S., Hammond M., Herrero J., Holland R., Howe K., Howe K., Johnson N., Kahari A., Keefe D., Kokocinski F., Kulesha E., Lawson D., Longden I., Melsopp C., Megy K., Meidl P., Ouverdin B., Parker A., Prlic A., Rice S., Rios D., Schuster M., Sealy I., Severin J., Slater G., Smedley D., Spudich G., Trevanion S., Vilella A., Vogel J., White S., Wood M., Cox T., Curwen V., Durbin R., Fernandez-Suarez X.M., Flicek P., Kasprzyk A., Proctor G., Searle S., Smith J., Ureta-Vidal, A. and Birney, E. (2007). "Ensembl 2007." Nucleic Acids Res. **35**: D610-D617.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S.H. and Linsley, P. S. (2001). "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." Nat. Biotechnol. **19**: 342-347.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., , Pagni, M. and Sigrist, C. J. A. (2006). "The PROSITE database." Nucleic Acids Res. **34**: D227-D230.

Huminiecki, L. and Wolfe, K. H. (2004). "Divergence of spatial gene expression profiles following species-specific gene duplicationsin human and mouse." Genome Res. **14**: 1870–1879.

Huson, D. H., Reinert, K., Kravitz, S.A., Remington, K.A., Delcher, A.L., Dew, I.M., Flanigan, M., Halpern, A.L., Lai, Z., Mobarry, C.M., Sutton, G.G. and Myers, E.W. (2001). "Design of a compartmentalized shotgun assembler for the human genome." Bioinformatics. 2001; **17**: S132-9.

Hyde, D. and Cunningham, J. R. C. (1995). "The short and long-term effects of temperature on the dynamic range of signalling in horizontal cells of carp retina." J. Therm. Biol.: 223-229.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome." Nature **409**: 860-921.

International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**: 931-945.

Iyer, V. N. and Szybalski, W. (1963). "A molecular mechanism of mitomycin action: linking of complementary DNA strands." Proc. Natl. Acad. Sci. U.S.A. **50**: 355–362.

Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C. and Lander, E.S. (2003). "Whole-genome sequence assembly for mammalian genomes: Arachne 2." Genome Res. **13**: 91-6.

Jain, A. N., Tokuyasu, T. A., Snijders, A. M., Segraves, R., Albertson, D. G. and Pinkel, D. (2002). "Fully automatic quantification of microarray image data." Genome Res. **12**: 325-32.

Jambu, M. and Lebeaux, M.-O. (1983). Cluster analysis and data analysis, North-Holland Publishing Company.

Johnson, J. M., Edwards, S., Shoemaker, D. and Schadt, E. E. (2005). "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments." Trends Genet. **21**: 93-102.

Johnston, I. A. and Temple, G. K. (2002). "Thermal plasticity of skeletal muscle phenotype in ectothermic vertebrates and its significance for locomotory behaviour." J. Exp. Biol. **205**: 2305-22.

Jolliffe, I. (2002). Principal Component Analysis, Springer.

Jordan, B. (2002). "Historical background and anticipated developments." Ann. N. Y. Acad. Sci. **975**: 24-32.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005). "Reactome: a knowledgebase of biological pathways." Nucleic Acids Res. **33**: D428-32.

Joshi-Tope, G., Vastrik, I., Gopinathrao, G., Matthews, L., Schmidt, E., Gillespie, M., D'Eustachio, P., Jassal, B., Lewis, S., Wu, G., Birney, E. and Stein, L. (2003). "The genome knowledgebase: A resource for biologists and bioinformaticists." Cold Spring Harb. Symp. Quant. Biol. **68**: 237-43.

Jurka, J. (2000). "Repbase update: a database and an electronic journal of repetitive elements." Trends Genet. **16**: 418-20.

Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002). "The KEGG databases at GenomeNet." Nucleic Acids Res. **30**: 42-46.

Kent, W. J. (2002). "BLAT—The BLAST-Like Alignment Tool." Genome Res. **12**: 656–664.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. and Hermjakob, H. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res. **35**: D561-5.

Kikuchi, K., Yamashita, M., Watabe, S. and Aida, K. (1995). "The warm temperature acclimation-related 65-kDa protein, Wap65, in goldfish and its gene expression." J. Biol. Chem. **270**: 17087-92.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. and Davidson, G. S. (2001). "A gene expression map for *Caenorhabditis elegans*." Science **293**: 2087-2092.

Kimura, T., Jindo, T., Narita, T., Naruse, K., Kobayashi, D., Shin-I, T., Kitagawa, T., Sakaguchi, T., Mitani, H., Shima, A., Kohara, Y. and Takeda, H. (2004). "Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics." Mech. Dev. **121**: 915–932.

Klausner, R. D., Rouault, T. A. and Harford, J. B. (1993). "Regulating the fate of mRNA: the control of cellular iron metabolism." Cell **72**: 19-28.

Klug, W. S., Cummings, M. R. and Spencer, C. A., Eds. (2005). Concepts of genetics,.

Knudsen, S. (2004). Guild to analysis of DNA microarray data. Hoboken, John Wiley & Sons, Inc.

Knudsen, S., Workman, C., Sicheritz-Ponten, T. and Friis, C. (2003). "GenePublisher: Automated analysis of DNA microarray data." Nucleic Acids Res. **31**: 3471-3476.

Kohonen, T. (1995). Self-Organizing Maps. Berlin, Springer.

Kondo, H., Morinaga, K., Misaki, R., Nakaya, M. and Watabe, S. (2005). "Characterization of the pufferfish *Takifugu rubripes* apolipoprotein multigene family." Gene **346**: 257-66.

Konno, H., Fukunishi, Y., Shibata, K., Itoh, M., Carninci, P., Sugahara, Y. and Hayashizaki, Y. (2001). "Computer-based methods for the mouse full-length cDNA encyclopedia: Real-time sequence clustering for consortruction of a nonredundant cDNA library." Genome Res. **11**: 281-289.

Koonin, E. V. (2001). "An apology for ortholgs - or brace new memes." Genome Biol. **2**: 1005.1–1005.2.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Rogozin, I.B., Smirnov, S., Sorokin, A.V., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D. A. (2004). "A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes." Genome Biol. **5**: R7.

Koutaniemi, S., Warinowski, T., Karkonen, A., Alatalo, E., Fossdal, C.G., Saranpaa, P., Laakso, T., Fagerstedt, K.V., Simola, L.K., Paulin, L., Rudd, S. and Teeri, T. H. (2007). "Expression profiling of the lignin biosynthetic pathway in Norway spruce using EST sequencing and real-time RT-PCR." Plant Mol. Biol. **65**: 311-28.

Kreil, D., Sykacek, P., Auburn, R., Meadows, L., Russell, R., Fischer, B., Matilla, S., Rana, D., Russell, S. and Micklem, G. (2007). "Microarray probe design, validation and calibration experiments." Genome Biol. **In press**.

Kreil, D. P., Russell, R. R. and Russell, S. (2006). "Microarray oligonucleotide probes." Meth. Enzymol. **410**: 73-98.

Krizman, D. B., Wagner, L., Lash, A., Strausberg, R. L. and Emmert-Buck, M. R. (1999). "The Cancer Genome Anatomy Project: EST sequencing and the genetics of cancer progression." Neoplasia **1**: 101–106.

Kulikova, T., Akhtar, R., Aldebert P., Althorpe N., A. M., Baldwin A., Bates K., Bhattacharyya S., Bower L., Browne P., Castro M., Cochrane G., Duggan K., Eberhardt R., Faruque N., Hoad G., Kanz C., Lee C., Leinonen R., Lin Q., Lombard V., Lopez R., Lorenc D., McWilliam H., Mukherjee G., Nardone F., Garcia-Pastor M.P., Plaister S., Sobhany S., Stoehr P., Vaughan R., Wu D., Zhu W. and Apweiler, R. (2007). "EMBL Nucleotide Sequence Database in 2006." Nucleic Acids Res. **35**: D16-D20.

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., Nagai, Y., Lian, J., Ito, T., Kanamori, M., Matsumaru, H., Maruyama, A., Murakami, H., Hosoyama, A., Mizutani-Ui, Y., Takahashi, N.K., Sawano, T., Inoue, R., Kaito, C., Sekimizu, K., Hirakawa, H., Kuhara, S., Goto, S., Yabuzaki, J., Kanehisa, M., Yamashita, A., Oshima, K., Furuya, K., Yoshino, C., Shiba, T., Hattori, M., Ogasawara, N., Hayashi, H. and Hiramatsu, K. (2001). "Whole genome sequencing of meticillin-resistant Staphylococcus aureus." Lancet. **357**: 1225-40.

Langfeld, K. S., Crockford, T. and Johnston, I. A. (1991). "Temperature acclimation in the common carp: force-velocity characteristics and myosin subunit composition of slow muscle fibres." J. Exp. Biol. **155**: 291-304.

Lederberg, J. and McCray, A. (2001). " 'Ome Sweet 'Omics - A genealogical treasury of words " Scientist **15**: 8.

Lee, J. A. and Cossins, A. R. (1988). "Adaptation of intestinal morphology in the temperature-acclimated carp, *Cyprinus carpio L.*" Cell Tissue Res. **251**: 451-6.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J., Holt, I., Liang, F., and Quackenbush, J. (2002). "Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA)." Genome Res. **12**: 493-502.

Lethimonier, C., Flouriot, G., Valotaire, Y., Kah, O. and Ducouret, B. (2000). "Transcriptional interference between glucocorticoid receptor and estradiol receptor mediates the inhibitory effect of cortisol on fish vitellogenesis." Biol. Reprod. **62**: 1763-71.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004). "SMART 4.0: towards genomic data integration." Nucleic Acids Res. **32**: D142-4.

Levine, D., Ramsey, P. and Smidt, R. (2001). Applied statistics for engineers and scientists, Prentice Hall.

Li, L., Stoeckert CJ, J. and Roos, D. S. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res. **13**: 2178-2189.

Li, W. and Godzik, A. (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**: 1658-9

Li, W., Olohan, L., Williams, D., Hughes, M., Gracey, A. and Cossins, A. (In press). Application of ESTs in microarray analysis. Bioinformatics protocol. Parkinson, J. and Blaxter, M.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). "An optimized protocol for analysis of EST sequences." Nucleic Acids Res. **28**: 3657-3665.

Lindstrom, C. D., Van do, T., Hordvik, I., Endresen, C. and Elsayed, S. (1996). "Cloning of two distinct cDNAs encoding parvalbumin, the major allergen of Atlantic salmon (*Salmo salar*)." Scand. J. Immunol. **44**: 335-344.

Lipman, D. J. and Pearson, W. R. (1985). "Rapid and sensitive protein similarity searches." Science **227**: 1435-41.

Lottaz, C. and Spang, R. (2005). "stam--a Bioconductor compliant R package for structured analysis of microarray data." BMC Bioinformatics **6**: 211.

Lu, F., Jiang, H., Ding, J., Mu, J., Valenzuela, J.G., , Ribeiro, J. M. and Su, X. Z. (2007). "cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome." BMC Genomics **8**: 255.

Luo, C. C., Li, W. H., Moore, M. N. and Chan, L. (1986). "Structure and evolution of the apolipoprotein multigene family." J. Mol. Biol. **187**: 325-40.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res. **1**: D54-8.

Mao, C., Cushman, J. C., May, G. D. and Weller, J. W. (2003). "ESTAP—an automated system for the analysis of EST data." Bioinformatics **19**: 1720–1722.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C.A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song. J. S., Thiessen, P. A., Yamashita, R. A., Yin, J. J., Zhang, D. and Bryant, S. H. (2005). "CDD: a Conserved Domain Database for protein classification." Nucleic Acids Res. **1**: D192-196.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. and Bryant, S. H. (2002). "CDD: a database of conserved domain alignments with links to domain three-dimensional structure." Nucleic Acids Res. **1**: 281-283.

Marsh, J. J. and Lebherz, H. G. (1992). "Fructose-bisphosphate aldolases: an evolutionary history." Trends Biochem. Sci. **17**: 110-3.

Martin, D. M., Berriman, M. and Barton, G. J. (2004). "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes." BMC Bioinformatics **5**: 178.

Martinez, M. J., Roy, S., Archuletta, A. B., Wentzell, P. D., Anna-Arriola, S. S., Rodriguez, A. L., Aragon, A. D., Quinones, G.A., Allen, C. and Werner-Washburne, M. (2004). "Genomic analysis of stationary-phase and exit in *Saccharomyces cerevisiae*: gene expression and identification of novel essential genes." Mol. Biol. Cell **15**: 5295-5305.

Mattick, J. S. (2007). "A new paradigm for developmental biology." J. Exp. Biol. **210**: 1526-47.

Maxam, A. M. and Gilbert, W. (1977). "A new method for sequencing DNA." Proc. Natl. Acad. Sci. U.S.A. **74**: 560-564.

McCarthy, J. E. and Kollmus, H. (1995). "Cytoplasmic mRNA-protein interactions in eukaryotic gene expression." Trends Biochem. Sci. **20**: 191-7.

McLachlan, G. J., Do, K. and Ambroise, C. (2004). Analyzing microarray gene expression data, A John Wiley & Sons, Inc., Publication.

Mendenhall, W. and Sincich, T. (1988). Statistics for the engineering and computer sciences. London, Collier Macmillan Publishers.

Mewe, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F. and Zollner, A. (1997). "Overview of the yeast genome." Nature **387**: 7-8.

MGC Project Team (2004). "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)." Genome Res. **14**: 2121-7.

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T. and Tateno, Y. (2004). "DDBJ in the stream of various biological data." Nucleic Acids Res. **32**: D31-4.

Montgomery, D. C. and Runger, G. C. (2003). Applied statistics and probability for engineers, John Wiley & Sons, Inc.

Moran, J. (2004). Analysis, Evaluation and Redevelopment of Expressed Sequence Tag (EST) pipeline Software. School of Biological Sciences, university of Manchester.

Mouse Genome Sequencing Consortium (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**: 520-562.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P.S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J.D., Sigrist, C.J., Thomas, P.D., Valentin, F., Wilson, D., , Wu, C. H. and Yeats, C. (2007). "New developments in the InterPro database." Nucleic Acids Res. **35**: D224-D228.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova,

M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J.A., Vaughan, R. and Zdobnov, E. M. (2003). "The InterPro database, 2003 brings increased coverage and new features." Nucleic Acids Res. **31**: 315-318.

Nathans, D. and Smith, H. (1975). "Restriction endonucleases in the analysis and restructuring of dna molecules." Annu. Rev. Biochem. **44** 273-93.

Naughton, P. and Schildt, H. (1999). Java TM 2: The complete reference. California, Brand A. Nordin.

Neiman, P. E., Burnside, J., Elsaesser, K., Hwang, H., Clurman, B.E., Kimmel, R. and Delrow, J. (2006). "Analysis of gene expression, copy number and palindrome formation with a Dt40 enriched cDNA microarray." Subcell Biochem. **40**: 245-56.

Nielsen, H. B., Wernersson, R. and Knudsen, S. (2003). "Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays." Nucleic Acids Res. **31**: 3491–3496.

Nilsson, G. E. and Renshaw, G. M. (2004). "Hypoxic survival strategies in two fishes: extreme anoxia tolerance in the North European crucian carp and natural hypoxic preconditioning in a coral-reef shark." J. Exp. Biol. **207**: 3131-9.

Nordberg, E. K. (2005). "YODA: selecting signature oligonucleotides." Bioinformatics **21**: 1365-70.

Ohno, S., J. Muramoto, L. Christian, and N. B. Atkin. (1970). Evolution by gene duplication. New York, NY, Springer Verlag.

Ohno, S., Muramoto, J., Christian, L. and Atkin, N. B. (1967). "Diploid-tetraploid relationship among old world members of the fish family *Cyprinidae*." Chromosoma **23**: 1–9.

Okazaki, Y., Furuno, M. and Kasukawa, T., *et al.* (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." Nature **420**: 563-73.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsuba, K. (1992). "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression." Nat. Genet. **2**: 173–179;.

Olohan, L. A., Li, W., Wolf, T., Jarme, H. O. and Cossins, A. R. (Submission). "Detection of anoxia-responsive genes in cultured cells of the rainbow trout, Oncorhynchus mykiss, using an optimised, genome-wide oligoarray." J. Fish Biol.

Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989). "A common language for physical mapping of the human genome." Science **245**: 1434-5.

Osato, N., Itoh, M., Konno, H., Kondo, S., Shibata, K., Carninci, P., Shiraki, T., Shinagawa, A., Arakawa, T., Kikuchi, S., Sato, K., Kawai, J. and Hayashizaki, Y. (2002). "A computer-based method of selecting clones for a full-length cDNA project: simultaneous collection of negligibly redundant and variant cDNAs." Genome Res. **12**: 1127-34.

Oualline, S. (1991). Practical C programming. Sebastopol, O'Reilly & Associates, Inc.

Palmer, L. E., O'Shaughnessy, A. L., Preston, R. R., Santos, L., Balija, V. S., Nascimento, L. U., Zutavern, T. L., Henthorn, P. S., Hannon, G. J. and McCombie, W. R. (2003). "A survey of canine expressed sequence tags and a display of their annotations through a flexible web-based interface." J. Hered. **94**: 15–22.

Paquola, A. C., Nishyiama. M.Y. Jr, Reis. E.M., da Silva. A.M. and Verjovski-Almeida. S. (2003). "ESTWeb: bioinformatics services for EST sequencing projects." Bioinformatics **19**: 1587-8.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans,U. and Brazma, A. (2007). "ArrayExpress--a public database of microarray experiments and gene expression profiles." Nucleic Acids Res. **35**: D747-50.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A. and Blaxter, M. (2004). "PartiGene--constructing partial genomes." Bioinformatics **20**: 1398-404.

Parkinson, J., Guiliano, D. B. and Blaxter, M. (2002). "Making sense of EST sequences by CLOBBing them." BMC Bioinformatics **3**: 31.

Parkinson, J., Whitton, C., Schmid, R., Thomson, M. and Blaxter, M. (2004). "NEMBASE: a resource for parasitic nematode ESTs." Nucleic Acids Res. **1**: D427-430.

Paschall, J. E., Oleksiak, M. F., VanWye, J. D., Roach, J. L., Whitehead, J. A., Wyckoff, G. J., Kolell, K. J. and Crawford, D. L. (2004). "FunnyBase: a systems level functional annotation of Fundulus ESTs for the analysis of gene expression." BMC Genomics **20**: 96.

Pearson, W. R. (2000). "Flexible sequence similarity searching with the FASTA3 program package." Methods Mol. Biol. **132**: 185-219.

Peng, F. Y., Reid, K. E., Liao, N., Schlosser, J., Lijavetzky, D., Holt, R., Martinez Zapater, J.M., Jones, S., Marra, M., Bohlmann, J. and Lund, S. T. (2007). "Generation of ESTs in Vitis vinifera wine grape (Cabernet Sauvignon) and table grape (Muscat Hamburg) and discovery of new candidate genes with potential roles in berry development." Gene **402**: 40-50.

Perham, R. N. (1990). "The fructose-1,6-bisphosphate aldolases: same reaction, different enzymes." Biochem. Soc. Trans. **18**: 185-7.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C. and Saccone, C. (2002). "UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002." Nucleic Acids Res. **30**: 335–340.

Peteri, A. (2007). Cultured aquatic species information programme - *Cyprinus carpio*. FAO Inland Water Resources and Aquaculture Service (FIRI). c2004- Cultured Aquatic Species Fact Sheets, FAO - Rome.

Petersen, R. (2002). Red Hat Linux: The Complete Reference, McGraw-Hill/Osborne.

Pierce, B. A. (2002). Genetics: a conceptual approach, W.H. Freeman and Company.

Postlethwait, J. H., Yan, Y. L. and Gates, M. A., *et al.* (1998). "Vertebrate genome evolution and the zebrafish gene map." Nat. Genet. **18**: 345–349.

Poulik, M. D. and Smithies, O. (1958). "Comparison and combination of the starch-gel and filter-paper electrophoretic methods applied to human sera: two-dimensional electrophoresis." Biochem. J. **68**: 636-43.

Pratap, R. (2002). Getting started with MATLAB A quick introduction for scientists and engineers. New York, Oxford Univesity Press.

Pruitt, K. D. and Maglott, D. R. (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res. **29**: 137-40.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001). "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species." Nucleic Acids Res. **29**: 159-64.

Rabinow, P. (1996). Making PCR: A story of biotechnology. Chicago, University of Chicago Press.

Rahmann, S. (2002). "Rapid large-scale oligonucleotide selection for microarrays." Proc IEEE Comput Soc Bioinform Conf. **1**: 54-63.

Rasooly, R. S., Henken, D., Freeman, N., Tompkins, L., Badman, D., Briggs, J., Hewitt, A.T. and National Institutes of Health Trans-NIH Zebra.sh Coordinating Committee (2003). "Genetic and genomic tools for zebrafish research: The NIH zebrafish initiative." Dev. Dyn. **228**: 490–496.

Rebeiz, M. and Lewin, H. A. (2000). "COMPASS of 47,747 cattle ESTs." Animal Biotechnology **11**: 75-241.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J. Mol. Biol. **314**: 1041-1052.

Rise, M. L., Schalburg, K. R., Brown, G. D., Mawer, M. A., Devlin, R. H., Kuipers, N., Busby, M., Beetz-Sargent, M., Alberto, R., Gibbs, A.R., Hunt, P., Shukin, R., Zeznik, J.A., Nelson, C., Jones, S.R.M., Smailus, D.E., Jones, S.J.M., Schein, J.E., Marra, M.A., Butterfield, Y.S.N., Stott, J.M., Ng, S.H.S., Davidson, W.S. and Koop, B. F. (2004). "Development and application of a aalmonid EST satabase and cDNA microarray: data mining and interspecific hybridization characteristics." Genome Res. **14**: 478-490.

Roest, C. H. and Weissenbach, J. (2005). "Fish genomics and biology." Genome Res. **15**: 1675-82.

Romualdi, C., Vitulo, N., Del Favero, M. and Lanfranchi, G. (2005). "MIDAW: a web tool for statistical analysis of microarray data." Nucleic Acids Res. **33**: W644-9.

Rouillard, J. M., Zuker, M. and Gulari, E. (2003). "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." Nucleic Acids Res. **31**: 3057-62.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson , A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. and Smith, M. (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**: 687-695.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A. **74**: 5463–5467.

Scheetz, T. E., Trivedi, N., Roberts, C.A., Kucaba, T., Berger, B., Robinson, N.L., Birkett, C.L., Gavin, A.J., O'Leary, B., Braun, T.A., Bonaldo, M.F., Robinson, J.P., Sheffield, V.C., Soares, M.B. and Casavant, T.L. (2003). "ESTprep: preprocessing cDNA sequence reads." Bioinformatics **19**: 1318-24.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**: 467-70.

Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E.M., Richmond, C.S., Blattner, F.R., Lockhart, D.J. and Church, G. M. (2000). "RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array." Nat. Biotechnol. **18**: 1262-8.

Shaw-Lee, R., Lissemore, J. L., Sullivan, D. T. and Tolan, D. R. (1992). "Alternative splicing of fructose 1,6-bisphosphate aldolase transcripts in *Drosophila melanogaster* predicts three isozymes." J. Biol. Chem. **267**: 3959-67.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. and Cherry, J. M. (2001). "The Stanford Microarray Database." Nucleic Acids Res. **29**: 152-155.

Shiokawa, K., Kajita, E., Hara, H., Yatsuki, H. and Hori, K. (2002). "A developmental biological study of aldolase gene expression in Xenopus laevis." Cell Res. **12**: 85-96.

Sigrist, C. J. A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. (2002). "PROSITE: a documented database using patterns and profiles as motif descriptors." Brief Bioinform. **3**: 265-274.

Slatter, J. G., Templeton, I. E., Castle, J. C., Kulkarni, A., Rushmore, T.H., Richards, K., He, Y., Dai, X., Cheng, O.J., Caguyong, M. and Ulrich, R. G. (2006). " Compendium of gene expression profiles comprising a baseline model of the human liver drug metabolism transcriptome." Xenobiotica **36**: 938-62.

Smith, H. and Nathans, D. (1973). "Letter: A suggested nomenclature for bacterial host modification and restriction systems and their enzymes." J. Mol. Biol. **81**: 419-23.

Smith, T. F. a. W., M.S. (1981). "Identification of Common Molecular Subsequences." Journal of Molecular Biology **147**: 195-197.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J. Mol. Biol. **98**: 503-517.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert, C.J. Jr and Brazma, A. (2002). "Design and implementation of microarray gene expression markup language (MAGE-ML)." Genome Biol. **3**: RESEARCH0046.

Stecyk, J. A. and Farrell, A. P. (2002). "Cardiorespiratory responses of the common carp (*Cyprinus carpio*) to severe hypoxia at three acclimation temperatures." J. Exp. Biol. **205**: 759-68.

Stecyk, J. A. and Farrell, A. P. (2006). "Regulation of the cardiorespiratory system of common carp (*Cyprinus carpio*) during severe hypoxia at three seasonal acclimation temperatures." Physiol. Biochem. Zool. **79**: 614-27.

Strausberg, R. L., Feingold, E. A., Klausner, R. D. and F.S., C. (1999). "The Mammalian Gene Collection." Science **286**: 455-457.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**: 249-255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B.L., , Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc. Natl. Acad. Sci. U.S.A. **102**: 15545-50.

Sun, H. W., Hui, C. F. and Wu, J. L. (1998). "Cloning, characterization, and expression in *Escherichia coli* of three creatine kinase muscle isoenzyme cDNAs from carp (*Cyprinus carpio*) striated muscle." J. Biol. Chem. **273**: 33774-80.

Sung, W. K. and Lee, W. H. (2003). "Fast and accurate probe selection algorithm for large genomes." Proc IEEE Comput Soc Bioinform Conf. **2**: 65-74.

Sutton, G. G., White, O., Adams, M. D. and Kerlavage, A. R. (1995). "TIGR Assembler: a new tool for assembling large shotgun sequencing projects." Genome Sci. Technol. **1**: 9-19.

Tagari, M., Tate, J., Swaminathan, G. J., Newman, R., Naim, A., Vranken, W., Kapopoulou, A., Hussain, A., Fillon, J., Henrick, K. and Velankar, S. (2006). "E-MSD: improving data deposition and structure quality." Nucleic Acids Res. **34**: D287-90.

Tang, H., Heeley, T., Morlec, R. and Hubbard, S. J. (2007). "Characterising alternate splicing and tissue specific expression in the chicken from ESTs." Cytogenet Genome Res. **117**: 268-77.

Tarantola, A. (2005). Inverse problem theory and the methods for models parameter estimation. Philadelphia, Society for Industrial and Applied Mathematics.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L, Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D. A. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. and Van de Peer, Y. (2003). "Genome duplication, a trait shared by 22,000 species of ray-finned fish." Genome Res. **13**: 382–390.

Taylor, J. S., Van de Peer, Y. and Meyer, A. (2001). "Genome duplication, divergent resolution and speciation." Trends Genet. **17**: 299-301.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). "The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res. **25**: 4876-4882.

Tiku, P. E., Gracey, A. Y., Macartney, A. I., Beynon, R. J. and Cossins, A. R. (1996). "Cold-induced expression of delta 9-desaturase in carp by transcriptional and posttranslational mechanisms." Science **271**: 815-8.

Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., Cho, J., Parvizi, B., Cheung, F. and Quackenbush, J. (2001). "RESOURCERER: a database for annotating and linking microarray resources within and across species." Genome Biol. **2**: SOFTWARE0002.

Ullman, L. and Liyanage, M. (2005). Visual quickstart guide C programming, Peachpit Press.

Van de Peer, Y., Taylor, J. S. and Meyer, A. (2003). "Are all fishes ancient polyploids?" J. Struct. Funct. Genomics **3**: 65-73.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995). "Serial analysis of gene expression." Science **270**: 484-7.

Venter, J. C., Adams, M. D. and Myers, E. W., *et al.* (2001). "The sequence of the human genome." Science **291**: 1304-51.

Verjovski-Almeida, S., Venancio, T. M., Oliveira, K. C., Almeida, G. T. and Demarco, R. (2007). "Use of a 44k oligoarray to explore the transcriptome of Schistosoma mansoni adult worms." Exp. Parasitol. **In Press**.

Wall, L., Christiansen, T. and Schwartz, R. L. (1996). Programming PERL. Sebastopol, O'Reilly & Associates, Inc.

Wang, G., Gong, Y., Anderson, J., Sun, D., Minuk, G., Roberts, M.S. and Burczynski, F. J. (2005). "Antioxidative function of L-FABP in L-FABP stably transfected Chang liver cells." Hepatology **42**: 871-9.

Wang, Y., Yang, C., Liu, G. and Jiang, J. (2007). "Development of a cDNA microarray to identify gene expression of Puccinellia tenuiflora under saline-alkali stress." Plant Physiol Biochem. **45**: 567-76.

Wei, C. and Brent, M. R. (2006). "Using ESTs to improve the accuracy of de novo gene prediction." BMC Bioinformatics **7**: 327.

Weisstein, E. W. (2006). ""Fisher's Exact Test" from MathWorld--A wolfram web resource." from http://mathworld.wolfram.com/FishersExactTest.html.

Wernersson, R., Schierup, M. H., Jorgensen, F. G., Gorodkin, J., Panitz, F., Staerfeldt, H.H., Christensen, O.F., Mailund, T., Hornshoj, H., Klein, A., Wang, J., Liu, B., Hu, S., Dong, W., Li, W., Wong, G.K., Yu, J., Wang, J., Bendixen, C., Fredholm, M., Brunak, S., Yang, H. and Bolund, L. (2005). "Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing." BMC Genomics **6**: 70.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. (2003). "Database resources of the National Center for Biotechnology." Nucleic Acids Res. **31**: 28–33.

Whitfield, C. W., Band, M. R., Bonaldo, M. F., Kumar, C. G., Liu, L., Pardinas, J. R., Robertson, H.M., Soares, M.B. and Robinson, G. E. (2002). "Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee." Genome Res. **12**: 555-66.

Wilhelm, J. E. and Vale, R. D. (1993). "RNA on the move: the mRNA localization pathway." J. Cell Biol. **123**: 269-74.

Williams, D. R., Epperson, L. E., Li, W., Hughes, M. A., Taylor, R., Rogers, J., Martin, S. L., Cossins, A. R. and Gracey, A. Y. (2005). "Seasonally hibernating phenotype assessed through transcript screening." Physiol. Genomics **24**: 13-22.

Wilson, C. L. and Miller, C. J. (2005). "Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis." Bioinformatics **21**: 3683-3685.

Wolfe, K. H. (2001). "Yesterday's polyploids and the mystery of diploidization." Nat. Rev. Genet. **2**: 333–341.

Wrobel, G., Chalmel, F. and Primig, M. (2005). "goCluster integrates statistical analysis and functional interpretation of microarray expression data." Bioinformatics **21**: 3575-3577.

Xia, J., Radford, C., Guo, X. and Magor, K. E. (2007). "Immune gene discovery by expressed sequence tag analysis of spleen in the duck (*Anas platyrhynchos*)." Dev. Comp. Immunol. **31**: 272-85.

Xie, W., Shao, N., Ma, X., Ling, B., Wei, Y., Ding, Q., Yang, G., Liu, N., Wang, H. and Chen, K. (2006). "Bacterial endotoxin lipopolysaccharide induces up-regulation of glyceraldehyde-3-phosphate dehydrogenase in rat liver and lungs." Life Sci. **79**: 1820-7.

Xu, H., He, L., Zhu, Y., Huang, W., Fang, L., Tao, L., Zhu, Y., Cai, L., Xu, H., Zhang, L., Xu, H. and Zhou, Y. (2003). "EST pipeline system: detailed and automated EST data processing and mining." Genomics Proteomics Bioinformatics **1**: 236-42.

Yamey, G. (2000). "Scientists unveil first draft of human genome." BMJ. **321**: 7.

Yang, L. and Gui, J. F. (2004). "Positive selection on multiple antique allelic lineages of transferrin in the polyploid Carassius auratus." Mol. Biol. Evol. **21**: 1264-77.

Yang, L., Zhou, L. and Gui, J. F. (2004). "Molecular basis of transferrin polymorphism in goldfish (*Carassius auratus*)." Genetica **121**: 303-13.

Yang, Y. H., Buckley, M. J. and Speed, T. P. (2001). "Analysis of cDNA microarray images." Brief Bioinform. **2**: 341-9.

Yusufi, F. N. K. (2004). Statistical analysis and data mining on microarray data sets. Computer Science Department. Manchester, Manchester University: 23-24.

Zar, J. H. (1996). Biostatistical analysis, Prentice-Hall.

Zdobnov, E. M. and Apweiler, R. (2001). "InterProScan - an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**: 847-8.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K,J., Riss, J., Barrett, J.C. and Weinstein, J. N. (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol. **4**: R28.

Zhong, S., Storch, K. F., Lipan, O., Kao, M. C., Weitz, C. J. and Wong, W. H. (2004). "GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space." Appl. Bioinformatics **3**: 261-4.

Zhou, L., Wang, Y., Yao, B., Li, C. J., Ji, G. D. and Gui, J. F. (2005). "Molecular cloning and expression pattern of 14 kDa apolipoprotein in orange-spotted grouper, *Epinephelus coioides*." Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **142**: 432-7.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. and Siebert, P. D. (2001). "Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction." BioTechniques **30**: 892-897.

APPENDICES

Appendix 2.1: A sample of Bad Repeat ESTs

```
>A sample of bad repeat ESTs
ATGCGCNNGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGACG
```

Appendix 3.1: GO annotation tables

a: Biological process mapping for sub-groups

Categories and sub-categories	Representation (Sub-groups)	% Representation of total
Biological process	1595	26.43%
behavior	11	0.18%
biological process unknown	31	0.51%
death	40	0.66%
developmental processes	146	2.42%
physiological processes	47	0.78%
cell communication	313	5.18%
cell adhesion	56	0.93%
cell recognition	4	0.07%
cell-cell signalling	20	0.33%
response to external stimulus	120	1.99%
signal transduction	152	2.52%
cell growth and-or maintenance	1391	23.06%
autophagy	1	0.02%
cell cycle	63	1.04%
cell growth	12	0.20%
cell motility	43	0.71%
cell organization and biogenesis	69	1.14%
cell proliferation	24	0.40%
cell-cell fusion	3	0.50%
homeostasis	19	0.31%
stress response	68	1.13%
transport	302	5.01%
metabolism	1049	17.38%
alcohol metabolism	8	0.13%
aldehyde metabolism	1	0.02%
amine metabolism	12	0.20%
amino acid and derivative metabolism	52	0.86%
aromatic compound metabolism	2	0.03%
biosynthesis	243	4.03%
carbohydrate metabolism	118	1.96%
catabolism	225	3.73%
coenzymes and prosthetic group metabolism	30	0.50%
electron transport	89	1.48%
energy pathways	103	1.71%
lipid metabolism	121	2.01%
neurotransmitter metabolism	1	0.02%
nitrogen metabolism	6	0.10%
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	240	3.98%
one-carbon compound metabolism	10	0.17%

organic acid metabolism	2	0.03%
oxygen and reactive oxygen species metabolism	17	0.28%
protein metabolism	13	0.22%
regulation of metabolism	9	0.15%
secondary metabolism	4	0.07%
sulfur metabolism	1	0.02%
toxin metabolism	1	0.02%
vitamin metabolism	10	0.17%
xenobiotic metabolism	18	0.30%

b: Cellular component mapping for sub-groups

Categories and sub-categories	Representation (Sub-groups)	% Representation of total
Cellular component	1363	22.59%
unlocalized	15	0.25%
cell	1230	20.39%
cell fraction	69	1.14%
Dendrite	3	0.05%
Intracellular	1044	17.30%
Membrane	452	7.49%
site of polarized growth	1	0.02%
extracellular	120	1.99%
extracellular matrix	22	0.36%
extracellular space	21	0.35%
Fibrinogen	3	0.50%
Virion	1	0.17%
external protective structure	1	0.17%
cell envelope	1	0.17%
Glycocalyx	1	0.17%

c: Molecular function mapping for sub-groups

Categories and sub-categories	Representation (Sub-groups)	% Representation of total
Molecular function	1775	29.40%
antioxidant	7	0.12%
cell adhesion molecule	1	0.02%
chaperone	56	0.93%
defense or immunity protein	22	0.36%
enzyme	789	13.08%
enzyme regulator	87	1.44%
ligand binding or carrier	783	12.98%
molecular function unknown	61	1.01%
motor	20	0.33%
signal transducer	157	2.60%
structural molecule	149	2.47%
transcription regulator	82	1.36%
transporter	260	4.31%

Appendix 4.1: Expression alignments for 23 interesting *K*-means groups (containing 1728 cDNAs in tissues of cooled fish).

The numbers in the left-hand side are the group numbers and the top indicate the tissue types.