THE DEVELOPMENT AND APPLICATION OF INFORMATICS-BASED SYSTEMS FOR THE ANALYSIS OF THE HUMAN TRANSCRIPTOME

JANET KELSO

Thesis presented in fulfilment of the requirements for the Degree of *Doctor Philosophiae* at the South African National Bioinformatics Institute, Department of Biochemistry, Faculty of Natural Sciences, University of the Western Cape

April 2003

Advisor: Prof. Winston Hide

Abstract

Despite the fact that the sequence of the human genome is now complete it has become clear that the elucidation of the transcriptome is more complicated than previously expected. There is mounting evidence for unexpected and previously underestimated phenomena such as alternative splicing in the transcriptome. As a result, the identification of novel transcripts arising from the genome continues. Furthermore, as the volume of transcript data grows it is becoming increasingly difficult to integrate expression information which is from different sources, is stored in disparate locations, and is described using differing terminologies. Determining the function of translated transcripts also remains a complex task. Information about the expression profile – the location and timing of transcript expression – provides evidence that can be used in understanding the role of the expressed transcript in the organ or tissue under study, or in developmental pathways or disease phenotype observed.

In this dissertation I present novel computational approaches with direct biological applications to two distinct but increasingly important areas of research in gene expression research. The first addresses detection and characterisation of alternatively spliced transcripts. The second is the construction of an hierarchical controlled vocabulary for gene expression data and the annotation of expression libraries with controlled terms from the hierarchies. In the final chapter the biological questions that can be approached, and the discoveries that can be made using these systems are illustrated with a view to demonstrating how the application of informatics can both enable and accelerate biological insight into the human transcriptome.

Declaration

I declare that *"The Development and Application of Informatics-based Systems for the Analysis of the Human Transcriptome"* is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Janet Kelso April 2003

Acknowledgements

First and foremost I would like to thank my advisor, Professor Winston Hide. His experienced opinion and sound advice have guided the course of this research to its successful conclusion, and with his support I have already been able to publish parts of this research. He has also provided me with innumerable opportunities to collaborate widely and to present my research at respected international conferences. It has been a privilege to work under his supervision.

I am grateful to my colleagues at SANBI and Electric Genetics – both past and present – who have assisted me with insightful comments, advice, technical support and helpful discussions during the course of this work. Vladimir Babenko, Rüdiger Braüning, Alan Christoffels, Junaid Gamieldien, Raphael Isokpehi, Andrey Ptitsyn and Cathal Seoighe have all been generous with their time and have contributed immeasurably to the completion of this thesis. I also owe thanks to Irvine Short for maintaining the computing facilities without which this work would not have been possible.

Without data and interesting problems there can be no bioinformatics research. Thanks therefore to my collaborators who have provided me with a never-ending stream of excellent data, and plenty of interesting problems – particularly to Andrew Simpson, Anamaria Camargo Aranha, Helena Samaia Brentani and Otávia Caballero of the Ludwig Institute for Cancer Research, Gregory Theiler, Dmitry Kuznetsov and Victor Jongeneel of the Office of Information Technology of the Ludwig Institute for Cancer Research / Swiss Institute of Bioinformatics, Minoru Ko, George Kargul, Yong Qian, Dawood Dudekula and Mark Carter in the Laboratory of Genetics of the

National Institute of Ageing at the NIH, and Mark McCarthy and Damian Smedley of Imperial College London .

Without the support and encouragement of my family I would never have started down this path. A huge thank you to my parents for providing me with every opportunity to pursue my interests, for their many sacrifices, enduring encouragement, and their unquestioning faith in my abilities.

Last but not least, I thank my husband and best friend, Johann for providing unconditional support, technical assistance, an excellent sounding board for my ideas, and world-class hot chocolate. Above all, for showing me that there is more to life than a PhD.

Publications arising from this thesis

Brentani, H. Caballero, O., Camargo, A., da Silva, A., Araújo da Silva, W., Dias Neto, E., Grivet, M., Gruber, A., Edson Moreira Guimaraes, P., Hide, W., Iseli, C., Jongeneel, C.V., Kelso, J., Nagai, M.A., Ojopi, E., Osorio, E., Reis, E., Riggins, G., Simpson, A., de Souza, S., Stevenson, B., Strausberg, R.L., Tajara, E., Verjovski-Almeida, S. **The Generation and Utilization of a Cancer Oriented Representation of the Human Transcriptome, Using Expressed Sequence Tags.** *Accepted: Proceedings of the National Academy of Sciences.*

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C.V., McCarthy, M.I., Hide, T., Hide, W. **eVOC: A Controlled Vocabulary for Unifying Gene Expression Data**. *Genome Research* 2003. 12:1222-1230.

Butcher, S., Kelso, J., Littlejohn, T. and Rubin, E. **Training and Support for Bioinformatics – theoretical and practical aspects.** Tutorial presented at *Intelligent Systems for Molecular Biology* 2003.

Vincent VanBuren, Yulan Piao, Dawood B. Dudekula, Yong Qian, Mark G. Carter, Patrick Martin, Carole A. Stagg, Uwem Bassey, Kazuhiro Aiba, Toshio Hamatani, George J. Kargul, Amber G. Luo, Janet Kelso, Winston Hide and Minoru S. H. Ko. **Assembly, verification, and initial annotation of NIA 7.4K mouse cDNA clone set.** *Genome Research* 2002. 12(12) 1999-2003.

Nembaware, V., Crum, K., Kelso, J and Seoighe, C. **Impact of the Presence of Paralogs on Sequence Divergence in a Set of Mouse-Human Orthologs.** *Genome Research* 2002. 12:1370-1376

Winston A.Hide, Vladimir N. Babenko, Peter A. van Heusden, Cathal Seoighe, and Janet F. Kelso. **The contribution of exon-skipping events on Chromosome 22 to Protein Coding Diversity.** *Genome Research* 2001. 11:1848-1853

Win Hide, Valerie Mizrahi, B. Venkatesh, Sydney Brenner, Andrew Simpson, Greg Blatch, Himla Soodyall, Katherine Denby, Mike Wingfield, Brenda Wingfield, Paul van Helden, Raj Ramesar, Rosemary Dorrington, Janet Kelso, Ekow Oppon, Elizabeth Goyvaerts, Michele Ramsay, Etienne de Villiers, Carel van Heerden, Basil Allsopp and Cathal Seoighe. **The National Genome Initiative**. *S.A Medical Journal.* 2001. 91(12): 1006-1007

Publications pending

Hide, W. and Kelso, J. (2003). **Application of eVOC: controlled vocabularies for unifying gene expression data.** *Under review: Proceedings of the French National Academy of Sciences.*

Imanishi, T. *et al.* (2003). **Systematic identification of human full-length cDNAs and standardized annotation of human protein functions.** *Under review: Nature.*

Kelso, J and Hide, W.A. **The Transcriptome** in *An Introduction to Bioinformatics*. (2003). Attwood T.A. and Jongeneel C.V. (eds) John Wiley & Sons Ltd. Chichester, UK. *Invited chapter submission.*

Seoighe, C. and Kelso, J. (2003) **Nucleotide polymorphisms alter transcript processing in a significant number of human genes.** *In preparation for submission to Proceedings of the National Academy of Sciences.*

Sharov, A., Piao, Y., Matoba, R., Dudekula, D., Qian, Y., VanBuren, V., Falco, G., Martin, P., Stagg, C., Bassey, U., Wang, Y., Carter, M., Hamatani, T., Aiba, K., Akutsu , H., Sharova, L., Tanaka, T., Kimber, W., Yoshikawa, T., Jaradat, S., Pantano, S., Nagajara, R., Boheler, K., Taub, D., Hodes, R., Longo, D., Schlessinger, D., Keller, J., Klotz, E., Kelsoe, G., Umezawa, A., Vescovi, A., Rossant, J., Kunath, T., Hogan, B., Curci, A., D'Urso, M., Kelso, J., Hide, W., Ko, M. (2003). **Transcriptome analysis of mouse stem cells and early embryos**. *Submitted: Public Library of Science: Biology.*

Smedley, D., Kelso, J., Hüsler, P., Visagie, J., Hide, W., McCarthy, M. (2003). **The Candidate Gene Profiler – a tool for identifying and prioritising disease gene candidates.** *In preparation for submission to Nucleic Acids Research.*

Tamura, T., Hishiki, T., Ogasawara, O., Murakawa, K., Kawamoto, S., Watanabe, S., de Souza, S., Hide, W., Kelso, J., Auffray, C., Imbeaud, S., Sudo, Y., Orikasa, A., Takagi, T., Gojobori, T., and Okubo, K. (2003). **An exhaustive overview of the human transcriptome.** *In preparation for submission to Nature.*

Table of Contents

List of Figures

Chapter 1

Chapter 3

Chapter 4

List of Tables

Chapter 2

Chapter 3

Chapter 4

Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	base pairs
CDS	Coding sequence
CGAP	Cancer Genome Anatomy Project
ePCR	Electronic Polymerase Chain Reaction
EST	Expressed Sequence Tag
GO	Gene Ontology
HGI	Human Gene Index
H-Inv	Human Full-Length cDNA Invitational
HUGE	Human Unidentified Gene-Encoded Large Proteins
Mb	Megabases
MGC	Mammalian Gene Collection
MGED	Microarray Gene Expression Database
MIAME	Minimum Information About Microarray Experiments
MPSS	Massively Parallel Signature Sequencing
ORESTES	Open Reading Frame ESTs
ORF	Open Reading Frame
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
STACK	Sequence Tag and Alignment Consensus Knowledgebase
TC	Tentative Consensus Sequence
THC	Tentative Human Consensus Sequence
TIGR	The Institute for Genomic Research
TOGA	TIGR Orthologous Gene Alignment database
UTR	Untranslated region

Preface

The production of complete genomic sequence for various organisms has accelerated rapidly in recent years. In large part these sequencing efforts have been driven by the need to identify and characterise their complete gene complement. While genomes contain a number of sequence elements, it is the expressed component which promises to provide insight into organism function, development and disease. While the genome is the total DNA complement of an organism, the transcriptome is that part of the genome which is transcribed into mRNA – the expressed genome. The translated products of mRNA give rise to the proteome. The relationship between proteome and transcriptome is poorly understood, and will only become clearer as more proteomic data become available.

Genome sequencing presents one route to gene discovery. In the case of prokaryotes this has proved relatively successful as genes are in close proximity (1 gene per kb), and are uninterrupted by introns. However, gene identification is significantly more complex in eukaryotes in which gene density is low (1 gene per 100kb), and where the coding sequence is interrupted by introns. In mammals between 1% and 2% of the genome is thought to be made up of coding sequence (Lander et al., 2001; Venter et al., 2001), and this low gene density makes gene finding relatively complex. The entire genome is present in each nucleated cell of the organism but only a fraction of the genes are expressed in each cell type, with significant variation in the number of transcripts of each gene present. Unique spatio-temporal patterns of gene expression allow the genome to provide the complexity required for life.

Prior to the advent of genome sequencing, the importance of characterising the transcriptome led to the development of various approaches to determine the identity, sequence, expression levels and expression patterns of genes. As enabling technologies have developed and improved there has been a progression from low throughput to high throughput data production and interpretation. Large-scale gene expression analysis provides a global view of gene function through the identification and quantification of gene expression products. In addition to its role in the identification and functional classification of gene products the large-scale investigation of gene expression is providing insight into the process of development, physiological response and disease in a way which is not possible using a gene-by-gene approach. The potential for significant and useful discoveries means that gene expression studies are at the very forefront of genomics research.

Advances in the technologies for monitoring gene expression and the large-scale production of gene expression data has resulted in significant informatics challenges, including those of data tracking, capture, analysis, visualisation, integration, mining and storage.

An overview of the sequence-based approaches which are commonly used for characterising and quantifying gene expression data, including methods of generation, the relevant databases used for storage, and the computational approaches to mining of gene expression data are discussed in chapter 1.

In chapters 2 and 3 I present two novel informatics-based approaches to addressing some of the challenges that continue to face those interested in large-scale gene expression analysis.

Chapter 2 addresses the detection and characterisation of alternative splicing, The recent recognition that alternative splicing may contribute more significantly to the diversity of the expressed gene complement than previously estimated led to the development of a novel computational approach to the detection of exon skipping, the most common form of alternative splicing. Chapter 2 describes this approach and its application to the first published human genome sequence – that of chromosome 22.

Closely tied to the quantification and characterisation of transcripts discussed in Chapters 1 and 2 is the ability to make biologically important functional inferences about identified transcripts based on the location and timing of their expression. For example, identifying differences in the location and/or timing of expression of alternatively spliced transcripts from the same gene is likely to be of biological or pharmaceutical relevance. Chapter 3 presents the development and application of controlled vocabularies for describing the biological source of materials used in gene expression experiments. These controlled vocabularies define a common terminology for sharing information about the gene expression knowledge domain, and define the relevant concepts and relationships between these concepts. The implementation of controlled vocabularies enables both humans and machines to share and reuse the domain knowledge which has been captured through the input of specialist curators. They also promote the rapid and accurate mining of the transcript databases. In summary, Chapter 4 demonstrates how the approaches taken in Chapters 2 and 3 can be applied to the preliminary identification of differentially expressed alternative spliceforms. Results obtained for an analysis of the differential expression of alternative transcripts in cancer and normal tissues are presented.

Chapter 1

Characterising and Quantifying Gene Expression

Numerous technologies are used to investigate gene expression. These techniques have varying applications, depending on whether they are high or low throughput, and on whether they provide gene-identification or gene-expression-level information. In this section, we discuss the strengths and weaknesses of these techniques, with particular focus on the informatics required to perform high-throughput analysis.

The primary objectives of gene expression analysis are two-fold:

- To identify the expressed gene complement (transcript characterisation);
- To quantify the expression level of these transcripts (transcript quantification).

Experimental methods for characterisation and quantification of the transcriptome can be broadly divided into sequence- and hybridisation-based techniques. Sequence-based methods include those that identify transcripts expressed in a given state (*e.g.*, generating ESTs, full-length mRNAs), and those that quantify the level of expression of the transcripts (*e.g.*, SAGE). Hybridisation-based methods may be low-throughput (*e.g.*, *in situ* hybridisation, Northern blots), or high-throughput (*e.g.*, cDNA/oligonucleotide arrays), but are limited to the analysis of previously identified transcripts.

Overview: Sequence-based approaches to gene identification and expression level quantification

Sequence-based approaches achieve the identification and/or quantification of expressed genes by sequencing either tag-level representations of transcripts (transcript fragments), or entire transcripts from starting materials of interest. The advent of high-throughput sequencing technologies has accelerated the generation of transcript sequence data, providing large amounts of transcript data for expression mining.

Gene expression is dynamic – differing between cells, tissues, developmental stages, physiological responses and disease states. While capturing the gene expression profiles in every possible state is not feasible using sequence-based methods, this approach can provide a "snapshot" of the gene expression in the substrate of interest, and as such provide valuable insight into the transcriptome of the substrate.

Sequence-based methods include techniques for both transcript identification and transcript quantification.

Techniques used in transcript identification:

- Expressed Sequence Tag (EST) sequencing
- Full-length mRNA sequencing

Gene identification via complete or partial transcript capture has been a significant contributor to the early understanding of the transcribed eukaryotic genome (Boguski and Schuler, 1995; Schuler et al., 1996). EST sequencing is often used in pilot gene identification projects as it can identify the more commonly expressed genes in a

system. However, only an exhaustive comparison between a completed genome and a very high coverage transcriptome can provide insight into the full complement of genes within a genome (Saha et al., 2002). The identification of transcripts using a sequencing approach is complicated by the wide variation in transcript abundance. Even in normalised cDNA libraries highly abundant transcripts tend to obscure rare transcripts making them difficult to capture and sequence. The time and expense involved in sequencing sufficiently large numbers of ESTs and / or mRNAs makes these approaches generally unsuitable for transcript quantification.

Techniques used in transcript quantification:

- Serial analysis of gene expression (SAGE)
- Massively Parallel Signature Sequencing (MPSS)

The quantification of expressed transcripts can be achieved through the sequencing and enumeration of short sequence tags that are uniquely associated with a gene.

While the results of the SAGE and MPSS approaches are similar, the method and depth of analysis differ significantly. The SAGE method generates 14 base pair tags that are concatenated and sequenced using a traditional DNA sequencing approach. Approximately 50 000 tags are generally produced per SAGE library largely due to cost and convenience constraints. (Velculescu et al., 1995; Velculescu et al., 2000). Using the more recent MPSS technology sequence tags of 20bp in length are cloned onto microbeads and sequenced in parallel to yield a measure of transcript abundance (Brenner et al., 2000). Using the MPSS approach in excess of 1 000 000 tags can be produced simultaneously. As a result of this increased tag production, and in contrast

to the traditional SAGE approach, MPSS allows those genes expressed at levels of lower than 20 copies per cell to be detected and accurately quantified.

1.1 Transcript identification using ESTs

Expressed sequence tags (ESTs) are partial fragments of expressed genes generated by single-pass sequencing from the 5' and 3' ends of a cDNA clone (Wolfsberg and Landsman, 1997) (Boguski et al., 1993; Lennon et al., 1996) (Figure 1). The large-scale sequencing of cDNA clones has proved to be a rapid and valuable route to gene discovery . Many groups are contributing thousands of ESTs representing numerous organisms and expression states to public databases. The data deposited in EST databases is generally unorganised, sparsely annotated, redundant and of poor quality. Using various approaches EST data can be organised and mined in such as way as to produce valuable information about gene expression.

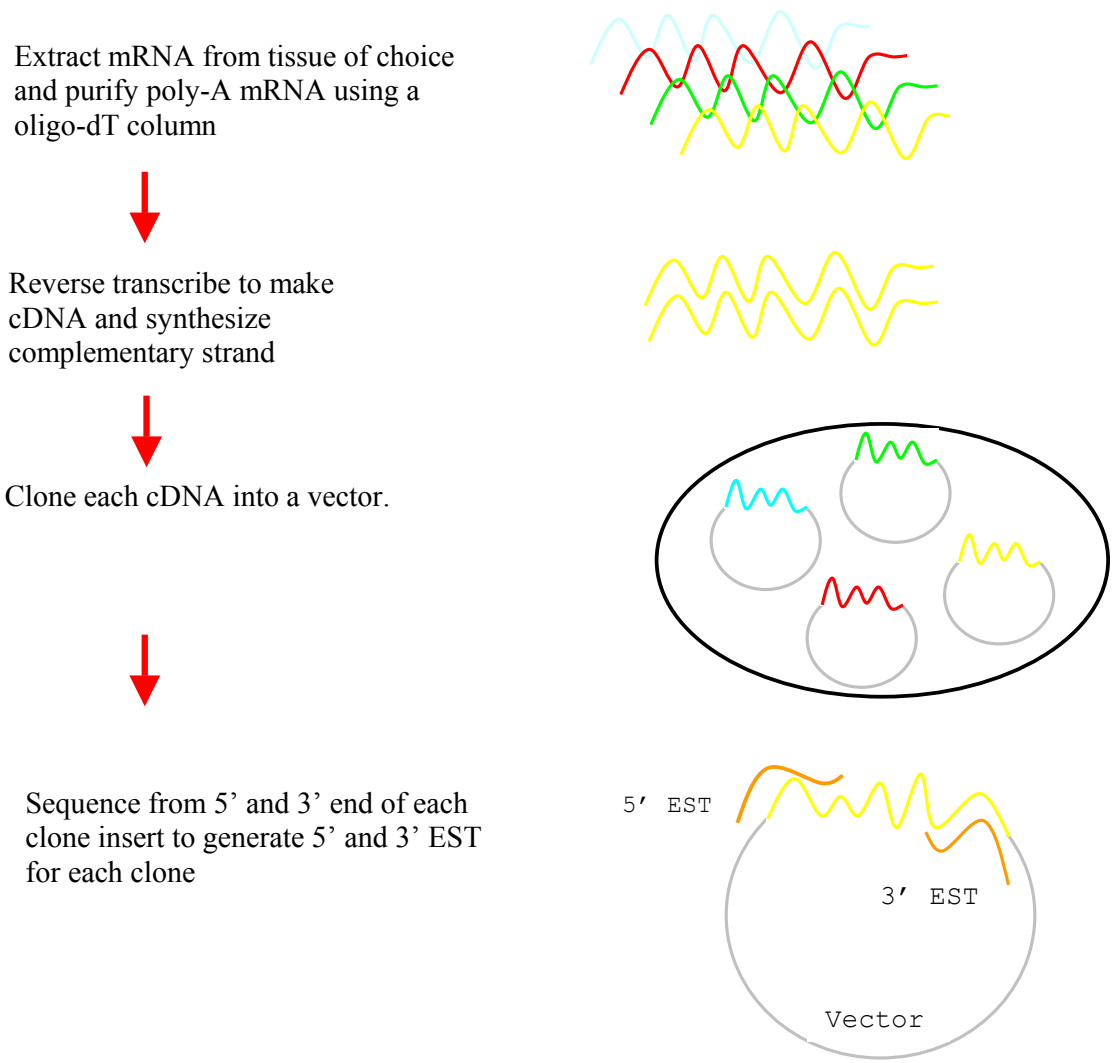


Figure 1. Construction of a cDNA library and EST production. mRNA is isolated and converted to double stranded cDNA which is ligated into a vector and cloned. Techniques vary and the full-length cDNA may, or may not, be ligated into the vector. Sequencing of the clone insert from either the 3'-end or 5'end using standard primers results in the production of 3' and 5' ESTs which may or may not overlap, depending on the lengths of the sequences.

1.1.1 cDNA library construction and EST sequencing

The production of ESTs requires the construction of a cDNA library that is representative of the transcriptome of the tissue or cell type of interest. Many of the libraries used for the generation of ESTs randomly sample the transcriptome of the material from which they were generated. In these libraries the relative abundance of clones derived from a unique mRNA is representative of the expression levels of that mRNA in the starting material. The advantage of these “non-normalised” libraries is that transcript quantification and comparison is possible. The disadvantage is that clones containing more rare transcripts are likely to be poorly represented or completely absent from the library. The rapid and cost-effective identification of novel sequences depends to a large extent on the cDNA libraries that are used. In a typical cell a small number of unique mRNAs make up more than 50-65% of the total mRNA mass. Random sequencing efforts aimed at novel gene identification are therefore confounded by the identification of redundant copies of genes of the prevalent and intermediate classes. Bonaldo et al. introduced normalisation and subtraction as two approaches to cDNA library construction that facilitate gene discovery by increasing the representation of less abundant transcripts, therefore accelerating the identification of novel genes and reducing the costs of sequencing redundant clones (Bonaldo et al., 1996).

Normalisation utilises the fact that rare cDNAs reanneal less rapidly than common transcripts, and that the single-stranded fraction of the cDNA therefore becomes progressively normalised as the reaction progresses. Subtraction involves the hybridization of a single-stranded cDNA library (the "tracer") with a collection of PCR–amplified cDNAs to be eliminated (the "driver"). Double-stranded molecules

are then removed from the sample by hybridisation to hydroyapatite, resulting in the creation of a single-stranded, "subtracted" library.

The construction of a cDNA library (Figure 1) begins with total RNA extraction from the tissue, developmental stage or pathological state of interest. Poly-adenylated (poly-(A)) RNA is isolated by passing the total RNA thorough a solid-phase matrix to which a complementary-polynucleotide is bound. The poly-adenylated mRNA binds selectively to the matrix and is later eluted. This isolated mRNA is converted to double-stranded mRNA/cDNA hybrid using reverse transcriptase, following which the RNA strand is selectively degraded leaving single-stranded cDNA which is used as a template to produce the complementary strand. The double-stranded cDNA can then be cloned into a vector. The time taken before the reverse transcriptase dissociates from the cDNA strand determines the length of the clone insert, and therefore what fraction of the mRNA is represented by the clone insert. Insert lengths therefore vary and may represent the entire mRNA or just a small part of the full-length sequence. Techniques which stabilise the reverse transcriptase, resulting in the production of clones with longer inserts, thereby providing increased coverage of the mRNA, have been developed (Carninci et al., 1998; Carninci and Hayashizaki, 1999; Carninci et al., 2001). The set of clones produced from the total RNA pool represents the clone library. Usually several hundred to several thousand clones are isolated at random from the cDNA library.

Clones undergo single-pass sequencing from one or both ends of the clone insert using vector-based primers to produce 3’ and/or 5’ sequences of varying length and quality. 3' tag sequencing dominated early EST studies because of the unique nature of the 3' untranslated regions (UTR) of genes, and because the poly-(A) tail of the

cDNA insert could be used for priming in sequencing reactions. Subsequent studies have been extended to generate 5' ESTs, producing pairs of sequence reads sharing the same parent clone. EST sequences are usually between 300-500 readable bases in length and 3’ and 5’ sequence reads may overlap with one another depending on the length of the insert. Since a full-length transcript may be several thousand bases in length ESTs are short sample tags which represent the source mRNA, enabling rapid transcript identification at the expense of sequence length and quality.

A limitation of the commonly used end-sequencing approach is that the central regions of long transcripts are not well-represented in the EST databases. The ORESTES (Open Reading Frame ESTs) project (Camargo et al., 2001) has undertaken the generation of cDNA libraries from the central, coding regions of transcripts using a randomly-primed RT-PCR-based approach. The addition of more than 700 000 ORESTES to the public databases has significantly increased the representation of the central protein coding regions of transcripts as well as further increasing the detection of novel transcripts.

1.1.2 EST quality

EST sequence data is considered to be of low quality. Only single-pass sequence reads are generally generated for each clone with no attention to the quality of these sequences. Compression and basecalling errors, which may result in frameshifts, occur approximately once every 100bp. Other errors present in EST data include lane tracking errors common when using slab-gel sequencing, internal priming and clone end reversal which are documented on the Washington University website at http://genome.wustl.edu/est/esthmpg.html. Further, the ligation of unrelated cDNAs results in the presence of chimeric clone inserts, and the presence of contaminating

sequence including genomic, vector, mitochondrial and ribosomal DNA, and cDNA from unrelated species are all common.

1.1.3 EST clustering and transcript reconstruction

ESTs offer a rapid and inexpensive route to gene discovery, reveal expression and regulation information and are instrumental in the detection of alternative splicing events. Unfortunately the short, unprocessed, error-prone nature of EST data means that full advantage cannot be taken of this valuable sequence information. However, the sheer volume of EST data generated by large-scale EST sequencing projects (Figure 2) means that a significant improvement in reliability can be gained by taking advantage of EST redundancy to reduce error and increase the length of represented transcripts (Jongeneel, 2000). EST clustering projects pre-process, cluster and post-process EST data to yield higher quality transcript information. An aim of these projects is the construction of gene indices, non-redundant catalogues where all available transcripts are partitioned into clusters such that transcripts are placed in the same cluster if they represent the same gene or gene isoform. Gene indices facilitate gene expression studies and novel transcript detection. Some groups also perform transcript reconstruction by using assembled clusters to build a consensus sequence that provides a longer and more accurate representation of the transcript represented by the cluster.

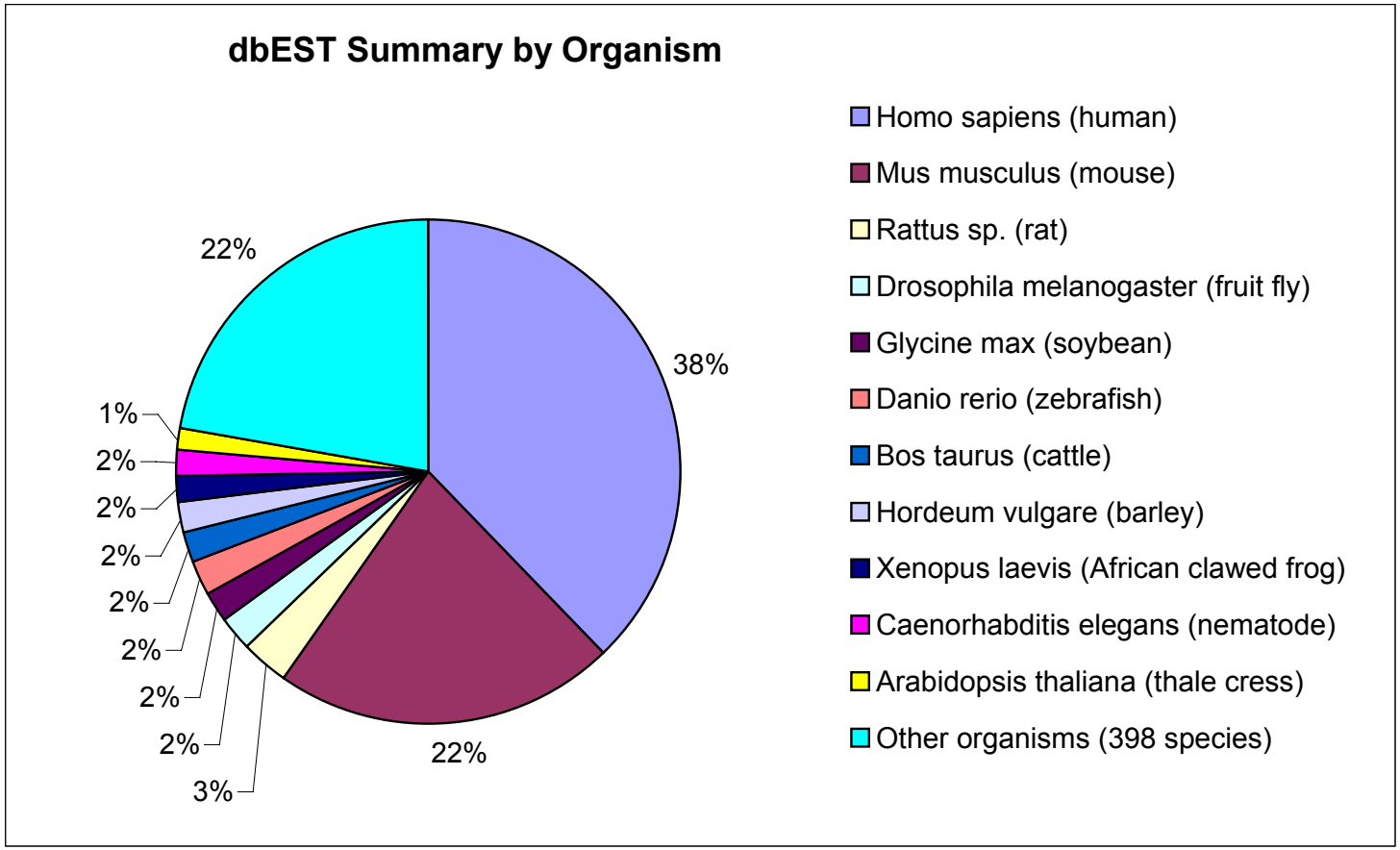


Figure 2. Representation of organism transcriptomes in dbEST. The May 2002 release of dbEST contains more than 11.5 million ESTs from 409 organisms. Human ESTs predominate, making up 38% of the data, with mouse ESTs making up 22% of the data. Organisms well-represented in dbEST are generally those which are well-studied model organisms.

1.1.3.1 What is an EST cluster?

An EST cluster is a collected set of ESTs which represent the same gene or gene isoform. Membership of the cluster is based on sequence similarity. Ideally each cluster should represent only one gene, and all ESTs from the same gene should be in a single cluster.

1.1.3.2 Overview of EST clustering

The grouping of transcripts based on sequence similarity forms the basis of EST clustering. Initially sequence identity is used to determine cluster membership. In addition ESTs that are annotated as having been sequenced from opposite ends of the same clone can be grouped together on the basis that they are from the same clone insert and therefore from the same gene. This shared annotation information provides a secondary (though less reliable) method of clustering.

A generalised clustering system is organised around the rapid initial grouping of sequences sharing significant similarity, followed by the accurate alignment of sequences within each cluster. Clustering can be performed with or without the generation of consensus sequences. The value of a consensus sequence is that it can be used as a single representative of the cluster as a whole.

1.1.3.2.1 Loose and stringent clustering

Depending on the aim of the clustering process either loose or stringent clustering algorithms can be implemented. Stringent clustering systems tend to sacrifice consensus length in favour of sequence fidelity, and result in lower coverage of expressed genes and the inclusion of fewer transcript isoforms. Loose clustering systems result in greater coverage of expressed genes and the inclusion various

trancript isoforms at the cost of the possible inclusion of paralogs and lower fidelity data. Stringent clustering is performed by systems such as TIGR_ASSEMBLER and looser clustering is implemented in Unigene and STACK_PACK. Individual clustering systems include various pre- and post- processing steps in order to manage the shortcomings of each of these approaches.

1.1.3.2.2 Supervised and unsupervised clustering

The aim of clustering is that each cluster should represent only one gene, and all ESTs from the same gene should be in a single cluster. In the absence of full-length mRNAs, or genomic sequence ESTs are clustered based on sequence similarity and clone of origin. This "unsupervised" clustering may result in ESTs representing the same gene being split into separate clusters if they do not share significant sequence identity. Available full-length sequence such as mRNA, or genomic sequence can be used as a scaffold upon which to cluster ESTs. The increased length, and therefore representation, provided by these scaffolds allows ESTs which may not have been clustered based on sequence similarity to be placed in the same cluster based on their identity to a common scaffold. The use of such scaffolds as the basis for clustering is known as "supervised" clustering.

1.1.3.3 Steps in EST clustering

Though specific implementations vary widely EST clustering generally proceeds through certain basic steps. (Figure 3)

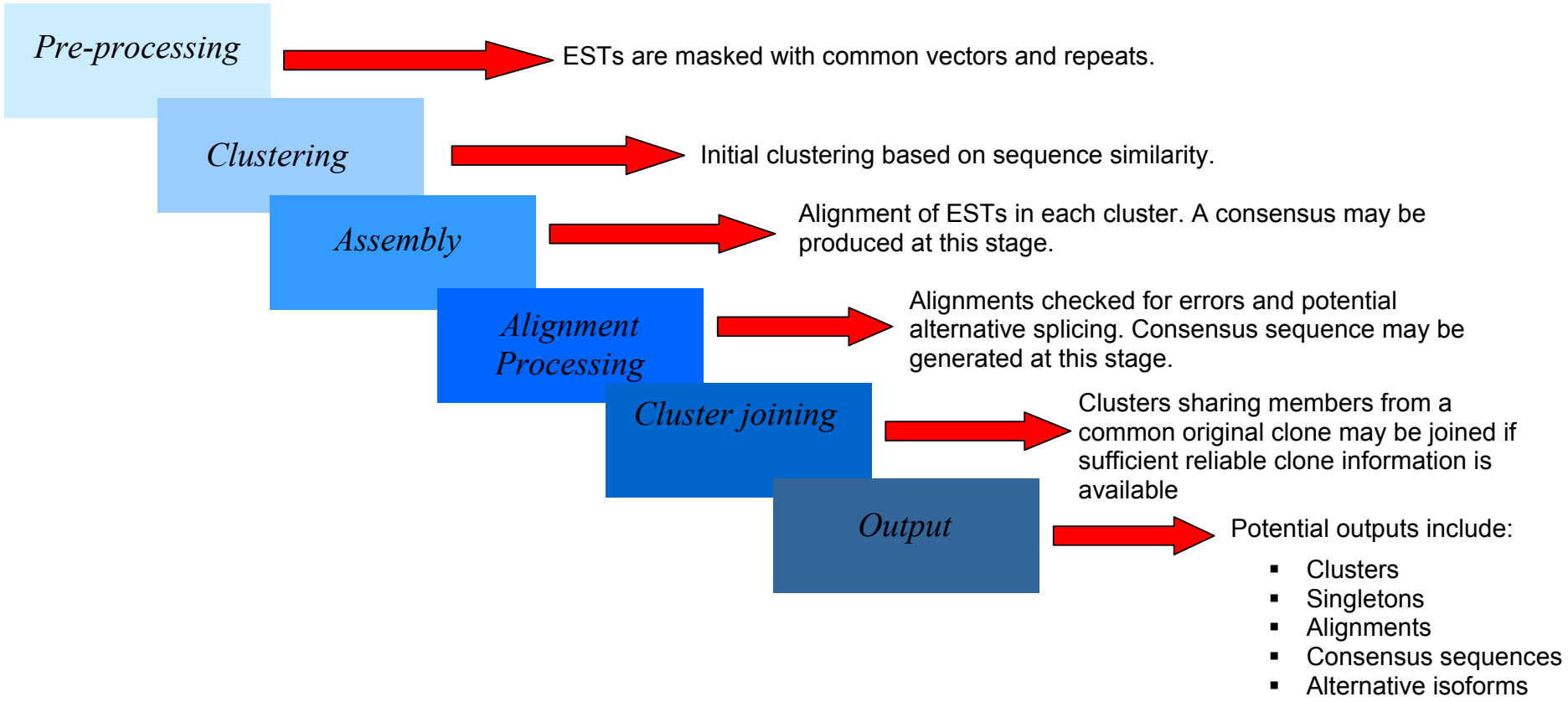


Figure 3. Steps in EST Clustering. EST clustering is performed by a number of groups each using different approaches. A generalised approach to EST clustering involves pre-processing to remove contaminants, clustering based on sequence similarity, alignment and consensus generation. The output of EST clustering is a set of groups (clusters) of ESTs where each cluster ideally represents a single gene. Many systems produce alignments and consensus sequences for each cluster. The consensus sequence is a single sequence which represents the member sequences of the cluster. Singletons are also produced, singletons are ESTs which share no similarity with any other sequence in the dataset and which may represent single genes. Some systems may identify potential alternative isoforms.

1.1.3.3.1 Pre-processing

Membership of an EST cluster is primarily determined by shared sequence similarity between cluster members. Sequence quality is of overriding importance in assigning ESTs to the correct clusters. Pre-processing in the form of masking, trimming and filtering is used to optimise sequence quality prior to clustering.

1.1.3.3.1.1 Masking

A common problem in EST clustering is the presence of contaminating sequence elements. ESTs generated from distinct genes but which share these contaminating sequence elements will be clustered together despite the fact that they represent distinct transcripts. It is therefore essential that these elements be removed by masking prior to clustering. Common contaminating sequence elements include:

a) Repeats

Repetitive elements are common features of EST data. These include the ubiquitous ALU, SINE and LINE elements common to human genes. Repeat databases such as RepBase (http://www.girinst.org/) are a valuable resource that can be used for masking of the input EST data.

b) Low Complexity Sequence

Low complexity sequence (microsatellite repeats such as (CA)n and poly-A tracts) may cause problems for clustering – particularly when sequence similarity based on sequence alignment is used for assigning cluster membership. Those clustering algorithms that use word-based cluster assignment approaches can be modified to assign low weight to these low complexity words.

The most effective method for removing contaminant sequence is to compare each sequence in the input dataset against a database of repeats, vectors and other sources of potential contamination. Fast, accurate algorithms such as XBLAST (NCBI) and cross_match, which implements the Smith-Waterman-Gotoh algorithm, have both been used successfully with cross_match providing greater flexibility and sensitivity than XBLAST. In cases where a direct identity is found with a sequence in the repeat or vector database a "mask residue" is inserted into the read (Figure 4). The resulting strings of NNNs or XXXs will be ignored by most clustering algorithms. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) is a program that screens DNA sequences for repeats and (optionally) low complexity DNA sequences. RepeatMasker invokes cross_match to perform the masking using various repeat databases, and provides a complete report on the repeats present in the input data as well as a modified input file in which the annotated repeats have been replaced with either Xs or Ns. While masked sequence is essential for accurate clustering it is important to revert to using the raw, unmasked sequence in the assembly step. The elements such as repeats and low-complexity regions which are removed during pre-processing are valuable in ensuring accurate sequence assembly.

In cases where sequence contamination is not detected prior to clustering sequences sharing a common contaminant should be placed into a single cluster. There is no automated method for detecting contaminated clusters, however, if a cluster containing an unusually large number of sequences is detected this should be carefully analysed for the presence of unmasked contaminants. Once contaminants have been identified and removed the cluster can be broken down into its individual sequence components and re-processed.

Raw unmasked EST

```
>Seq 1
CTTGGATCCTCTAGAGCGGCCGCCCTTTTTTTTTTTTTTTTGGTATAGCCCTGGCTGTC
CTGGAACTCACTTTGTAGACCAGGCTGGCCTCGAACTCAGAAATCCGCCTGCCTCTGCCT
CCCAAGTGCTGGGATTAAAGGCATGCACCACCACGGCCGTTTTGGAAGCATTTTCTTTTT
TCTTTTGTTTTTTTGTTTTTCAAATCTTTGTATTTTATTGTGAAAAATATTTGATGTGAG
AAGCATTTTCTTAACTGGGGTTCTTGCCTCTCAAAGGATTCTAGCCCATGCCAAATTAAC
ATAAAGTTAGATAGAACACTGATTAAAAAGATGCTCACTCTGAAAAACAATGTCCATCAT
TTCCTTCAAAGCTGTAAGGCTTTCTCACAGGTACGTATCTTGACCCTGTGTGTGTGTGCG
CGCGCGTGCACCCCACAAAAAAAATACACGTCATTTTTCTTCATTTCTCTCAGCCTGTTA
TTTTTCAAGATGGACAGACTCGCTTTGTGGTCTAGCTGGTCCAAAATTCACTCTGTTGCT
CATACTGGCTNTTTTGATTCTCAA
```

Masked EST

```
>Seq1A
CTTGGATCCTCTAGAGCGGCCGCCCXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXTTTTGGAAGCATTTTCTTTTTTCTTTTGTTTTTTTGTTTTT
CAAATCTTTGTATTTTATTGTGAAAAATXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXCTGATTAAAAAGATGCTCACTCTGAAAAACAA
TGTCCATCATTTCCTTCAAAGCTGTAAGGCTTTCTCACAGGTACGTATCT
TGACCCTGTGTGTGTGTGCGCGCGCGTGCACCCCACAAAAAAAATACACG
TCATTTTTCTTCATTTCTCTCAGCCTGTTATTTTTCAAGATGGACAGACT
CGCTTTGTGGTCTAGCTGGTCCAAAATTCACTCTGTTGCTCATACTGGCT
NTTTTGATTCTCAA
```

Figure 4. An EST before and after masking for vectors and common repeats. Common vectors and repeat sequences in the raw EST data are identified by comparison to a file containing common vectors and repeats. A string of X's is inserted to replace the contaminating sequence before clustering is begun.

1.1.3.3.1.2 Filtering

The pre-processing of data for clustering may also include a sequence length and quality assessment step. Using chromatogram data sequence quality cut-off values can be imposed with only sequences above a certain quality threshold being submitted for clustering. PHRED (http://www.phrap.org/phrap.docs/phred.html), software which uses sequencer-produced tracefiles to perform basecalling and the assignment of quality values to each base, can be used to filter the EST sequence data for sequence above a given quality threshold. Raw tracefiles for a large number of EST sequences are available via FTP from the Washington University Genome Sequencing Center (http://genome.wustl.edu/). Extremely short sequences can also be discarded at this stage.

1.1.3.3.1.3 Trimming

Before EST sequences are deposited in public databases the vector sequence is generally trimmed from the ends of the sequence read. However, even short vector fragments can cause spurious clustering and must be removed. The VecBase database (ftp://ncbi.nlm.nih.gov/blast/db/vector.Z) is a valuable source of common vector sequence which can be used for masking. If a custom vector has been used for cloning the sequence of this vector should be used for masking the input data.

1.1.3.3.2 Clustering

Briefly, sequences are grouped using a fast measure of sequence identity, and ESTs sharing significant sequence similarity are placed in a single cluster. This cluster assignment is then subject to further verification.

1.1.3.3.2.1 EST clustering and assembly tools.

Tools for clustering and assembly vary in their aims and approaches as outlined below.

a) Using Common Homology Based Tools. The well-known tools for sequence comparison (Smith-Waterman, BLAST and FASTA) are designed for homology searching, the purpose of which is to detect and quantitate the similarity (distance) between any two sequences. Although not developed specifically for clustering these packages are generally widely available, and the default parameters can be modified to enable clustering. Since the distance measure used in EST clustering is reduced to a binary it is only necessary to detect near or perfect matches. For this reason it is possible to select for speed over sensitivity in the initial pairwise comparison. The complexity of an EST clustering task is dependant on the number of ESTs in the input dataset. Datasets of a few hundred to a few thousand ESTs can be clustered efficiently using standard tools for multiple sequence alignment and assembly. However, these approaches are untenable for larger projects. Obtaining even a binary distance between potentially millions of ESTs is far from trivial – even using modern supercomputers.

b) Purpose-built Alignment Based Clustering Methods. A number of dedicated alignment-based clustering algorithms have been developed, though few have been implemented for large-scale clustering. For single seed clusters a dynamic-build based strategy which uses iterative BLAST searches to build clusters from single 'seed' ESTs is feasible. This method is not generally implemented for large-scale database building. A second example of purpose-built alignment-based clustering software is the JESAM package used at EBI to build the alignments and clusters for the

EuroGeneIndices (Parsons and Rodriguez-Tome, 2000). JESAM first finds and stores the alignment between sequences, following which clusters are built using these alignments. Since these two steps are separate, different algorithms can be implemented for the clustering step. Alignment-based tools are often intolerant of sequencing error.

c) Non-Alignment Based Clustering Methods. Word-based agglomerative algorithms and pre-indexing methods fall into this category. Agglomerative clustering means that each EST starts out in a unique cluster, and that the final clustering is generated through a series of merges. Merges are made using transitive closure rules whereby any two sequences with a given level of similarity will be placed into a single cluster. Hence, dissimilar sequences A and B will be placed in a single cluster if they both share similarity to sequence C. Word-based (rather than alignment-based) similarity is used. Clusters are merged when two sequences share word identity and multiplicity above a set threshold within a specified window size. Non-alignment based methods, while more tolerant of sequencing error than alignment-based methods, tend to capture gene variants and contaminating sequence (Burke et al., 1999). These artefacts can be identified by post-processing of the clusters using assembly and analysis tools.

1.1.3.3.3 Assembly

Assembly may be a part of the clustering step, or may be performed later by specialist assembly software such as PHRAP or CAP3. A consensus sequence may be derived directly from the assembly.

Commonly used sequence assemblers include PHRAP (Green, 1996) TIGR_Assembler and CAP3. PHRAP and CAP3 have been reported to be more tolerant of sequence error, and are therefore more suited to EST assembly. CAP3 has been demonstrated to produce fewer assemblies per gene and may produce a higher quality consensus sequence.

1.1.3.3.4 Alignment processing

Aligned clusters, particularly those generated as part of a loose clustering strategy, should be processed to detect errors and alternative splice forms. Consensus sequences may be generated as a part of this step, or may be accepted directly from the assembly.

1.1.3.3.5 Cluster joining

Clusters or cluster consensi can be further grouped using annotation information. Clone-linking utilises the fact that 3' and 5' reads from the same clone share a clone id. It should be noted that linking based on clone annotation is entirely dependent on the accuracy of clone annotation in the EST database and is therefore subject to error.

1.1.3.4 Overview of gene indices produced by clustering ESTs

A number of gene indices have been produced using publically available EST data. These aim to reduce the redundancy present, and thereby enhance the information which can be gleaned from EST data. The TIGR and UniGene databases have focused on reconstruction of the gene complement of genomes and their technological developments have been directed towards achieving that goal. The STACK database has focussed on the detection and visualisation of transcript variation in the context of tissue, developmental stage and pathological states.

1.1.3.4.1 Unigene

Unigene, based at the National Centre for Biotechnology Information (NCBI), is one of the earliest and most enduring efforts for the automatic production of gene indices from Genbank sequences. Each Unigene cluster contains mRNA and EST sequences which represent a unique gene. Additional information such as the identity of the gene, chromosomal map location, and tissue types in which the gene is expressed (from SAGE and EST data) is also provided. NCBI does not generate contigs and/ or consensus sequences for Unigene clusters. The HumanInfoBase (http://www.mips.biochem.mpg.de/proj/human/) database at MIPS provides assembled, annotated Unigene clusters.

Unigene databases are available for 11 organisms (at time of writing): human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), zebrafish (*Danio rerio*), Cow (*Bos taurus*), Clawed frog (*Xenopus laevis*), Arabidopsis (*Arabidopsis thaliana*), wheat (*Triticum aestivum*), rice (*Oryza sativa*), barley (*Hordeum vulgare*) and maize (*Zea mays*). Databases are updated weekly with new ESTs, and bimonthly with newly characterised sequences. All Unigene databases are available for download from: ftp://ncbi.nlm.nih.gov/repository/UniGene/

Unigene clusters can be searched by gene name, Unigene cluster ID, chromosomal location, cDNA library, accession number, and text terms. Sequence-based searches against Unigene human, rat and mouse databases are available from the Swiss Institute of Bioinformatics at: http://www.ch.embnet.org/

Unigene has been used for the selection of unique transcripts for the construction of a cDNA microarray for the large-scale analysis of gene expression, and as the candidates for the production of a human gene map.

For information on the construction of Unigene see http://www.ncbi.nlm.nih.gov/UniGene/build.html

1.1.3.4.2 TIGR gene indices

The Institute for Genome Research (TIGR) produces gene indices for more than 40 organisms including various animal, plant, protist and fungal species (Quackenbush et al., 2000; Quackenbush et al., 2001). The TIGR indices incorporate both ESTs sequenced at TIGR, ESTs from dbEST and mRNAs from Genbank. Each TIGR cluster contains a fasta formatted consensus sequence with a unique accession as well as additional information including details of the assembly, tissues in which the gene is expressed and putative gene identification. Related databases generated by TIGR provide additional information about TIGR TCs. The Genomic Maps database provides genomic mapping for a subset of organisms for which TCs are available. The TIGR Orthologous Gene Alignment database (TOGA) (Lee et al., 2002) provides information about orthologous sequences between TCs for the organisms for which TIGR Gene Indices have been generated.

Each TIGR cluster is represented by a Tentative Consensus sequence (TC, or THC in the case of Tentative Human Consensi). The TIGR databases are freely available to researchers at non-profit organisations at http://www.tigr.org/tdb/tgi.shtml. The TIGR Human Gene Index (HGI) is produced annually. The frequency of new releases varies between species, and depends on the accumulation of new transcripts. The TIGR gene indices can be searched by nucleotide or protein sequence, EST, transcript or consensus identifiers, tissue, cDNA library name or library identifier, gene product name, functional classification according to Gene Ontology (GO) terms (Ashburner et al., 2000). Various publications on the TIGR Gene indices are available.

1.1.3.4.3 STACK

The STACK human gene index is generated by clustering EST and mRNA data, and offers human transcript consensus sequences that reflect gene expression forms and alternate expression variants within 15 tissue-based and one disease category (Miller et al., 1999; Christoffels et al., 2001). This organisation of transcript by expression site presents the opportunity to explore transcript expression in specific tissues or subsets such as disease related sequences.

Each STACK cluster contains alignments, consensus sequences and assembly information, and is dynamically linked to the UniGene database. Web-based software allows for the visualisation of clusters and alignments, and highlights transcript variation. STACK database releases are made available with varying frequency – on average twice a year. STACKdb and the stackPACK toolset used to generate STACK are freely available to academic groups and can be downloaded from http://www.sanbi.ac.za/CODES. Sequence-based searching of STACKdb is available at http://juju.egenetics.com/stackpack/webblast.html.

STACKdb has been used to support the detection of a novel retinal-specific gene responsible for retinitis pigmentosa. The STACKpack toolset has been used in the production of various gene indices and for the survey of genes in the malarial genome.

1.1.3.5 Gene indices incorporating genome data

1.1.3.5.1 Ensembl

The Ensembl database at EBI (http://www.ensembl.org/) provides an automatic annotation of a number of completed genomes including human, mouse, fly, and fugu. The system combines automated gene predictions with external data derived from

both experimental and computational sources to provide an integrated source of gene, transcript and protein sequence data, as well as functional information. An open SQL database and a query interface, Ensembl Mart, provide users with the ability to access all the available information and to perform biologically useful queries of the stored data.

1.1.3.5.2 RefSeq

The RefSeq project at NCBI aims to produce a single set of curated reference sequences for each genomic region, transcript and protein (Pruitt et al., 2000). http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html. RefSeq provides a stable and non-redundant set of reference sequences for gene characterisation, including expression studies, mutation analysis and the detection of polymorphisms.

RefSeq is made up of three primary projects: (i) curated RefSeq (http://www.ncbi.nlm.nih.gov/LocusLink/build.html), (ii) genome annotation (http://www.ncbi.nlm.nih.gov/genome/guide/build.html) and (iii) complete genomes, each of which are generated using different methods.

1.1.3.5.3 AllGenes

AllGenes (http://www.allgenes.org/) provides access to an integrated database of known and predicted genes in mouse and human . The major strength of AllGenes is its structured approach to the integration of ESTs, genomic sequence, expression information and, functional annotation. A relational database and controlled vocabulary provides users with the ability to perform useful biological queries in a uniform manner.

1.1.4 Finding coding regions using EST data

The detection of novel proteins using EST data is complicated by the fact that searching EST databases with commonly used six-frame translation using software such as tBLASTn yields largely sequences which have no coding region, or which are in the incorrect reading frame, or the incorrect strand.

The short length and poor sequence quality of ESTs means that finding the open reading frame (ORF) is not generally feasible.

Successful approaches to the detection of the coding sequence (CDS) in EST data have been in two major areas:

- Detection of similarities to known protein sequences or sequence motifs. This method requires that there is similarity to known proteins or protein motifs. This restricts this method to the identification of sequences with similarity to previously identified proteins.

- Detection of statistical biases in the CDS nucleotide sequence associated with codon frequency usage. This approach assumes no similarity to known proteins, but requires modification for each species to which it is applied to account for differences in the codon usage. This method has been applied in the program ESTScan (Iseli et al., 1999) (http://www.ch.embnet.org/software/ESTScan.html) which is able to detect and correct sequence errors resulting in frameshifts within the CDS.

1.1.5 Expressed sequence tag databases

1.1.5.1 dbEST

dbEST, distributed by NCBI, and the EST divisions of Genbank and EMBL, are EST repositories which contain sequence and annotation information for publically available EST data. More than 10 million ESTs representing in excess of 375 organisms have been deposited in dbEST since its inception. This EST data is available by anonymous ftp from ftp://ncbi.nlm.nih.gov/genbank/. Individual sequences and small batches can be obtained using Entrez (http://www.ncbi.nlm.nih.gov/entrez/).

The EST highly redundant data in these databases are not clustered or assembled, and may or may not be grouped by species of origin. Unrestricted homology searches against dbEST will therefore commonly return numerous sequences which represent the same gene as the query, paralogous genes, and sequences from related species. Both NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) and SIB (http://www.ch.embnet.org/software/aBLAST.html) offer the ability to search subsets of dbEST restricted by species, with NCBI offering human, mouse and "other" divisions, and SIB offering the ability to select one or more from a large number of divisions including plants, prokaryotes, fungi, invertebrates, zebrafish, human, mouse and rat.

Searching clustered EST collections such as Unigene will result a more concise report than searching dbEST. Homology searching against clustered databases which provide contigs and consensus sequences for each cluster is very rapid, though the accuracy of the contig production and consensus sequence generation may affect the

quality of the matches obtained. TIGR, SANBI and MIPS offer BLAST searching of their gene indices; TIGR Gene indices, STACK and HIB on their respective websites.

1.2 Transcript identification using full-length mRNAs

Despite the abundance of EST data, the production of full-length cDNA libraries and transcripts remains an important priority in the elucidation of complete transcriptomes as the best evidence for an expressed gene is a fully sequenced transcript. The complete transcriptome cannot accurately be deduced from the genome sequence alone, owing to the complexities of transcription and transcript processing. While the capture of expressed transcripts from a large number of expression states remains impractical, the availability of full-length sequences provides coverage of the entire length of the transcript; something which is not available from EST data. Coverage across the length of the gene provides valuable information about exon usage and the position of exon-intron boundaries which, in turn, contributes to an increase in the accuracy with which transcripts can be mapped to the genome sequence. Full-length transcripts can be used as scaffolds/organisers of EST data – improving the accuracy of EST clustering, and providing further information about the occurrence of alternative splicing. Complete sequence can also be used more efficiently than ESTs for the prediction of protein structure and function and the isolation of the cognate protein.

Both the libraries of clones containing full-length cDNA inserts, and sequences of the full-length cDNAs are important components of a full-length resource. The availability of full-length clones in an organised public collection is a critical resource for ongoing genetic research.

For 90% of gene predictions, the true 5'-end of the coding sequence is not correctly annotated. For this reason the amplification of ORFs based on these predictions does not provide a reliable transcript of the full-length protein-coding gene. Full-length cDNA cloning and sequencing therefore remains the method of choice in the identification of full transcripts. However, this approach will not capture all transcripts, but is specific to the site, cell type, developmental stage and treatment/pathology of the material from which the library was prepared. Full length sequencing of mRNAs can be interpreted as sequencing of the complete protein coding sequence, or sequencing of the mRNA from cap site through to final polyadenylation site. The latter is the preferred definition.

Ideally the production of full-length sequence for all transcripts would represent the transcriptome, however, the routine production of full-length transcript sequences for all possible transcript isoforms in every tissue, developmental stage, and environmentally affected condition – and every combination of these factors is not a feasible undertaking.

1.2.1 Full-length cDNA library production and sequencing

Until relatively recently technologies for the large-scale production of stable full-length cDNA libraries did not exist or were not feasible given the time and cost involved. A number of recent advances, including the use of stable enzymes and non PCR-based techniques, have allowed for the production and selection of long insert cDNA libraries. Additionally, the advances in sequencing technologies during the progress of the Human Genome Project have increased the capability for the high-throughput production of sequence data while simultaneously reducing the associated costs. There are various limitations in the production of full-length cDNA libraries.

The preparation of full-length cDNAs is simpler for shorter mRNAs, while longer transcripts are more difficult to clone and propagate. This leads to a bias in the insert size in full-length cDNA libraries. In order to identify rare genes – particularly where these transcripts are longer than average – it is necessary to develop techniques which overcome limitations on insert size and clone propagation.

1.2.2 Full-length mRNA databases

1.2.2.1 Human Full-Length cDNA Annotation Invitational (H-Inv)

A number of groups have been involved in the production of full-length human cDNAs. To co-ordinate these efforts and thus provide a highly annotated, unified set of high quality human transcripts the Japanese Biological Information Research Centre and the DNA Database of Japan initiated the establishment of a core transcriptome database; the Human Full-Length cDNA Annotation Invitational project is an international collaboration to produce a unique set of high quality full-length cDNA clones by automatic annotation and human curation under unified criteria. The H-Invitational Database (H-InvDB) provides annotation of biological, structural, functional and evolutionary information for each transcript.

The cDNAs included in H-InvDB were obtained from eight groups involved in the production of full-length human cDNAs.

1. Full-Length Human cDNA Sequencing Project by NEDO

2. Full Length cDNA by Institute of Medical Science, University of Tokyo

3. Hunt: Human Novel Transcripts by Helix Research Institute, Inc.

4. HUGE: Human Unidentified Gene-Encoded Large Proteins by Kazusa DNA Research Institute

5. NEDO Database at Kazusa DNA Research Institute

6. Mammalian Gene Collection (MGC) by NCI/NIH

Initiated in 1999 as a collaborative effort between various institutes of the NIH, the Mammalian Gene Collection project aims to provide a catalogue of full-length mammalian genes (Strausberg et al., 1999). The project has focussed initially on the production of full-length cDNAs for human and mouse, and will later extend to include other mammals. Clones produced by the project are prepared from high quality mRNA extracted from cell lines or tissues. Clones are made available through the IMAGE Consortium, while 3' and 5' ESTs are generated and released to public databases. An ongoing informatics challenge is the selection of clones likely to represent full-length transcripts. In the initial phases of the project clones with inserts of up to 3 to 4kb were sequenced using techniques such as shotgun sequencing, primer walking, and concatenation. Sequence data is generated to the same standards as those specified by the Human Genome Project – finished sequence is therefore 99.99% accurate. Annotation of the sequence data is also performed.

The goals of the project include the development and improvement of supporting technologies, including: (i) improving the preparation of full-length cDNA libraries

from small quantities of starting material, (ii) identification of rare transcripts, long transcripts, or transcripts with complex structures (iii) algorithms and software for identifying clones containing uncharacterised full-length inserts; and (iv) techniques for faster, less expensive sequence production.

As of January 2002 a non-redundant set of more than 20 000 putative full-length human and mouse clones have been identified and full sequences for 9000 human and 4000 mouse clones have been produced. 75% of the selected clones contain full-length ORFs.

Clone library lists, clone lists and insert sequences in fasta format are available for download from http://mgc.nci.nih.gov/. Sequenced clones can be searched using BLAST at the same site. Additionally, the genes represented by MGC clones can be searched by gene name or keyword at the website.

7. German Human cDNA Project by DKFZ

The German cDNA Consortium is the largest European full-length cDNA generation and sequencing project. A major objective is the functional characterisation of the full-length cDNAs identified by the project. The sequences produced by the project undergo comprehensive manual and automated annotation and curation and data is available for homology searching at http://mips2.gsf.de/proj/cDNA/blast_search.html

Clones produced by the project are freely available for research from (http://www.rzpd.de), or by request to: clone@rzpd.de

8. Human cDNAs produced by Chinese National Human Genome Center (CHGC)

1.2.2.2 RIKEN Mouse Gene Encyclopedia Project

The RIKEN Mouse Gene Encyclopedia Project aims to identify and sequence all full-length transcripts for the mouse genome (Kawai et al., 2001). In addition information regarding expression locations, chromosomal mapping and annotation are collected and presented at http://genome.rtc.riken.go.jp/

A key feature of the project has been the development of technologies for the generation of full-length cDNA libraries and for high-throughput template preparation and sequencing. Developments in informatics for data management and annotation have been undertaken and provide the project with the ability to do real-time clustering of the 3' ends of 5'-end validated clones in order to provide a continually updated, non-redundant encyclopedia.

A full-length cDNA microarray constructed from a set of clones representing 19 000 full-length cDNAs has been prepared and used to examine developmental and metabolic pathways in 49 tissues.

In 2000 experts in biology and bioinformatics gathered for the Functional Annotation of Mouse (FANTOM) meeting, the aim of which was to collaboratively annotate approximately 21 000 full-length sequences generated by the RIKEN group. All published RIKEN sequence data including 3' ESTs, 5' ESTs and full-length sequences are available from the public DNA databases (http://genome.gsc.riken.go.jp/homology/about.html#release). Homology searching against this RIKEN Mouse Gene Encyclopedia data is possible from the RIKEN BLAST website at http://genome.gsc.riken.go.jp/homology/blast.html. In addition the

full-length mouse encyclopedia sequences, annotation data, and predicted amino acid sequences for the FANTOM dataset are available for download from the RIKEN website at: http://genome.gsc.riken.go.jp/resource.html#archive. The approximately 21 000 full-length clones are available from RIKEN upon request.

1.3 Transcript quantification using SAGE

The use of sequence-based methods for the quantification of gene expression is more recent than their use in transcript identification, but has proved a rapid and valuable method to determine the distribution of transcripts in a sample of interest.

Serial Analysis of Gene Expression (SAGE) was developed in 1995 to take advantage of high-throughput sequencing technologies for the rapid identification and quantification of expressed gene transcripts (Velculescu et al., 1995; Velculescu et al., 2000). The advantage of SAGE over other methods of gene expression analysis is that it requires no prior knowledge about the identity of the genes of interest, and it provides quantitative expression information for the transcriptome under study. While sequencing ESTs from non-normalised cDNA libraries can also provide quantitative expression information the cost of obtaining a depth of sequencing comparable to SAGE is prohibitive. Consequently SAGE is generally a more cost-effective method for detecting low-abundance transcripts.

The SAGE technique has numerous applications including the identification of disease-related genes, the elucidation of developmental and disease pathways, and the analysis of treatments on cell lines or tissues.

1.3.1 Description of the SAGE method

The SAGE technique does not in fact measure the expression of a gene, but quantifies a "tag" which represents a gene transcript.

SAGE tags are nucleotide sequences of defined length directly adjacent to the 3'-most restriction enzyme site for a specified restriction enzyme. Original SAGE tags were 9bp in length, but more recent protocols generate 10 to 14 bp tags. NlaIII is the most commonly employed restriction enzyme though other 4-bp cutters may be used.

The generation of SAGE tags involves the conversion of extracted mRNA to cDNA followed by the digestion of the cDNA using a 4bp cutter enzyme (usually NlaIII). Digested cDNA is divided into two pools and different linker/adaptor sequences are ligated to the cDNAs in each pool. Each linker contains the docking site for a second restriction enzyme (BsmF1). The second restriction enzyme is then used to cleave the cDNA molecule a short distance (~20bp) downstream, resulting in short tags consisting of the linker sequence and about 20bp of the adjacent cDNA. Ligation of tags from the two pools results in the production of ditags which are PCR amplified. The linkers are then cleaved off before the ditags are concatenated. Concatamers containing 25 or more ditags are ligated into sequencing vectors before being cloned and sequenced.

The data produced by SAGE is a list of tags and the count value for each, thus representing the expression level of each tag (and therefore its corresponding gene) in the sample.

1.3.2 Disadvantages

While SAGE provides a rapid and inexpensive means of quantifying gene expression levels there is the potential for error in the assignment of tags to genes and in the absolute quantification of transcripts. A 10bp tag is not necessarily a specific or unique representation of a transcript. Instances in which one tag can be assigned to more than one gene (ambiguous tag to gene assignment), and in which one gene has more than one tag (non-specific tag to gene assignment due to alternative polyadenylation or polymorphism) can and do occur. This is compounded by the fact that even acceptable levels of sequencing error can have a significant effect on a short tag.

There are therefore two major issues to be addressed:

1. Ensuring that the tags and tag counts are a valid representation of the genes and their expression levels.

Sequencing error has the greatest impact on the validity of the tags and their counts. A sequencing error of 1% per base translates to a 10% chance of one or more errors occurring in a 10bp tag. The results of this sequencing error could be to:

   a. Decrease the count for that tag by one

   b. Increase the count of another unrelated tag by one

   c. Establish counts for a tag that does not represent any gene

In cases where the tag count is high, decreasing it by one has relatively little effect. However, tag counts of 1 need to be treated with suspicion – and are routinely discarded from any analysis, as their accuracy cannot be easily verified.

2. Providing accurate tag to gene mapping

The valid and useful assignment of tags to genes is made difficult by the ambiguous and non-specific nature of tag to gene mapping and compounded by sequencing error. Tags are derived from transcribed sequences usually from incompletely characterised transcriptomes. The set from which tags can be derived is therefore incomplete. The specificity and unambiguity of the tag to gene mapping can therefore not be confirmed. A sequencing error rate of 1% per base means that one or more errors may occur in the average 10bp tag, compounding the unspecific and ambiguous tag to gene mapping. Electronic quantification of the non-specific and ambiguous tag to gene mappings can be estimated by extracting SAGE tags from well-characterised, low-error mRNA sequences in public databases and matching these to defined gene units.

Using EST data for SAGE tag extraction and quantification is complicated by the sequencing error present in ESTs. EST tag to gene assignments can be corrected for the estimated 10% of tags likely to be due to error. This correction is accomplished by removing 10% of the most rarely occurring tags for a particular gene, and removing 10% of the most rarely occurring genes for a particular tag. Of course, this method may remove evidence for rare, naturally occurring transcripts.

1.3.3 SAGE databases and tools

SAGE tags and counts for various libraries are available for online query through the SAGEmap website at NCBI http://www.ncbi.nlm.nih.gov/SAGE/. Using this site researchers are able to retrieve SAGE data by tag, sequence, gene and by library. User are also able to perform SAGE-based differential expression analyses for any pair of

libraries for which SAGE tags have been generated, or to compare the relative expression levels of selected SAGE tags in two libraries.

Integration of SAGE tags with genomic information is made possible using the MapViewer display tool also available from the NCBI SAGE website. Using this tool the genomic position of individual SAGE tags can be viewed and total tag counts and tag distributions obtained.

1.3.4 CGAP SAGE libraries

The Cancer Genome Anatomy Project (CGAP) has invested in producing SAGE libraries for human colon and brain tissues. Differential expression between cancer and normal samples can be performed using xProfiler at the NCBI SAGE site mentioned above.

1.3.5 SAGEtag to UniGene mapping

The mapping of SAGE tags to UniGene is an automated process which results in the assignment of a UniGene cluster identifier to each SAGE tag via the following steps:

1. Extraction of human sequences from Genbank
2. Assignment of a SAGE tag to each sequence
    a. Assessment of orientation based on poly-adenylation signal (aTTAAA or AATAA), polyA-tail and sequence annotation
    b. Extraction of a 10bp tag 3'-adjacent to the 3'-most NlaIII site (CATG)
3. Assignment of a UniGene identifier to each human sequence with a SAGE tag

Both “reliable” and “full” Tag to gene assignments for all Genbank transcripts and for UniGene clusters have been constructed and are publically available from the SAGEmap ftp site. The reliable mappings are corrected for EST sequencing error whereas the full mappings are not.

1.4 Limitations

It is important to note that the methods discussed in this chapter deal with identifying and quantifying expressed genes using transcript data, and that there is not necessarily a direct correlation between the expression level of transcripts and the production of their corresponding protein. Complex transcriptional and post-transcriptional regulation of gene expression operate in a broad range of eukaryotes. Post-transcriptional mechanisms such as mRNA degradation (Bevilacqua et al., 2003) and post-transcriptional gene silencing (Cogoni and Macino, 2000; Pickford and Cogoni, 2003) have been described, and will determine the relationship between the transcript and protein complements of a cell. The experimental approaches described in this chapter deal with direct measurement of transcript abundance and will therefore not accurately reflect the ultimate identity and levels of protein expression.

Appendix I

Further Reading Resources

1. ESTs

Using EST data

Jongeneel, C. V. 2000. Searching the expressed sequence tag (EST) databases: panning for genes. *Brief.Bioinform.* 76-92.

Common errors and contaminants in EST data

Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S., and Elliston, K. O. 1996. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.* 829-845.

NCBI's VecScreen website: http://www.ncbi.nlm.nih.gov/VecScreen/contam.html

The Open Reading Frame ESTs (ORESTES) Project:

Dias Neto, E., Garcia-Correa, R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da, Silva W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., Matsukuma, A., Baia, G. S., Simpson, D. H., Brunstein, A., de Oliveira, P. S., Bucher, P., Jongeneel, C. V., O'Hare, M. J., Soares, F., Brentani, R. R., Reis, L. F., de Souza, S. J., and Simpson, A. J. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc.Natl.Acad.Sci.U.S.A.* 3491-3496.

The public EST database: dbEST

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. 1993. dbEST--database for "expressed sequence tags". *Nat.Genet.* 332-333.

2. cDNA library construction

Making normalised cDNA libraries

Bonaldo, M. F., Lennon, G., and Soares, M. B. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 791-806.

Construction of full-length cDNA libraries

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., and Schneider, C. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*. 327-336

Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C., and Hayashizaki, Y. 1997. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res*. 61-66.

Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 1998. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc.Natl.Acad.Sci.U.S.A*. 520-524

Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol*. 19-44

Carninci, P., Shibata, Y., 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*. 1617-1630.

High-throughput template preparation and sequencing

Shibata, K., Itoh, M., Aizawa, K., Nagaoka, S., Sasaki, N., Carninci, P., Konno, H., Akiyama, J., Nishi, K., Kitsunai, T., Tashiro, H., Itoh, M., Sumi, N., Ishii, Y., Nakamura, S., Hazama, M., Nishine, T., Harada, A., Yamamoto, R., Matsumoto, H., Sakaguchi, S., Ikegami, T., Kashiwagi, K., Fujiwake, S., Inoue, K., and Togawa, Y. 2000. RIKEN integrated sequence analysis (RISA) system--384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res.* 1757-1771

3. Constructing gene indices

TIGR Human Gene Index

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. 2000. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* 141-145.

Unigene

Boguski, M. S. and Schuler, G. D. 1995. ESTablishing a human transcript map. *Nat.Genet.* 369-371.

STACKPACK

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., and Hide, W. A. 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* 1143-1155.

EuroGeneIndices

Parsons, J. D. and Rodriguez-Tome, P. 2000. JESAM: CORBA software components to create and publish EST alignments and clusters. *Bioinformatics.* 313-325.

HumanInfoBase
Geier, B., Kastenmuller, G., Fellenberg, M., Mewes, H. W., and Morgenstern, B. 2001. The HIB database of annotated UniGene clusters. *Bioinformatics.* 571-572.

4. Sequence assembly software

Review of sequence assembly
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.* 3657-3665.

PHRAP
Green, P. 1996. PHRAP.

http://www.genome.washington.edu/uwgc/analysistools/phrap.htm

phg@u.washington.edu.

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

5. SAGE

SAGE technique and applications
Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995. Serial analysis of gene expression. *Science*. 484-487.

Appendix II

Useful Links

1. Raw EST resources

Raw EST data: dbEST http://www.ncbi.nlm.nih.gov/dbEST/

Download all sequences	ftp://ncbi.nlm.nih.gov/genbank/
Download individual sequences and small batches	http://www.ncbi.nlm.nih.gov/entrez/
BLAST searchable dbEST	http://www.ncbi.nlm.nih.gov/BLAST/ http://www.ch.embnet.org/software/aBLAST.html

EST Tracefile archives

Washington University Traces Viewer	http://genome.wustl.edu/est/est_search/nci_viewer.html
NCBI Trace Archive	http://www.ncbi.nlm.nih.gov/Traces/

General information about ESTs

Washington University Genome Sequence Center http://genome.wustl.edu/est/

2. Sequence processing resources

Sequence contamination masking resources

Repeat and vector databases

Repbase, a database of common repeats	http://www.girinst.org/
Vecbase, a database of common vectors	ftp://ncbi.nlm.nih.gov/blast/db/vector.Z

Tools for performing masking

XBLAST	http://bioweb.pasteur.fr/docs/man/man/xblast.1.html
RepeatMasker	http://ftp.genome.washington.edu/RM/RepeatMasker.html

Sequence quality assessment resources

PHRED website http://www.phrap.org/phrap.docs/phred.html

3. Gene indices

Unigene

Unigene build information	http://www.ncbi.nlm.nih.gov/UniGene/build.html
Download Unigene	ftp://ncbi.nlm.nih.gov/repository/UniGene/
BLAST searchable Unigene	http://www.ch.embnet.org/

TIGR

TIGR information and download	http://www.tigr.org/tdb/tgi.shtml
BLAST searchable TIGR Gene Indices	http://tigrblast.tigr.org/tgi/

STACK

STACK information and download	http://www.sanbi.ac.za/Dbases.html
BLAST searchable STACKdb	http://juju.egenetics.com/stackpack/webblast.html

Annotated Unigene Clusters:

The HumanInfobase (HIB)	http://www.mips.biochem.mpg.de/proj/human/

4. Gene indices incorporating genome data

RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
AllGenes	http://www.allgenes.org/

5. Detecting open reading frames in ESTs

ESTScan	http://www.ch.embnet.org/software/ESTScan.html
FrameFinder	http://www.hgmp.mrc.ac.uk/~gslater/estateman/framefinder.html

6. Full-length mRNA databases

Mammalian Gene Collection (MGC) http://mgc.nci.nih.gov/

RIKEN http://genome.rtc.riken.go.jp/

Download sequences	http://genome.gsc.riken.go.jp/resource.html#archive
BLAST search sequences:	http://genome.gsc.riken.go.jp/homology/blast.html
Browse annotated sequences:	http://genome.gsc.riken.go.jp/homology/about.html#release

German cDNA Consortium http://mips2.gsf.de/proj/cDNA/

Homology searching	http://mips2.gsf.de/proj/cDNA/blast_search.html.
Clone ordering	http://www.rzpd.de

7. SAGE

NCBI SAGEmap website	http://www.ncbi.nlm.nih.gov/SAGE/
Download tag to gene mappings for Genbank and Unigene	ftp://ftp.ncbi.nih.gov/pub/sage/

Chapter 2

The Contribution of Exon-Skipping Events on Chromosome 22 to Protein Coding Diversity

2.1 Abstract

Completion of the human genome sequence provides evidence for a gene count with lower bound 30,000–40,000. Significant protein complexity may derive in part from multiple transcript isoforms. Recent EST based studies have revealed that alternate transcription, including alternative splicing, polyadenylation and transcription start sites, occurs within at least 30–40% of human genes. Transcript form surveys have yet to integrate the genomic context, expression, frequency, and contribution to protein diversity of isoform variation. We determine here the degree to which protein coding diversity may be influenced by alternate expression of transcripts by exhaustive manual confirmation of genome sequence annotation, and comparison to available transcript data to accurately associate skipped exon isoforms with genomic sequence. Relative expression levels of transcripts are estimated from EST database representation. The rigorous *in silico* method accurately identifies exon skipping using verified genome sequence. 545 genes have been studied in this first hand-curated assessment of exon skipping on chromosome 22. Combining manual assessment with software screening of exon boundaries provides a highly accurate and internally consistent indication of skipping frequency. 57 of 62 exon skipping events occur in the protein coding regions of 52 genes. A single gene, (FBXO7) expresses an exon repetition. 59% of highly represented multi-exon genes are likely to express exon-

skipped isoforms in ratios that vary from 1:1 to 1:>100. The proportion of all transcripts corresponding to multi-exon genes that exhibit an exon skip is estimated to be 5%.

2.2 Introduction

Gene expression products can have variable forms, characterized by alternate start sites of transcription and polyadenylation (Gautheret et al., 1998), exon skipping and alternate donor and acceptor sites at exon boundaries(Mironov et al., 1999a; Mironov et al., 1999b; Brett et al., 2000; Croft et al., 2000). Exon skipping in transcript isoforms is the most frequent event altering the protein coding sequence of genes, (Lander et al., 2001) (http://industry.ebi.ac.uk/~thanaraj/gene.html). Surveys of the incidence of alternative splicing, including exon skipping, have been performed (Andreadis et al., 1987; Iida, 1997; Valentine, 1998; Thanaraj, 1999), and a growing number of anecdotal observations confirm the utilization of exon-skipped transcripts in developmental (Dufour et al., 1998; Lim et al., 1999; Lambert de Rouvroit et al., 1999; Unsworth et al., 1999; Kawahara et al., 2000) tissue-specific (Zacharias et al., 1995), and disease-specific (Jiang and Wu, 1999; Mercatante and Kole, 2000; Strehler and Zacharias, 2001) states.

Several approaches have successfully used hybridization experiments both in silico (Wolfsberg and Landsman, 1997; Gautheret et al., 1998; Thanaraj, 1999; Mironov et al., 1999b; Brett et al., 2000; Croft et al., 2000; Beaudoing et al., 2000; Schweighoffer et al., 2000; Lander et al., 2001) and in vitro (Schweighoffer et al., 2000; Strehler and Zacharias, 2001) to assess alternate transcript diversity. Nevertheless there exist difficulties with interpretation of the results that include (1) the existence of gene families, paralogs, gene copies and pseudogenes that have similar DNA sequences,

providing false positive hybridization; (2) the existence of orphan genes that are located in the complementary strand of intronic or flanking regions (Mironov et al., 1999a); (3) insufficient representation of expressed sequence data in public EST databases to identify all transcript isoforms.

We have taken an exhaustive approach to the detection of exon skipping from carefully annotated, protein-confirmed genes in order to maximize the accurate assessment of the degree of isoform diversity.

2.3 Results

In order to develop an unambiguous assessment of the degree to which exon skipping contributes to expressed transcript isoform diversity, and to assess the impact on protein coding of exon skipping events within coding regions of transcripts from known genomic loci, we have compared ESTs to 545 annotated genes on chromosome 22. Although no standard measure of relative spliceform frequencies for human genes exists, coverage of exon boundaries by ESTs provides a measure of the diversity of isoforms for a particular gene. The incidence of captured ESTs spanning exon junctions may also provide a reasonable, though uncomprehensive, view of transcript diversity and expression. Detection of transcripts displaying exon skipping was performed using novel software, j_explorer, which reduced the complexity of the gene sequences to a set of possible splice junctions which were used to search public EST databases to identify ESTs spanning the annotated exon-exon junctions. The software employs standard data format (EMBL sequence format) and visualization tools (ARTEMIS (Rutherford et al., 2000)) in the analysis (www.sanbi.ac.za/exon_skipping/). Removal of single and double exon genes reduced the set to 347 multi–exon genes (Table 1), of which 10 were previously annotated in

literature or public databases as having experimentally confirmed exon skips (Table 2). Exon skipping events were recorded when all original junctions involved in the skipping event, including flanking exons, were confirmed by EST sequences. All ESTs supporting exon skipping events were subsequently confirmed to be unambiguous transcripts of the corresponding gene and not products of paralogous genes, pseudogenes, or related members of an extended gene family by BLAST searches against the non-redundant (nr) database at NCBI. Highly specific identification of exon skipping and exon repetition events has resulted.

Sensitivity was assessed using the 10 genes with experimentally confirmed exon skipping. J_explorer accurately identified the previously reported skipped exons in 4 of the genes (NF2, ADSL, CLTCL and GGT1). Novel isoforms were detected in EWSR1, PLA2G6 and GGT1 (Table 2), while previously described exon skipping events in 4 genes (CACNA1I, BZRP, MTMR3, SEP3) were not detected because ESTs mapping to these exon junctions were not available in the public EST databases. The approach has a zero false positive rate, as confirmed by available mRNA and genomic data, and provides a solid basis for the development of models of transcript diversity that can be generated from a single gene.

We have discovered 62 exon skipping events in 52 genes (Table 2); 57 of the 62 (92%) exon-skipping events occur within the protein coding region. The remainder occur in either the 3'(1/62) or 5' (4/62) UTR. In 31/62 (50%) of cases the reading frame is maintained but regions are deleted. In 18/62 cases (29%) the introduction of a skip destroys the reading frame resulting in a frame shift. Proteins for the remaining 8/62 (13%) could not be reconstructed. In 5 cases an alternative stop codon is used, while in 4 cases there is an alternative start codon introduced.

Gene transcripts were scanned for exon repetition using similarity searching of repeated exon constructs against public EST data. A single tandem repetition of exon 2 of the F-box protein (NM_012179) was detected with high identity to EST AA569698. Exon repetition has previously been reported in a number of eukaryotes (De Lange et al., 1983; Boylan et al., 1990; Frantz et al., 1999).

Ratios of transcript isoforms are difficult to resolve using only EST data, however using the relative capture frequency of skipped exons as a measure provides an indication of the incidence of more commonly occurring isoforms (4 or more ESTs confirm the isoform with exon skipping in: CLTCL1, ADSL GGT1, GSTT1, HMG2L1, MFNG, dJ222E13.1) as compared to rarer isoforms. In 47/62 (76%) cases, the reference isoform, constructed from the genomic EMBL entry, is more frequently represented than a skipped exon isoform (Table 2).

The degree to which the level of gene expression, and hence database representation, affects the probability of finding a skipped transcript was assessed using the number of EST exon-exon junction captures per gene as a relative measure of transcript representation. Three categories comprising equal gene numbers were selected: low capture, which corresponds to less than 14 EST matches per gene, medium capture, those from 14 to 50 EST matches per gene, and high capture, those with 50 or more EST matches per gene (Table 3). 44 genes had no matches to ESTs. We found that over 60% of genes that demonstrate exon skipping have large numbers of ESTs matching to them. Although no relationship between degree of gene expression and extent of skipping can be determined from this study, the degree to which exon junctions are represented in transcripts reveals that highly represented genes demonstrate skipping more frequently. 10 of the 17 (58.8%) most highly represented

multi-exon genes show exon skipping and of these, 3 (18%) express more than one isoform (Table 2, www.sanbi.ac.za/exon_skipping).

2.4 Discussion

Our approach precisely identifies exon skipping when EST transcript data that spans exon boundaries is available. A possible limitation is that the detection of exon skipping using j_explorer is sensitive to the gene structure provided as input. Selection of the mRNA used to determine the gene structure will affect whether a skip can be detected. The number of ESTs that cover an exon-exon boundary determines the likelihood of discovering an exon skip, but capture of exon skipping events are dependent on the ratio of low abundance to high abundance isoforms of transcripts from the gene. The depth of transcript representation in EST databases, level of expression, and number and length of exons all contribute to the complexity of estimation of the number of genes which may have exon-skipped expressed transcripts. Estimation of the genome-wide extent of exon skipping is supported here by 52 of 347 multi-exon genes (~15%). This conservative estimate reflects the fact that only 68% of exon-exon junctions have EST coverage, and that this coverage is skewed towards over-representation of the 3' untranslated regions. In contrast, 58.8% of multi-exon, highly EST-represented genes present exon skipping. More sensitive transcript capture techniques may discover exon skipping to be far more widespread than the previous estimates of 10 to ~20% (Croft et al., 2000) (Mironov et al., 1999a) (http://industry.ebi.ac.uk/~thanaraj/gene.html) which have been based on EST frequency-independent measures. Expression studies will clarify the relationship between level of expression and degree of exon skipping in transcripts. The diversity of skipped exon transcript forms is likely to contribute significantly to the diversity of

protein products encoded by the genome, especially since the ratio of skipped isoforms of transcripts appears to vary widely, which is likely to have significant functional impact on the proteins for which they code. At least 50% of exon skips that we have detected result in in-frame deletions in the predicted protein products. In 29% of cases exon skipping results in a disruption of the reading frame which may change or disrupt the function of the protein product. Functional roles for these protein isoforms remain to be explored experimentally.

2.5 Methods

J-explorer (available for download from http://www.sanbi.ac.za/exon_skipping) was used to assemble exon-constructs from mRNA-annotated genomic sequences produced by the Human Chromosome 22 Sequencing Group at the Sanger Centre (Chr22.genes.dna file at http://www.sanger.ac.uk/HGP/Chr22/cwa_archive/Nature_02-12-1999/Chr22Genes.tar.gz). Using a 50bp tag from the 3' terminus of the preceding exon and a 50bp tag from the 5' terminus of all downstream exons a set of all consecutive and non-consecutive exon-exon junctions for each gene was created. Each junction was submitted for similarity searching against dbEST (human) using BLAST 2.0 (Altschul et al., 1990). By combining junctions in a consecutive (ie: exon 1 - exon 2 junction) and non-consecutive (ie: exon 1 – exon 3 junction) manner the incidence of exon skipping was assessed. A skipping event is reported when an EST is detected which does not contain the exon(s) in question, but does contain an uninterrupted tag made up of 50bp from each of the flanking exons. In cases where a flanking exon was less than 50bp in length j_explorer uses the entire short exon in building the construct. Exon repetition was investigated by creating splice junctions

composed of the concatenation of the 3' and 5' 50bp splice junctions of the same exon. ESTs showing significant ($P<1e^{-40}$) homology to an exon junction were extracted and aligned to the corresponding genomic sequence using sim4 (Florea et al., 1998). In order to exclude the possibility that ESTs confirming exon skipping events were the products of paralogous genes or members of gene families all ESTs identifying exon skipping were confirmed to be unique to a single target gene from Chromosome 22. Both interchromosomal and intrachromosomal specificity of the transcripts was confirmed using BLAST with a cut-off score of $1e^{-30}$. sim4 was employed where ambiguous matches were encountered. The resulting 'unambiguous transcripts', can therefore be unambiguously assigned to the correct gene of origin. The effect of these transcripts on the reading frame of the protein for which they code was assessed for frameshifts and in-frame deletions. The estimate that there exist zero false positives was arrived at by manual analysis of all cases of possible misalignments. During this process no false instances of reported exon skips were found. False positives could occur in genes with tandemly repeated exon structure, j_explorer provides an additional method for confirming that ESTs confirming the skip are not homologous to the reference gene structure. The identity and genomic location of each of the ESTs was converted into EMBL format and added as annotation to the relevant EMBL sequence file. Sequences were then analysed using ARTEMIS (Rutherford et al., 2000) and are presented together with supplemental information, annotated EMBL entries and links to ENSEMBL genes and transcripts at http://www.sanbi.ac.za/exon_skipping. All exon structure annotations for the genes used (both confirmed and predicted) were confirmed to be correct. In order to prevent the detection of skips as a result of incorrectly annotated exon boundaries we required that an EST spanning consecutive (or linear) exon boundaries was present in addition

to the EST/s confirming the skip. All linear junctions which could not be confirmed by ESTs resulted in that junction being excluded from further analysis. To address data consistency, we confirmed that in EMBL release 64 (GenBank 119) and 65 (GenBank 121) about 68% of splice junctions are covered with an EST. This figure does not vary significantly between the two releases.

2.6 Acknowledgements

The experimental approach taken in this study was developed by Dr Vladimir Babenko and implemented in software (j_explorer) by Peter van Heusden. Dr Cathal Seioghe was responsible for the statistical analysis of the results. I shared responsibility for data management including collating the gene set, addressing data quality issues, data processing, and extraction of and reporting on the output.

Table 1: Selection and Exon Structure of Genes for Study.

Number of multiple-exon genes selected for study	347
Number of exons	3240
Number of exon junctions	2893
Mean exon length	254 bp
Minimum exon length observed	8 bp
Maximum exon length observed	7660 bp
Maximum number of exons observed in one gene	54

We selected 347 multiple-exon genes of a total 545 genes present on chromosome 22 for study. Those removed included 134 single-exon genes and 64 double-exon genes that could not be assessed for exon skipping.

Table 2: Identification of Chromosome 22 genes with unambiguous transcripts of exon-skipped isoforms.

Locus name	Skipped exon(s)	Effect of exon skip	No. ESTs confirming exon skip	Average no. ESTs confirming reference isoform	Database annotation
J_explorer identified an experimentally confirmed isoform					
CLTCL1‡	29	C+	6	4.0	Clathrin heavy polypeptide-like 1
ADSL‡ *	12	3’	11	63.0	Adenylosuccinate lyase
NF2‡	2-3	C f/s	1	5.7	Neurofibromatosis 2 (bilateral acoustic neuroma)
J_explorer identified an experimentally confirmed isoform and a novel isoform					
GGT1‡	7 3	C f/s 5’	2 4	- -	Gamma-glutamyltransferase 1
J_explorer identified a novel isoform and not the experimentally confirmed isoform					
PLA2G6‡	3 +5	C f/s	2	1.5	Phospholipase A2 group VI
EWS‡ *	6	C+	1	48.0	Ewing sarcoma breakpoint region 1
Novel exon skipping events identified by J_explorer					
ATP6E *	2 5-7	C+ C+3’t	1 1	73.5 38.0	ATPase H+ transporting lysosomal (vacuolar proton pump) 31kD
MIL1	3	C+	2	13.0	Homo sapiens MIL1 protein

UFD1L *	2-3	C+3't	1	21.7	Ubiquitin fusion-degradation 1 like
TR	13	C+	1	6.0	Thioredoxin reductase beta
ARVCF	19	C+	1	1.5	Armadillo repeat gene deletes in velocardiofacial syndrome
AC005500.4	2-3	C+	1	5.4	Zinc finger protein
PIK4CA *	36-42	C f/s	1	5.6	Phosphatidylinositol 4-kinase catalytic alpha polypeptide
BCR	20	C f/s	1	5.5	Active BCR-related gene
AP000350.2	5	C+	2	1.0	Similar to glucose transporters SW:P22732
	2	C f/s	4	7.5	
GSTT1	2-3	C f/s	1	8.0	Glutathione S-transferase theta 1
	3-4	C+	1	9.3	
AC004997.1†	5	C N/A	1	4.0	GATS protein
	5-6	C+	1	4.0	
SEC14L2	10	C f/s	2	4.5	SEC14 (S. cerevisiae)-like 2
SMTN†	14-15	C N/A	1	5.7	Smoothelin
dJ858B16.1	27	C+ 3't	1	2.3	Homo sapiens mRNA for KIAA0542 protein complete cds.
AC005004.1	22-23	C+	2	1.7	Homo sapiens mRNA for KIAA0645 protein complete cds
HMG2L1	2	5'	4	1.5	High-mobility group protein 2-like 1
			1		

	5	5’	1	4.0	
	2 + 5	5’	1	4.6	
CE132D12.1	6	C f/s	1	21.5	Similar to RAS-related protein RAB-5A (HS)
MFNG† *	7	C f/s	5	46.0	Manic fringe (Drosophila) homolog
	2	C+	1	16.0	
LGALS1 *	3	C f/s	5	>100.0	Lectin galactoside-binding soluble 1 (galectin 1)
GCAT	2-3 + 5	C N/A	1	2.8	Glycine C-acetyltransferase (2-amino-3-ketobutyrate-CoA ligase
dJ1014D13.1 *	12	C+	1	>100.0	Weakly similar to casein kinase I homologue HRR25
GTPBP1	2	C+ 5’t	8	17.0	GTP binding protein 1
dJ508I15.1	2	C + 5’t	1	8.0	Novel human gene mapping to chromosome 22
dJ508I15.4	3	C N/A	1	1.5	Homo sapiens mRNA for KIAA0668 protein
RPL3† *	8	C+	7	>100.0	Ribosomal protein L3
dJ1042K10.2	2	C+ 5’t	1	10.5	Similar to C.elegans predicted protein with probable rabGAP domains and src homology
SLC25A17	2-4	C f/s	1	9.0	Solute carrier family 25 (mitochondrial carrier-peroxisomal membrane protein 34kD) member 17
	3-4	C+	2	9.0	
ST13 *	8	C f/s	1	17.5	Suppression of tumorigenicity 13 (Hsp70-interacting protein)

RBX1 *	2	C f/s	3	50.5	Ring-box 1
	3-4	C+ 3't	1	92.0	
PMM1	4	C f/s	1	24.5	Phosphomannomutase 1
TCF20 (AR1)†	3	C N/A	1	5.0	Transcription factor 20 (AR1)
dJ222E13.3†	7	C f/s	2	12.0	Weak match to Arabidopsis RNA and export factor binding protein
dJ222E13.1	8-9	C f/s	5	2.0	Novel protein with some similarity to Drosophila KRAKEN
bK1191B2.3†	3	C+	4	2.0	Weakly similar to dJ1118 COA-ACYL carrier protein transacylase
dJ796I17.2 *	3	C+	1	27.0	CGI-51
NPAP60L	4	C N/A	1	11.5	Nuclear pore-associated protein 60L
dJ355C18.1	9	C+	1	1.5	Matches KIAA0027 gene with weak similarity to GTPase activating protein
ECGF1	5	C N/A	1	3.0	Endothelial cell growth factor 1 (platelet-derived)
GTSE1 (B99)	8	C f/s	1	13.5	Homo sapiens G-2 and S-phase expressed 1 (GTSE1),
dJ1163J1.4†	3	C+	1	1.0	Novel protein similar to C. elegans B0035.16 and bacterial tRNA (5-Methylaminomethyl-2-thiouridylate)-Methyltransferases

DGCR2	2-3	C+	1	2.3	DiGeorge syndrome critical region gene 2
AC007050.6	2	C N/A	1	11.0	Homo sapiens mRNA- from clone DKFZp434G1017
UBE2L3 *	2	C+ 5't	1	71.0	Ubiquitin-conjugating enzyme E2L3
DJ756G23.3	5	C+	1	2.0	Similar to Tr:Q24191 Drosophila TRANSCRIPTIONAL REPRESSOR PROTEIN
bK212A2.1	2	C+ 3't	1	-	TNF-inducible protein CG12-1 mRNA
G22P1 *	3	C+	1	>100.0	Thyroid autoantigen 70kD (Ku antigen)

We tested 347 multiple-exon genes on Chromosome 22 for exon-skipping events using J_explorer and EST sequences from GenBank 119. Genes in which novel exon skipping events have been identified are ordered according to their relative physical organization along chromosome 22. Genes are identified using the HUGO name if one exists. In the absence of a HUGO identifier, the accession number of the sequence or the Sanger Centre clone name is used. Exon numbering is based on the exon structure of the original EMBL entries obtained from the Sanger Centre. ESTs confirming a skip were required to span both the 3' and 5' flanks of the skipped exon. To calculate the average number of ESTs confirming the reference isoform, the exon flanking ESTs in the reference isoform were totalled and the sum divided by corresponding averaged number of junctions. In cases where the reference isoform was not represented in the public EST databases, the sequence was confirmed using a corresponding experimentally-determined mRNA. Skip location and context is denoted as follows: (C) skip occurs in protein coding region; (+) ORF remains unchanged; (3') skip occurs in 3' UTR; (5') skip occurs in 5' UTR; (f/s) frameshift is introduced by skip; (5't) alternative start codon is used; (3't) alternative stop codon is used; (N/A) not possible to reconstruct a protein; (†) genes (eight entries total) with an already-annotated exon skip in EMBL entries; (‡) (six entries total) with experimentally-confirmed notation: 2–4 indicates that exons 2, 3 and 4 skipped exons; 2–3 +5 indicates that exons 2, 3 and 5 are skipped. Experimentally confirmed skipping events in the genes CACNA1I, BZRP, MTMR3, and SEP3 had no EST matches and are not included.

Table 3. Capture of exon skipping relative to expression representation.

Number (n) of ESTs matching exon junctions per gene (interval)	Number of genes	Number of genes with skips detected by j_explorer
0<n<14	101 (33%)	4 (8%)
14≤n<50	101 (33%)	16 (31%)
n≥50	101 (33%)	32 (61%)
Total	303	52

Chapter 3

eVOC: A Controlled Vocabulary for Unifying Gene Expression Data

3.1 Abstract

Expression data contribute significantly to the biological value of the sequenced human genome, providing extensive information about gene structure and the pattern of gene expression. The EST databases have been a central repository for increasing amounts of expression data since 1991. Together with SAGE libraries and microarray experiment information these provide a broad and rich view of the transcriptome. However, it is difficult to query data generated by these diverse experimental approaches, and even more difficult to perform large-scale mining of expression data. Not only is it stored in disparate locations using different platforms and with different conceptual organisation, but there is also frequent ambiguity in the meaning of basic terms used to describe the biological source of the material used for the experiment. Untangling semantic differences between the data provided by different resources is therefore largely reliant on the skill and domain knowledge of a human user. We have developed a system which associates sample reagents such as labelled target cDNAs for microarray experiments, or cDNA libraries and their associated transcripts or genes with controlled terms in a set of hierarchical vocabularies. We present eVOC - four orthogonal controlled vocabularies to describe domains of human gene expression data including Anatomical System, Cell Type, Pathology and Developmental Stage. We have manually translated the inconsistent terminology used

to describe the library source into controlled terms in the four orthogonal ontologies, and have curated and annotated 7016 cDNA libraries represented in dbEST, as well as 104 SAGE libraries, with expression information. We provide this as an integrated, public resource that allows the linking of transcripts and libraries with expression terms. Downstream applications include the analysis of tissue expression profiles and specificity, gene expression levels, and the comparison of expression between species. Both the vocabularies and the vocabulary-annotated libraries can be retrieved from http://www.sanbi.ac.za/evoc/ and are applied within Ensembl (http://www.ensembl.org/Homo_sapiens/martview) to provide a standard for linking expression phenotype information with the genome sequence. Several groups are working to provide shared development of this resource such that it is of maximum use in unifying transcript expression information.

3.2 Introduction

Mining of large volumes of transcriptome data is currently frustrated by an inability to relate sequence and descriptive information. In part this is due to the absence of a common structured vocabulary to describe the source of the biological sample materials.

Recent years have seen a growing trend towards the adoption of ontologies for the management of biological knowledge. In Computer Science an ontology is defined as an "explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest, and the relationships that hold among them" (The Free Online Dictionary of Computing http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi?query=ontology).

Biological ontologies aim to overcome the semantic heterogeneity commonly encountered in molecular biology databases, and to provide a common terminology for the description of a focussed aspect of biology. One such resource, TAMBIS (Stevens et al., 2000), implements ontologies for both bioinformatics tasks and molecular biology to provide users with transparent access to multiple heterogeneous bioinformatics resources. The Gene Ontology Consortium (GO) - the group largely responsible for raising the profile of ontologies in biology - provides a set of generic ontologies for the description of core biological functions (Ashburner et al., 2000). Extensive functional conservation of proteins across the eukaryotes means that a common set of terms can be applied to the description of genes and gene products in order to provide information on the roles of orthologous proteins in novel organisms. The GO ontologies provide terms describing three aspects of biological function: biological process, molecular function and cellular compartment. These can be applied independently, and it is recognised that the relationship between a gene product and the ontologies is one to many – as proteins may function in several processes, or carry out a multitude of molecular functions in alternate cellular locations. Other biological ontologies include the EcoCyc ontology (Karp et al., 2002b) which represents important metabolic and signal-transduction events in *E.coli,* and the MetaCyc (Karp et al., 2002a) and KEGG (Kanehisa et al., 2002) ontologies which describe aspects of the relationships between the chemical reactants, catalysts, substrates and products. Numerous other ontologies representing a wide array of biological phenomena exist or are under development.

Although several ontologies for the formal description of sample materials exist or are under development (Table 1), these are not suitable for querying gene expression data. For example; clinical ontologies including anatomical, pathological and

developmental stage-specific concepts have been available for some time (ICD-9-CM, SNOMED, GALEN, MeSH) but these have not been widely adopted for describing human gene expression profiles. A major reason why clinical ontologies are not widely used for describing gene expression is that they are extremely detailed and often tangled (Rector et al., 2001), with distinct concepts with varying relationship types mixed together, making them unwieldy and difficult to adopt for general use. An example is the mixing of anatomical and pathological terms in ICD-9-CM *eg:* "benign neoplasm of the stomach". The complexity of the concepts represented by these ontologies makes them unsuitable for the computational interrogation of gene expression data to determine simple and complex expression profiles.

Implementing multiple ontologies with simple concepts in orthogonal domains provides a preferable solution as it enables users to produce logical ontology cross-products. Cross-products are hybrid ontologies which can be constructed through the combination of simple ontologies. For example: the ICD-9-CM term mentioned above could have been constructed through the combination of terms from an anatomical and a pathological ontology by producing the cross-product of the terms "stomach" and "neoplasm | benign" from the respective ontologies.

Ideally, ontologies for gene expression should reflect a level of detail appropriate to the data being classified and the level at which queries are likely to be performed, while simultaneously providing sufficient flexibility to enable regular updating without needing to significantly restructure the hierarchies.

For the extensive description of gene expression and to provide maximum flexibility in querying we have developed eVOC - four orthogonal ontologies which aim to provide an appropriately detailed set of terms for describing the sample source of

cDNA and SAGE libraries and labelled target cDNAs for microarray experiments. We have taken a data-driven approach to determining the level of granularity required.

We have annotated all publicly available human cDNA and SAGE libraries as extensively as possible. This is achieved by the assignment of terms from each of the four ontologies to the libraries. Initial assignment of terms to libraries was performed computationally, with curators who are domain experts performing assessment of annotation quality and further manual assignment. Where information was lacking in the library record the original submitters were contacted where possible to provide more extensive information.

The most widely used ontology for keywording human SAGE and EST libraries is the CGAP/UniLib vocabulary (ftp://ftp.ncbi.nih.gov/pub/bioannot/info/keys) currently used by the NCI to categorise libraries for CGAP (http://www.ncbi.nlm.nih.gov/CGAP/).

CGAP provides a single integrated hierarchy of keywords which includes terms from multiple classification domains (including tissues, developmental stage, library preparation and chemical agents among others). There are many different relationships between parent and child terms in different sections of the hierarchy. eVOC, by contrast, provides completely orthogonal ontologies covering four distinct domains. There is a single implied type of relationship between the terms within each of the eVOC ontologies.

The structure of the CGAP ontology enables rapid keyword searching, whereas the eVOC data structure, by incorporating the rigorous separation of classification terms into orthogonal domains and the formalisation of relationships between terms, allows

for a degree of computer reasoning to be applied. This facilitates a wide range of query types. For example, a comparison of eVOC and UniLib querying shows clearly that both eVOC and UniLib allow querying for multiple terms combined with "AND" (the intersection set), and yield comparable results in terms of the libraries returned. However, UniLib is unable to support more complex queries incorporating "OR" and "NOT" which are possible with eVOC. eVOC therefore provides users with greater flexibility as more complex biological queries can be formulated. While this may be a simple implementation issue, it is one which directly affects the user interaction with the data.

A major distinction between CGAP and eVOC is that the CGAP hierarchy is cancer-specific by design. The terms included are therefore those of interest in cancer whereas eVOC is designed for more general application. Specifically, CGAP lacks the comprehensive pathology terminology which is necessary for a broadly applicable human expression ontology.

3.3 Methods and discussion

The design and creation of the expression ontologies is distinct from the annotation of cDNA and SAGE libraries using each of the ontologies. These processes will be discussed separately.

3.3.1 Development of a data structure for expression ontologies

The expression ontologies have been developed in four orthogonal (mutually exclusive) knowledge domains including Anatomical System, Cell Type, Developmental Stage and Pathology (Appendix I). Anatomical System and Cell Type describe where a gene is expressed, Developmental Stage describes the timing of gene

expression during development, and Pathology describes the disease state in which the gene is expressed. These four ontologies were found to represent the vast majority of the expression data currently under classification. The addition of further ontologies may be appropriate in the future.

The expression ontologies are independent pure hierarchies (or trees). In a pure hierarchy, each node has only one parent but may have multiple children. Each node is associated with a specific concept in the knowledge domain represented by the hierarchy through the association of each node with one or more synonymous terms. For example, the terms “nasal” and “nose” are synonyms attached to a single node in the anatomy ontology.

In these pure hierarchies there is only a single type of relationship between the nodes in each hierarchy, although the nature of the relationship is not explicitly defined. For each ontology, the nature of the expression domain imposes an implicit type on the relationship between the nodes. For instance, in the “Anatomical System” ontology, the relationships are of the “part-of” type. In the “Cell Type” and “Pathology” ontologies, they are of the “subclass” type, and in the “Developmental Stage” ontology, the relationships are of the “is-a” variety.

Pure hierarchies have a number of advantages over the more complex data structures often used to represent ontologies (Rector et al., 2001). They are easy to maintain and expand and they can be visualised easily. Moreover, it is possible to construct a simple yet extremely powerful and flexible mechanism to query data across multiple hierarchies.

In cases where terms appear to have more than one parent, two options are available; migration to a directed acyclic graph (DAG), or untangling of the hierarchy to yield a

pure hierarchy (Figure 1). In order to handle multi-parent terms and different parent-child relationships the GO project (Gene Ontology Consortium, 2001) has implemented a DAG structure. During the development of the eVOC ontologies, and based on the available cDNA and SAGE libraries, we have found that where it appears that there is a need to represent multiple relationship types in one hierarchy it is possible to untangle the hierarchy further by splitting it into separate hierarchies with more narrowly defined relationship types.

The disadvantage of maintaining untangled orthogonal ontologies is that the volume of work involved in curation increases linearly with the number of hierarchies. It is therefore necessary to strike a balance between keeping the number of ontologies manageable, and representing relationships in as fine-grained a fashion as possible. The sort of queries the ontologies are required to accommodate dictates where this balance is found. In other words, the ontology design should be data-driven.

Each of the terms in the ontologies has a numeric identifier which uniquely identifies the term and which can be used as an unambiguous database cross-reference. Definitions of each of the terms are to be provided as part of the ongoing development. The source of each definition will be made available, along with the definition.

3.3.1.1 Development of the four expression ontologies

The four expression ontologies (Figure 2) currently implemented are:

3.3.1.1.1 Anatomical System ontology

The Anatomical System ontology provides a controlled vocabulary for the description of the anatomical system or organ in which a gene is expressed. It is based on the

controlled vocabulary used in the Computational Biology and Informatics Laboratory's (CBIL) databases (www.cbil.upenn.edu/anatomy.php3) but with modifications including the removal of all references to tissue type, cell type or developmental stage. Organisation of the Anatomical System hierarchies is currently systems-based. Examples of broad Anatomical Systems are "digestive system" or "nervous system", with more specific anatomical terms within these systems being "pancreatic islets" or "retina". Future developments to eVOC will include the creation of an Anatomical Site ontology that extends the current Anatomical System ontology by dividing anatomical parts according to their spatial position, rather than according to the system to which they belong. This is of particular value in describing libraries from spatially distinct anatomical sites containing multi-system anatomical sites. For example "head" is a distinct anatomical site, but includes both nervous and circulatory systems. The Anatomical System ontology contains 372 terms.

3.3.1.1.2 Cell Type ontology

The Cell Type ontology provides a fine-grained description of where a gene is expressed. It is a listing of human cell types extracted from Gray's Anatomy (Gray et al., 1995). The Cell Type ontology includes 153 different cell types.

Since various cell types are represented across many anatomical systems, cell types could have been included in the Anatomical Site ontology, with cell type terms having multiple parents. Instead we have separated the Anatomical System and Cell Type ontologies in order to maintain pure trees. This separation provides users with greater flexibility, as they can query on specific cell types, regardless of the anatomical location, and can also perform combined queries across Cell Type and Anatomical System terms to yield results for a cell type in a specified location.

3.3.1.1.3 Developmental Stage ontology

The Developmental Stage ontology provides an ordered timeline of human development for the description of gene expression in temporal space. Examples of terms in the current hierarchy include "embryo" and "adult". Embryogenesis is further divided into the standard Carnegie stages (www.ana.ed.ac.uk/anatomy/database/humat/) which define the first two months of human development. Each of the major stages of development is further divided into appropriate weekly and yearly categories (Supplementary Table 1c). The Developmental Stage ontology contains 132 distinct terms.

3.3.1.1.4 Pathology ontology

The Pathology ontology is loosely based on the World Health Organisation's ICD-9-CM (www.mcis.duke.edu/standards/termcode/icd9/1tabular.html). ICD-9-CM is designed for the classification of morbidity and mortality information for statistical purposes and for the indexing of hospital records by disease and surgical operations. We have implemented a modified version of the first two levels of this hierarchy, and have incorporated terms that are widely used in sample description, but which are not present in ICD-9-CM *e.g*. Wilm's tumor. We have also removed terms that refer to systems, organs, tissues and cell types as these are already included in the Anatomical System and Cell Type ontologies. The Pathology ontology contains 141 terms.

3.3.1.2 Species-specific considerations

The broad domains covered by eVOC's four orthogonal hierarchies are sufficiently generic to be applicable to a wide and diverse variety of eukaryotic organisms. However, given that each organism has unique tissue organisation, development and disease processes, organism-specific ontologies are appropriate for expression data.

For instance, an extensive mouse-specific expression ontology, the Mouse Anatomical Dictionary, has been collaboratively developed by the Jackson Laboratories and the Edinburgh Mouse Atlas project (http://www.informatics.jax.org/searches/anatdict_form.shtml).

There is significant value in being able to identify and relate "equivalent" tissues in different species, and to compare gene expression patterns in these tissues. While it is not clear that it will always be possible to identify these equivalent tissues in the model organisms, the production of species-specific ontologies to form the basis of these comparisons is the first step. To facilitate interoperability between species-specific ontologies these need to be in a compatible, accessible format (Bard and Winter, 2001). The eVOC human expression ontologies are provided in a format which promotes easy adoption and which will facilitate the interrogation of cross-species ontologies from different sources.

3.3.1.3 Curation of the eVOC ontologies

We maintain a central, versioned database of eVOC ontologies which are updated, modified and released publicly, by domain-experts on an ongoing basis. The curators have the ability to add or delete terms and synonyms and to make changes to the hierarchies.

Groups that choose to modify the ontologies for their own purposes are encouraged to contribute their modifications and corrections to the curators for inclusion. A mailing list: evoc@sanbi.ac.za has been established for this purpose.

3.3.2 Annotation of cDNA and SAGE libraries using eVOC

The ontologies presented here are independent of the expression data that they are used to annotate. We have already annotated publicly available cDNA and SAGE libraries using these expression ontologies, Supplementary Tables 1a-d (available from http://www.sanbi.ac.za/evoc/) provide statistics for the number of libraries and ESTs annotated with specific terms in each of the ontologies. Figure 3 provides an example of the annotation of cDNA libraries in a subsection of the Pathology ontology. The eVOC ontologies are also highly appropriate for the annotation of labelled target cDNAs for microarray experiments.

cDNA and SAGE libraries are collections of the transcribed sequences expressed in the biological sample material from which the library is prepared. Information about the source of the sample is stored with the library information. The amount and quality of the source information provided varies depending on the source of the library. Libraries submitted to public databases are described using highly inconsistent terminology. Here curators have manually translated the unstructured terms used in the library records into standardised terms selected from the four ontology domains, and have applied these to each of the libraries. Ideally an ontology-based form would guide submitters in selecting appropriate terms for the description of their libraries. This would reduce the curation required and facilitate querying of the public databases in a manner not currently possible.

Each of the cDNA and SAGE libraries was assigned computationally to the most specific possible terms in each of the four ontologies. Manual curation and annotation of the computational assignments was then performed. Libraries are annotated with terms in each of the four hierarchies if sufficient information is available in each of

the ontology domains. Annotation of a library in one ontology is completely independent of annotation in another ontology. Each annotation is transferred from the library information provided by the original submitter. While the curators exercise domain expertise in assigning libraries to specific terms within each hierarchy, they derive no new information. This process is therefore largely objective. Evidence for annotations is primarily based on the original submission record for both cDNA and SAGE libraries.

In most instances annotation of data from existing databases is performed following the development of ontologies. Appropriate terms are assigned to data points on the basis of information already present in the database. This “post-facto” approach results in an often-imperfect mapping between data and terms as much of the sample information is not provided in the original submission and is therefore lost. The Ontologies Working Group of the MGED Consortium is providing ontologies which supply standardised terms for the annotation of microarray experiments. These will be offered as resources for meeting the guidelines laid down by MIAME. The MIAME guidelines specify the minimum information required about a microarray experiment in order to interpret, analyse and verify microarray data. The MIAME specification lists a broad range of information about a microarray experiment (called MIAME 'concepts') that should be captured. One such concept is that of the Biomaterial – the biological material from which the nucleic acids were extracted for subsequent labelling and hybridisation. eVOC is appropriate for use in the description of aspects of the Experimental Sample, or BioSource used on the array, when the source is human tissue. The philosophy of MGED is to provide a number of ontologies and allow the user to select the most appropriate ontology for their application. The eVOC ontologies are being offered as choices for describing the MIAME BioSource

properties. A challenge is ensuring that the depth of the terms provided by the eVOC ontologies is sufficiently detailed to cover the requirements of microarray experiments, while remaining of a size and complexity that is appropriate for human browsing and use. In addition there may in future exist a need for further ontologies addressing aspects of sample treatment and preparation. The implementation of a similar ontology-based data entry system for the public nucleotide databases would be of immense value for the submission of cDNA and SAGE library information.

The clone libraries annotated here are generated from biological sample materials representing specific expression states (e.g.: infant lung). These libraries represent collections of each of the transcripts expressed in the original sample. The transcripts expressed in the original biological sample can therefore be sequenced as ESTs from the clone library. By mapping the clone libraries to a set of controlled terms (the ontologies) all the ESTs from each clone library can be transitively linked to these same standardised terms in the relevant ontology via their association with their parent clone library. In the case of ESTs we maintain a database for the bi-directional accession to clone library lookup which in turn allows us to link vocabulary terms directly to ESTs (Figure 4).

We have annotated 7016 human cDNA and 104 human SAGE libraries with the eVOC expression ontologies. These represent all the human cDNA and SAGE libraries that were publicly available in April 2002. The amount of information provided for each library varies widely. In some cases extensive information about the anatomical system, developmental stage and pathological state of the sample source is provided, while in other cases only a subset of this information is provided. The majority of the cDNA libraries (94.8%) have the information required for

classification in the Anatomical System ontology and most have information required for annotation with Pathology (82.7%) and Developmental Stage (89.9%) terms (Table 2). Where libraries were unable to be annotated this was because the library information provided by submitters did not capture the relevant information. As a result of the fact that cDNA and SAGE libraries are largely derived from whole organs and tissues rather than from individual cell types the majority of the libraries (94.2%) could not be annotated using the Cell Type ontology.

3.3.3 Using the ontologies

3.3.3.1 Querying

Untangled hierarchies allow for the implementation of a very simple query schema. A query for a particular term returns the node with which that term is associated, as well as all the nodes in the entire subtree (branch) rooted at that node. For instance, a query for the term "neoplasia" returns a particular node in the Pathology ontology, as well as all of its children, recursively. The next step in building a useful querying system lies in utilising the mappings from nodes to public databases (for example, cDNA libraries). In this way, a query for a particular term is translated first to a node, then expanded to a set of nodes, and then translated to a set of cDNA libraries. The set of libraries includes all the libraries associated with all the nodes in the branch rooted at the node which was originally associated with this node.

This simplistic query methodology can be the basis of an enormously powerful query infrastructure if the ability to perform basic set algebra (union and intersection) operations on the returned sets of cDNA libraries is used.

Consider, for instance, the query "liver AND neoplasia" (Figure 5). A query on "liver" resolves to a node in the Anatomical System ontology, which in turn results in a set of cDNA libraries (all the libraries associated with the "liver" node and all its subnodes). Similarly, a query on "neoplasia" returns the set of cDNA libraries associated with a subtree of the Pathology ontology. The combined query – "liver AND neoplasia" – returns the intersection of these two sets of cDNA libraries. In other words, it will return only libraries which were constructed from neoplastic liver samples.

3.3.3.2 Example applications

The ontologies and the associated annotated cDNA and SAGE libraries have a wide array of applications.

Through simply curating dbEST using the eVOC ontologies users are provided with the ability to perform queries based on location, state and timing of expression on human ESTs or cDNA libraries. Querying using terms from any combination of the ontologies, both libraries and transcripts can be selected from the database based on their expression patterns. Moreover, the differential expression of genes or gene isoforms based on EST data can be determined swiftly and accurately by providing a list of EST accessions and analysing the distribution of terms attached to each EST.

Laboratory based applications of eVOC include the selection of clone libraries relevant to laboratory research projects; for example: a simple query which returns the total number of publicly available retinal cDNA libraries yields 22 results (Figure 6). To select suitable libraries for the comparison of gene expression in adult and fetal retina further refined queries can be used to show that 7 libraries are derived from

adult retina, 3 are derived from fetal retina, and that 12 libraries do not have information about the developmental stage from which the retinal issue was isolated.

Similarly the number of cDNA libraries available for pancreatic tissue yields 31 results. To determine how many of these are pancreatic islet libraries a second query is performed and yields a total of 10 pancreatic islet libraries which have source descriptions as diverse as "Human insulinoma" and “"HR85 islet".

Additionally, the ability to identify cDNA and SAGE libraries from similar expression states provides access to an increased resource for data-mining, and allows users to identify and analyse genes which are differentially expressed both in their expression location and their expression level. We have used the system to identify neoplastic and normal cDNA libraries, and have identified differential gene expression and alternative splicing in these expression states.

To illustrate the power of expression ontologies in determining the tissue-specificity of alternatively spliced transcripts we have analysed the data produced by Xu *et al*. (Xu et al., 2002) who performed a genome-wide detection of alternatively spliced transcripts and identified those which show tissue-specificity. To determine the tissue-specificity of the spliceforms Xu *et al.* classified 4271 (~60%) of the publicly available cDNA libraries according to a flat list of 46 human tissue classes. This classification was used to determine the tissue distribution of alternatively spliced transcripts, identifying 667 tissue-specific alternative spliceforms. Since in the eVOC system cDNA libraries are classified according to a more detailed hierarchical vocabulary and because the classification is according to four orthogonal ontologies it is possible to extend the information already derived regarding the tissue-specific isoforms identified by Xu *et al.*.

We submitted the isoform-specific EST lists provided for a subset of the genes identified by Xu *et al.* as having tissue-specific isoforms to eVOC in order to determine the expression profile of each isoform according to each of the four eVOC ontologies: Anatomical System, Cell Type, Developmental Stage and Pathology (Table 3). We were able to duplicate the tissue-specificity results described previously by comparing the expression profiles of each isoform delivered by eVOC with the published tissue-specificity. Additionally we were able to derive more information about the Pathology and Developmental Stage specificity of these isoforms. For example: IRP3 was described by Xu *et al.* as having a brain-specific isoform. Additional information provided by the Developmental Stage ontology in eVOC showed that this isoform has only been observed in the infant brain.

By implementing a set of orthogonal, hierarchical controlled vocabularies eVOC provides a detailed and flexible system for the detection of expression-state-specific spliceforms. eVOC can be used to identify not only tissue-specific spliceforms, but also splicing which is specific to certain developmental stages, cell types and pathological states, or any combination of these states.

3.3.3.3 Future applications

The eVOC ontologies have been implemented as part of a candidate disease gene profiling tool which uses expression information in conjunction with other evidence to prioritise disease gene candidates within specified regions of the genome (manuscript in preparation).

3.3.4 Availability and interfaces (editing and graphical browsing)

eVOC is provided under a BSD-style licence and is available for download free of charge from http://www.sanbi.ac.za/evoc/ and can be used and modified without restriction. From the website users are also able to download the annotated datasets, join the Expression Vocabulary Consortium and sign-up to use the eVOC mailing list.

Although genomic information is not directly integrated into eVOC, users have the ability to integrate the expression information within eVOC with human genome information through the transitive mapping of ESTs (generated from the clone libraries which are mapped to eVOC) to the genome. This functionality is being provided through the integration of eVOC with the EnsemblMart data mining resource which is part of the Ensembl Project at EBI. The eVOC ontologies will be available in the January 2003 release of the EnsemblMart database (http://www.ensembl.org/Homo_sapiens/martview). EnsemblMart is a data retrieval tool which provides users with the ability to build queries of the biological data (including genome sequence and annotation data) present in the Ensembl genome database. Since ESTs have been mapped to the genome by Ensembl, eVOC terms can be linked transitively (via their parent clone library which is mapped to the eVOC ontologies) to the genomic sequence. As a result users will be able to perform expression-based queries in the context of genomic data and will be able to extract transcripts and genes based on the location, state and timing of their expression.

A graphical interface for querying eVOC has been developed by Electric Genetics (Figure 3) and is available from info@egenetics.com. This interface provides users with the ability to view the ontologies, browse the hierarchical trees and to perform set operations on the annotated cDNA library data. Using this interface it is possible

to obtain the list of cDNA libraries or ESTs returned by a query, or to provide a list of libraries or EST accessions and obtain the associated expression profile. The interface will be extended to include curation facilities, simplifying the users ability to modify the existing eVOC ontologies or create de novo ontologies of their own. In addition Electric Genetics has developed an API which provides the ability to develop custom software to interface eVOC with external data repositories, and to perform complex ontological queries on that data.

3.4 Summary

We have presented here a set of ontologies for the description of gene expression data, and have provided a database of the mappings between these ontologies and public cDNA and SAGE libraries. These have been applied successfully in retrieving expression information about ESTs from public databases, selecting clone libraries from particular expression states and in the detection of expression state-specific alternative spliceforms.

The simple orthogonal ontologies are flexible and extensible, making them applicable to real data and allowing them to be both machine and human-readable. The ontologies are under continual development; existing ontologies are extended and altered, appropriate new ontologies are added, and the annotation of expression libraries is regularly updated. Both the ontologies and the annotated expression libraries are publicly available and able to be freely adopted, modified and integrated for both novel and existing applications. The wide number of potential applications makes eVOC a valuable resource for the biologist.

3.5. Acknowledgements

I was responsible for co-developing the ontologies, for ongoing modification of the terms and structure of the ontologies, for the manual annotation of the EST and SAGE libraries with terms from each of the four ontologies, as well as for the querying and example applications presented in the paper. Many others were involved in the development of eVOC. Specifically, Gregory Theiler developed scripts to perform the automated annotation of cDNA libraries and contributed alterations to the ontologies. Development of the eVOC ontology data structure and query language was performed by Johann Visagie. Dr Soraya Bardien and Dr Alan Christoffels were responsible for the initial set of eVOC ontology terms and early annotations. Damian Smedley was responsible for the integration of eVOC with EnsMart.

Table 1. Existing ontologies which are relevant to human expression data.

	Website	Scope
CBIL	http://www.cbil.upenn.edu/anatomy.php3	Adult anatomy
Cytomer	http://www.biobase.de/pages/products/cytomer.html	Human developmental anatomy
HUMAT	http://www.ana.ed.ac.uk/anatomy/database/humat/	Human developmental anatomy
EPOdb	http://www.cbil.upenn.edu/EpoDB/release/version_2.2/controlled.vocab.html	Human anatomy, developmental stage, cell type,
GeneX	http://www.ncgr.org/genex/	Human gene expression
MeSH	http://www.nlm.nih.gov/mesh/meshhome.html	Clinical ontology
UMLS	http://www.nlm.nih.gov/research/umls/umlsmain.html	Clinical ontology
GALEN	http://www.opengalen.org/	Clinical ontology
SNOMED	http://www.snomed.org/main.html	Clinical ontology
ICD-9-CM	http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm	Clinical ontology

Table 2. **Total number of annotated cDNA and SAGE libraries in each ontology. Most libraries can be annotated with Anatomical System terms as these are generally present in the library record. Less information is available for Cell Type and Developmental Stages as these are not consistently captured during the capture of library information.**

	Total Libraries	Annotated Libraries	Not Annotated
Anatomical System	7120	6752	5.2%
Cell type	7120	410	94.2%
Developmental Stage	7120	5891	17.3%
Pathology	7120	6401	10.1%

Table 3. **eVOC extends the expression information that can be obtained from other sources. IRP3, described by Xu et al. as having a brain-specific isoform, was shown to be infant brain specific by combining information gathered from the eVOC ontologies. The ESTs for each isoform were submitted to eVOC and the associated terms in each of the four ontologies were examined to identify expression state specificity. Five of the six ESTs from distinct cDNA libraries were found to support the brain-specificity reported by Xu et al. Further, using eVOC four of the six libraries had been annotated with developmental stage information and this was used to confirm that isoform 1 of IRP3 is only observed in infant libraries.**

Gene Name	Isoform 1		Isoform 2	
	Xu et al.	**eVOC**	**Xu et al.**	**eVOC**
IRP3	Brain-specific	5 nervous >brain 1 respiratory >lung	No specificity	2 urogenital >genital >female >uterus 1 urogenital >genital >female >placenta 1 haematological >blood
		4 infant		3 adult
WNK1	Kidney-specific	7 urinary >kidney	No specificity	2 urogenital >genital male >penis 1 alimentary >pancreas

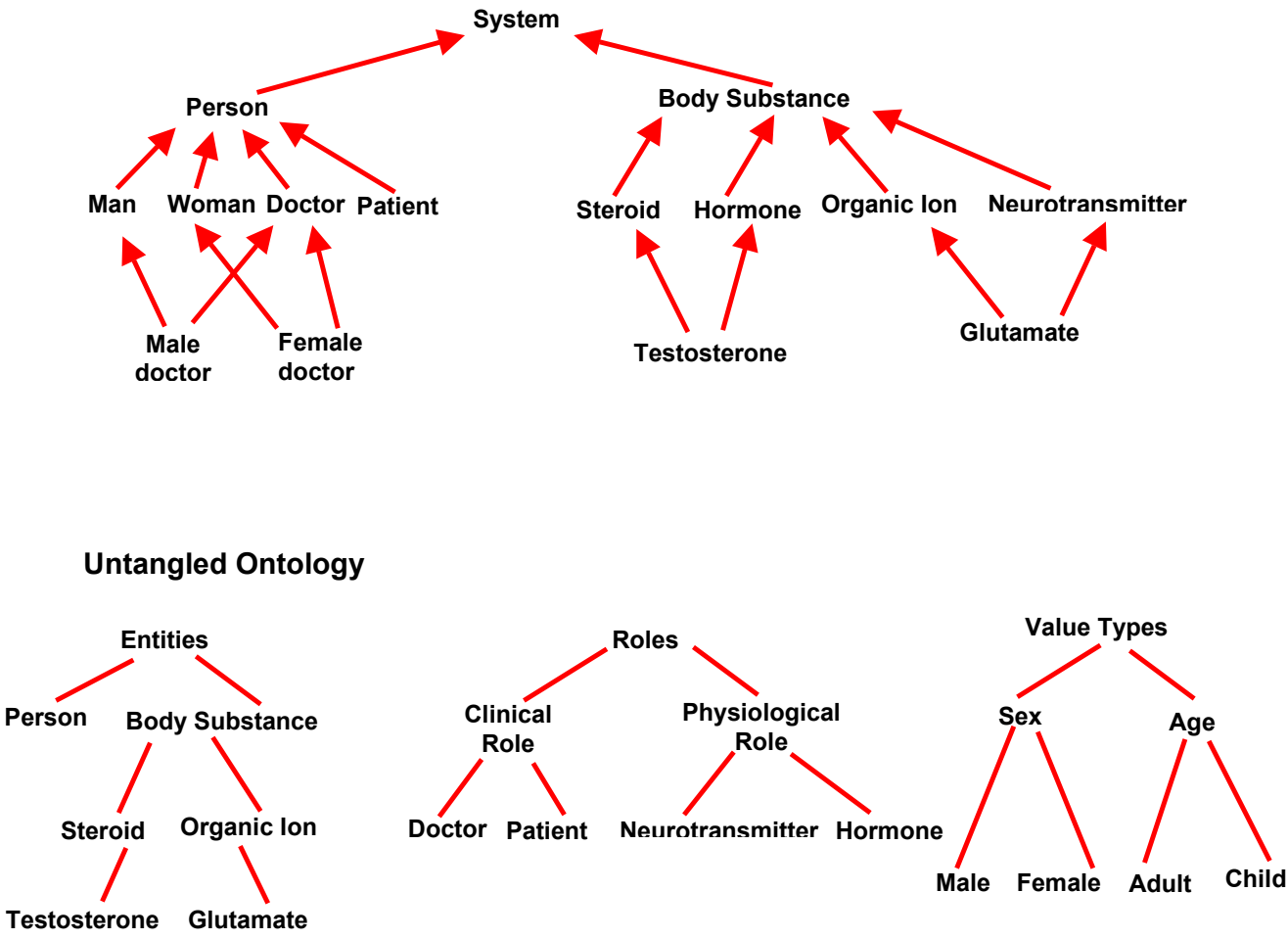


Figure 1. **Untangling a tangled ontology (modified from (Kemp and Gray, 2002)) A complex mixed ontology can be simplified by creating simpler ontologies representing distinct domains.**

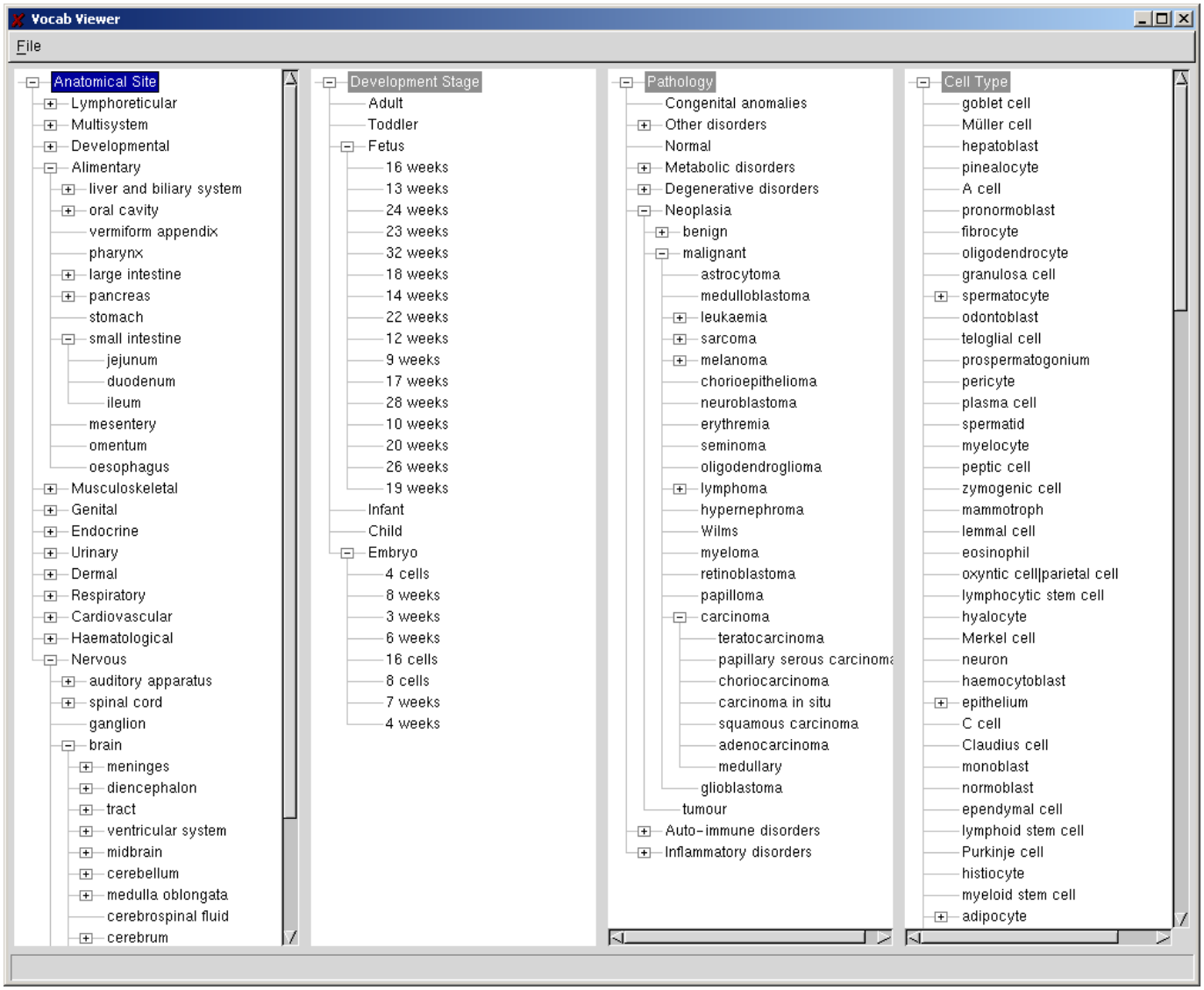
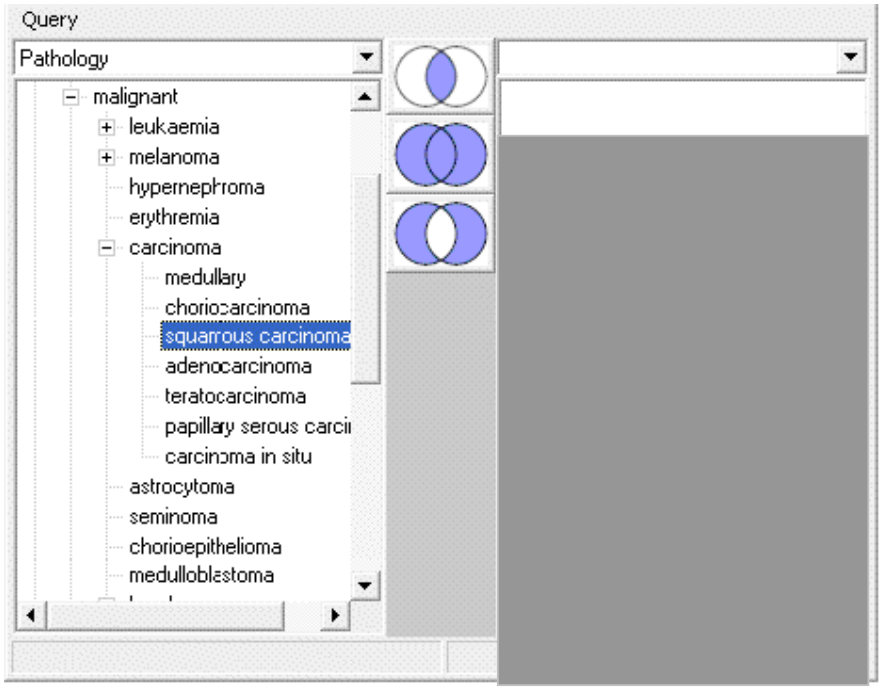


Figure 2. **A screenshot of the 4 ontologies. Anatomical System, Developmental Stage, Pathology and Cell Type hierarchies are displayed with indications where the tree can be expanded.**



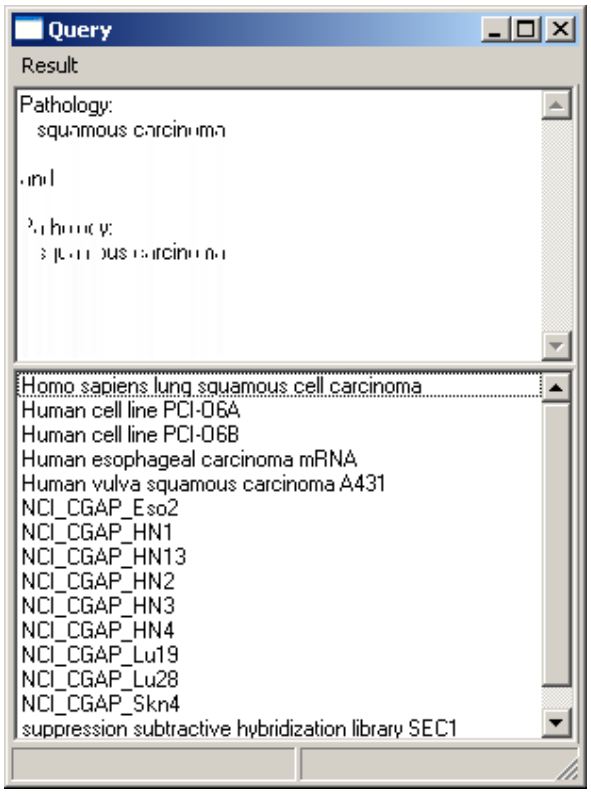


Figure 3. **A screenshot of the Pathology ontology with the term "squamous cell carcinoma" selected. Selection of a term displays the libraries which are annotated with that term (squamous cell carcinoma in this case) in the lower window. Using this GUI (developed by Electric Genetics), users can view the ontologies, browse the hierarchical trees and perform set operations on the annotated cDNA library data. The user is able to obtain the list of cDNA libraries or ESTs returned by a query, or provide a list of libraries or EST accessions and obtain the associated expression profile.**

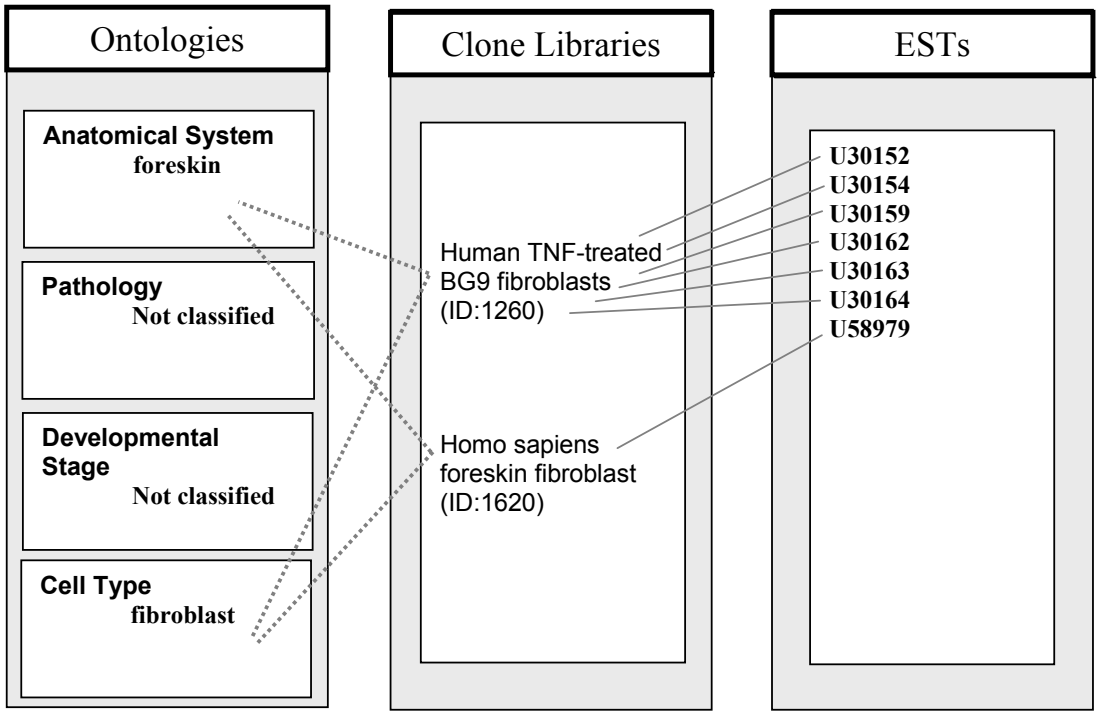


Figure 4. **The four expression ontologies are used to annotate cDNA clone libraries. ESTs can be transitively associated with ontology terms via their association with a unique clone library. Clone libraries are generated from biological sample materials representing specific expression states (e.g.: human foreskin fibroblasts). All the genes/transcripts expressed in the original biological sample are captured in the clone library and can be sequenced as ESTs from the library. By mapping the clone libraries to a set of controlled terms (the ontologies) all the ESTs from each clone library can be transitively linked to these same standardised terms in the relevant ontology via their association with their parent clone library.**

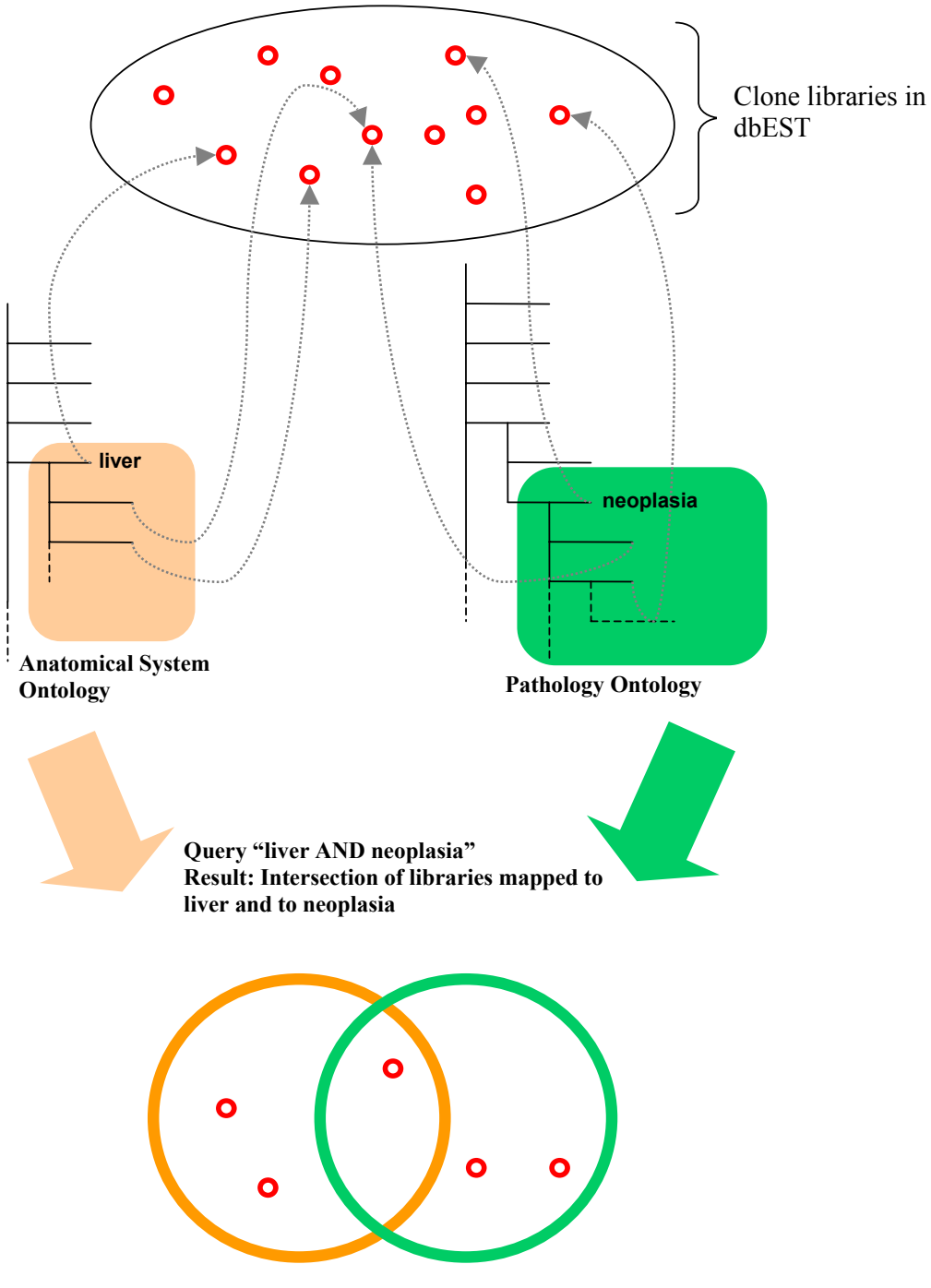


Figure 5. **Schematic of query system. Libraries are attached to terms which are nodes in the ontology trees. Boolean queries such as “liver AND neoplasia” are translated into set operations on the libraries below the nodes matching the query terms. The result is a list of libraries which meet the criteria set by the query.**

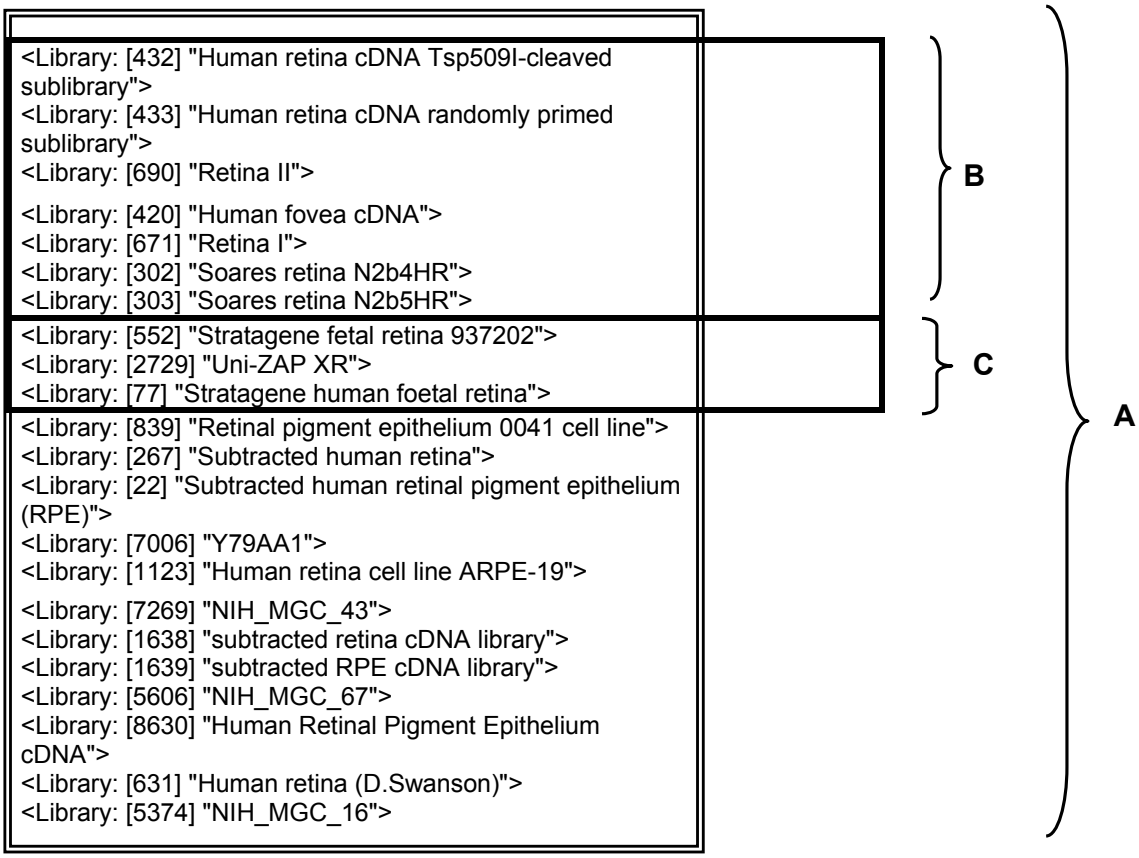


Figure 6 **Sample query to determine suitable libraries for laboratory research project on differential gene expression between adult and fetal retina. (A) The query: "retina" results in a list of the 22 libraries associated with the term "retina" in the Anatomical System ontology. (B) Further refining the query to: "retina & adult" results in a list of the 7 libraries associated with the terms "retina" in the Anatomical System ontology and also with the term "adult" in the Developmental Stage ontology. (C) A list of the 3 libraries which represent fetal retina can be obtained using the query: "retina & fetus".**

Appendix I The eVOC Ontologies

The Anatomical System Ontology

```
developmental
        notochord
        ectoderm
                neuroectoderm
        endoderm
        mesoderm
                mesenchyma
multisystem
        head and neck
        thorax
        abdomen
        pelvis
        perineum
        upper limb
        lower limb
        pooled tissues
        whole body
cardiovascular {vascular}
        heart
                atrium
                ventricle
                endocardium
                myocardium
                pericardium
                cardiac valve
                cardiac conducting system
        artery
                aorta
                arterial intima
                arterial media
                arterial adventitia
        vein {venae}
                venous intima
                venous media
                venous adventitia
        capillary
respiratory
        nose
        sinus
        larynx
        trachea
        bronchus
        lung
                small cell lung {small cell}
        alveolus
        pleura
haematological {hematopoietic}
        bone marrow
        blood
                peripheral blood leukocyte {PBL; peripheral blood leucocyte}
lymphoreticular {lymphoid tissue}
        lymph
        lymph node
                germinal center {germinal centre}
        tonsil
                lingual tonsil
                pharyngeal tonsil
        spleen
alimentary {digestive}
        oral cavity
                tongue
                salivary gland
                        parotid gland
                        submandibular gland {submaxillary gland}
```

```
                        sublingual gland
                tooth
                gum {gingiva}
        pharynx
                nasopharynx
                oropharynx
                hypopharynx
        oesophagus {esophagus}
        stomach
        intestine {gut}
                small intestine
                        duodenum
                        jejunum
                        ileum
                vermiform appendix
                large intestine
                        colorectal
                                colon
                                rectum
                        anus
        mesentery
        omentum
                greater omentum
                lesser omentum
        peritoneum {peritonaeum}
        liver and biliary system
                liver
                gall bladder {gallbladder}
                bile duct
        pancreas
                exocrine pancreas
urogenital {genitourinary}
        Urinary
                kidney
                ureter
                bladder
                urethra
        Genital
                male genitals
                        testis
                        epididymis
                        prostate
                        vas deferens
                        penis
                                glans
                                foreskin
                female genitals
                        ovary
                        uterine tube
                        uterus {womb}
                                cervix
                                endometrium
                        vagina
                        vulva
                        placenta
                        trophoblast
                        chorion
                        amnion
                        amniotic fluid
                        breast
                                mammary gland
                        milk
endocrine
        endocrine pancreas
                islets of Langerhans {pancreatic islet}
        pineal gland {pineal}
        pituitary gland {pituitary}
        thyroid
        parathyroid {parathyroid gland}
        adrenal gland {adrenal}
                adrenal cortex
                adrenal medulla
        thymus {thymus gland}
musculoskeletal
        bone
        cartilage
        joint
```

```
                synovium
                ligament
                meniscus
        muscle
                skeletal
                smooth
        tendon
        fascia
dermal
        skin {cuticle}
                epidermis
                dermis
        appendages
                hair follicle
                hair
                nail bed
                nail
                sweat gland
                sebaceous gland
nervous
        central nervous system {CNS}
                brain
                        cerebrum {hemisphere}
                                cerebral cortex {white matter; grey matter}
                                        frontal lobe {frontal cortex}
                                        parietal lobe {parietal cortex}
                                        temporal lobe {temporal cortex}
                                        occipital lobe {occipital cortex}
                                        visual
                                        insula
                                        olfactory bulb
                                        anterior olfactory nucleus
                                        lateral olfactory stria
                                        medial olfactory stria
                                        olfactory tubercle
                                        primary olfactory cortex
                                        secondary olfactory cortex
                                        hippocampus
                                        parahippocampal gyrus
                                basal nuclei
                                        amygdaloid nucleus
                                        central amygdaloid nucleus
                                        medial amygdaloid nucleus
                                        cortical amygdaloid nucleus
                                        claustrum
                                        corpus striatum
                                        caudate nucleus
                                        lentiform nucleus
                                        putamen
                                        globus pallidus
                        diencephalon
                                thalamus
                                        anterior thalamic nuclei
                                        anterior dorsal thalamic nucleus
                                        anterior medial thalamic nucleus
                                        anterior ventral thalamic nucleus
                                        medial thalamic nuclei
                                        medial dorsal thalamic nucleus
                                        parafascicular thalamic nucleus
                                        submedial thalamic nucleus
                                        paracentral thalamic nucleus
                                        central lateral thalamic nucleus
                                        ventral thalamic nuclei
                                        ventral anterior thalamic nucleus
                                        ventral intermediate thalamic nucleus
                                        ventral posterior thalamic nucleus
                                        lateral thalamic nuclei
                                        lateral dorsal thalamic nucleus
                                        lateral posterior thalamic nucleus
                                        pulvinar
                                        reticular thalamic nucleus
                                        centromedian thalamic nucleus
                                        limiting thalamic nucleus
                                metathalamus
                                        medial geniculate nucleus
                                        lateral geniculate nucleus
                                epithalamus
```

```
                pineal body
                habenular nucleus
        subthalamus
                subthalamic nucleus
        hypothalamus
                preoptic nucleus
                supraoptic nucleus
                suprachiasmatic nucleus
                paraventricular nucleus
                infundibular nucleus
                anterior nucleus
                dorsomedial nucleus
                ventromedial nucleus
                lateral nucleus
                posterior nucleus
                premamillary nucleus
                tuberomamillary nucleus
                medial mamillary nucleus
                lateral mamillary nucleus
                lateral tuberal nucleus
brain stem
        midbrain
                crus cerebri
                colliculi
                        superior colliculi
                        inferior colliculi
                substantia nigra
                red nucleus
                periaqueductal grey matter
                oculomotor nucleus
                trochlear nucleus
                mesencephalic trigeminal nucleus
        pons
                vestibular nuclei
                        medial
                        lateral
                        superior vestibular nuclei
                        inferior vestibular nuclei
                        interstitial
                cochlear nuclei
                        dorsal
                        ventral
                superior olivary nucleus
                trapezoid nucleus
                nucleus of the lateral lemniscus
                abducent nucleus
                facial nucleus
                salivatory nuclei
                        superior salivatory nuclei
                        inferior salivatory nuclei
                trigeminal nucleus
                        nucleus of the spinal tract
                        motor
                        principal sensory
        medulla oblongata
                olivary nuclei
                        inferior olivary nuclei
                        medial accessory
                        dorsal accessory
                nucleus gracilis
                nucleus cuneatus
                supraspinal nucleus
                spinal nucleus of the accessory nerve
                nucleus of the spinal tract of the trigeminal

                accessory cuneate nucleus
                nucleus of the hypoglossal nerve
                dorsal vagal nucleus
                nucleus of the tractus solitarius
                nucleus parasolitarius
                arcuate nuclei
                nucleus intercalatus
                nucleus ambiguus
cerebellum
        cerebellum cortex
                anterior lobe of the cerebellum
                middle lobe of the cerebellum
```

```
                        flocculonodular lobe
                        vermis
                cerebellum nuclei
                        dentate nucleus
                        nucleus emboliformis
                        nucleus globosus
                        nucleus fastigii
        tract
                corpus callosum
                olfactory
        ventricular system
                lateral ventricle
                third ventricle
                cerebral aqueduct
                fourth ventricle
        cerebrospinal fluid
        meninges
                dura mater
                arachnoid
                pia mater
spinal cord
        dorsal column
        substantia gelatinosa
        nucleus proprius
        nucleus thoracicus
        visceral column
        lateral column
        intermediolateral column
        intermediomedial column
        sacral parasympathetic nucleus
        ventral column
        retrodorsolateral column
        dorsomedial column
        dorsolateral column
        phrenic nucleus
        ventrolateral column
        ventromedial column
        accessory nucleus
        lumbosacral nucleus
peripheral nervous system {PNS}
spinal cord
        visual apparatus {eye}
                globe
                eyelid
                lacrimal gland
                conjunctiva
                cornea
                sclera
                lens
                vitreous humor
                iris
                ciliary body
                choroid
                retina
                optic nerve
                trabecular meshwork
        auditory apparatus
                external ear
                auricle
                external acoustic meatus
                middle ear
                tympanum
                auditory tube
                auditory ossicle
                internal ear
                utricle
                saccule
                semicircular canal
                cochlea
                spiral organ of Corti
        olfactory apparatus
        peripheral nerve {nerve}
        ganglion
                spinal ganglion {dorsal root ganglion}
        sympathetic chain
```

The Pathology Ontology

```
congenital anomalies
genetic disorders
        Charcot-Marie-Tooth disease
        Denys-drash
        Down's syndrome {Down syndrome}
        Huntington's disease
        Wiskott-Aldrich syndrome
        fragile X syndrome
infectious disorders
        viral
                AIDS
                cytomegalovirus {CMV}
                Epstein-Barr virus {EBV}
                hepatitis
                        hepatitis A virus {HAV}
                        hepatitis B virus {HBV}
                        hepatitis C virus {HCV}
                poliomyelitis
        bacterial
                botulism
                staphyloccocus
                streptoccocus
                syphilis
                tuberculosis
        fungal
                candidiasis
        parasitic
                helminthiasis
                pediculosis
inflammatory disorders
        arteritis
        arthritis
        autoimmune disorders
                Crohn's disease
                ulcerative colitis
                alopecia areata {patchy baldness}
                lupus
                        discoid lupus
                        systemic lupus erythematosus {systemic lupus}
                multiple sclerosis {MS}
                rheumatoid arthritis
        encephalitis
neoplasia {leukoplakia}
        tumour {tumor; neoplasm}
                benign tumour
                        adenoma
                        angioma
                        carcinoid
                        fibroma
                        fibrothecoma
                        glioma
                        haemangioma
                        insulinoma
                        leiomyoma {fibroid}
                        lipoma
                        lymphangioma
                        meningioma
                        phaeochromocytoma {pheochromocytoma}
                        Schwannoma
                        teratoma
                malignant tumour {cancer}
                        astrocytoma
                        carcinoma
                                carcinoma in situ
                                adenocarcinoma
                                choriocarcinoma
                                papillary serous carcinoma
                                teratocarcinoma
                        chorioepithelioma
                        erythremia
                        glioblastoma
```

```
                        histiocytoma
                                fibrous histiocytoma
                        hypernephroma
                        leukaemia {leukemia}
                                lymphoblastic
                                lymphocytic
                                        B-cell
                                        T-cell
                                promyelocytic
                                erythroleukaemia {erythroleukemia; polycythemia rubra;
Vasquez disease}
                                lymphosarcoma cell
                                megakaryocytic
                                myeloid {myelogenous; nonlymphocytic}
                                        monocytic
                        lymphoma
                                Hodgkin's
                                non-Hodgkin's
                                        Burkitt's
                        medulloblastoma
                        melanoma
                        myeloma
                        neuroblastoma
                        oligodendroglioma
                        retinoblastoma
                        sarcoma
                                liposarcoma
                                fibrosarcoma
                                gliosarcoma
                                osteosarcoma
                                Ewing's {peripheral neuroectodermal tumor; PNET}
                                Kaposi's
                                leiomyosarcoma
                                lymphosarcoma
                                reticulosarcoma
                                rhabdomyosarcoma
                        seminoma
                        Wilms
metabolic disorders
        Cushing's disease
        diabetes insipidus
        diabetes mellitus
                type I {juvenile; insulin-dependent}
                type II {type 2; NIDDM; noninsulin-dependent; adult-onset}
        Grave's disease
degenerative disorders
        atherosclerosis
        osteoarthritis
        encephalopathy
                Alzheimer's disease
                Creutzfeldt-Jakob disease
other disorders
        aneurysm
        systemic sclerosis
        cirrhosis
        dystrophies
                facioscapulohumeral {FSHD}
        growth disorders
                atrophy
                hypertrophy
                hyperplasia
                        psoriasis
                        goitre {bocio tumor}
                dysplasia
        malignant hyperthermia
        mental disorders
                schizophrenia
                affective disorders
                        bipolar disease
                        depression
normal
```

```
alpha cell
acidophil cell
acinar cell
adipoblast
adipocyte {fat cell}
        brown
        white
amacrine cell
beta cell
capsular cell
cementocyte
chief cell
chondroblast
chondrocyte
chromaffin cell
chromophobic cell
Claudius' cell {Claudius cell}
corticotroph
delta cell
dendritic cell
        follicular dendritic cell
        Langerhans cell
enterochromaffin cell {enteroendocrine cell; Kulchitsky cell; argentaffin cell}

ependymocyte {ependymal cell}
epithelium
        basal cell
        squamous
                endothelium {endothelial cell}
        transitional
erythroblast
erythrocyte {red blood cell}
fibroblast
fibrocyte
follicular cell
germ cell
        oocyte
                primary oocyte
                secondary oocyte
        spermatid
        spermatocyte
                primary spermatocyte
                secondary spermatocyte
        gamete
                ovum
                spermatozoon {spermatozoid; spermatozoa; sperm cell}
germinal epithelium
giant cell
glial cell {neuroglia}
        astroblast
        astrocyte
        oligodendroblast
        oligodendrocyte
glioblast
goblet cell
gonadotroph
granulosa cell
haemocytoblast
hair cell
Hensen cell
hepatoblast
hepatocyte
hyalocyte
interstitial cell {Leydig cell}
juxtaglomerular cell
keratinocyte
keratocyte {prickle cell}
lemmal cell
leukocyte {leucocyte; white blood cell}
        granulocyte {polymorphonuclear leukocyte; polymorphonuclear leucocyte}
                basophil
```

```
                eosinophil
                neutrophil
        lymphoblast
                B-lymphoblast
                T-lymphoblast {T lymphoblast}
        lymphocyte
                B-lymphocyte {B lymphocyte; B-cell; B cell}
                T-lymphocyte {T lymphocyte; T-cell; T cell}
                natural killer cell {NK cell}
        macrophage
                alveolar macrophage
                foam cell
                histiocyte {tissue macrophage}
                Kupffer cell
luteal cell
lymphocytic stem cell
lymphoid cell
lymphoid stem cell
macroglial cell
mammotroph
mast cell
medulloblast
megakaryoblast
megakaryocyte
melanoblast
melanocyte
Merkel cell
mesangial cell
mesothelium
metamyelocyte
monoblast
monocyte
mucous neck cell
Müller cell
muscle cell {muscle}
        cardiac muscle {heart muscle}
        smooth muscle
        striated muscle {skeletal muscle}
myelocyte
myeloid cell
myeloid stem cell
myoblast
myoepithelial cell
myofibrobast
neuroblast
neuroepithelium
neuron
normoblast
odontoblast
osteoblast
osteoclast
osteocyte
oxyntic cell {parietal cell}
Paneth cell
parafollicular cell
paraluteal cell
parietal cell
peptic cell
pericyte
phaeochromocyte
phalangeal cell {Deiters' cell}
pinealocyte
pituicyte
plasma cell
platelet
podocyte
prickle cell
proerythroblast
promonocyte
promyeloblast
promyelocyte
pronormoblast
Purkinje cell
reticulocyte
retinal pigment epithelium {pigmented retinal epithelium}
retinoblast
Schwann cell
```

Sertoli cell
somatotroph
stem cell
sustentacular cell
teloglial cell
zymogenic cell

The Developmental Stage Ontology

```
embryo
        4 cells
        8 cells
        16 cells
        3 weeks {3 wk}
        4 weeks {4 wk}
        6 weeks {6 wk}
        7 weeks {7 wk}
        8 weeks {8 wk}
fetus {foetus}
        9 weeks {9 wk}
        10 weeks {10 wk}
        12 weeks {12 wk}
        13 weeks {13 wk}
        14 weeks {14 wk}
        16 weeks {16 wk}
        17 weeks {17 wk}
        18 weeks {18 wk}
        19 weeks {19 wk}
        20 weeks {20 wk}
        22 weeks {22 wk}
        23 weeks {23 wk}
        24 weeks {24 wk}
        26 weeks {26 wk}
        28 weeks {28 wk}
        32 weeks {32 wk}
infant {newborn; neonate}
        0 year
        1 year {1 yr; age 1}
child
        toddler
                2 years {2 year; 2 yr; age 2}
                3 years {3 year; 3 yr; age 3}
        4 years {4 year; 4 yr; age 4}
        5 years {5 year; 5 yr; age 5}
        6 years {6 year; 6 yr; age 6}
        7 years {7 year; 7 yr; age 7}
        8 years {8 year; 8 yr; age 8}
        9 years {9 year; 9 yr; age 9}
        10 years {10 year; 10 yr; age 10}
        11 years {11 year; 11 yr; age 11}
        12 years {12 year; 12 yr; age 12}
adolescent {teenager; teens}
        13 years {13 year; 13 yr; age 13}
        14 years {14 year; 14 yr; age 14}
        15 years {15 year; 15 yr; age 15}
        16 years {16 year; 16 yr; age 16}
        17 years {17 year; 17 yr; age 17}
adult
        18 years {18 year; 18 yr; age 18}
        19 years {19 year; 19 yr; age 19}
        20 years {20 year; 20 yr; age 20}
        21 years {21 year; 21 yr; age 21}
        22 years {22 year; 22 yr; age 22}
        23 years {23 year; 23 yr; age 23}
        24 years {24 year; 24 yr; age 24}
        25 years {25 year; 25 yr; age 25}
        26 years {26 year; 26 yr; age 26}
        27 years {27 year; 27 yr; age 27}
        28 years {28 year; 28 yr; age 28}
        29 years {29 year; 29 yr; age 29}
        30 years {30 year; 30 yr; age 30}
        31 years {31 year; 31 yr; age 31}
        32 years {32 year; 32 yr; age 32}
        33 years {33 year; 33 yr; age 33}
        34 years {34 year; 34 yr; age 34}
        35 years {35 year; 35 yr; age 35}
        36 years {36 year; 36 yr; age 36}
        37 years {37 year; 37 yr; age 37}
        38 years {38 year; 38 yr; age 38}
        39 years {39 year; 39 yr; age 39}
```

```
40 years {40 year; 40 yr; age 40}
41 years {41 year; 41 yr; age 41}
42 years {42 year; 42 yr; age 42}
43 years {43 year; 43 yr; age 43}
44 years {44 year; 44 yr; age 44}
45 years {45 year; 45 yr; age 45}
46 years {46 year; 46 yr; age 46}
47 years {47 year; 47 yr; age 47}
48 years {48 year; 48 yr; age 48}
49 years {49 year; 49 yr; age 49}
50 years {50 year; 50 yr; age 50}
51 years {51 year; 51 yr; age 51}
52 years {52 year; 52 yr; age 52}
53 years {53 year; 53 yr; age 53}
54 years {54 year; 54 yr; age 54}
55 years {55 year; 55 yr; age 55}
56 years {56 year; 56 yr; age 56}
57 years {57 year; 57 yr; age 57}
58 years {58 year; 58 yr; age 58}
59 years {59 year; 59 yr; age 59}
60 years {60 year; 60 yr; age 60}
61 years {61 year; 61 yr; age 61}
62 years {62 year; 62 yr; age 62}
63 years {63 year; 63 yr; age 63}
64 years {64 year; 64 yr; age 64}
65 years {65 year; 65 yr; age 65}
66 years {66 year; 66 yr; age 66}
67 years {67 year; 67 yr; age 67}
68 years {68 year; 68 yr; age 68}
69 years {69 year; 69 yr; age 69}
elderly {geriatric}
       70 years {70 year; 70 yr; age 70}
       71 years {71 year; 71 yr; age 71}
       72 years {72 year; 72 yr; age 72}
       73 years {73 year; 73 yr; age 73}
       74 years {74 year; 74 yr; age 74}
       75 years {75 year; 75 yr; age 75}
       76 years {76 year; 76 yr; age 76}
       77 years {77 year; 77 yr; age 77}
       78 years {78 year; 78 yr; age 78}
       79 years {79 year; 79 yr; age 79}
       80 years {80 year; 80 yr; age 80}
       81 years {81 year; 81 yr; age 81}
       82 years {82 year; 82 yr; age 82}
       83 years {83 year; 83 yr; age 83}
       84 years {84 year; 84 yr; age 84}
       85 years {85 year; 85 yr; age 85}
       86 years {86 year; 86 yr; age 86}
       87 years {87 year; 87 yr; age 87}
       88 years {88 year; 88 yr; age 88}
       89 years {89 year; 89 yr; age 89}
       90 years {90 year; 90 yr; age 90}
       91 years {91 year; 91 yr; age 91}
       92 years {92 year; 92 yr; age 92}
       93 years {93 year; 93 yr; age 93}
       94 years {94 year; 94 yr; age 94}
       95 years {95 year; 95 yr; age 95}
       96 years {96 year; 96 yr; age 96}
       97 years {97 year; 97 yr; age 97}
       98 years {98 year; 98 yr; age 98}
       99 years {99 year; 99 yr; age 99}
```

Chapter 4

The Application of Ontologies to the Identification of Alternatively Spliced Transcripts with Unique or Restricted Expression

Two novel informatics approaches to the detection of variant transcripts, and to the classification of when and where transcripts are expressed have already been presented. In this final chapter a preliminary study combining these techniques is provided. Specifically, I present preliminary results for the application of these approaches to the identification of differentially expressed transcript isoforms.

As described in Chapter 3, eVOC can be used to identify the conditions under which expressed human transcripts have been detected. The identification of when and where a transcript is expressed provides potentially valuable functional information. Transcripts which show expression which is specific or restricted to particular tissues, diseases or developmental stages (expression states) are of particular interest as these are frequently regarded as having important functional roles in that particular state.

The recent suggestion that regulated alternative splicing may play a role in increasing the functional genetic repertoire of an organism (Kriventseva et al., 2003), (Brett et al., 2002) has led to an increased interest in investigating the extent and impact of alternative splicing on the production of functionally distinct proteins. Several alternatively spliced isoforms have already been demonstrated to show specificity to particular tissues, diseases and developmental stages. Besides their application as potential therapeutic targets, transcripts which show specificity in their expression can

be used as markers of developmental progression (Mouchel et al., 2003), (Patel et al.) or of disease diagnosis, prognosis and progression (Caballero et al., 2001).

As described in Chapter 2, large-scale mining of expressed sequence tags (ESTs) and comparison with available human genome sequence has allowed the detection of variations in the exon composition of the mature transcripts of genes. The occurrence of exon skipping, the most common form of alternative splicing, has been linked to various disease phenotypes, including cancer (Caballero et al., 2001), (Kwabi-Addo et al., 2001). Alternative splicing may play a major role in tumourigenesis, and cancer-specific transcript isoforms could prove to be useful diagnostic markers or therapeutic targets. However, it remains difficult to link alternatively spliced transcripts with specific expression states. Despite large amounts of available expression data, biological discovery is currently hampered by the variable quality, non-standardised annotation terms and dispersed nature of this information. Using eVOC, a controlled vocabulary, which partitions expression information extracted from cDNA library annotation into four categories: anatomical site, cell type, developmental stage and pathological state we have attempted to determine the specificity of the expression state of both skip and constitutive transcript isoforms.

I present here the development of two complementary systems which can be integrated to detect alternatively spliced transcripts and link these with specific expression states. The first system detects exon skipping using genomic sequence and ESTs, while the second determines the expression of the spliced and constitutive isoforms according to the four eVOC expression categories. An example implementation of this integrated approach - the detection and characterisation of alternatively spliced transcripts which are unique to cancer - is described. We have

successfully detected and characterised 323 exon skipping events in 241 genes. Preliminary computational results indicate that in three of these genes the constitutive isoforms are uniquely associated with normal tissues, and the exon-skipped transcripts are uniquely associated with cancer tissues.

Identification and characterisation of exon skipped transcripts specifically expressed in cancer tissues

4.1 Aim

In this study we applied the previously described j_explorer and eVOC tools to EST data in order to identify genes that produce cancer-specific alternative spliceforms.

4.2 Background

In excess of 2 million ESTs from human tumours and corresponding normal tissues have been deposited in public transcript databases, largely as the result of the work of two major consortia – the Cancer Genome Anatomy Project (CGAP) and the Open Reading Frame ESTs (ORESTES) projects. These ESTs provide significant insight into the cancer transcriptome (Strausberg et al., 2002), (Camargo et al., 2001) offering evidence for 25,000 genes, of which about 3,000 are only represented by EST data. Less than 1% of the known cancer-related genes do not have corresponding ESTs, indicating that the representation of genes associated with commonly studied tumors is high (Camargo, personal communication). The careful recording of the biological source of all the ESTs that have been produced by these sequencing projects enables detailed analysis of where the genes they represent are expressed in the human body.

The incidence of alternative splicing of human genes has been estimated to be between 30% and 60% (Modrek and Lee, 2002). Exon skipping, the most common form of alternative splicing (Mironov et al., 1999a), produces distinct transcript isoforms, and may result in the generation of protein products with distinct functions. Splice variants associated with the induction of cell death, regulation of cellular

proliferation and differentiation, cell signaling, and angiogenesis are present in a variety of cancers (Mercatante and Kole, 2000). These alternatively spliced isoforms associated specifically with cancer have the potential to be useful as prognostic or diagnostic markers. Cancer-specific, functionally distinct proteins produced by alternative splicing can be used as targets for novel therapeutic treatments targeted at cancerous cells. Modification of spliceform ratios using antisense oligonucleotides also shows promise in restoring the normal phenotype (Mercatante et al., 2001).

In order to identify these potential targets the genetics community requires tools which link emerging genome sequence information and expression data with disease phenotype. Successful mining of expression information can be facilitated by the use of a standardised nomenclature. This nomenclature should capture and present available data in an appropriate manner in order to allow for the extraction of expression profiles relevant to disease phenotypes. The system we have developed integrates transcript information and genomic sequence for the identification of alternatively spliced cancer genes, and incorporates a controlled vocabulary for the description of the expression state of alternatively spliced candidates.

4.3 Methods and results

4.3.0.1 Data sources

The analysis presented here is based on data from four major sources: a set of cancer-related genes, an early assembly of the human genome, human ESTs from EMBL, and human cDNA libraries mapped to the eVOC controlled vocabularies. A non-redundant set of 1011 cancer-related genes was manually selected by expert curators based on querying GenBank (http://www.ncbi.nlm.nih.gov/GenBank/index.html), GeneCards (http://bioinfo.weizmann.ac.il/cards/index.html) and the Harvard

University (http://sbweb.med.harvard.edu/sgc/gene_list.htm). The gene list, available at http://madhatter.fmrp.usp.br/jamborestes/, represents an incomplete yet representative set of the well-characterised genes implicated in human cancers, and includes well-characterised genes such as BRCA1, RB1 and TP53. Human genomic contigs for the April 2001 "Golden Path", an early "pre-finished" assembly produced by the University of California, Santa Cruz, were downloaded from ftp://genome.cse.ucsc.edu/goldenPath/. Human ESTs were obtained from the human EST division of EMBL 67 to which additional open reading frame ESTs (ORESTES) generated by the Brazilian Human Cancer Genone Sequencing Project (Camargo et al., 2001) were added. The eVOC controlled vocabularies used were those for the August 2001 version which was based on the EST division of EMBL 70 with additional ORESTES cDNA libraries added.

4.3.0.2 Mapping cDNAs to the genome

The most accurate approach to identifying alternative splicing is based on alignment of full-length cDNAs to their cognate genomic sequence as this provides information about the transcripts produced. In the absence of significant numbers of full-length cDNAs from a broad range of tissues, developmental stages and pathologies, ESTs have been widely used to identify and characterise alternative splicing (Kan et al., 2002), (Xu et al., 2002), (Hide et al., 2001), (Modrek et al., 2001), (Kan et al., 2001), (Kan et al., 2000).

In order to identify alternative splicing using transcribed sequences it was necessary to determine the complete gene structure of each of the selected cancer genes. The human genome sequence was first masked for repeats and low complexity regions using RepeatMasker (Smit,

1999)(http://ftp.genome.washington.edu/RM/RepeatMasker.html). To increase accuracy and reduce the time required, the representative cDNA selected for each of the 1011 cancer-related genes was then mapped to the Golden Path assembly in three stages (Figure 1). In the initial step cDNAs were assigned to the correct chromosome using ePCR against genemap99, txmap and genethon markers. cDNAs which could be assigned to more than one marker, and different cDNAs which mapped to the same marker were removed. Subsequently, MegaBLAST (Zhang et al., 2000) was used to identify the genomic contig to which each cDNA, including those not mapped using ePCR, could be mapped. In total 944 genes could be unambiguously assigned to a chromosome using this method. The remaining 67 genes could not be assigned to any chromosome with any confidence and were excluded from further analysis.

In the final stage of processing SPIDEY (Wheelan et al., 2001) was used to obtain a spliced alignment of each transcript against the identified genomic contig. The alignment was required to have a coverage of 97% and at least 98% identity. Using these criteria exon-intron structures for all 944 genes were determined. 845 genes were found to consist of three or more exons, while 99 genes consisted of either a single or double exon (Table 1).

An EMBL formatted sequence record containing the sequence of the genomic contig and with the exon positions annotated was produced for each gene. Transcripts that could not be mapped accurately to the genome, and those composed of fewer than three exons were excluded from further analysis.

4.3.0.3 Identification of exon skipping

Alternative splicing leads to transcript variability and may play a significant role in the development (Stoilov et al., 2002), progression (Assimakopoulos et al., 2002),

prognosis, diagnosis and treatment (Mercatante et al., 2001) (Mercatante and Kole, 2000) of human cancers.

J-explorer (Hide et al., 2001) was used to assemble exon-constructs from the mRNA-annotated genomic sequences. A set of all consecutive and non-consecutive exon-exon junctions for each of the mapped cancer genes was created, and each junction was submitted for similarity searching against EMBL 67 (human EST division with extra ORESTES added) using BLAST 2.0 (Altschul et al., 1990). A skipping event was recorded when an EST which did not contain the exon(s) in question, but did contain an uninterrupted tag made up of 50bp from each of the flanking exons. ESTs showing significant ($P<1e^{-40}$) homology to an exon junction were extracted, aligned to the corresponding genomic sequence using SPIDEY and manually inspected. In order to exclude the possibility that ESTs confirming exon skipping events were the products of paralogous genes or members of gene families all ESTs identifying exon skipping were confirmed to be unique to a single target gene, Both interchromosomal and intrachromosomal specificity of the transcripts was confirmed using BLAST with a cut-off score of $1e^{-30}$. In order to prevent the detection of skips as a result of incorrectly annotated exon boundaries we required that for every instance of an EST-predicted exon skipping event, at least one EST spanning the consecutive (or linear) exon boundaries was also present. All consecutive junctions which could not be confirmed by ESTs resulted in that junction being excluded from further analysis.

This approach precisely identifies exon skipping when EST transcript data that spans exon boundaries is available. A potential limitation of j_explorer is that it is reliant on the quality of the gene structure presented for skipping analysis. Since the mRNAs used to determine the exon-intron boundaries were potentially non-canonical isoforms

the estimate of exon skipping presented here is a conservative one. The depth of transcript representation in the EST databases, and the level of transcript expression both influence the likelihood of discovering an exon skip. The cancer gene set was found to be relatively well represented in the combined CGAP and ORESTES EST datasets, with each gene being represented by an average of 107 ESTs. 97% of the genes in the list have at least one corresponding CGAP or ORESTES transcript, while only 18 of the genes were found to have no EST coverage at all. Additionally, the ratio of low abundance to high abundance transcript isoforms also influences the detection of exon skipping.

In the 845 genes examined a total of 323 exon skipping events were detected in 241 genes (Table 2). This figure of 29% agrees well with previously published estimates of the overall incidence of exon skipping in human genes based on EST data (Mironov et al., 1999a), but is lower than the figure of 59% obtained for a small set of highly expressed genes in the chromosome 22 analysis (Hide et al., 2001). This is as a result of the requirement that both the consecutive flanking junctions be represented in the EST data before a non-consecutive junction covered by one or more transcripts was accepted as accurate.

4.3.0.4 Identification and analysis of candidate cancer-specific isoforms

The production of in excess of 100 000 ORESTES for both normal and cancer pathologies in each of seven major organs (brain, head and neck, colon, lung, breast, uterus and kidney) provides a deep survey of the transcripts expressed in these states. The capture of both the anatomical system and pathological information, and its association with the sequence data through eVOC, provides an opportunity to mine for transcripts specific to restricted expression states. Transcripts which show

restricted or unique expression are potentially valuable markers or therapeutic targets for human cancer.

In order to determine whether the isoforms of the genes demonstrating alternative splicing also show distinct expression patterns (where one isoform is unique to cancer while the other is uniquely seen in normal tissues) we first identified ESTs which uniquely identify the respective isoforms ("isoform-specific ESTs"). These isoform specific ESTs were then classified with respect to their pathological state of origin using the eVOC ontologies. When the isoform-specific ESTs were enriched in cancer libraries by more than three fold compared to normal libraries, the corresponding isoforms were defined as being cancer-specific. As this was a preliminary investigation no further statistical analysis was performed.

In three of the genes identified as showing exon skipping, the alternatively spliced isoforms were only detected in cDNA libraries derived from cancer tissues, while the constitutive isoforms were found only in normal cDNA libraries (Table 3), suggesting that the splice variants may be tumour-specific and could be potential markers or therapeutic targets. Since there was generally more sequence data available for cancer-derived tissues than for normal tissues, and because both normal and cancer samples are often highly heterogeneous, containing a mixture of different cell types, the identified candidates were submitted for verification using experimental methods.

In all three genes the skipped exon was within the protein-coding region. Analysis of the potential effect of the skip on the predicted protein product showed that in all three cases the reading frame remained unaltered. An investigation of the protein motifs which may be impacted showed that in the case of *PTPN13* the skipping of exon 26 results in the complete removal of the PDZ domain which is involved in intracellular

signalling (Ponting et al., 1997) and may play a role in modulating cell death (Li et al., 2002). There are no known functional motifs affected by the exon skipping in *CD53* and *GOLGA4*.

The genes involved, *CD53*, *PTPN13* and *GOLGA4* were prioritised for verification in the laboratory of the Ludwig Institute for Cancer Research, São Paulo, Brazil. GOLGA4 was subsequently excluded from testing as sufficiently unique primers to the skipped exon could not be designed. The skipping of central, coding exons in *CD53* (NM_00560) and *PTPN13* (NM_006264) was tested using a strategy that enhanced the amplification of the transcript isoform containing the skipped exon. Using an RT-PCR primer that spanned the novel splice site formed, it was possible, in each case, to detect the alternative transcript in both normal and tumor derived samples. Thus, although carefully documented alternative splicing events can almost always be readily validated, tumor specific alternative splicing may be rare. The imbalance between normal and tumor derived transcript sequences in the EST databases is likely to lead to many apparently tumor specific transcripts being false positives.

4.4 Discussion and opportunities

Mining large collections of expression data, such as EST collections, provides valuable and extensive information about the incidence and extent of transcript variation as well as about the location and timing of gene expression. The EST databases capture a broad, yet incomplete and uneven representation of the total human transcriptome.

A number of studies have used the quantification of ESTs to identify genes which show expression which is unique or restricted to specific tissues or diseases including colon cancer genes (Brett et al., 2001) or genes which are differentially expressed between cancer and normal tissues (Schmitt et al., 1999), but few have taken into account the extent of alternative splicing, or have tried to systematically describe differential expression of alternatively spliced transcripts arising from a single gene.

As a result there is still relatively little known about how the specific expression of alternatively spliced transcripts is regulated, and relatively few large-scale efforts have been carried out to identify large numbers of isoforms showing tissue specific expression patterns.

We have performed a stringent, large-scale analysis of a set of genes implicated in human cancer. Results indicate that while alternative splicing can be readily detected, the unique differential expression of transcript isoforms is rare.

Answering questions about the fraction of alternative splicing which shows tissue or disease specificity, and at what level it is most appropriate to analyse this specificity (ie: whether the spliceforms are associated with whole systems, organs, tissues, or cells) will rely on methods of large-scale expression analyses such as those possible using microarray technologies. A promising approach is that taken by Shoemaker et al. (Shoemaker et al., 2001) who have used ‘exon’ or ‘tiling’ arrays to identify full-length spliceforms on chromosome 22. Using this technology it appears that it will be possible to perform large-scale identification of both transcript variants, and their disease or tissue-specific expression profiles simultaneously.

It is hoped that such analyses will provide some insight into the mechanisms controlling the production of state-specific alternative transcripts, and will lead to an

increased ability to predict computationally the transcripts and expression patterns of genes implicated in disease.

4.5 Acknowledgements

I carried out the exon skipping analysis presented in this chapter independently. The cancer gene set which was assessed for exon skipping, and the ORESTES sequences used for the study were provided as part of a research collaboration with the Ludwig Institute for Cancer Research. I acknowledge the input of, among others, Dr Andrew Simpson, Dr Anamaria Camargo, and Dr Helena Brentani. Dr. Otavia. Caballero was responsible for the RT-PCR-based laboratory verification of the genes which were identified *in silico* as having cancer-specific exon skipping.

Figure 1. **Experimental workflow: The exon-intron structure of the genes was determined by mapping the cDNAs to genomic contigs using progressively more refined techniques. Single and double exon genes were removed from further analysis, and multi-exon genes were submitted to j_explorer for the detection of exon skipping. The expression of transcripts from genes showing exon skipping was assessed using eVOC and transcripts from the same gene showing diferential expression were prioritised for experimental verification.**

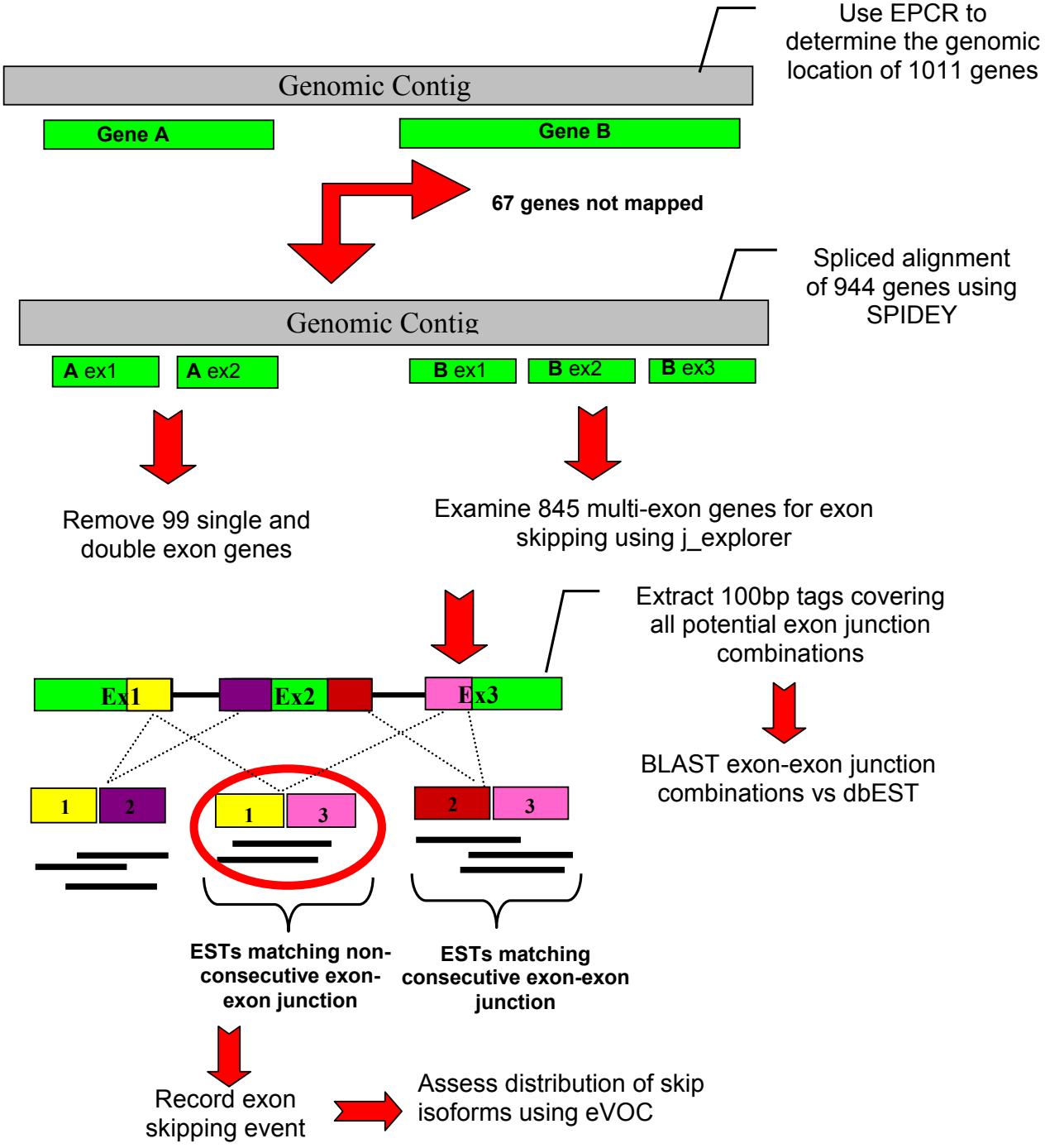


Table 1 **Processing of cancer-related genes selected for alternative splicing analysis.**

Total number of cancer-related genes	1011
Genes mapped to Golden Path assembly	944
Genes not mapped to Golden Path assembly	67
Number of Multi-exon genes	845
Number of single and double exon genes	99

Table 2 **Exon structure and exon skipping information for the 845 cancer-related genes determined using j_explorer.**

Number of multiple exon genes	845
Number of exons	10770
Total number of consecutive junctions	9925
Mean exon length	207 bp
Maximum number of exons in one gene	71
Number of genes with exon skipping	241 (29%)
Number of skipping events	323
Number of ESTs covering consecutive junction	16.14
Number of ESTs covering non-consecutive junction	3.20
Average number of exon junctions per multi-exon gene	11.75
Probability that a multi-exon mRNA has at least one non-consecutive junction	0.07 (7%)

Table 3 **Three genes were found to show exon skipped transcripts in cDNA libraries prepared from cancer tissues, while the constitutive product was only observed in libraries prepared from normal tissues. Translation and motif analysis provided information about the potential effect of the skip on the protein product.**

Gene	Accession	Function	Splice Variant	Protein Modification
PTPN13	NM_006264	Intracellular signalling, amino acid dephosphorylation, hydrolase	Exon 26 skipped	Reading frame remains intact PDZ domain removed
CD53	NM_00560	Signal transduction	Exon 7 skipped	Reading frame remains intact. Prenyl group removed
GOLGA4	U41740	Vesicle-mediated transport	Exon 7 skipped	Reading frame remains intact No known motif affected

Conclusions

I have presented a summary of the commonly used techniques for quantifying and characterising gene expression and have added to this the development and implementation of novel informatics-based approaches to transcript analysis. Through the application of informatics approaches to understanding gene expression I have provided novel methods to identify and characterise transcript variation, and have gained insight into the identity, structure, and expression patterns of expressed transcripts.

In the first instance I have applied a computational approach to the detection and characterisation of alternatively spliced transcripts. Through the development of a novel software tool and the use of publicly available genome and transcript data I have shown that at least 15% of human genes demonstrate exon skipping. Additionally, I have show that exon skipping can be detected in at least 58% of the genes which are well represented by transcript data. This has demonstrated that the detection of alternative splicing is heavily dependent on the coverage of genes by available transcript data. This finding underlines the need for ongoing transcript sequencing – particularly from those tissues and expression states that are under-represented in current transcript databases. Increased, directed transcript sequencing is likely to provide the data required to refine these current estimates of the extent of alternative splicing, and are likely to result in the discovery that this phenomenon is more widespread than has been estimated to date.

Furthermore, I have shown that 92% of exon-skipping events occur within the protein coding region of the gene and that 50% of cases the reading frame is maintained.

While this is not definitive evidence that these spliceforms are functional, it suggests that these transcripts potentially encode proteins and these may have distinct functions from those produced by the constitutive product. While the *in vitro* validation and functional analysis of these candidates would have to be performed on a case-by-case basis, the computational approach taken here provides qualitative and quantitative predictions of alternative splicing which are valuable in selecting targets for further experimental validation and characterisation. It is expected that such informatics approaches, when combined with evidence from *in vitro* experiments will ultimately lead to the elucidation of the mechanisms that regulate alternative splicing and will result in the ability to perform predictive assessment of alternative splicing and the expected biological impact.

Additionally, through the construction of four ontologies of appropriate granularity for describing the source of the biological materials used in gene expression experiments I have provided a novel means to integrate expression data from diverse experimental approaches, including EST, SAGE and microarray experiments, based on transcript expression information.

The incorporation of eVOC into the Ensembl DataMart has provided the ability to integrate both the phenotype and sequence information from these expression experiments with genomic sequence and annotation information including functional information obtained from the gene ontologies. For the first time it is now possible to perform genome-wide queries integrating genomic information with transcript expression and functional information.

The application of these ontologies to the annotation of 7016 cDNA and 104 SAGE libraries has already provided the ability to cross-query data from these sources in

order to characterise and quantify the transcripts identified using these approaches. The incorporation of microarray data will increase the utility of the system by providing large-scale information on expression levels and transcript variants which may not have been captured previously. Additionally, by taking advantage of the hierarchical structure of the ontologies the question of specific and/or differential expression of transcripts can be addressed at various levels. For example, genes expressed in a single organ or even a single cell can be identified.

Through the combined application of the approaches described for the detection of exon skipping and for the description of gene expression patterns using controlled vocabularies, I have shown that it is possible to mine the EST datasets for transcript isoforms which are differentially expressed in cancer and normal tissues. In the three identified candidates laboratory verification was able to confirm the alternative splicing events, but showed that the spliceforms were not unique to distinct pathologies. However, the approach described in Chapter 4 is generic enough that it can be applied to a variety of datasets, and it is expected that such studies will provide powerful insights into transcripts showing restricted expression and into the regulation of such expression, given sufficiently large amounts of data.

Bibliography

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

Andreadis,A., Gallego,M.E., and Nadal-Ginard,B. (1987). Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. Annu. Rev. Cell Biol. *3*, 207-242.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M., and Sherlock,G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25* , 25-29.

Assimakopoulos,D., Kolettas,E., Patrikakos,G., and Evangelou,A. (2002). The role of CD44 in the development and prognosis of head and neck squamous cell carcinomas. Histol. Histopathol. *17*, 1269-1281.

Bard,J. and Winter,R. (2001). Ontologies of developmental anatomy: their current and future roles. Brief. Bioinform. *2*, 289-299.

Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M., and Gautheret,D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res. *10*, 1001-1010.

Bevilacqua,A., Ceriani,M.C., Capaccioli,S., and Nicolin,A. (2003). Post-transcriptional regulation of gene expression by degradation of messenger RNAs. J. Cell Physiol *195*, 356-372.

Boguski,M.S., Lowe,T.M., and Tolstoshev,C.M. (1993). dbEST--database for "expressed sequence tags". Nat. Genet. *4*, 332-333.

Boguski,M.S. and Schuler,G.D. (1995). ESTablishing a human transcript map. Nat. Genet. *10*, 369-371.

Bonaldo,M.F., Lennon,G., and Soares,M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. *6*, 791-806.

Boylan,K.B., Ayres,T.M., Popko,B., Takahashi,N., Hood,L.E., and Prusiner,S.B. (1990). Repetitive DNA (TGGA)n 5' to the human myelin basic protein gene: a new form of oligonucleotide repetitive sequence showing length polymorphism. Genomics *6*, 16-22.

Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M., Roth,R., George,D., Eletr,S., Albrecht,G., Vermaas,E., Williams,S.R., Moon,K., Burcham,T., Pallas,M., DuBridge,R.B., Kirchner,J., Fearon,K., Mao,J., and Corcoran,K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. *18*, 630-634.

Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J., and Borka,P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. *474*, 83-86.

Brett,D., Kemmner,W., Koch,G., Roefzaad,C., Gross,S., and Schlag,P.M. (2001). A rapid bioinformatic method identifies novel genes with direct clinical relevance to colon cancer. Oncogene *20*, 4581-4585.

Brett,D., Pospisil,H., Valcarcel,J., Reich,J., and Bork,P. (2002). Alternative splicing and genome complexity. Nat. Genet. *30*, 29-30.

Burke,J., Davison,D., and Hide,W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Res. *9*, 1135-1142.

Caballero,O.L., de Souza,S.J., Brentani,R.R., and Simpson,A.J. (2001). Alternative spliced transcripts as cancer markers. Dis. Markers *17*, 67-75.

Camargo,A.A., Samaia,H.P., Dias-Neto,E., Simao,D.F., Migotto,I.A., Briones,M.R., Costa,F.F., Nagai,M.A., Verjovski-Almeida,S., Zago,M.A., Andrade,L.E., Carrer,H., El Dorry,H.F., Espreafico,E.M., Habr-Gama,A., Giannella-Neto,D., Goldman,G.H., Gruber,A., Hackel,C., Kimura,E.T., Maciel,R.M., Marie,S.K., Martins,E.A., Nobrega,M.P., Paco-Larson,M.L., Pardini,M.I., Pereira,G.G., Pesquero,J.B., Rodrigues,V., Rogatto,S.R., da Silva,I.D., Sogayar,M.C., Sonati,M.F., Tajara,E.H., Valentini,S.R., Alberto,F.L., Amaral,M.E., Aneas,I., Arnaldi,L.A., de Assis,A.M., Bengtson,M.H., Bergamo,N.A., Bombonato,V., de Camargo,M.E., Canevari,R.A., Carraro,D.M., Cerutti,J.M., Correa,M.L., Correa,R.F., Costa,M.C., Curcio,C., Hokama,P.O., Ferreira,A.J., Furuzawa,G.K., Gushiken,T., Ho,P.L., Kimura,E.,

Krieger,J.E., Leite,L.C., Majumder,P., Marins,M., Marques,E.R., Melo,A.S., Melo,M., Mestriner,C.A., Miracca,E.C., Miranda,D.C., Nascimento,A.L., Nobrega,F.G., Ojopi,E.P., Pandolfi,J.R., and Pessoa,L.G. (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc. Natl. Acad. Sci. U. S. A *98*, 12103-12108.

Carninci,P. and Hayashizaki,Y. (1999). High-efficiency full-length cDNA cloning. Methods Enzymol. *303*, 19-44.

Carninci,P., Nishiyama,Y., Westover,A., Itoh,M., Nagaoka,S., Sasaki,N., Okazaki,Y., Muramatsu,M., and Hayashizaki,Y. (1998). Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. Proc. Natl. Acad. Sci. U. S. A *95*, 520-524.

Carninci,P., Shibata,Y., Hayatsu,N., Itoh,M., Shiraki,T., Hirozane,T., Watahiki,A., Shibata,K., Konno,H., Muramatsu,M., and Hayashizaki,Y. (2001). Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. Genomics *77*, 79-90.

Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T., and Hide,W. (2001). STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Res. *29*, 234-238.

Cogoni,C. and Macino,G. (2000). Post-transcriptional gene silencing across kingdoms. Curr. Opin. Genet. Dev. *10*, 638-643.

Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P., and Mattick,J.S. (2000). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. *24*, 340-341.

De Lange,T., Liu,A.Y., Van der Ploeg,L.H., Borst,P., Tromp,M.C., and Van Boom,J.H. (1983). Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? Cell *34*, 891-900.

Dufour,C., Weinberger,R.P., Schevzov,G., Jeffrey,P.L., and Gunning,P. (1998). Splicing of two internal and four carboxyl-terminal alternative exons in nonmuscle tropomyosin 5 pre-mRNA is independently regulated during development. J. Biol. Chem. *273*, 18547-18555.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M., and Miller,W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. *8*, 967-974.

Frantz,S.A., Thiara,A.S., Lodwick,D., Ng,L.L., Eperon,I.C., and Samani,N.J. (1999). Exon repetition in mRNA. Proc. Natl. Acad. Sci. U. S. A *96*, 5400-5405.

Gautheret,D., Poirot,O., Lopez,F., Audic,S., and Claverie,J.M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res. *8*, 524-530.

Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. Genome Res. *11*, 1425-1433.

Gray,H.L., Bannister,L.H., Williams,P.L., Collins,P., and Berry,M.M. (1995). Gray's Anatomy., H.L.Gray, L.H.Bannister, P.L.Williams, P.Collins, and M.M.Berry, eds. (Britain: Churchill Livingstone Inc.).

Green, P. PHRAP. http://www.genome.washington.edu/uwgc/analysistools/phrap.htm phg@u.washington.edu. 1996.
Ref Type: Generic

Hide,W.A., Babenko,V.N., van Heusden,P.A., Seoighe,C., and Kelso,J.F. (2001). The contribution of exon-skipping events on chromosome 22 to protein coding diversity. Genome Res. *11*, 1848-1853.

Iida,Y. (1997). A mechanism for unsplicing and exon skipping in human alpha- and beta-globin mutant pre-mRNA splicing. Nucleic Acids Symp. Ser. 183-184.

Iseli,C., Jongeneel,C.V., and Bucher,P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 138-148.

Jiang,Z.H. and Wu,J.Y. (1999). Alternative splicing and programmed cell death. Proc. Soc. Exp. Biol. Med. *220*, 64-72.

Jongeneel,C.V. (2000). The need for a human gene index. Bioinformatics. *16*, 1059-1061.

Kan,Z., Gish,W., Rouchka,E., Glasscock,J., and States,D. (2000). UTR reconstruction and analysis using genomically aligned EST sequences. ISMB. *8*, 218-227.

Kan,Z., Rouchka,E.C., Gish,W.R., and States,D.J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. *11*, 889-900.

Kan,Z., States,D., and Gish,W. (2002). Selecting for Functional Alternative Splices in ESTs. Genome Res. *12*, 1837-1845.

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

Karp,P.D., Riley,M., Paley,S.M., and Pellegrini-Toole,A. (2002a). The MetaCyc Database. Nucleic Acids Res. *30*, 59-61.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C., and Gama-Castro,S. (2002b). The EcoCyc Database. Nucleic Acids Res. *30*, 56-58.

Kawahara,H., Kasahara,M., Nishiyama,A., Ohsumi,K., Goto,T., Kishimoto,T., Saeki,Y., Yokosawa,H., Shimbara,N., Murata,S., Chiba,T., Suzuki,K., and Tanaka,K. (2000). Developmentally regulated, alternative splicing of the Rpn10 gene generates multiple forms of 26S proteasomes. EMBO J. *19*, 4144-4153.

Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H., Adachi,J., Fukuda,S., Aizawa,K., Izawa,M., Nishi,K., Kiyosawa,H., Kondo,S., Yamanaka,I., Saito,T., Okazaki,Y., Gojobori,T., Bono,H., Kasukawa,T., Saito,R., Kadota,K., Matsuda,H.A., Ashburner,M., Batalov,S., Casavant,T., Fleischmann,W., Gaasterland,T., Gissi,C., King,B.,

Kochiwa,H., Kuehl,P., Lewis,S., Matsuo,Y., Nikaido,I., Pesole,G., Quackenbush,J., Schriml,L.M., Staubli,F., Suzuki,R., Tomita,M., Wagner,L., Washio,T., Sakai,K., Okido,T., Furuno,M., Aono,H., Baldarelli,R., Barsh,G., Blake,J., Boffelli,D., Bojunga,N., Carninci,P., de Bonaldo,M.F., Brownstein,M.J., Bult,C., Fletcher,C., Fujita,M., Gariboldi,M., Gustincich,S., Hill,D., Hofmann,M., Hume,D.A., Kamiya,M., Lee,N.H., Lyons,P., Marchionni,L., Mashima,J., Mazzarelli,J., Mombaerts,P., Nordone,P., Ring,B., Ringwald,M., Rodriguez,I., Sakamoto,N., Sasaki,H., Sato,K., Schonbach,C., Seya,T., Shibata,Y., Storch,K.F., Suzuki,H., Toyo-oka,K., Wang,K.H., Weitz,C., Whittaker,C., Wilming,L., Wynshaw-Boris,A., Yoshida,K., Hasegawa,Y., Kawaji,H., Kohtsuki,S., and Hayashizaki,Y. (2001). Functional annotation of a full-length mouse cDNA collection. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation meeting. Nature *409*, 685-690.

Kemp, G. and Gray, P. Modelling Biological Data in Hierarchies. 2002. Ref Type: Conference Proceeding

Kriventseva,E.V., Koch,I., Apweiler,R., Vingron,M., Bork,P., Gelfand,M.S., and Sunyaev,S. (2003). Increase of functional diversity by alternative splicing. Trends Genet. *19*, 124-128.

Kwabi-Addo,B., Ropiquet,F., Giri,D., and Ittmann,M. (2001). Alternative splicing of fibroblast growth factor receptors in human prostate cancer. Prostate *46*, 163-172.

Lambert de Rouvroit,C., Bernier,B., Royaux,I., de,B., V, and Goffinet,A.M. (1999). Evolutionarily conserved, alternative splicing of reelin during brain development. Exp. Neurol. *156*, 229-238.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J., LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W., McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T., Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H., Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P., Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L., Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A., Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R.,

Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R., McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglou,S., Birney,E., Bork,P., Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A., Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E., Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., and Szustakowki,J. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

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.

Lennon,G., Auffray,C., Polymeropoulos,M., and Soares,M.B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics *33*, 151-152.

Li,W., Srinivasula,S.M., Chai,J., Li,P., Wu,J.W., Zhang,Z., Alnemri,E.S., and Shi,Y. (2002). Structural insights into the pro-apoptotic function of mitochondrial serine protease HtrA2/Omi. Nat. Struct. Biol. *9*, 436-441.

Lim,S., Naisbitt,S., Yoon,J., Hwang,J.I., Suh,P.G., Sheng,M., and Kim,E. (1999). Characterization of the Shank family of synaptic proteins. Multiple genes, alternative splicing, and differential expression in brain and development. J. Biol. Chem. *274*, 29510-29518.

Mercatante,D. and Kole,R. (2000). Modification of alternative splicing pathways as a potential approach to chemotherapy. Pharmacol. Ther. *85*, 237-243.

Mercatante,D.R., Sazani,P., and Kole,R. (2001). Modification of alternative splicing by antisense oligonucleotides as a potential chemotherapy for cancer and other diseases. Curr. Cancer Drug Targets. *1*, 211-230.

Miller,R.T., Christoffels,A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R., and Hide,W.A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res. *9*, 1143-1155.

Mironov,A.A., Fickett,J.W., and Gelfand,M.S. (1999a). Frequent alternative splicing of human genes. Genome Res. *9*, 1288-1293.

Mironov,A.A., Koonin,E.V., Roytberg,M.A., and Gelfand,M.S. (1999b). Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucleic Acids Res. *27*, 2981-2989.

Modrek,B. and Lee,C. (2002). A genomic view of alternative splicing. Nat. Genet. *30*, 13-19.

Modrek,B., Resch,A., Grasso,C., and Lee,C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. *29*, 2850-2859.

Mouchel,N., Broackes-Carter,F., and Harris,A. (2003). Alternative 5' exons of the CFTR gene show developmental regulation. Hum. Mol. Genet. *12*, 759-769.

Parsons,J.D. and Rodriguez-Tome,P. (2000). JESAM: CORBA software components to create and publish EST alignments and clusters. Bioinformatics. *16*, 313-325.

Patel,D.D., Hale,L.P., Whichard,L.P., Radcliff,G., Mackay,C.R., and Haynes,B.F. Expression of CD44 molecules and CD44 ligands during human thymic fetal development: expression of CD44 isoforms is developmentally regulated.

Pickford,A.S. and Cogoni,C. (2003). RNA-mediated gene silencing. Cell Mol. Life Sci. *60*, 871-882.

Ponting,C.P., Phillips,C., Davies,K.E., and Blake,D.J. (1997). PDZ domains: targeting signalling molecules to sub-membranous sites. Bioessays *19*, 469-479.

Pruitt,K.D., Katz,K.S., Sicotte,H., and Maglott,D.R. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. *16*, 44-47.

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*, 159-164.

Quackenbush,J., Liang,F., Holt,I., Pertea,G., and Upton,J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. *28*, 141-145.

Rector, A. L., Wroe, C., Rogers, J., and Roberts, A. Untangling taxonomies and relationships: personal and practical problems in loosely coupled development of large ontologies. Gil, Y., Musen, M., and Shavlik, J. K-CAP'01 , 139-146. 10-23-2001. Victoria, British Columbia, Canada, ACM. 10-22-0010. Ref Type: Conference Proceeding

Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A., and Barrell,B. (2000). Artemis: sequence visualization and annotation. Bioinformatics. *16*, 944-945.

Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W., and Velculescu,V.E. (2002). Using the transcriptome to annotate the genome. Nat. Biotechnol. *20*, 508-512.

Schmitt,A.O., Specht,T., Beckmann,G., Dahl,E., Pilarsky,C.P., Hinzmann,B., and Rosenthal,A. (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res. *27*, 4251-4260.

Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E., Bentolila,S., Birren,B.B.,

Butler,A., Castle,A.B., Chiannilkulchai,N., Chu,A., Clee,C., Cowles,S., Day,P.J., Dibling,T., Drouot,N., Dunham,I., Duprat,S., East,C., Hudson,T.J., and . (1996). A gene map of the human genome. Science *274*, 540-546.

Schweighoffer,F., Ait-Ikhlef,A., Resink,A.L., Brinkman,B., Melle-Milovanovic,D., Laurent-Puig,P., Kearsey,J., and Bracco,L. (2000). Qualitative gene profiling: a novel tool in genomics and in pharmacogenomics that deciphers messenger RNA isoforms diversity. Pharmacogenomics. *1*, 187-197.

Shoemaker,D.D., Schadt,E.E., Armour,C.D., He,Y.D., Garrett-Engele,P., McDonagh,P.D., Loerch,P.M., Leonardson,A., Lum,P.Y., Cavet,G., Wu,L.F., Altschuler,S.J., Edwards,S., King,J., Tsang,J.S., Schimmack,G., Schelter,J.M., Koch,J., Ziman,M., Marton,M.J., Li,B., Cundiff,P., Ward,T., Castle,J., Krolewski,M., Meyer,M.R., Mao,M., Burchard,J., Kidd,M.J., Dai,H., Phillips,J.W., Linsley,P.S., Stoughton,R., Scherer,S., and Boguski,M.S. (2001). Experimental annotation of the human genome using microarray technology. Nature *409*, 922-927.

Stevens,R., Baker,P., Bechhofer,S., Ng,G., Jacoby,A., Paton,N.W., Goble,C.A., and Brass,A. (2000). TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. *16*, 184-185.

Stoilov,P., Meshorer,E., Gencheva,M., Glick,D., Soreq,H., and Stamm,S. (2002). Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA Cell Biol. *21*, 803-818.

Strausberg,R.L., Camargo,A.A., Riggins,G.J., Schaefer,C.F., de Souza,S.J., Grouse,L.H., Lal,A., Buetow,K.H., Boon,K., Greenhut,S.F., and Simpson,A.J. (2002).

An international database and integrated analysis tools for the study of cancer gene expression. Pharmacogenomics J. *2*, 156-164.

Strausberg,R.L., Feingold,E.A., Klausner,R.D., and Collins,F.S. (1999). The mammalian gene collection. Science *286*, 455-457.

Strehler,E.E. and Zacharias,D.A. (2001). Role of Alternative Splicing in Generating Isoform Diversity Among Plasma Membrane Calcium Pumps. Physiol Rev. *81*, 21-50.

Thanaraj,T.A. (1999). A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures. Nucleic Acids Res. *27*, 2627-2637.

Unsworth,B.R., Hayman,G.T., Carroll,A., and Lelkes,P.I. (1999). Tissue-specific alternative mRNA splicing of phenylethanolamine N- methyltransferase (PNMT) during development by intron retention. Int. J. Dev. Neurosci. *17*, 45-55.

Valentine,C.R. (1998). The association of nonsense codons with exon skipping. Mutat. Res. *411*, 87-117.

Velculescu,V.E., Vogelstein,B., and Kinzler,K.W. (2000). Analysing uncharted transcriptomes with SAGE. Trends Genet. *16*, 423-425.

Velculescu,V.E., Zhang,L., Vogelstein,B., and Kinzler,K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., Amanatides,P., Ballew,R.M.,

Huson,D.H., Wortman,J.R., Zhang,Q., Kodira,C.D., Zheng,X.H., Chen,L., Skupski,M., Subramanian,G., Thomas,P.D., Zhang,J., Gabor Miklos,G.L., Nelson,C., Broder,S., Clark,A.G., Nadeau,J., McKusick,V.A., Zinder,N., Levine,A.J., Roberts,R.J., Simon,M., Slayman,C., Hunkapiller,M., Bolanos,R., Delcher,A., Dew,I., Fasulo,D., Flanigan,M., Florea,L., Halpern,A., Hannenhalli,S., Kravitz,S., Levy,S., Mobarry,C., Reinert,K., Remington,K., Abu-Threideh,J., Beasley,E., Biddick,K., Bonazzi,V., Brandon,R., Cargill,M., Chandramouliswaran,I., Charlab,R., Chaturvedi,K., Deng,Z., Di,F., V, Dunn,P., Eilbeck,K., Evangelista,C., Gabrielian,A.E., Gan,W., Ge,W., Gong,F., Gu,Z., Guan,P., Heiman,T.J., Higgins,M.E., Ji,R.R., Ke,Z., Ketchum,K.A., Lai,Z., Lei,Y., Li,Z., Li,J., Liang,Y., Lin,X., Lu,F., Merkulov,G.V., Milshina,N., Moore,H.M., Naik,A.K., Narayan,V.A., Neelam,B., Nusskern,D., Rusch,D.B., Salzberg,S., Shao,W., Shue,B., Sun,J., Wang,Z., Wang,A., Wang,X., Wang,J., Wei,M., Wides,R., Xiao,C., Yan,C., Yao,A., Ye,J., Zhan,M., Zhang,W., Zhang,H., Zhao,Q., Zheng,L., Zhong,F., Zhong,W., Zhu,S., Zhao,S., Gilbert,D., Baumhueter,S., Spier,G., Carter,C., Cravchik,A., Woodage,T., Ali,F., An,H., Awe,A., Baldwin,D., Baden,H., Barnstead,M., Barrow,I., Beeson,K., Busam,D., Carver,A., Center,A., Cheng,M.L., Curry,L., Danaher,S., Davenport,L., Desilets,R., Dietz,S., Dodson,K., Doup,L., Ferriera,S., Garg,N., Gluecksmann,A., Hart,B., Haynes,J., Haynes,C., Heiner,C., Hladun,S., Hostin,D., Houck,J., Howland,T., Ibegwam,C., Johnson,J., Kalush,F., Kline,L., Koduru,S., Love,A., Mann,F., May,D., McCawley,S., McIntosh,T., McMullen,I., Moy,M., Moy,L., Murphy,B., Nelson,K., Pfannkoch,C., Pratts,E., Puri,V., Qureshi,H., Reardon,M., Rodriguez,R., Rogers,Y.H., Romblad,D., Ruhfel,B., Scott,R., Sitter,C., Smallwood,M., Stewart,E., Strong,R., Suh,E., Thomas,R., Tint,N.N., Tse,S., Vech,C., Wang,G., Wetter,J., Williams,S., Williams,M., Windsor,S., Winn-Deen,E., Wolfe,K.,

Zaveri,J., Zaveri,K., Abril,J.F., Guigo,R., Campbell,M.J., Sjolander,K.V., Karlak,B., Kejariwal,A., Mi,H., Lazareva,B., Hatton,T., Narechania,A., Diemer,K., Muruganujan,A., Guo,N., Sato,S., Bafna,V., Istrail,S., Lippert,R., Schwartz,R., Walenz,B., Yooseph,S., Allen,D., Basu,A., Baxendale,J., Blick,L., Caminha,M., Carnes-Stine,J., Caulk,P., Chiang,Y.H., Coyne,M., Dahlke,C., Mays,A., Dombroski,M., Donnelly,M., Ely,D., Esparham,S., Fosler,C., Gire,H., Glanowski,S., Glasser,K., Glodek,A., Gorokhov,M., Graham,K., Gropman,B., Harris,M., Heil,J., Henderson,S., Hoover,J., Jennings,D., Jordan,C., Jordan,J., Kasha,J., Kagan,L., Kraft,C., Levitsky,A., Lewis,M., Liu,X., Lopez,J., Ma,D., Majoros,W., McDaniel,J., Murphy,S., Newman,M., Nguyen,T., Nguyen,N., and Nodell,M. (2001). The sequence of the human genome. Science *291*, 1304-1351.

Wheelan,S.J., Church,D.M., and Ostell,J.M. (2001). Spidey: A Tool for mRNA-to-Genomic Alignments. Genome Res. *11*, 1952-1957.

Wolfsberg,T.G. and Landsman,D. (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. *25*, 1626-1632.

Xu,Q., Modrek,B., and Lee,C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. *30*, 3754-3766.

Zacharias,D.A., Dalrymple,S.J., and Strehler,E.E. (1995). Transcript distribution of plasma membrane Ca2+ pump isoforms and splice variants in the human brain. Brain Res. Mol. Brain Res. *28*, 263-272.

Zhang,Z., Schwartz,S., Wagner,L., and Miller,W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. *7*, 203-214.