Generation of a human gene index and its application to disease candidacy

by

Alan Christoffels

Thesis presented in fulfillment 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

September 2001

Supervisor: Prof. Winston Hide

ABSTRACT

Generation of a human gene index and its application to disease candidacy.
Alan Christoffels
PhD Thesis, South African Bioinformatics Institute, Department of Biochemistry, University of Western Cape.

With easy access to technology to generate expressed sequence tags (ESTs), several groups have sequenced from thousands to several thousands of ESTs. These ESTs benefit from consolidation and organization to deliver significant biological value. A number of EST projects are underway to extract maximum value from fragmented EST resources by constructing gene indices, where all transcripts are partitioned into index classes such that transcripts are put into the same index class if they represent the same gene. Therefore a gene index should ideally represent a non-redundant set of transcripts. Indeed, most gene indices aim to reconstruct the gene complement of a genome and their technological developments are directed at achieving this goal. The South African National Bioinformatics Institute (SANBI), on the other hand, embarked on the development of the sequence alignment and consensus knowledgebase (STACK) database that focused on the detection and visualisation of transcript variation in the context of developmental and pathological states, using all publicly available ESTs. Preliminary work on the STACK project employed an approach of partitioning the EST data into arbitrarily chosen tissue categories as a means of reducing the EST sequences to manageable sizes for subsequent processing. The tissue partitioning provided the template material for developing error-checking tools to analyse the information embedded in the error-laden EST sequences. However, tissue partitioning increases redundancy in the sequence data because one gene can be expressed in multiple tissues, with the result that multiple tissue partitioned transcripts will correspond to the same gene. Therefore, the sequence data represented by each tissue category had to be merged in order to obtain a comprehensive view of expressed transcript variation across all available tissues. The need to consolidate all EST information provided the impetus for developing a STACK human gene index, also referred to as a whole-body index.

In this dissertation, I report on the development of a STACK human gene index represented by consensus transcripts where all constituent ESTs sample single or multiple tissues in order

to provide the correct development and pathological context for investigating sequence variation. Furthermore, the availability of a human gene index is assessed as a disease-candidate gene discovery resource.

A feasible approach to construction of a whole-body index required the ability to process error-prone EST data in excess of one million sequences (1,198,607 ESTs as of December 1998). In the absence of new clustering algorithms, at that time, we successfully ported D2_CLUSTER, an EST clustering algorithm, to the high performance shared multiprocessor machine, Origin2000. Improvements to the parallelised version of D2_CLUSTER included:
(i) ability to cluster sequences on as many as 126 processors. For example, 462000 ESTs were clustered in 31 hours on 126 R10000 MHz processors, Origin2000.
(ii) enhanced memory management that allowed for clustering of mRNA sequences as long as 83000 base pairs.
(iii) ability to have the input sequence data accessible to all processors, allowing rapid access to the sequences.
(iv) a restart module that allowed a job to be restarted if it was interrupted.

The successful enhancements to the parallelised version of D2_CLUSTER, as listed above, allowed for the processing of EST datasets in excess of 1 million sequences. An hierarchical approach was adopted where 1,198,607 million ESTs from GenBank release 110 (October 1998) were partitioned into "tissue bins" and each tissue bin was processed through a pipeline that included masking for contaminants, clustering, assembly, assembly analysis and consensus generation. A total of 478,707 consensus transcripts were generated for all the tissue categories and these sequences served as the input data for the generation of the whole-body index sequences. The clustering of all tissue-derived consensus transcripts was followed by the collapse of each consensus sequence to its individual ESTs prior to assembly and whole-body index consensus sequence generation.

The hierarchical approach demonstrated a consolidation of the input EST data from 1,198607 ESTs to 69,158 multi-sequence clusters and 162,439 singletons (or individual ESTs). Chromosomal locations were added to 25,793 whole-body index sequences through assignment of genetic markers such as radiation hybrid markers and généthon markers. The whole-body index sequences were made available to the research community through a sequence-based search engine (http://ziggy.sanbi.ac.za/~alan/researchINDEX.html).

The accuracy of the whole-body index was assessed using the genomic sequence for the annotated 599 chromosome 22 genes. A total of 63.3% of the chromosome 22 genes had significant identity to whole-body index sequences. In addition, 25 whole-body index sequences matched regions of chromosome 22 that were not previously annotated. Alignment of whole-body indices to chromosome 22 genes demonstrated a 0.96 fold redundancy in STACK, similar to the radiation hybrid mapping data. A total of 84,387 genes in the human genome were estimated from the chromosome 22 verified whole-body index sequences. Two novel splice variants were identified in the whole-body index clusters corresponding to neurofibromatosis2 gene and fibulin1 gene. A detailed report for the characterised events in 25 known alternatively spliced genes, present in EST assemblies, can be viewed at http://www.sanbi.ac.za/~alan/twentyfive_splicegenes.htm). In addition, 493 indices that mapped onto chromosome 22 genes were analysed and classified as exon sequence (349/493; 5 exon-skips), intron sequences (3/493), gapped exons (8/493; 3 exon skips) and combined intron-exon transcripts (133/493; 8 exon skips).

The STACK human gene index was applied to a human genetic project aimed at identifying the causative gene for progressive familial heartblock1 (PFHB1) on chromosome 19. The work presents an integration of the genetic and physical maps for the *PFHB1* locus, STACK and BodyMap transcripts, mouse developmental ESTs and RefSeq contigs. Potential novel microsatellites were identified in 29 out of 36 BAC and cosmid clones. PHRAP assembly reduced the 1184 chromosome 19 genomic fragments to 370 contigs and 874 singletons. The assemblies were annotated by mapping 119 STACK transcripts, 24 BodyMap transcripts, 54 mouse ESTs and six RefSeq contigs. Seven positional candidates, previously demonstrated to be expressed in heart tissue, have been identified including GLTSRC2, DKF2P761A179, Kaptin, T-elongation factor 4, nucleobindin, CHI-123 protein and CD37-antigen.

DECLARATION

I declare that *Generation of a human gene index and its application to disease candidacy* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have or quoted have been indicated and acknowledged by complete references.

Alan Christoffels September 2001

Signed

BIOGRAPHICAL SKETCH

Alan Christoffels was born in Kensington, Cape Town on 16 November 1970. He attended Bridgeville Primary and matriculated in 1988 from Spes Bona High School. Alan registered for a Bachelor of Science degree at University of Cape Town (UCT) in 1989 and completed the BSc (Microbiology and Biochemistry) in 1992 before enrolling for an Honours degree in Pharmacology the following year. Upon completion of the BSc(hons) degree at UCT in 1993, Alan worked for a year as a research assistant at Tygerberg medical campus, University of Stellenbosch. In 1995, he commenced his studies towards a MSc (Medical Biochemistry) degree at the University of Stellenbosch where he worked in a human genetics laboratory. He completed his MSc degree in 1997 which was entitled “Identification of of novel CA repeat markers for use in fine mapping of the *PFHB1* gene and screening of the *HRC* candidate gene”. Alan furthered his career in genetics at another level by enrolling for a PhD degree at the South African National Bioinformatics Institute, University of Western Cape (UWC) where he worked exclusively in an *in silico* environment on a project entitled “The development of a human gene index and its application to disease candidacy.

PUBLICATIONS ARISING FROM THIS THESIS

Papers:

John E. Carpenter, **Alan Christoffels**, Yael Weinbach and Winston A. Hide. (2001) Assessment of the Parallelization approach of *d2_cluster* for high performance sequence clustering. *Journal of Computational Chemistry*. Accepted.

Christoffels, A.G., A. van Gelder, G. Greyling, R.T. Miller, T. Hide and W. Hide. (2001) STACK: Sequence tag alignment and consensus knowledge base. *Nucleic Acids Res.* **29:** 234-238.

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

Hide W., J. Burke., **A. Christoffels** and R. Miller. (1997) A novel approach towards a comprehensive consensus representation of the expressed human genome. *Genome Informatics. Universal Academy Press Inc. Tokyo Japan. Ed. S. Muiyano and T. Takagi.* 187-195.

ACKNOWLEDGEMENTS

This study was performed in the South African National Bioinformatics Institute, University of Western Cape.

All praise and honour goes to God for sustaining me and providing me with wisdom and strength to reach the culmination of a 12 year period of study.

A special thanks goes to my study leader, Prof. Winston Hide, for his enthusiasm and support during the course of my study. Thanks for your willingness to support my international exposure as it enriched my training.

My heartfelt thanks to my friends and colleagues for their moral support and technical assistance over the past five years especially Junaid Gamieldien, Robert Miller, Ekow Oppon and Peter van Heusden.

I am grateful to Dr. Valerie Corfield for her constructive criticisms of part of my thesis.

Thanks to Mr. Antoine van Gelder for his assistance in deciphering D2_CLUSTER.

Thanks for teaching me how to use the computer facility at the National Center for Supercomputing Applications.

Thanks to Melvine de Vos and Bruce Andrews for their assistance in generating high quality graphics.

I have to acknowledge the support of my parents who have allowed me the freedom to pursue my dreams. To Mervyn, Liesl, Melita and Bruce, thanks for all your help financially, spiritually and emotionally.

Suzanne, thanks for your encouragement during the times when my thesis felt overwhelming.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Literature Review and thesis rationale

Literature Review and thesis rationale

List of Figures

1 The development of a human gene index

1.1 What is a gene index?

A gene index is ideally a collection of information about genes in which all the information pertaining to a particular gene is organised into a single gene class, and each gene class is distinct from all other gene classes (Aaronson et al., 1996; Burke et al., 1999; Jongeneel 2000). The information needed to generate such an index (or catalog) of human genes has been provided by the high sequence coverage resulting from the recently completed draft sequence of the human genome together with the ongoing high throughput efforts to sequence the transcribed portions of the human genome, hereafter referred to as the transcriptome (International Genome Consortium 2001; Celera 2001). However, despite this wealth of information, there remains disagreement as to the number of genes in the human genome (Ewing and Green 2000; Liang et al., 2000a; International Genome Consortium 2001). The varying gene numbers derived for the human genome potentially arise from the different methods used to analyse the transcriptome, specifically the analysis of expressed sequence tags (ESTs).

Expressed sequence tags (ESTs) represents partial cDNA sequences that have been sequenced once (also referred to as single-pass) and have provided the most comprehensive window into the transcriptome (Bortoluzzi et al., 2000; Rezvani et al., 2000). The partial and error-prone nature of ESTs have complicated the definition of a set of gene classes that form an index. However, the ability to deal with partial reads and low quality sequences result in more accurate gene indexing that facilitate expression studies (Schmitt et al., 1999; Piétu et al., 1999; Claverie 1999), highlight gene sequence diversity and splicing (Wolfsberg and Landsman 1997; Miller et al., 1999; Christoffels et al., 2001) and accelerate gene discovery (Bortoluzzi et al., 2000). The objective of this review is two-fold namely, (1) to describe the characteristics of EST data and how it influences the approaches taken to develop gene indices and (2) the utility of EST resources to accelerate candidate disease gene discovery.

1.2 Expressed sequence tags (ESTs)

1.2.1 Generation of ESTs

Random sequencing of cDNA clones has been used for nearly 20 years as a method for gene discovery (Costanzo et al., 1983). This technique more recently has been termed expressed sequence tag analysis and has resulted in the discovery of a variety of human genes (Adams et al., 1991, 1995; Wilcox et al., 1991; Okubo et al., 1992; Houlgatte et al., 1995). An EST, is a fragment of a cDNA clone that has been sequenced and represents a tiny portion of an entire gene. The process by which ESTs are manufactured requires the construction of a cDNA library (Figure 1.1; Bonaldo et al., 1996).

Bonaldo et al. (1996) have provided a detailed description of how cDNA libraries are constructed and how normalisation and library subtraction can be used to increase relative representation of less abundantly transcribed mRNAs. Briefly, the starting material for the construction of a cDNA library includes total RNA from a specific tissue or specific developmental stage of embryogenesis. From this material, poly(A) mRNA is extracted by specifically binding to a complementary polynucleotide which is bound to a solid matrix (eg., oligo(dT) bound to cellulose). The poly(A) mRNA selectively binds to the oligo(dT) components and can then be eluted using buffers of high ionic strength to dissociate the hydrogen bonding. The isolated poly(A) mRNA can then be converted to a double-stranded mRNA/cDNA hybrid using reverse transcriptase. The double-stranded hybrid is rendered single-stranded by the addition of RNase H, which specifically digests RNA that is bound to DNA. The single-stranded cDNA can be used as a template to synthesis a complementary strand using DNA polymerase and the double-stranded cDNA can then be cloned (Figure 1.1). The collection of clones produced for the total mRNA represents the cDNA library.

Each clone contains cDNA whose sequence length varies depending on the time taken for the reverse transcriptase to terminate the production of cDNA. The varying cDNA length is an important factor for development of coverage for each mRNA template of an available gene. Clones are sequenced once, from one or both ends of the DNA insert, using universal primers that are complementary to the vector at the multiple cloning site. The M13 forward primer may be located near the 5’ or the 3’ end of the cloned insert, depending on how the inserts were directionally cloned. Certain EST generation techniques use random primers, which results in production of fragments without direction, originating from different non-overlapping parts of the same mRNA (Kapros et al., 1994). Irrespective of the type of protocol used to generate the ESTs, approximately 300-500 readable bases are produced from each sequencing read, yet a full gene transcript may be several thousands of bases long. ESTs therefore serve as a tag for an expressed gene sequence, trading quality and total sequence length for the high quantity of genes that can be tagged in a given amount of time.

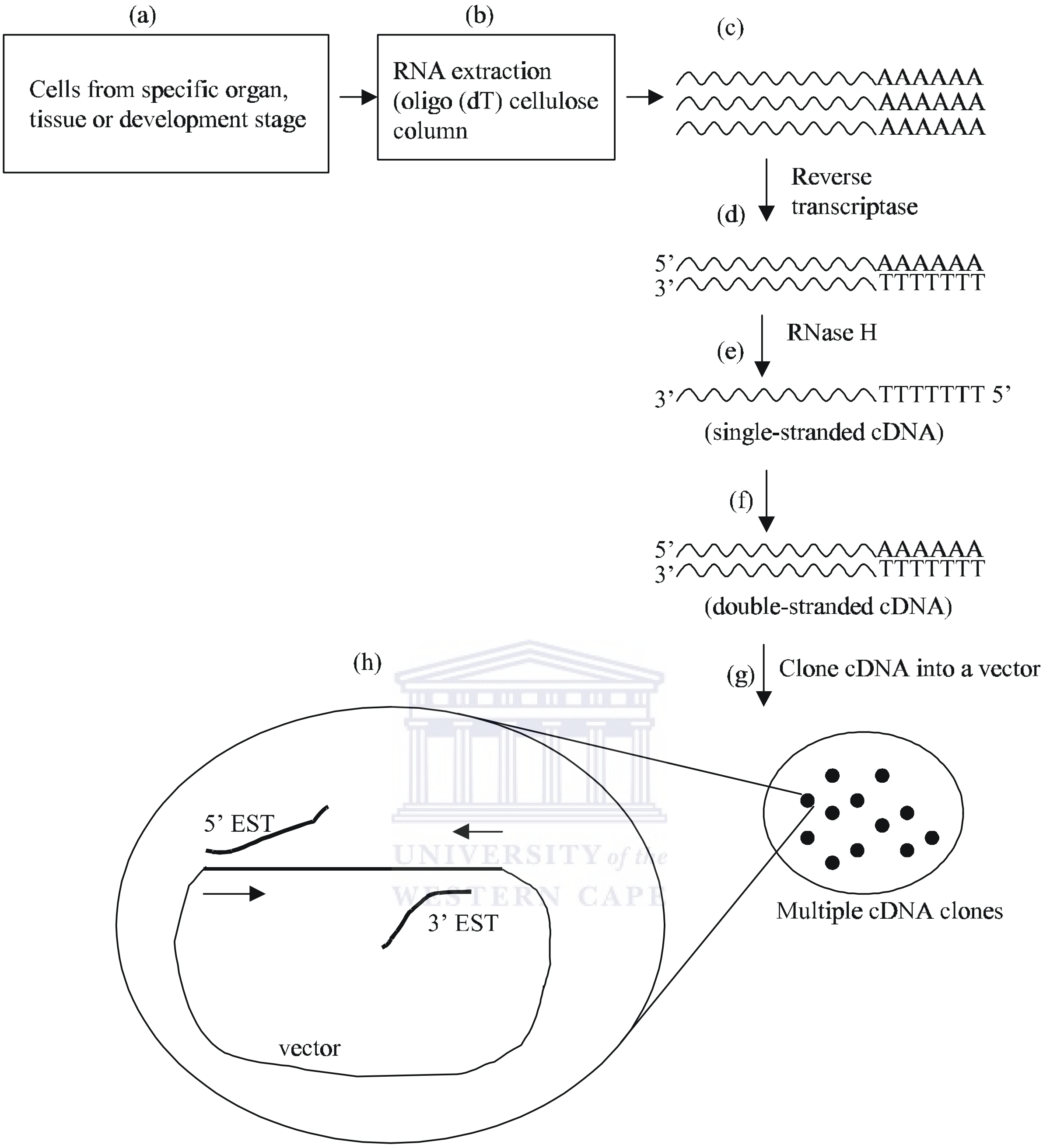


Figure 1.1 Diagram illustrating the manufacture of ESTs from poly-T primed mRNA cDNA library construction (a-g). Total RNA is extracted from a specific tissue or specific developmental stage (a-b). From this material, poly(A) mRNA is extracted by specifically binding to a complementary polynucleotide which is bound to a solid matrix (eg., oligo(dT) bound to cellulose) (c). The poly(A) mRNA selectively binds to the oligo(dT) components and can then be eluted using buffers of high ionic strength to dissociate the hydrogen bonding. The isolated poly(A) mRNA can then be converted to a double-stranded mRNA/cDNA hybrid using reverse transcriptase (d). The double-stranded hybid is rendered single-stranded by the addition of RNase H, which specifically digest RNA that is bound to DNA (e). The single-stranded cDNA can be used a a template to synthesis a complementary strand using DNA polymerase and the double-stranded cDNA can then be cloned (f). The collection of clones produced for the total mRNA represents the cDNA library (g). An EST is generated by sequencing a clone insert once from the 5' and/or 3' end (h). The clone insert lengths vary from clone to clone. ESTs are generated that range approximately from 300-500 bases.

1.2.2 EST quality

1.2.2.1 Errors arising from the EST manufacture process

Generation of EST data results in ‘low quality’ sequence information. A single read is generated for each EST, and as such will contain errors from its generation at each step including basecalling and compression errors that result in frameshifts (http://genome.wustl.edu/est/esthmpg.html). In addition, Aaronson et al (1996) assessed a variety of error types in EST data including (a) lane-tracking errors, (b) insert lengths, (c) clone end reversal and (d) internal priming. The aforementioned EST error classes are detailed below.

Lane-tracking errors

The rate of sequences exhibiting lane-tracking errors was approximately 0.5% (Aaronson et al., 1996). The accuracy with which sequences are mapped onto clones has implications for generating a human gene index as it provides a level of confidence that all sequences with the same cloneID are derived from the same transcript and should be present in the same gene class if there exist some degree of sequence overlap.

Insert lengths

Accurate insert size data for cDNA clones enables the selection of the longest clone for each index class which is valuable for performing assemblies within a class. However, the average error over most EST clones was reported to be between 15-20% (Aaronson et al., 1996).

Clone end reversal and internal priming

Reversed clones and internal priming can result in the incorrect identification of sequences as 3’ ends of genes. The error rates for reversed clones and internal priming were estimated at 5% and 2-3% respectively (Aaronson et al., 1996).

The EST error types outlined above contribute to regions of high quality very close to regions of low quality, where quality can be defined as the number of correctly sequenced bases within a known window of reference. It is possible to utilise poor quality sequence as long as relevant strategies for maximising their utility are taken.

1.2.3 EST clustering

With easy access to technology to generate ESTs (Figure 1.1), several groups have sequenced from thousands to several thousands of ESTs (Adams et al., 1991, 1995). The fragmented nature of ESTs (section 1.2.2) hinders the discovery of full-length cDNAs for each human gene. However, in the absence of a reference sequence for each human gene, increased value is added to the redundant, low quality fragmented EST data by attempting to piece together the gene sequences from which the ESTs were derived.

1.2.3.1 What is an EST cluster?

An EST cluster has been defined by Burke et al (1999) as “fragmented EST data (DNA or protein) and (if known) gene sequence data, consolidated, placed in correct context and indexed by gene such that all expressed data concerning a single gene is in a single index class, and each index class contains the information for only one gene”. However, there are EST clustering systems that deviate slightly from the above definition or add additional criteria to the concept of an EST cluster. For example, earlier releases of the STACK database have ignored mRNA and genomic DNA information even though this information was available. STACK clusters were identified according to the tissue of origin for each EST (Miller et al., 1999). There are other databases that do not use tissue of origin as one of the clustering criteria but instead makes this tissue information available (Bouck et al., 1999). The definition of an EST cluster is further complicated by the way in which alternative gene variants are handled. For example, TIGR human gene index separates each alternative variant to a separate cluster (Adams et al., 1995; Quackenbush et al., 2001). UniGene keeps all variants in one group as long as they have some common part.

1.2.3.2 Overview of EST clustering

The earliest reported implementation of an EST clustering procedure was two-fold: (i) ESTs were first submitted to a fast pair-wise sequence comparison to build seed clusters and (ii) These initial clusters were then treated by a slow but accurate alignment procedure (Hiller et al., 1996). This general concept of EST clustering has been adopted by a number of EST projects for the initial stages of data preparation (Cariaso et al., 1999; Quackenbush et al., 2001).

EST clustering is performed as a process that utilises ‘clustering information’ that is less and less definitive. Initially sequence identity provides a good guide to cluster membership.

Shared annotation provides joining information that can be of more variable quality. Thus the number of accurately clustered ESTs is heavily dependent on a strategy that can assign cluster membership based on verifiable criteria; sequence identity is currently the most useful of these. Clustering can be performed with or without sequence consensus generation as detailed in a modern clustering procedure (Hide et al., 1999; http://www.sanbi.ac.za/submissionl.PDF). It is preferable, although more difficult, to manufacture a consensus sequence from each cluster.

A brief description of a modern clustering procedure is outlined below as defined by Hide et al. (1999).

- **Preprocessing**

EST data are known to contain a variety of contaminating sequences that will alter the outcome of a clustering procedure intended to group together sequences that share identical regions. For this reason, all input sequences are masked for repeats and vectors, and formatted for the clustering engine. Sequence quality is often assessed at this step. A minimum number of residues are accepted above a known quality threshold. For example, the South African National Bioinformatics Institute (SANBI) implementation of STACK accepts only masked sequence data above 50 bases in length. The National Center for Biotechnology and Information (NCBI) discards ESTs with a window of less than 100 bases of 'clean' data.

- **Initial clustering**

An initial clustering is performed based on a fast measure of high sequence identity like D2_CLUSTER (Burke et al., 1999). ESTs having a high degree of similarity, detected by such fast, although rough measure are grouped in one cluster. Clusters, formed at this stage require further verification.

- **Assembly**

Assembly is either part of the initial clustering (as used in TIGR_ASSEMBLER (Sutton et al., 1995)) or separated into clustering followed by assembly performed by an assembly package such as PHRAP (P. Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/phrap.htm) or CAP3 (Huang and Madan 1999).

- **Alignment Processing**

Aligned clusters, particularly those generated by a loose clustering engine, need to be processed for errors and alternate forms of expressed sequences. Consensus generation may be a result of this step or a consensus can be accepted directly from the assembly step where the consensus sequence is determined by a majority rule for each nucleotide position. Consensus sequences are chosen based on maximal length.

- **Clone linking**

Cluster consensus sequences can be linked by available information contained in annotation such as cloneID. Clone linking utilises the physically shared cloneID between 3' and 5' EST fragments sequenced from the same starting clone. Linking by clone annotation is an error-prone step as it relies entirely on the accuracy of the sequence annotation and the uniqueness of cloneIDs if data from disparate sources is to be used (Miller et al., 1999).

1.2.3.3 Supervised and unsupervised clustering

EST clustering methods can be divided into two general classes, supervised and unsupervised clustering. In supervised EST clustering, sequences are classified with respect to known reference sequences. In unsupervised clustering, no pre-defined sequences are used and the number of resulting clusters is typically unknown until the end of the clustering procedure. Some EST clustering systems are strictly or partly supervised, like TIGR Gene Index (Quackenbush et al., 2001) and IMAGene (Cariaso et al., 1999), some are totally unsupervised like STACK (Miller et al., Christoffels et al., 2001) and some use a combination of two approaches, like UniGene (Wagner et al., 2000). In the systems using some form of supervision, a genomic DNA and/or a mRNA sequence is used as a core to assemble ESTs.

1.2.3.4 EST clustering approaches

EST clustering methods used in contemporary EST projects can be classified into and not restricted to (i) contig assembly tools (implemented in TIGR's gene index reviewed in section 1.4.1.1), (ii) alignment scoring methods such as FASTA and BLAST (implemented in IMAGene reviewed in section 1.4.2 and UniGene reviewed in section 1.4.1.2), and (iii) non-alignment based scoring methods such as D2_CLUSTER (implemented in STACK reviewed in section 1.4.1.3)

1.2.4 EST contamination

EST data are known to contain a variety of contaminating sequences that will alter the outcome of a clustering procedure intended to group together sequences that share identical regions. These sequence contaminants include (i) vector, bacterial and mitochondrial sequences, (ii) repeats, (iii) microsatellites, (iv) low-complexity regions and (v) chimeric clones

(i) Vector and mitochondrial sequence contamination

The presence of contaminating sequences in the public database, GenBank, was first reported by Lamperti et al (1992). In this study, vector fragments were found in 0.23% of all sequences available at the time. Miller et al (1998) identified slightly more vector contamination in a study that focused on vector contamination dynamics in GenBank from 1992 to 1996. In addition, Miller et al (1998) showed that the percentage of contaminated ESTs were lower than the average contamination for GenBank for the period from 1992 to 1996. Insight into the level of EST contamination was obtained from Hiller et al (1996), who provided a quantitative estimation of the level of contamination in EST libraries. In this study, EST sequences were screened against databases of bacterial sequences, mitochondrial sequences and vector sequences. All libraries contained mitochondrial sequences ranging from a high of 16% of ESTs to as low as 1% of ESTs. Some EST libraries were found to contain as much as 20% of bacterial contamination.

The ability to "clean" EST data from contaminants such as vector sequences require a collection of possible contaminating fragments that can be used as a reference database for identifying the specific contaminant. A collection of cloning vectors, specifically prepared for this purpose is available from NCBI ftp site (ftp://ncbi.nlm.nih.gov/blast/db/vector.Z).

The human genome comprise two genomes: a complex nuclear genome and a simple mitochondrial genome. The bulk of the mitochondrial polypeptides are encoded by the nuclear genes and are synthesized on cytoplasmic ribosomes, before being imported into the mitochondria (Strachan and Read 1997). The presence of mitochondial transcripts in the nuclear genome can result in EST capturing of mitochondrial genes during the sequencing of a cDNA library.

(ii) Repeat sequences

Unlike cloning vector fragments, repeats cannot be regarded as contaminating sequences. Repetitive sequences include LINES (Fanning and Singer 1987), SINES (Singer 1982a), ALU (Deininger 1989) and satellite repeats (Singer 1982b). Despite recent evidence for ALU repetitive sequences as functional binding sites for retinoic acid reponse elements and estrogen receptors (Vansant et al., 1995; Norris et al., 1995 respectively), the functional significance of repetitive sequences remains unclear. However, the unique portion of the genome is thought to comprise the functional constituents of the human genome, including exons, introns and regulatory DNA elements. It is the unique DNA pieces within EST sequences that are being brought together in any effort to cluster ESTs. The unique portions of an EST can be obscured by repetitive sequence and represent a serious challenge for EST clustering.

In newly sequenced genomes, such as plant and other eukaroyte systems, repeat sequences represent a common and frustrating clustering problem. Repeat databases provide a resource against which repeats can be detected. The repeat databases are dependent on continuing curation and detection of novel repeats in genomes and thus provide a valuable resource. Since the early 1990's, the most comprehensive repeat collection, Repbase, has been supported by Genetic Information Research Institute (http://www.girinst.org) (Jurka et al., 1992; 1998).

(iii) **Microsatellites**

Microsatellite DNA families comprise tandem repeats that have repeating units of length 1-6 base pairs, which are interspersed throughout the genome (Tóth et al., 2000). They have been used extensively for genetic mapping and population studies (Gyapay et al., 1994). However, much remains unknown about the possible functions microsatellites may have in the genome. Microsatellite repeats are remarkably variable by number of copies, small deletions, insertions and single base mutations inside the repeat (Bull et al., 1999). The variability and multiplicity of the microsatellite repeats makes their recognition by comparison to a sample sequence, stored in a database, ineffective.

(iv) Low complexity sequences

Low complexity sequence is a more general term for stretches of DNA of low complexity with or without detectable repetitive structure (Jurka 1998). Lack of a certain consensus makes them impossible to detect by comparison to a sample. But like interspersed repeats and microsatellite repeats, low complexity regions also have the potential to provide an artifactual basis for cluster membership. The problem is more significant for strategies that employ alignable similarity in the first pass cluster assignment. Word-based cluster assignment can be modified to provide low weight to low complexity words. The latest version of BLAST (Altschul et al., 1997) widely used as a sequence comparison engine in EST clustering, is capable of filtering out low complexity sequences. In a new EST clustering algorithm developed at SANBI (Ptitsyn 2001), there is no need for masking of low complexity DNA as such regions tend to have a highly redundant oligonucleotide composition. The new sequence comparison algorithm (Ptitsyn 2001) scale oligonucleotides according to their potential information content with the result that highly redundant oligos are given very low weight and low complexity regions are excluded from consideration.

(v) Chimeric clones

Cloning artifacts, i.e., co-ligation of two different restriction fragments, can produce cells that contain two non-contiguous pieces of DNA from the desired genome (Larionov et al., 1994). 5’ and 3’ ends sequenced from a chimeric clone will result in EST pairs sampled from different portions of the genome. EST clustering tools that rely on the clone information of ESTs to determine cluster membership will generate false EST clusters. For example, UniGene assigns ESTs to a cluster based on cloneID. Any 5’and 3’ ESTs that originate from a chimeric clone will result in one EST cluster representing more than one gene. The STACK implementation uses cloneIDs in its clone linking step when EST clusters are joined to generate linked clusters. The occurrence of ESTs from chimeric clones will produce linked clusters that do not sample the same gene.

1.3 Masking strategies

The most effective method to remove contaminants is to compare each read against a reference database of repeats such as RepBase (Jurka et al., 1998) and vector sequences (VecBase, http://vectordb.atcg.com or vector collection at NCBI, ftp://ncbi.nlm.nih.gov/blast/db/vector.Z) using an algorithm that is reasonably fast and accurate. XBLAST (NCBI tools, ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools) and CROSS_MATCH, an implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green (Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/swat.htm) has been used successfully, with CROSS_MATCH demonstrating greater flexibility than XBLAST (Miller et al., 1999). DUST is used for masking repetitive sequences at NCBI (unpublished; http://www.ncbi.nlm.nih.gov/UniGene/Build.html). DUST is able to reveal and mask low-complexity sequences as well. Another recent development, RepeatMasker (Smit and Green 1999), available from Washington University (http://ftp.genome.washington.edu/RM/RepeatMasker.html) is able to mask huge amounts of data and recognise low-complexity DNA, for example, regions of DNA consisting of greater than 84% of CA's or greater than 87% GT's (default settings). Recently, MaskerAid was developed as an enhancement to RepeatMasker where CROSS_MATCH was replaced with WUBLAST (W. Gish, unpublished, http://blast.wustl.edu). MaskerAid was written in PERL and represents a software wrapper around WUBLAST (Korf et al., 2000).

1.4 Implementation strategies of gene indexing projects

1.4.1 Overview of TIGR human gene index, UniGene and STACK

Over the past five years a number of gene indices have been produced that were aimed at addressing the problems associated with ESTs as described in section 1.2.2. At the start of 1996 there were essentially two gene indices being developed namely (a) UniGene at NCBI and (b) the human gene index at the Institute for Genome Research (TIGR). The protocols for the generation of UniGene were not available at that time, but over the past 12 months there has been a release of publications pertaining to the UniGene build in the form of poster presentations (Wagner et al., 2000), manuscripts (Zhang et al., 2000) and internet webpages (http://www.ncbi.nlm.nih.gov/UniGene/build.html). Clues to the approach adopted by TIGR, at that time, to generate its human gene index came from the release of the TIGR_ASSEMBLER (Sutton et al., 1995) and its subsequent application to EST sequences (TIGR_ASSEMBLER-EST). The protocols for the development of gene indices by NCBI and TIGR suggested that strict approaches were being implemented in order to reconstruct the expressed gene complement of the human genome. For example, TIGR_ASSEMBLER approach grouped sequences into a cluster if they shared a minimum of 95% identity over a 40 nucleotide or longer region with fewer than 20 bases of mismatched sequence at either end (Sutton et al., 1995). Besides the use of TIGR_ASSEMBLER, many academic institutions were using PHRAP (P. Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/phrap.htm), an alignment-based assembly tool that is based on the Smith-Waterman algorithm, for the assembly of shotgun sequences.

During 1996, the South African National Bioinformatics Institute embarked on the Sequence Tag Alignment and Consensus Knowledgebase (STACK) project aimed at capturing transcript variation in the context of developmental and pathological states. This approach required additional steps tolerant of sub-sequence diversity, the ability to perform assembly analysis and the ability to handle the exponential increase in EST data (Benson et al., 1999).

The gene indexing projects described below implement a combination of data preparation, clustering, assembly, alignment analysis, consensus generation, clone linking and visualisation.

1.4.1.1 The Institute for Genome Research (TIGR) human gene index

The protocol described below focuses specifically on the TIGR human gene index but can be applied to any of its 29 indices (Liang et al., 2000b; Quackenbush et al., 2001).

a. Data preparation

TIGR human gene index (HGI) incorporates both ESTs and annotated gene sequences that have been submitted to dbEST and GenBank respectively. The first step in the process involves the construction of a database of annotated gene sequences. All sequences from GenBank are downloaded and the CDS and CDS join features for full-length genes and mRNA sequences are parsed from the records. One representative sequence is chosen for each redundant entry but all links to alternative GenBank records are maintained. The annotation of all the human expressed transcripts (HT) are checked for consistency before being loaded into the Expressed Gene Anatomy Database (EGAD; http://www.tigr.org/tdb/egad/egad.html). ESTs are downloaded from dbEST daily and screened for contaminating sequences as outlined in section 1.1.3.

b. Clustering

Clean ESTs, HT sequences from EGAD, tentative human consensus sequences (THCs) from a previous build and singletons are compared pairwise to identify overlaps using a program called FLAST that is based on DDS (Huang et al., 1997). Sequences are grouped into a cluster if they share a minimum of 95% identity over a 40 nucleotide or longer region with fewer than 20 bases of mismatched sequence at either end. THCs are collapsed to their component ESTs and HT sequences prior to cluster assembly using CAP3 (Huang and Madan 1999). Recent articles have reported that a strict pairwise comparison using WUBLAST (W. Gish, unpublished, http://blast.wustl.edu) is being implemented to generate TIGR's gene index clusters (Liang et al., 2000b). All newly constructed THCs are passed through a second round of clustering and assembly to identify and eliminate most of the redundancy introduced during the first phase. The resulting THCs are loaded into the Gene Index database for annotation.

c. Annotation and display

THCs containing a known gene are assigned the function of the gene whereas THCs without assigned functions are searched using DPS (Huang et al., 1997) against a non-redundant protein database. The highest-scoring hits are assigned a putative function. The THC is presented as a FASTA-formatted consensus sequence together with a graphical representation of each component sequence within the cluster, links to GenBank and functional and mapping information where available. A novel assembly, caused by joining or splitting a previous THC assembly, is assigned a new unique identifier. Previously used identifiers are never reused and information regarding previous assemblies is never lost. Database queries using a THC identifier from a previous build return the most current version of that assembly.

d. TIGR's human gene index availability

TIGR maintains gene indices for 33 organisms as of 1st July 2001**.** All of the TIGR databases can be accessed from the TIGR Database page at http://www.tigr.org/tdb/tdb.html. HGI has been generated every 12 months (current release, version 6.0, 30th June 2000) (http://www.tigr.org/tdb/hgi/index.html). Current development at TIGR is focused on a scheduled release where each gene index is updated every three months (http://www.tigr.org/tdb/gifaq.html). The TIGR human gene index (HGI) in particular, can be queried with a sequence or a text string at http://www.tigr.org/. TIGR offers sequence searches at both nucleotide and protein level using WU-BLAST2.0 (http://www.tigr.org/docs/tigr-scripts/nhgi_scripts/tgi_blast.pl?organism=Human).

e. TIGR's human gene index utility

The HGI can be queried for tissue expression information such as (i) tissue specific transcripts, (ii) cDNA libraries by keyword and (iii) cDNA libraries by catalog (http://www.tigr.org/tdb/hgi/searching/xpress_search.html).

1.4.1.2 UniGene

The National Center for Biotechniology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. NCBI provides data retrieval systems and computational resources for the analysis of GenBank data (i.e., nucleic acid sequence database) and a variety of other biological data. UniGene (http://www.ncbi.nlm.nih.gov/UniGene; Schuler 1997) represents an example of a computational resource, developed at NCBI, for automated partitioning of GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, and is linked to related information such as tissue types in which the gene is expressed, model organism protein similarities and its chromosomal map location (Wheeler et al. 2001).

a. Data preparation

ESTs and annotated mRNA sequences obtained from dbEST and GenBank respectively, are utilised in the UniGene build as described below (section 1.4.1.2b). Sequences are screened for contaminants such as mitochondrial, ribosomal, and vector sequences as outlined in section 1.3. Cleaned sequences have to meet the requirement of at least 100 informative nucleotides before it is can be considered for inclusion into UniGene.

b. Clustering

mRNA sequences are compared to each other and significantly similar sequences are grouped to form the initial clusters. ESTs are compared with these initial clusters using megablast (http://www.ncbi.nlm.nih.gov/UniGene/build.html; unpublished). Megablast incorporates a greedy algorithm (Zhang et al., 2001) as suggested on the UniGene website. No documentation exists for the implementation of "the greedy algorithm" in megablast but certain inferences can be made from a similar implementation (Zhang et al., 2001). In their publication, Zhang et al (2001) provides an example of the use of their "greedy algorithm" in one of the BLAST tools (unnamed) at NCBI. In this example, the genomic sequences of two strains of *M. tuberculosis* sequences were aligned. The unnamed NCBI program begins by making a table of 12-mers in the one genomic sequence (a typical BLAST -like approach). The second sequence is scanned for matching 12-mers that are included in a 30bp exact match. At this point the "greedy algorithm" was applied to knit the 30bp matches together into long gapped alignments by recursively picking a longest exact match that does not intersect one of the earlier gapped alignments while extending in both directions.

After comparing ESTs to mRNAs, any links that join two gene clusters that were not joined before the addition of ESTs are discarded. The resulting clusters are kept if they contain a sequence with a polyadenylation site or at least two ESTs that are labelled as 3' ends. These "anchor" clusters are merged if they share at least two identical cloneIDs with both 5' and 3' ends. ESTs that do not belong to an anchored cluster are rechecked at a lower megablast stringency (unpublished). ESTs that find matching anchor clusters, as a result of the lower level of stringency, are added to the cluster that showed the best match. Singletons are compared against the rest of the UniGene sequences at a lower level of stringency (unpublished) and added to the cluster that contains the most similar sequence. Clusters are then compared with the previous week's build and renumbered.

c. UniGene availability

UniGene databases exist for human, mouse, rat and zebrafish sequences and are updated weekly with new ESTs and bimonthly with newly characterised sequences (Wheeler et al., 2001). Clustered sequences can be download via the ftp (ftp://ncbi.nlm.nih.gov/repository/UniGene).

UniGene clusters may be searched by gene name, chromosomal location, cDNA library, accession number and ordinary text words. Sequence-based searching against the UniGene database is available at the Swiss Institute for Bioinformatics (http://www.ch.embnet.org).

d. UniGene utility

The UniGene collection has been used as a source of mapping candidates for the construction of a human gene map (Deloukas et al., 1998). In this study, 3’ untranslated regions (UTRs) are converted to sequence tag sites (STSs; see section 2.1.1) that are then placed on physical maps and integrated with pre-existing genetic maps of the genome (Deloukas et al., 1998).

The UniGene collection has been used as a source of unique sequences in micro-array chip design for large-scale study of gene expression (Ermolaeva et al., 1998).

1.4.1.3 Sequence Tag Alignment and Consensus Knowledgebase (STACK)

STACK is a tool for detection and visualisation of expressed transcript variation in the context of developmental and pathological states. The data system organises and reconstructs human transcripts from available public data in the context of expression state that is the captured expression of a transcript such as developmental state, pathological association, site of expression and/or isoform of expressed transcript. Comprehensive capture of transcript variants is achieved by the use of a novel clustering approach that is tolerant of sub-sequence diversity and does not rely on pairwise alignment (Christoffels et al., 2001). The STACK database represents a consolidation of all publicly available EST data through clustering after characterising ESTs into arbitrary tissue bins. The clustering procedure includes subpartitioning, masking, clustering, assembly, alignment analysis, consensus partitioning and clone linking (Miller et al., 1999).

a. Subpartitioning

The first step involves selection of human ESTs from GenBank and their partitioning into arbitrary-selected tissue bins. The tissue-partitioning step generates EST data sets that have a managable size for input into the clustering and assembly engines. Sequences that are annotated as derived from a disease-related tissue are duplicated and placed in a single set to facilitate exploration of cross-tissue similarities between these ESTs. The "tissue_type" subkey of the "FEATURES" key is only provided sometimes with nonstandardised terms in the data field. As a result, the assignment of an output file name for each sequence is based on (1) FEATURES/tissue_type, (2) FEATURES/cell_type, (3) FEATURES/clone_lib or SOURCE/library, (4) FEATURES/chromosome or (5) FEATURES/map.

b. Masking

Sequences are masked as outlined in section 1.3.

c. Clustering

SANBI has implemented a clustering approach that differs from the TIGR gene index and UniGene. The use of D2_CLUSTER for the clustering step in STACK is central to the detection of sequence variation. D2_CLUSTER (Hide et al., 1994, 1997) is a word multiplicity comparison method that utilizes an agglomerative algorithm that has been specifically developed for rapidly and accurately partitioning transcript sequences into index classes by clustering ESTs and full-length sequences according to minimal linkage or "transitive closure" rules. Agglomerative clustering method means that every sequence begins in its own cluster and the final clustering is constructed through a series of merges that may be described in terms of minimal linkage, sometimes called single linkage or "transitive closure". The term transitive closure refers to the property that any two sequences with a given level of similarity will be in the same cluster, hence A and B are in the same cluster even if they share no similarity but there exists a sequence C with enough similarity to both A and B (Burke et al., 1999). The criterion for joining clusters is the detection of two sequences that share a window of (Window_Size) bases that is (Stringency) percent or more identical. The only criterion for clustering is sequence overlap and source or annotation information is not used. To detect the overlap criterion, the d^2 algorithm is used with parameters and threshold values as described in (Torney et al, 1990; Hide, et al, 1994; Wu et al, 1997). The initial and final state of the algorithm is a partition of the input sequences where each sequence is in a cluster and no sequence appears in more than one cluster. D2_CLUSTER uses an approach of word matching within a window, together with a measure of the multiplicity (if any) of that word within a window. The principal concept is that it doesn't attempt an alignment, not even in a reduced form. The results of comparison are derived directly from the comparison of word composition (word identity and multiplicity) of two sequence windows. Thus, the algorithm can be significantly faster than BLAST. Speed comes with a price: to collect significant statistics, the fragments must be long enough (about 50 bp) and only very high similarities can be detected (above 96% identity within a window). D2_CLUSTER is used to produce initial loose clusters in the STACK clustering system. The results of D2_CLUSTER alone are between 8% and 20% less fragmented than Unigene (Burke et al., 1999) and the STACK data system produces clean clusters that are 16% less fragmented than Unigene (Miller et al., 1999).

d. Assembly

PHRAP (P. Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/phrap.htm) is used for the assembly of all clusters.

SANBI and TIGR implement strategies that generate consensus sequences during or after the sequence assembly step. STACK does not rely on the consensus sequences generated by the PHRAP assembly and instead a consensus sequence is generated during its assembly analysis phase. Recent reports by TIGR have documented the use of CAP3 (Huang and Madan 1999) to generate the consensus sequences for the TIGR gene indices (Liang et al., 2000b). UniGene on the other hand does not produce alignments and therefore does not generate consensus sequences. The absence of alignments in UniGene can be explained by the presence of non-overlapping sequences that have been assigned to a cluster based on the sequence origins from the same clone.

e. Alignment Analysis

To distinguish alternative splicing from problematic alignments introduced by low quality sequence or partially divergent members of otherwise closely related families, cluster alignments in STACK undergo additional processing that produces sub-clusters (Miller et al., 1999; Christoffels et al., 2001). CRAW and STACK_analysis have been developed to address post-clustering and assembly artifacts. CRAW is used to maximize consensus length, partition subassemblies and provide a simple means to view clusters (Burke et al. 1998). CRAW checks the agreement along the columns of a multiple sequence alignment and uses this information to sort related sequences within each cluster and generates a consensus sequence for each sub-cluster. A sub-cluster is generated if 50% or more of a 100-base window differs from the remaining sequences of a cluster, excluding the initial 100 bases of any read. The approach depends fundamentally on the alignment quality of each assembly generated by the assembly tool. For example, a poor alignment will yield erroneous sub-clusters, and too low a gap penality may yield too many columns in agreement and thus not create subclusters where they would be appropriate.

f. Consensus Partitioning

STACK_analysis independently partitions the aligned sequences generated from the CRAW consensus sequences then ranks the consensus sequences according to the number of assigned sequences and number of called bases. The best ranking consensus sequence, defined as the sequence with the most contributing ESTs, is taken as the primary representative of a cluster, whereas the remaining consensus sequences are logged with the best consensus sequence in Genetic data environment (GDE, Smith et al., 1994) file format (Miller et al., 1999; Christoffels et al., 2001). The 5' or 3' orientation of each cluster is determined by a vote of the individual EST annotations and all output consensus sequences are arranged to read 5' to 3'. Low-quality regions defined as 2 N's followed by at least thirteen IUPAC codes with four or less clear A, T, C or G calls are replaced by a single run of 10 N's.

g. Clonelinking

The clone information is used to extend the length of the cluster consensus sequences by joining clusters containing ESTs with shared cloneIDs. Clone-links are accepted if two independent clones link the same two clusters (Figure 3.3). Each EST from GenBank is searched for clone information to trace the transcripts corresponding to the same gene. Clone-linked consensus sequences are ordered 5'-unassigned-3' based on a majority rule from the EST annotations in each cluster (Miller et al., 1999).

h. STACK availability

A new release of the STACK database is made available at least four times a year. STACK is freely available to academia and is distributed via the Web at http://www.sanbi.ac.za/CODES. The stackPACK tool set performs clustering, clustering management, alignment processing and analysis and is freely available to academic institutions and is distributed from http://www.sanbi.ac.za/CODES.

i. STACK utility

The STACK database can be queried via the Web at http://www.sanbi.ac.za/stacksearch.html using a sequence as input. The BLAST search algorithms implemented in the search engine allow for both DNA and protein queries. The results of a blast query are hyperlinked to the STACK viewer, which allows for the extraction of detailed information pertaining to the matching STACK sequence. STACK consensus sequences matched to *Drosophila* sequences are searchable on the Drosophila Related Expressed Sequences (DRES) home page at the Telethon Institute of Genetics and Medicine (http://www.tigem.it). Alternately, all clustered data for a specific STACK tissue category can be accessed via WebProbe (http://www.sanbi.ac.za/stackpack/webprobe.html). A query from this page returns a summary report with links to detailed information for all clusters and linked clusters contained within the specified tissue category.

1.4.2 Other gene index implementations

IMAGene

A large scale and systematic public effort to isolate all human genes began in 1993 when the Integrated Molecular Analysis of Genomes and their Expression (I.M.A.G.E) consortium was formed to create, collect and characterize cDNA libraries from various tissues and states of normalization (Lennon et al., 1996). The ultimate goal of the I.M.A.G.E consortium was to provide a collection of clones that best represent the mRNAs found in GenBank for use in re-arraying of clones into minimal redundant micro-arrays (Cariaso et al., 1999). To this end, the IMAGene suite of tools was designed to analyse and organise ESTs associated with I.M.A.G.E clones, which constitute approximately 75% of all human dbEST sequences. The first release of IMAGene (IMAGene1) was focused on clustering I.M.A.G.E clones associated with mRNAs obtained from GenBank (Cariaso et al., 1999). A recent poster publication suggested the availability of IMAGene3, a program to cluster human ESTs against NCBI's reference set of genes (Refseq) (Prange et al., 2000). The absence of information for the implementation of IMAGene2 and IMAGene3 has limited this review to IMAGene1, which is outlined below.

a. Data preparation

Human ESTs derived from the I.M.A.G.E Consortium clones are extracted from dbEST and screened for poor text annotation and low quality regions. Key features such as cloneID, library name, EST orientation (i.e., 5' or 3') and sequence are identified and formatted into an annotated FASTA record prior to entry into IMAGene. The FASTA records are indexed by a GenBank accession number and I.M.A.G.E cloneID and made blast searchable. Human genes are extracted from the mRNA records in GenBank and redundant entries are removed. The gene set together with the EST data are generated with each build of IMAGene. The first release of IMAGene contained repeats that were not masked.

b. Clustering

For the clustering procedure, IMAGene uses a combination of BLAST and FASTA with wraparound scripts. mRNA sequences are compared against all ESTs using BLAST. The default parameters for BLAST are used as a theshold for candidate EST selection and the 50 best hits are extracted from the indexed EST file. These candidates are copied to a temporary database and examined by FASTA. The speed of BLAST is balanced by the quality of FASTA, as only matches, confirmed by FASTA are accepted. For example, clones are selected from sequences that match a FASTA opt score of 1300 and ESTs derived from those clones are included in the cluster. A number of factors could affect the incorrect assignment of ESTs to a IMAGene cluster. For example, ESTs sharing a conserved domain in different genes could result in multiple assignment of the same EST. Sequencing artifacts where the 5' and 3' ends originate from different clones and alternative splicing could affect a cluster membership.

c. Alignment

The FASTA tool used in the clustering stage generates alignments but this may not find the optimal overall match since FASTA uses a heuristic to locate regions of high similarity. Alignments are therefore generated in a separate step using SIM4 (http://globin.cse.psu.edu; Florea et al., 1998). SIM4 has been written for the purpose of aligning a cDNA sequence to its genomic counterpart under the assumption that the only differences between the two sequences are (1) introns in the genomic sequnece and (2) sequencing errors in either sequence (Florea et al., 1998). Each EST is locally aligned to its associated gene, using SIM4, and the coordinates of the regions that align well are reported. Where necessary the matching regions are extended to ensure full coverage of the EST. These alignments are constructed into a multiple alignment table in which the known gene serves as a consensus sequence.

c. Sorting

Since IMAGene is intended as a tool for re-arraying (see above), its ability to pick the best clone is crucial. All clones within a cluster are sorted by preference; the highest one is considered the tentative candidate for a master array. The factors affecting the preference are: coverage of the coding area, reliability rating of the library and the length of the clone (i.e., clone coverage). The ranking of clones based on clone coverage has demonstrated that genes that average 1580 bases in length are represented in cDNA clones covering the entire coding region, while genes represented by partial length clones average 3063 bases in length (Cariaso et al., 1999). This suggests that the current methods for cDNA clone construction are insufficient to reliably produce clones long enough to fully represent many genes.

d. Display

A web-based user interface allows for a clone search based on the GenBank accession number of a gene or EST, an I.M.A.G.E cloneID or a sequence comparison (http://www-bio.llnl.gov/imagene/bin/search). Initial queries return a table containing information on each cluster that matches the search criteria. Each row of the table of results contains the geneID for that cluster, a description of the gene and the number of full coding and partial length clones contained within the cluster. The geneID is linked to a detailed description that provides a tabular description of each clone on the top of the page and the alignments of each clone or ESTs with the gene on the bottom. A master listing and a candidate_gold listing are generated for public use. The master listing contains the top ranked clone for each known gene cluster and the candidate_gold listing is a subset of the master list containing only clones that cover the coding region.

Merck gene index

In September 1994 Merck announced their plan to develop a collection of ESTs of human genes with associated cDNA clones as a publicly available resource. One year later, the Merck gene index project (MGIP) was initiated as a multicenter collaboration organised and managed by Merck where the sequence data would be generated and processed by the I.M.A.G.E consortium and the Genome Sequencing Center. The design of the MGIP incorporated the use of normalised cDNA libraries and sequencing clones from the 5' and 3' end (Williamson 1999). The use of normalised cDNA libraries provided equal representation of each expressed gene with the result that rare and common transcripts would be present at a similar frequency. These libraries reduced the sequencing redundancy and increased the efficiency as demonstated by Hillier at al (1996). Hiller and coworkers confirmed the reduction in redundancy of cDNA clones due to normalisation and provided evidence for the detection of rarely expressed genes ranging from 0-11.9% novel ESTs.

Overview

cDNA libraries are produced by M. Bento Soares and arrayed, through collaboration with the I.M.A.G.E laboratory of G. Lennon. The arrayed cDNA clones are sent to the Genome Sequence Center for single pass sequencing from each end and then submitted to dbEST. The original ABI sequence trace files for these ESTs are made available from http://genome.wustl.edu/est/est_search/ftp_guide.html. The computational biology and informatics laboratory at the University of Pennsylvania School of Medicine manages the LENS database that provides integration of data and monitors consistency. To date, the MGIP has contributed 79% of the human ESTs deposited into dbEST.

Merck generates and supports the Merck gene index (MGI) for use by in-house researchers. The MGI is a non-redundant set of clones and sequences, each representing a distinct gene, constructed from the 3' EST sequences present in dbEST. The construction of the index is iterative from all index-quality 3' ESTs. Poor quality sequences are eliminated and ESTs have to be at least 100 bases in length to be accepted. Each 3' EST is compared against the index and if it is equivalent to an index entry then that EST is added to that index class. However, if the EST is novel with respect to the index then a new class is created where the 3' EST becomes the representative sequence. Incremental runs are preformed nightly on any new EST data and the results are loaded into a relational database that underlies the MGI browser.

The MGI browser integrates data from different sources including LENS cDNA clones and ESTs, dbEST protein and non-EST nucleic acid similarity data, Washington University sequence chromatograms, Entrez sequence, Medline entries and UniGene gene clusters.

1.4.3 Gene indices incorporating genome data

National Center for Biotechnology and Information

The National Center for Biotechnology and Information has started a project, RefSeq, aimed at providing a single set of reference sequences for each gene, with a consistent and curated annotation (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html). There are currently three primary RefSeq projects namely; curated RefSeq (http://www.ncbi.nlm.nih.gov/LocusLink/build.html), genome annotation (http://www.ncbi.nlm.nih.gov/genome/guide/build.html) and complete genomes (no documentation available as of 8 June 2001). Predicted genes are represented in a minority of RefSeq records and only if strong inferences exist for gene structure and corresponding protein. RefSeq captures the redundancies and inconsistencies in the GenBank/EMBL/DDBJ repositories but it does not provide an index to genes for which there are very fragmented data such as a few ESTs.

AllGenes

The AllGenes website (http://www.allgenes.org) provides access to an integrated database of every identified and predicted human and mouse gene. The AllGenes effort focuses on integrating various types of data including EST sequences, genomic sequence, expression data and functional annotation. The backbone to the AllGenes database is a relational database that uses controlled vocabularies and ontologies to ensure that biologically meaningful queries can be addressed (Stoeckert et al., 2001). The AllGenes database is built around a gene index comprising predicted human and mouse genes. The predicted genes are obtained from (1) the Database of Transcribed Sequences (DoTs), which represents clustered and assembled ESTs and mRNAs, and (2) genes predicted by running gene finders GRAIL-EXP and GENSCAN on available human and mouse genomic sequences.

Swiss Institute of Bioinformatics

The most recent attempt at generating a gene index has focused on the goal of obtaining "true mRNA 3' ends" as reported in a recent editorial (Jongeneel 2000; ftp://ftp.licr.org/pub/databases/tags). A small region proximal to the poly(A) is extracted from each of the ESTs and mapped to the genome data. These mapped locations provide the starting point from which transcripts can be reconstructed with a genome scaffold (Jongeneel 2000). The uniqueness of each 3' tag as an index for each transcript is hampered by alternative polyadenylation (Gautheret et al., 1998; Beaudoing et al., 2000), which allows for more than one tag per transcript. The generation of a database of 3' tags that are extracted proximal to the poly(A) tails of ESTs require the use of trace files since many of the publicly available ESTs have had their poly(A) tails removed (http://genome.wustl.edu/est).

1.4.4 The need for ESTs in the wake of genome data

The interpretation of EST data in the context of genomic sequence provides added value to gene finding projects as it confirms the identification of exon/intron boundaries for plausible candidate genes (eg., Ensembl (http://www.enseml.org). In addition, mapped EST data provide immediate gene expression patterns for a defined chromosomal region. The mapping of processed EST data onto genomic sequence is being incorporated into the STACK project. The expression information gleaned from the EST libraries have been organised into a dictionary of physiological terms that allows the user to retrieve data relevant to a specific developmental or disease stage. A similar implementation has been incorporated into the database of transcribed sequences (http://www.allgenes.org). The STACK and AllGenes projects have developed their expression tissue vocabulary independently. The Allgenes database allows for the retrieval of map coordinates for ESTs positioned on the genome assemblies generated at the University of California Santa Cruz genome center (http://genome.ucsc.edu). The STACK project uses the genome assemblies from the Santa Cruz genome center and provides a genome mapping data and querying tool as part of the STACK clustering system.

1.6 Utilising ESTs as a disease candidate gene discovery resource.

A variety of protocols have incorporated the use of ESTs as a disease-candidate gene discovery resource. These strategies include (a) positional candidate gene approach (Levy-Lahad et al., 1995b; Brown et al., 1998; Hey et al., 1998; Semple et al., 2000), (b) functional candidate gene approach (Hwang et al., 1997; reviewed in Rezvani et al., 2000) and (c) *in-silico* differential display (Ji et al., 1997; Schmitt et al., 1999; Bortoluzzi et al., 2000).

1.6.1 Positional candidate gene approach

The approaches used to identify disease genes generate a list of candidate genes, which have to be tested individually to see if there is evidence that they are associated with a disease in question (reviewed in Keating 1992; Putnam et al., 1995). Candidate gene approaches may be based on particular properties of the product of the candidate genes that are consistent with their involvement in pathogenesis. Confidence in a particular candidate disease gene is increased substantially if it can be shown to map to the same subchromosomal region as the disease gene. This approach is referred to as a positional candidate gene approach and has been used successfully in a number of disease gene studies including, the search for the Alzheimer's disease genes (Goate et al., 1991; Schellenberg et al., 1992; Sherrington et al., 1995) and the hereditary nonpolyposis colon cancer gene (Wildrick and Boman 1988; Sarraf et al., 1999), Retinitis pigmentosa (Dryja et al., 1990; Sullivan et al., 1999) and Waardenburg syndrome type1 (Tassabehji et al., 1994). The positional candidate gene approach has become the choice for disease gene discovery because of the accelerated chromosomal mapping of genes and EST sequences (Schuler et al., 1996; Genome Consortium 2001).

1.6.1.1 ESTs as mapping reagents

The availability of a small amount of DNA sequence for a specific clone allows the synthesis of oligonucleotide primers, complementary to a short region of the DNA sequence, that can be used in the development of a PCR assay (Strachan and Read 1998). The site on the original genomic DNA from which the sequence was derived is described as a sequence tagged site (STS) (Olson et al., 1989). Mapping methodologies have centered around the use of STSs as unique landmarks across the genome (Olson et al., 1989). Wilcox et al (1991) demonstrated that single-pass sequences (ESTs) provide suitable templates for the design of gene-based STSs. Several reasons have been put forward for the use of 3' transcripts for the purpose of generating STSs (review Schuler 1997). For example, several mapping technologies incorporate the use of rodent material as the carrier of human genomic fragments. The use of rodent material could potentially show cross-reactivity with human STSs. However, the sequences near the 3' end of transcripts lie within untranslated regions and it has been demonstrated that such regions show less cross-species conservation than do coding regions (Makalowski et al., 1996).

An international consortium was established to develop STSs from ESTs for mapping studies (Schuler et al., 1996). In summary, a non-redundant set of 3' end sequences was selected from UniGene and distributed to participating laboratories where STSs were developed and mapped using primarily radiation hybrid (RH) techniques. The first report of the RH consortium provided map positions for 16000 genes (Schuler et al., 1996). Approximately 1000 genetic markers from the Genethon map were included in the analysis to serve as a mapping framework and to allow gene positions to be related to genetic linkage information. The latest version of Genemap99 incorporates approximately 52,000 RH markers that serves as a resource for identification of candidate genes once a disease gene has been localized to a chromosomal position.

1.6.1.2 Disease gene discovery

Mapping of ESTs to genomic regions containing disease-linked genes can help to identify mutations in candidate genes that may lead to disease susceptibility. Examples of discoveries made through identity to EST data include a novel gene, presenilin-2, associated with Alzheimer's Disease (Rogaev et al., 1995), *LRP5* and *LRP6* associated with type 1 diabetes (Brown et al., 1998; Hey et al., 1998) and candidate genes for bipolar disorder (Semple et al., 2000).

1.6.1.2.1 Alzheimer's Disease

Multiple loci were identified to be responsible for Alzheimers disease (AD) (Goate et al., 1991; Schellenberg et al., 1992; Sherrington et al., 1995). The Volga-german kindreds represent a group of seven AD families that showed autosomal dominant inheritance (Bird et al., 1988). AD-related loci on chromosome 21 and 14 were excluded but recently these families showed linkage to chromosome 1 (Levy-Lahad et al., 1995a).

A candidate gene for chromosome 1 AD locus (*AD4*) was identified with the aid of ESTs (Levy-Lahad et al., 1995b). YACs were used to clone the region as defined by linkage analysis. A total of 60 YACs were isolated and flanked by D1S229 and D1S103. Genes residing in this region were tested as candidates for the *AD4* locus. In the same report, a candidate gene for alzheimer’s disease subtype 3 (*AD3)* was used to fish out an EST from dbEST namely T03796. T03796 was identified in the translated EST database and was 80.5% identical in amino acid sequence to S182 (*AD3* candidate gene). The chromosome 1 map position for T03796 was confirmed by using a radiation hybrid panel followed by its (T03796) localisation to a single YAC. Brain and fibroblast cDNA libraries were screened using T03796 as a probe. Clones of 2.3kb were obtained and sequencing identified a 448 amino acid open reading frame (L43964) that showed 65% identity to S182. Affected people were screened for mutations in the entire coding region of this gene and resulted in the identification of an isoleucine substitution for an asparagine (N141I mutation).

1.6.1.2.2 Type 1 diabetes

Type 1 diabetes (insulin dependent diabetes mellitus, *IDDM*) is an autoimmune disease that develops as a consequence of the interaction of both genetic and environmental factors (Atkinson and Maclaren 1994). The genetic component of type 1 diabetes involves multiple genes (Tisch and McDevitt 1996; Todd and Farrall 1996 and Todd 1996) including the *MHC* locus on chromosome 6p21 (*IDDM1*) and the insulin gene in the *IDDM2* locus on chromosome 11p15 (Bell et al., 1974; Bennett and Todd 1996; Tisch and McDevitt 1996 and Vyse and Todd 1996). The *IDDM4* locus on chromosome 11q13 was one of 18 chromosomal regions that showed genetic linkage to type 1 diabetes and was identified by genome-wide scans of affected sib-pairs (Davies et al., 1994; Hashimoto et al., 1994). Independent studies confirmed the evidence for a susceptibility locus on chromosome 11q13 in a region that spanned 15 centimorgans (cM) (Field et al., 1994 and Luo et al., 1996). A portion of the *IDDM4* locus was shown to be in linkage disequilibrium with the disease and this refined region was used as a target for high-throughput DNA sequence analysis. (Hey et al., 1998).

A microsatellite marker (D11S1337) was used to screen a human BAC library resulting in the identification of a BAC clone (HBAC 14-1-15). The clone was shotgun sequenced and all fragments were used to retrieve matching ESTs from GenBank. ESTs were grouped into two UniGene clusters and one singleton EST (accession number F07016) not present in UniGene. Assembly of the shotgun sequences indicated that EST-F07016 was located between the two UniGene clusters. The observation that the two UniGene clusters showed similarity to the LDL-receptor family suggested that these clusters were part of the same gene. PCR primers were designed to the two UniGene clusters and EST F07016 and used to amplify products from human liver cDNA. This PCR strategy resulted in the isolation of 4.8kb of the low-density lipoprotein receptor related protein 5 (*LRP5*) cDNA sequence. This sequence did not encode the entire gene because the complete open reading frame lacked the potential to encode a N-terminal signal peptide for protein export, a common property of the members of the LDLR family (Herz et al., 1988).

A mouse ortholog of *LRP5* was isolated from a mouse liver cDNA library and provided the remaining 5' end sequence needed to encode the entire gene (Herz et al., 1988). *LRP5* serves as a positional candidate for the *IDDM4* locus and is supported by evidence from association studies using markers that map to the same cosmid containing a portion of the *LRP5* gene. Additional genetic analysis is required to determine whether this gene is a diabetes susceptibility gene. For example, specific polymorphisms must be identified that are associated with the disease and have the potential to alter the level of *LRP5* biological activity.

1.6.1.2.3 Bipolar disorder

Semple et al (2000) described the identification of transcripts and associated SNPs in an 11cM region of 4p (D4S394-D4S403), that showed linkage to bipolar affective disorder (BPAD) (Blackwood et al., 1996; Kennedy and Macciardi, 1998). In their study, one hundred and ninety publicly available marker sequences that mapped within the D4S394-D4S403 interval were retrieved from GeneMap99, Genethon and the collection of STS sequences at the Whitehead Institute and the Stanford Human Genome Center. These marker sequences were masked and then searched against Unigene, TIGR gene index, dbEST and STACK. The dbEST database and the consensus sequences of TIGR and STACK were searched using BLASTN (Altschul et al., 1997). UniGene was searched via a text query (GenBank EST/mRNA accessions) and the longest sequences from each UniGene cluster were searched using BLASTN to confirm the text-based search. BLAST searches of clustered EST databases were deemed significant at an arbitrary level ($1e^{-100}$) and the resulting alignments were verified manually (Semple et al., 2000).

Semple and colleagues (2000) identified a higher number of transcripts in UniGene. Contrary to the author's reasoning that UniGene has less strict clustering criteria, it has been demonstrated that STACK employs a loose clustering approach (Miller et al., 1999; Burke et al., 1999; Christoffels et al., 2001). The difference in transcript numbers could reflect the fact that the three databases were not in-synch with respect to the GenBank release. For each of the three databases under investigation, a small number of ESTs were assigned to clusters when the other two databases failed to do so. STACK represents the earliest GenBank release (relative to UniGene and TIGR's gene index) and therefore the absence of some EST sequences could be explained. Examination of the sequences unique to STACK should highlight the quality of data accepted for clustering where the other two databases failed to cluster. The transcripts identified by the different indices are being followed up as candidates for bipolar disorder.

1.6.2 Functional candidate gene approach

A functional candidate gene approach can be described as method that selects candidate genes based on the particular properties of the product of a candidate gene that is consistent with its involvement in pathogenesis. This approach has been used to search for apoptosis-related genes expressed in cardiac development and disease (Hwang et al., 1997; reviewed in Rezvani et al., 2000).

1.6.2.1 Apoptosis-related genes expressed in cadiovascular development and disease

Apoptosis (programmed cell death) is an important process, which in conjunction with cell proliferation, maintains cell number homeostasis. Recently, apoptosis has been suspected as a significant contributor to both disease and normal development of the cardiovascular system (Colucci 1996; Narula et al., 1996 and Olivetti et al., 1997). Identifying key genes that are involved in the regulation of apoptosis in the cardiovascular system serves as a basis for understanding how cardiac development is modulated (reviewed in Rezvani et al., 2000). The search for these apoptosis regulatory genes was undertaken with large-scale sequencing of ESTs from cardiovascular cDNA libraries (Hwang et al., 1997). In excess of 5000 genes from the cardiovascular system were characterised further as either effectors, suppressors or intermediate regulators of apoptosis depending on the functional classifications that are described in the literature. Examples of apoptotic regulators include (a) the interleukin-converting enzyme (ICE) family and (b) the Bcl-2 family.

a. Interleukin-converting enzyme (ICE) family

Two death effector genes namely ced-3 and ced-4 were discovered in *C. elegans* and provided the basis for an understanding of apoptosis in higher organisms. Recently caspases have been found to play an important role in regulating apoptosis in the cardiovascular system particularly in vascular smooth muscle cells (Horiuchi et al., 1999) and cardiac cells (Loppnow et al., 1998). The ced-3 protein was found to be homologous to the mammalian cystein protease, interleukin-1beta-converting enzyme (ICE), that is considered as a prototype of the caspase family of proteins. There are at least 14 known members of this family in human (Davis and Wells 1999), and they can be inhibited to block apoptosis (Fraser et al., 1996). Caspase proforms are proteolytically cleaved to generate activated forms of the enzyme. Two members of the ICE family, Ich-1L and Ich-1S, have been uncovered through random sequencing of a heart cDNA library (Hwang et al., 1997). Ich-1L is a gene encoding a 435 amino acid protein that induces programmed cell death and Ich-1S is a truncated version of Ich-1L that suppresses apoptosis when it is overexpressed (Wang et al., 1994). These two genes might provide a link in the future between apoptosis regulation and cardiac development.

b. Bcl-2 family

A number of reports have been published indicating the crucial roles of Bcl-2 and its family members in the progression of apoptosis during ischemia (Park et al., 1996; Saikumar et al., 1998; von Harsdorf et al., 1999; Maulik et al., 1999). A homologue to the death suppressor gene ced-9, Bcl-2, was identified in mammals and was found to be required for cell survival in human B-cell lymphoma (i.e., anti-apoptotic activity). Bcl-2 is able to form homodimers and heterodimers, a trait that is significant for its role in controlling apoptosis (Kroemer 1997; MacLellan and Schneider 1997). Rezvani et al. (2000) identified several of the Bcl-2 family members through EST sequencing of heart cDNA libraries. In addition, genes including Bcl-x, Bcl-2 binding components, BID and Bak were identified (Rezvani et al., 2000b). Bak and Bcl-x have been implicated in cytokine-induced cardiac myocyte apoptosis (Ing et al., 1999) whereas BID binds to other protein family members to induce apoptosis (Wang et al., 1998).

A number of apoptosis-related genes were identified through heart cDNA library sequencing as expressed in the cardiovascular system that were previously only characterised in other tissues or organisms. These include MA-3 a novel mouse gene (Shibahara et al., 1995), the Nip family of proteins that are involved in cell survival (Boyd et al., 1994) and DAD-1 that was shown to be an apoptotic suppressor in hamster cell lines (Nakashima et al., 1993). In addition, DAD-1 was found to be more highly expressed in cardiac hypertrophy compared to normal adult heart and therefore may play a role in controlling cell numbers during disease (Rezvani et al., 2000b). Understanding the involvement of novel cardiac cell modulators is important considering that in humans, myocytes irreversibly exit the cell cycle just before birth. Cardiomyocytes are particularly prone to abnormal imbalances in cell numbers, such as the case of myocardial infarction, where prolonged deprivation of oxygen leads to local necrosis of cardiomyocytes. This is damaging to the organism because of the inability of cardiomyocytes to re-enter a proliferating mitotic cell cycle thus preventing replacement of lost tissue. The damage is patched up with non-contractile fibroblasts that form fibrous scar tissue.

1.6.3 *In-silico* differential display

Gene expression is a process composed of several different steps. First, in the nucleus, genomic DNA serves as a template for RNA synthesis during transcription. The product of this process is various kinds of RNA, synthesised by RNA polymerases. The messenger between DNA in the nucleus and protein synthesis in the cytoplasm is messenger RNA (mRNA), a short-lived RNA that is easily degraded by nucleases. Proteins are then synthesised in a translation process according to instructions given by the mRNAs (Rawn 1989). Gene expression level is tightly regulated at several different levels, such as transcription, mRNA processing, transport and stability, translation and post-translational modification (Lewin 1997). However, transcription is the major control point for many genes (Lewin 1997) and therefore measuring the amount of a mRNA transcript is a way of quantifying the expression of the gene.

Gene expression levels can be measured for one gene at a time, using traditional methods such as Northern Blots (Alwine et al., 1977) or nuclease protection assays (Berk and Sharp 1977). However, large-scale EST sequencing provides for example, complete mRNA populations to compare gene expression patterns (or profiles), for multiple genes, between two tissues showing different disease states. Using the example of two tissues, ESTs can be used to create transcript expression profiles for the two tissue libraries. The frequencies at which the transcripts appear in different libraries can be re-calculated into expression patterns since they reflect the actual composition of a mRNA pool. If large amounts of transcripts are sequenced, the frequencies become statistically significant (Okubo et al., 1992). However, recent reports have shown that small amounts of ESTs are sufficient to provide statistically significant results (Hwang et al., 1997; Bortoluzzi et al., 2000). In this section, I review examples of (i) *in-silico* based studies that identify candidate differentially expressed genes in breast (Ji et al., 1997; Schmitt et al., 1999), prostate (Vasmatzis et al., 1998), heart (Reszvani et al., 2000) and muscle (Bortoluzzi et al., 2000) and (ii) the use of co-ordinated expression of genes as a method of assigning functions to novel genes.

1.6.3.1 Breast cDNA libraries

A breast cancer specific candidate gene 1 (BSCG1) was identified as a molecular marker for infiltrating breast carcinoma by comparing a normal breast library with that of a disease library and subsequent sequencing of specific clones (Ji et al., 1997). ESTs were generated from the breast cancer library and a matching normal breast library. Overlapping ESTs were merged into one group and the list of non-overlapping EST groups were compared for the quantity of EST members originating from the normal and cancer breast libraries. Cathepsin D was sampled by more ESTs from the breast cancer library than the normal breast library. This suggested a role for Cathepsin D in breast cancer metastasis. This finding supported previous studies that suggested a role for Cathepsin D in breast cancer (Rochefort et al., 1987; Capony et al., 1990; Cavailles et al., 1991).

The average size of EST libraries ranges from between 1000 and 10000 entries and therefore an EST library cannot be regarded as faithfully representing the gene expression pattern of a tissue (Vingron and Hoheisel 1999). It has been estimated that between 10000 and 30000 different genes are expressed in a given cell with an average of about 300000 mRNA molecules per cell (Bishop et al., 1974; Axel et al., 1976). However, the availability of EST libraries derived from the same type and state of tissue led to the pooling of equivalent EST libraries. Schmitt et al (1999) used such pools of libraries with the assumption that EST numbers that reach tens of thousands for a library pool would be a proportional representation of all abundant and moderately expressed genes. Schmitt and coworkers (1999) carried out mRNA analysis on non-normalised libraries and provided a procedure for mining of EST libraries for differentially expressed genes. In summary, for a given tissue a pool of ESTs from both tumour and normal tissues was created and a minimum of 10000 ESTs was required for each pool. A non-redundant set of ESTs was generated for the EST library under investigation using BLAST. These sequences were searched against dbEST, GenBank and two propriety databases to try to extend the length of the sequences. The non-redundant set of sequences retrieved from GenBank for a specific library (eg., NCI_CGAP_Br1.1) was compared against the EST pool of a normal and disease state for the same tissue (eg., ESTs derived from breast tissue).

The relative abundance of a gene was defined by Schmitt et al (1999) as the ratio of the number of homologous ESTs to the total number of ESTs in the corresponding pool. Relative abundance figures were determined for normal and tumour pools separately, and the ratio of the normal and tumour relative abundances was used as a measure for the down or up-regulation of a gene in tumour tissue with respect to normal tissue. Fisher's exact test (see section 2.3.6 for a description) was used to assess the distribution of hits between the normal and affected tissues observed in a BLAST search.

1.6.3.2 Prostate cDNA libraries

Vasmatzis and coworkers (1998) used a crude assembly system employing BLAST to identify prostate specific ESTs. Fifteen prostate-specific ESTs were identified that had no database homologs. Seven of fifteen prostate-specific ESTs were screened by northern blot hybridisation, of which three ESTs showed no hyridisation signal to any tissue other than prostate.

There are problems associated with identifying tissue-specific genes using EST data. Firstly, The EST database is incomplete and there is a possibility that there is not enough sampling of a specific gene transcript to identify other tissue locations. Secondly, tissue-specific ESTs could represent false positives. An EST might not match other ESTs from different tissues but they could all belong to the same gene. The probe sequence length might be too short to match the target sequence and this will give the impression of distinct genes.

1.6.3.3 Heart cDNA libraries

The generation of ESTs from human heart cDNA libraries has been used to characterise gene expression in various developmental and pathological states of the cardiovascular system (Liew 1993; Liew et al., 1994; Hwang et al., 1994; Hwang et al., 1995). Hwang et al (1997) reported on the analysis of 84904 ESTs from 13 cDNA libraries of the cardiovascular system. The entire cardiovascular data set could be divided into three classes, (i) ESTs with significant identity to known sequences in the non-redundant nucleotide and peptide databases (55%), (ii) ESTs that match other ESTs in dbEST but do not match any published gene sequences (33%) and (iii) ESTs that represent novel transcripts (12%). There were approximately 4575 previously identified genes in this data set. The relative level of expression for each gene identified in the EST data was calculated as the total number of ESTs that match a particular gene divided by the total number of ESTs matching all the genes in the data set. After assigning each gene and it corresponding ESTs to one of seven functional classes, expression patterns for each functional class were reported. For example, expression of cell motility genes in adult hypertrophic hearts was significantly diminished compared to normal adult hearts, whereas expression of transcripts involved in cell defense was slightly increased in both hypertrophic heart libraries compared to other cardiac libraries. These expression patterns and other patterns documented in the same study verified previous work by Hwang et al. (1995), who established that fetal heart exhibited fewer transcripts representing contractile proteins and more transcripts representing signal transduction and cell regulatory proteins than adult heart.

In-silico northern analysis was used to identify genes potentially over-expressed in cardiac hypertrophy compared with normal myocardium (Hwang et al., 1997). A small sample of ESTs was generated from each of two independent hypertrophic heart cDNA libraries (1089 EST and 474 ESTs respectively). Potentially differentially expressed genes were grouped as strong, good and weak candidates based on Poisson probabilities (P-value < 0.05 for a strong differentially expressed candidate gene). Myoglobin was identified as strong candidate for differential expression in the hypertrophic heart. For example, myoglobin was sampled by five ESTs in a total of 34736 ESTs from normal adult and fetal heart cDNA libraries whereas seven myoglobin ESTs were present in 474 ESTs from a hypertrophic heart cDNA library. The expected number of observations of myoglobin in a set of 474 ESTs is (474 x (5/34736) = 0.068 ESTs). Using the expected number of observations for myoglobin, the Poisson probability of observing 7 or more ESTs in the hypertrophic heart cDNA library by chance alone was calculated as $1.29x1e^{-12}$.

At least 10 genes out of 23 genes identified as strong candidates in the above experiment have previously been demonstrated to be involved or elevated in cardiac hypertrophy including atrial natriuretic factor (Buttrick et al., 1994; Poulos et al., 1996), brain natriuretic factor (Nakagawa et al., 1995), myosin light chain-2 (Doud et al., 1996), desmin (Collins et al., 1996; Watson et al., 1996) and superoxide dismutase (Kirshenbaum et al., 1995; Gupta and Singal 1989). The combined *in-silico* and experimental evidence suggest that a combined Poisson probability cutoff of P < 0.05 is appropriate for screening EST data sets with relatively low numbers and that this method has potential for genome-wide searching for novel genes involved in cardiovascular disorders. The application of this technique to ESTs without a parent mRNA match represents a problem because of the uncertainty whether two ESTs represent non-overlapping segments of the same gene. The use of a non-redundant gene index such as STACK that contains entries for clusters joined by their cloneIDs could help overcome the problem of non-overlapping ESTs from the same gene. A large number of published disease genes were identified in the cardiovascular EST data set. To assist in the identification of novel cardiovascular genetic disorders, 1048 out of 22,623 ESTs were mapped to their chromosomal loci. These ESTs, if mapped to a documented disease locus, would serve as a candidate gene for that disorder.

1.6.3.4 Adenomatous Polyposis Coli (APC) gene product in human cardiac development and disease

Sequence analysis of over 50000 ESTs generated from 11 cDNA heart libraries revealed several cDNA clones significantly matching (E < 10^{-10}) Adenomatous polyposis coli (APC) and its interacting protein, beta-catenin (Rezvani and Liew 2000). Digital nothern analysis indicated a differential expression of these genes during cardiac development and disease as they were tagged in different frequencies in fetal, adult and hypertrophic heart libraries. The frequency of gene expression was calculated as a percentage where the total number of ESTs representing a specific gene sampled from a heart library was divided by the total number of ESTs sampled from that heart library. Subsequent experimental evidence, using reverse transcriptase polymerase chain reaction analysis (RT-PCR), showed that APC was expressed at higher levels in adult heart compared with fetal heart in human and mouse while having no effect on the beta-catenin. This same effect of APC expression was observed in different developmental stages of a mouse heart. However, western blot analysis revealed higher levels of beta-catenin protein in a fetal and hypertrophic heart compared with adult heart. The up-regulation of APC in adult heart suggest that APC plays a role in the cardiomyocytes withdrawal from the cell cycle. The protein analysis suggests that APC plays has a regulatory role on beta-catenin.

1.6.3.5 Human skeletal muscle transcriptional profiles

The Genexpress knowledge base represents the integration of expression, sequencing and mapping data with the goal of identifying positional and functional muscular disease candidate genes (Houlgatte et al., 1995; Auffray et al., 1995; http://idefix.upr420.vjf.cnrs.fr/IMAGE/Page_unique/welcome_muscles.html). The value of integrating sequence, map and expression information has been illustrated by the identification of a gene responsible for a form of limb-girdle muscular dystrophy through positional cloning and subsequent confirmation by functional candidate studies (Fougerousse et al., 1994; Chiannikulchai et al., 1995; Richard et al., 1995). Expression profiles of human skeletal muscle involving less than 1000 genes have been produced with a combination of wet bench and *in-silico* aproaches (Pietu et al., 1996; Lanfranchi et al., 1996; Murano et al., 1997; Bortoluzzi et al., 1998). Recently, an *in-silico* approach was applied to human skeletal muscle where each transcript was classified according to its level of expression (Bortoluzzi et al., 2000). The aim of this approach was to provide a method to describe the transcriptional profile of single human tissues using data extracted from UniGene (4080 transcripts). A number of assumptions were made by the authors (Bortoluzzi et al. 2000) including:

(i) A redundancy of 1.3% was calculated for the 4080 UniGene clusters. The redundancy was based on the following calculation: 52 (1.3%) clusters contained 5' ESTs only. Since these 5' ESTs could refer to transcripts already identified by clusters containing 3' ESTs (90% of the UniGene clusters), it was assumed that the 5' ESTs (1.3%) represent the maximum redundancy in the data.

(ii) The level of expression for each cluster was quantitatively estimated as the number of skeletal muscle ESTs corresponding to a cluster over the total number of skeletal muscle ESTs for all the clusters. The use of EST number per cluster (or gene) to quantify gene expression has been adopted by a number of tools that are aimed at detecting differences in gene expression activity including X-Profiler (http://ww.ncbi.nlm.nih.gov/ncicgap/cgapxsetup.cgi) and digital differential display (http://www.ncbi.nlm.nih.gov/cgi-bin/UniGene/ddd?ORG=Hs; Strausberg et al., 1997). Highly expressed genes were classified as clusters containing nine or more ESTs (i.e., >= 0.0363% of the total ESTs). Moderately expressed genes were classified as clusters containing between three and nine ESTs, and weakly expressed genes were classified as clusters containing one or two ESTs.

In summary, three non-normalised and non subtracted skeletal muscle cDNA libraries with the maximum assigned ESTs were extracted from UniGene and accounted for 4080 clusters. A set of 417 transcripts was identified as the skeletal muscle transcriptional profile and included 370 highly expressed skeletal muscle transcripts and 47 putatively skeletal muscle-specific transcripts. The results were validated using 120 genes that were sampled in both the *in-silico* reconstructed transcriptional profile and the Rochester SAGE catalog (http://www.urmc.rochester.edu/smd/crc/Swindex.html; Welle et al., 1999). Additional validation for 13 of the 417 transcripts came from an independent study by Pietu et al. (1999), who reported the expression profiles of 910 genes expressed in skeletal muscle. These independent studies (Welle et al., 1999; Pietu et al., 1999) validated the detection of highly expressed skeletal muscle genes by Bortoluzzi et al. (2000), illustrating the effectiveness the methodology in estimating the expression levels of highly expressed genes. Given that each tissue is characterised by a relatively small number of highly expressed genes (Bishop et al., 1974; Axel et al., 1976), this approach might be applied to situations where only a small number of transcripts are available.

1.6.3.6 Fisher's exact test

Statistical methods which depend on the parameters of populations or probability distributions are referred to as parametric tests. Parametric tests includes t-test, ANOVA, Regression and Correlation. These tests are only meaningful for data that is sampled from a population with an underlying normal distribution or whose distribution can be rendered normal by mathematical transformation. A normal distribution, also referred to as a bell-shaped distribution, describes a data set that is characterised by many independent random factors acting in an additive manner to create variability.

Nonparametric methods require fewer assumptions about a population or probability distribution, i.e., neither the values obtained nor the population from which the sample was drawn need to have a normal distribution. A nonparametric test such as the Chi-squared test assumes that no single data point is zero and that at least 80% of the expected frequencies are five or more. These assumptions cannot be guaranteed in EST sampling and therefore the fisher's exact test is preferred. Fisher's exact test is a non-parameteric test used for testing the hypothesis that there is a statistically significant difference between two groups. It has the advantage that it does not make any approximations and so is suitable for small sample sizes.

Fisher's exact test is used to evaluate representations of yes/no outcomes obtained from two disjoint samples. The outcome of the 'Fisher's exact test' is a significance value, *P*, ranging between 0 and 1 that describes the likelihood of the null hypothesis being true: "The frequency of an event is the same in either of two samples" or as applied to differential expression: "The frequency of a gene is the same in normal and disease tissue”. A *P* value closer to 0 is indicative of significant differential expression of the gene under consideration. However, Fisher's exact test is a conservative test as compared to other statistical tests (Audic and Claverie 1997) and it has been suggested that the selection of genes for further investigation based upon the criterion of small *P* values can be considered restrictive (Schmitt et al., 1999).

1.6.4 Coordinated gene expression

A cDNA microarray approach provide an efficient technique to monitor expression levels of many different genes simultaneously (Welsh et al., 1992). The microarray approach is not based on sequencing, but on hybridisation between nucleic acids. Large-scale expression data, generated by hybridisation of DNA to microarrays, is potentially a rich source of information on gene function and regulation. However, the use of EST data to monitor gene expression patterns or profiles has been demonstrated (Ewing et al., 1999). For example, by clustering genes according to their expression patterns (profiles), groups of genes involved in the same pathways or sharing common regulatory mechanisms may be identified (reviewd in Claverie 1999). Using publically available ESTs, Ewing et al. (1999) generated 'digital expression profiles" by counting the frequency of tags for different genes sequenced from different cDNA libraries. A statistical test was used to associate genes having similar expression profiles. This approach was extended to using larger EST samples from UniGene projects (mouse, man and rat ESTs) where Ewing and Claverie (2000) showed that genes clustered on the basis of expression profile may represent genes implicated in similar pathways or coding for different subunits of multi-component enzyme complexes.

Expression profile clustering in the context of disease candidate gene selection could be applied to unraveling regulatory pathways that are affected by unannotated disease candidate genes. In particular, uncharacterized genes have been identified for their involvement in sudden cardiac death in a mouse model. Pathways implicated in sudden cardiac death could be identified by clustering the expression profiles of characterized genes with the novel mouse genes (Nguyên-Tran et al., 2000).

1.7. Thesis rationale

Unlike most gene indices, aimed at reconstructing the gene complement of the human genome (see chapter 1 section 1.4), the South African National Bioinformatics Institute (SANBI) has embarked on the development of the sequence alignment and consensus knowledgebase (STACK) database that focused on the detection and visualisation of transcript variation in the context of developmental and pathological states, using all publicly available ESTs. Preliminary work on the STACK project employed an approach of arbitrarily partitioning the EST data into tissue categories as a means of reducing the EST sequences to managable sizes for subsequent processing. The tissue partitioning provided the template material for the development of error-checking tools to analyse the information embedded in the error-laden EST sequences. However, tissue partitioning increases redundancy in the sequence data because one gene can be expressed in multiple tissues, with the result that multiple tissue partitioned transcripts correspond to the same gene. Therefore, the sequence data represented by each tissue category had to be merged in order to obtain a comprehensive view of expressed transcript variation.

This dissertation reports on the development of a human gene index where all EST sequences have been processed irrespective of tissue origins in order to provide the correct developmental and pathological context for investigating sequence variation. Furthermore, the availability of a human gene index was assessed as a disease candidate gene discovery resource. The development of a STACK human gene index as a disease gene discovery resource required (i) the ability to cope with the deluge of EST data in the public arena (1,198,607 ESTs GenBank 110, release 15th October 1998 and increasing) (ii) processing of all EST data through a pipeline that would generate consensus sequence transcripts from the fragmented EST data, (iii) validating the accuracy of the gene indices the ability to capture sequence variation, and (iv) an application of the human gene index to a disease gene discovery project to verify the utility of the STACK gene index as a disease gene discovery resource. In order to provide a comprehensive report of the above-mentioned considerations, the dissertation was organised into six chapters as follows:

- **Chapter One** provides the background to the approaches taken to generate gene indices and the use of EST data for disease gene discovery. The approaches implemented in four gene index projects are outlined including the approach developed during the course of this thesis.

- Chapters two, three, four and five are experimental chapters that are divided into five sections i.e., Introduction, Methods, Results, Discussion and References. Chapters two, three and four deals with different aspects of the generation of a human gene index. Chapter five explores the use of such a gene index for the identification of disease candidate genes.

- Tools available at the start of this thesis were inadequate to deal with the large volume of EST data put into the public arena. **Chapter Two** describes the optimisation of an EST clustering tool, D2_CLUSTER, for clustering large EST data sets.

- **Chapter Three** focuses on the generation of a human gene index (Sequence tag alignment and consensus knowledgebase (STACK)). A pipeline is described that includes tissue partitioning of ESTs, subsequent cleaning, clustering, assembly analysis, consensus generation, integration of genetic markers and clone linking.

- **Chapter Four** describes the accuracy of the STACK human gene index by comparing it to human chromosome 22. In addition, the ability to detect alternate splice events within the EST assemblies is illustrated.

- **Chapter Five** explores an approach to disease gene candidate discovery using the STACK human gene index

- The conclusions drawn from each aspect of this project are discussed at the end of the relevant chapters. **Chapter Six** provides a final comment and includes (i) future development prospects of the STACK gene index and (ii) future research prospects for the identification of the *PFHB1* gene.

1.8 References

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.* **6:** 829-845.

Adams M.D., Kelley J.M., Gocayne J.D., Dubnick M., Polymeropolous M.H., Xiao H., Merril C.R., Wu., Olde B., Moreno R.F., Kerlavage A.R., McCombie W.R. and Venter J.C. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **52:** 1651-1656.

Adams M.D., Kerlavage A.R., Fleischmann R.D., Fuldner R.A., Bult C.J., Lee N.H., Kirkness E.F., Weinstock K.G., Gocayne J.D., White O et al. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3-174;

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

Alwine J.C., Kemp D.J. and Stark G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridisation with DNAprobes. *Proc Natl Acad Sci USA.* **74:** 5350-5354.

Atkinson M.A. and Maclaren N.K. (1994) The pathogenesis of insulin-dependent diabetes mellitus. *New England J. Med.* **331:** 1428-1436.

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

Axel R., Feigelson P. and Schutz G. (1976) Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell* **7:** 247-254.

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.

Bell G., Horrta S. and Karem J. (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33:** 176-183.

Bennett S. and Todd J. (1996) Human type I diabetes and the insulin gene: principles of mapping polygenes. *Annu. Rev. Genet.* **30:** 343-370.

Benson D.A., Boguski M.S., Lipman D.J., Ostell J., Ouetelle B.F., Rapp B.A., and Wheeler D.L. Release Notes for NCBI-GenBank Flat File Release 112.0 (1999) *Nucleic Acids Res.* **27:** 12-17.

Berk A.J. and Sharp P.A. (1977) Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12:** 721-732.

Bird T.D., Lampe T.H., Nemens E.J. (1988) Familial Alzheimer's disease in American descendants of the Volga Germans: probable genetic founder effect. *Ann Neurol.* **23:** 25-31.

Bishop J. O., Morton J.G., Rosbash M. and Richardson M. (1974) Three abundance classes in HeLa cell messenger RNA. *Nature* **250:** 199-204.

Blackwood D.H., He L., Morris S.W., McLean A., Whitton C., Thomson M., Walker M.T., Woodburn K., Sharp C.M., Wright A.F., Shibasaki Y., St Clair D.M., Porteous D.J. and Muir W.J. (1996) A locus for bipolar affected disorder on chromosome 4p. *Nature Genet.* 12: 427-430.

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

Bortoluzzi S., Rampoldi L., Simionati B., Zimbello R., Barbon A., d'Alessi F., Tisco N., Pallavicini A., Toppo S., Cannata N., Valle G., Lanfranchi G. and Danieli G.A. (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* **8:** 817-825.

Bortoluzzi S., d'Alessi F., Romualdi C. and Danieli G.A. (2000) The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.* **10:** 344-349.

Bouck J., Yu W., Gibbs R. and Worley K. (1999) Comparison of gene indexing databases. *Trends Genet.* **15:** 159-162.

Boyd J.M., Malstrom S., Subramanian T., Venkatesh L.K., Schaeper U., Elangovan B,. D'SaEipper C. and Chinnadurai G. (1994) Adenovirus E1B 19kDa and Bcl-2 proteins interact with a common set of cellular proteins. *Cell* **79:** 341-351.

Brown S.D., Twells R.C.J., Hey P.J., Cox R.D., Levy E.R., Soderman A.R., Metzker M.L., Caskey C.T., Todd J.A. And Hess J.F. (1998) Isolation and characterisation of *LRP6*, a novel member of the low density lipoprotein receptor gene family. *Biochem.Biophys. Res. Commun.* **248:** 879-888.

Bull L.N., Pabon-Pena C.R. and Freimer N.B. (1999) Compound microsatellite repeats: practical and theoretical features. *Genome Res.* **9:** 830-838.

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

Burke J., Wang H., Hide W. and Davidson D. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276-290.

Buttrick P.M., Kaplan M., Leinwand L.A. and Scheuer J. (1994) Alterations in gene expression in the rat heart after chronic pathological and physiological loads. *J. Mol. Cell Cardiol.* **26:** 61-67.

Capony F., Rougeot C., Cavailles V. and Rochefort F. (1990) Estradiol increases the secretion by MCF-7 cells of several lysosomal pro-enzymes. *Biochem. Biophys. Res. Commun.* **171:** 972-978.

Cariaso M., Folta P., Wagner M., Kuczmarski T. and Lennon G. (1999) IMAGEne I: clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes. *Bioinformatics* **15:** 965-973.

Cavailles V., Augereau P. and Rochefort H. (1991) Cathepsin D gene in human MCF-7 cells contains estrogen-responsive sequences on its 5' proximal flanking region. *Biochem. Biophys. Res. Commun.* **174:** 816-824.

Chiannikulchai N.P., Pasturaud I., Richard C., Aufrfray C. and Beckmann J.S. (1995) A primary expression map of the chromosome 15q15 region containing the recessive form of limb-girdle muscular dystrophy (LGMD2A) gene. *Hum. Mol. Genet.* **4:** 717-725.

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.

Claverie J-M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8:** 1821-1832

Collins J.F., Pawloski-Dahm C., Davis M.G., Ball N. and Dorn G.W. and Walsh R.A. (1996) The role of the cytoskeleton in left ventricular pressure overload hypertrophy and failure. *J Mol Cell Cardiol.* **28:** 1435-1443.

Colucci W.S. (1996) Apoptosis in the heart. *New England J of Medicine* **335:** 1224-1226.

Costanzo F., Castagnoli L., Dente L., Arcari P., Smith M., Costanzo P., Raugel G., Izzo P., Pietronaolo T.C., Bougueleret L., Cimino F., Salvatore F. and Cortese R. (1983) Cloning of several cDNA segments coding for human liver proteins. *EMBO J.* **2:** 57-61.

Davies J., Kawaguchi Y., Bennett S., Coperman J., Cordell H., Reed P.L., Gough S., Jenkins S., Palmer S., Balfour K., Rowe B., Farrall M., Barnett A., Bain S. and Todd J. (1994) A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371:** 130-136.

Davis D. and Wells S. (1999) Caspase 3-like activity appears in buffer containing live intact cells. *BioRadiations.* **103:** 42-43.

Deininger P. (1989) In: Mobile DNA (eds DE Berg, MM Howe), pp. 619-636. American Society of Microbiology, Washington, DC.

Deloukas P., Schuler G.D., Gyapay G., Beasley E.M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T.C., McKusick K.B., Beckmann J.S. et al. (1998) A physical map of 30,000 human genes. *Science* **282:** 744-746.

Doud S.K., Pan L.X., Carleton S., Marmorstein S. and Siddiqui M.A. (1996) Adaptational response in transcription factors during development of myocardial hypertrophy. *J Mol Cell Cardiol.* **27:** 2359-2372.

Dryja T.P., McGee T.L., Reichel E., Hahn L.B., Cowley G.S., Yandell D.W., Sandberg M.A. and Berson E.L. (1990) A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* **343:** 364-366.

Ermolaeva O., Rastogi M., Pruitt K.D., Schuler G.D., Bittner M.L., Chen Y., Simon R., Meltzer P., Trent J.M. and Boguski M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet.* **20:** 19-23.

Ewing R.M., Kahla A.B., Poirot O., Lopez F., Audic S. and Claverie J-M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9:** 950-959.

Ewing B.and Green P. (2000) Analysis of expressed sequence tags indicates 35000 human genes. *Nat. Genet.* **25:** 232-234.

Ewing R.M. and Claverie J-M. (2000) EST databases as multi-conditional gene expression datasets. *Pac Symp Biocomput.* 430-432.

Fanning T.G. and Singer M.F. (1987) LINE- 1: a mammalian transposable element. *Biochim. Biophys. Acta* **910:** 203-212.

Ferrer J., Wasson J., Schoor K.D., Mueckler M. and Donis-Keller H. (1997) Mapping novel pancreatic islet genes to human chromosomes. *Diabetes* **46:** 386-392.

Field L.L., Tobias R. and Magnus T. (1994) A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* **8:** 189-194.

Fisher R.A. (1973) Statistical Methods and Scientific inferences. 3rd Edn. Macmillan Hafner, New York, NY.

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.

Fraser, A., McCarthy N and Evan G.I. (1996) Biochemistry of cell death. *Curr Opin Neurobiol.* **6:** 71-80.

Fougerousse F., Broux O., Richard I., Allamand V., de Souza A.P., Bourg N., Brenguier L., Devaud C., Pasturaud P., Roudaut C. et al. (1994) Mapping of a chromosome 15 region involved in limb-girdle muscular dystrophy. *Hum. Mol.Genet.* **2:** 285-293.

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.

Goate A.M., Chartier-Harlin M.C., Mullan M., Brown J., Crawford F., Fidani L., Giuffra L., Haynes A., Irving N., James L et al. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349:** 704.

Gress T.M., Muller-Pillasch F., Geng M., Zimmerhackl F., Zehetner G., Friess H., Buchler M., Adler G. and Lehrach H. (1996) A pancreatic cancer-specific expression profile. *Oncogene* **13:** 1819-1830.

Gupta M. and Singal P.K (1989) Higher antioxidative capacity during a chronic stable heart hypertrophy. *Circ Res.* **64:** 398-406.

Gayapay G., Morrissette J., Vignal A., Dib C., Fizames C., Millasseau P., Marc S., Bernadi G., Lathrop M and Weissenbach J. (1994). The 1993-1994 Genethon human genetic linkage map. *Molecular Cell Biology.* 6: 3010-3013.

Hashimoto L., Habita C., Beressi J., Delepine M., Bess C., Cambon-Thomsen A., Deschamps A., Rotter I., Djoulah S., James M., Froguel P., Weissenbach J., Lathrop G. and Julier C. (1994) Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. *Nature* **371:** 161-164.

Herz J., Hamann U., Rogne S., Myklebost O., Gausepohl H. and Stanley K.K. (1988) Surface location and high affinity for calcium of a 500-kd liver membrane protein closely related to the LDL-receptor suggest a physiological role as lipoprotein receptor. *EMBO J.* **7:** 4119-4127.

Hey P.T., Twells R.C.J., Phillips M.S., Nakagawa Y., Brown S.D., kawaguchi Y., Cox R., Xie G., Dugan V., Hammond H., Metzker M.L., Todd J. A. and Hess J.F. (1998) *Gene* **216:** 103-111.

Hide W., Burke J. and Davidson D. (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol*. **1:** 199-215.

Hide W., Burke J., Christoffels A. and Miller R. (1997) A novel approach towards a consensus representation of the expressed human genome. *Genome Informatics* Universal Academy Press Inc. Tokyo, Japan. ISSN 0919-9454, 187-196.

Hide W., Miller R., Ptitsyn A., Kelso J., Gopallkrishnan C. and Christoffels A. (1999) EST clustering tutorial. Intellegent Systems for Molecular Biology (ISMB1999). Germany.

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

Horiuchi M., Yamada H., Akishita M., Ito M., Tamura K. and Dzau VJ. (1999) Interferon regulatory factors regulate interleukin-1beta-converting enzyme expression and apoptosis in vascular smooth muscle cells. *Hypertension* **33:** 162-166.

Houlgatte R, Mariage-Samson R., Duprat S., Tessier A., Bentolila S., Lamy B and Auffray C. (1995) The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5:** 272-304.

Huang X., Adams M.D. Zhou H and Kerlavage A.R. (1997) A tool for analyzing and annotating genomic sequence. *Genomics* **46:** 37-45.

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

Hwang DM., Hwang WS. and Liew CC. (1994) Single pass sequencing of a unidirectional human fetal heart cDNA library . *Journal of Mol. Cell Cardiol.* **26:**1329-1333.

Hwang DM., Fung YW., Wang RX., Laurenssen CM., NG SH., Lam WY., Tsui KW., Fung KP., Waye M., Lee CY., Liew C-C. (1995) Analysis of expressed sequence tags from the fetal human heart cDNA library. *Genomics* **30:** 293-298.

Hwang DM., Dempsey AA., Wang RX., Rezvani M., Barrans JD., Dao K-S., Wang H-Y., Ma H., Cukerman E., liu Y-Q., Gu J-R., Zhang J-H., Tsui S., Waye M.M.Y., Fung K-P., Lee C-Y and Liew C-C. (1997) A genome-based resource for molecular cardiovascular medicine: toward a compendium of cardiovascular genes. *Circulation* **96:** 4146-4203.

Ing D.J., Zang J., Dzau V.J., Webster K.A., Bishopric N.H. (1999) Modulation of cytokine-induced cardiac myocyte apoptosis by nitric oxide, Bak and Bcl-x. *Circ Res.* **84:** 21-33.

Ji H., Liu Y.E., Jia T., Wang M., Liu J., Xiao G., Joseph B.K., Rosen C. and Shi Y.E. (1997) Identification of a breast cancer-specific gene, *BCSG1* by direct differential cDNA sequencing. *Cancer Research* **57:** 759-764.

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

Jurka J. (1998) Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* 8:333-337. http://charon.girinst.org/~server/repbase.html.

Jurka J., Klonowski P., Dagman V. and Pelton P. (1996) CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry* **20:** 119-122.

Kapros T., Robertson A.J. and Waterborg J.H. (1994) A simple method to make better probes from short DNA fragments. *Mol. Biotechnol.* **2:** 95-8.

Kennedy J.L. and Macciardi F.M. (1998) Chromosome 4 workshop. *Psychiatric Genetics*. **8:** 67-71.

Keating M.(1992) Linkage analysis and long QT syndrome. Using genetics to study cardiovascular disease. *Circulation* 85: 1973-1986.

Kirshenbaum L.A., Hill M. and Singal P.K. (1995) Endogenous antioxidants in isolated hypertrophied cardiac myocytes and hypoxia-reoxygenation injury. *J Mol Cell Cardiol.* **27:** 263-272.

Kroemer G. (1997) The proto-oncogene Bcl-2 and its role in regulating apoptosis. *Nat Med.* **3:** 614-620.

Korf I., Bedell J.A. and Gish W. (2000) MaskerAid: High Performance repeat identification. Genome Sequencing and Biology. Cold Spring Harbor Laboratories, Cold Spring Harbor. New York. 10-14 May.

Lamperti E.D., Kittelberger J.M., Smith T.F. and Villa-Komaroff L. (1992) Corruption of genomic databases with anomalous sequence. *Nucleic Acids Res.* **20:** 2741-7.

Langfranchi S.J., Murrano T., Caldara F., Pacchioni B., Pallavicini A., Pandolfo D., Toppo S., Trevisan S., Scarso S. and Valle G. (1996) Human skeletal muscle transcript map. *Genome Res.* **6:** 35-42.

Larionov V., Kouprina N., Nikolaishvili N. and Resnick M.A. (1994) The role of recombination and RAD52 in mutation of chromosomal DNA transformed into yeast. *Nucleis Acids Res.* **20:** 4154-4162.

Lennon G.G., Auffray G., 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.

Levy-Lahad, E., Wijsman E.M., Nemens E., Anderson L., Goddard K.A.B., Weber J.L., Bird T.D. And Schellenberg G.D. (1995a) A familial alzheimer's disease locus on chrmosome 1. *Science* **269:** 970-972.

Levy-Lahad, E., et al. (1995b) Candidate gene for the chromosome 1 familial alzheimer's disease locus. *Science* **269:** 973-977.

Lewin B. (1997) Genes VI. Oxford University Press Inc., New York pp1260.

Liang F., Holt I., Perea G., Karamycheva S., Salzberg S. L. and Quackenbush J. (2000a) Gene index analysis of the human genome estimates approximately 120000 genes. *Nat. Genet*. **25:** 239-240.

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

Liang P and Pardee A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257:** 967-971

Liew CC. A human heart cDNA library: the development of an efficient and simple method of automated DNA sequencing. (1993). *Journal of Mol. Cell. Cardiology*. **25:** 891-894.

Liew CC., Hwang DM., Fung YW., Laurenssen C., Cukerman E., Tsui S and Lee CY. (1994) A catalogue of genes in the cardiovascular system as identified by expressed sequence tags (ESTs). *Proc. Natl Acad. Sci. USA.* **91:** 10645-10649.

Livesey F.J. and Hunt S.P. (1996) Identifying changes in gene expression in the nervous system: mRNA differential display. *Trends Neurosci.* **19:** 84-88.

Loppnow H., Werdan K., Reuter G. and Flad H.D. (1998) The interleukin-1 and interleukin-1 converting enzyme families in the cardiovascular system. *Eur Cytokine Netw.* **9:** 675-680.

Lou D., Buzette R., Botter J., Maclaren N., Raffel L., Nistico L., Giovannini C., Pozzilli P., Thomson G. and She J. (1996) Confirmation of three susceptibility genes to insulin-dependent diabetes mellitus: IDDM4, IDDM5 and IDDM8. *Hum. Mol. Genet.* **5:** 693-698.

MacLellan W.R. and Schneider M.D. (1997) Death by design: Programmed cell death in cardiovascular biology and disease. *Circ. Res.* **81:** 137-144.

Makalowski W., Zhang Z. and Boguski M.S. (1996) Comparative analysis of 1196 orthologous mouse and human full length mRNA and protein sequences. *Genome Ress* **6:** 846-857.

Maulik N., Goswami S., Galang N. and Das DK. (1999) Differential expression of Bcl-2, AP-1 and NF-kappaB on cardiomyocyte apoptosis during myocardial ischemic stress adaptation. *FEBS Lett.* **443:** 331-336.

Miller G., Fuchs R. amd Lai E. (1997) IMAGE cDNA clones, UniGene clustering, and ACeDB: an integrated resource for expressed sequence information. *Genome Res.* 1027-32.

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

Murano T., Stephan D., Pallavicini A., Tiso N., Zimbello R., Danieli G.A., Hoffman E.H., Valle G. and Lanfranchi G. (1997) Chromosomal assignment of 115 expressed sequence tags (ESTs) from human skeletal muscle. *Cytogenet. Cell Genet.* **76:** 144-152.

Nakagawa O., Ogawa Y., Itoh H., Suga S., Komatsu Y., Kishimoto I., Nishino K., Yoshimasa T. and Nakao K. (1995) Rapid transcriptional activation and early mRNA turnover of brain natriuretic peptide in cardiocyte hypertrophy: evidence for brain natriuretic peptide as an ‘emergency’ cardiac hormone against ventricular overload. *J Clin Invest.* **96:** 1280-1287.

Nakashima T., Sekiguchi T., Kuroaka A., Fukushima K., Shibata Y., Komiyama S. and Nishimoto T. (1993) Molecular cloning of a human cDNA encoding a novel protein, DAD-1, whose defect causes apoptotic cell death in hamster BHK21 cells. *Mol Cell Biol.* **13:** 6367-6374.

Narula J., Haider N., Virmani R., DiSalvo T.G., kolodgue F.D., Hajjar R.J., Schmidt U., Semigran M.J., Dec G.W. And Khaw B.A. (1996) Programmed myocyte death in end-stage heart failure. *New Eng. J. of Med.* **335:** 1182-1189.

Neto E.D., Harrop R., Correa-Oliveira R., Wilson R.A., Pena S.D. and Simpson A.J. (1997) *Gene* **186:** 135-142.

Neto E.D., Garcia C., Verjovski-Almeida S., Briones M.R.S., Nagai M.A., de Silva W., Zago M.A., Bordin S., Costa F.F., Goldman G.H. et al. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA.* **97:** 3491-3496.

Nguyȇn-Tran Van TB., Kubalak S.W., Minamisawa S., Fiset C., Wollert K.C., Brown A.B., Ruiz-Lozano P., Barrere-Lemaire S., Kondo R., Norman L.W., Gourdie R.G., Rahme M.M., Feld G.K., Clark R.B., Giles W.R. and Chien K.R. (2000) A novel pathway for sudden cardiac death via defects in the transition between ventricular and conduction system cell lineages. *Cell* **102:** 671-682.

Norris J., Fan D., Aleman C., Marks J., Futrea P., Wiseman R., Iglehard J., Deininger P. and McDonnell D. (1995) Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J. Biol. Chem.* **270:** 22777-22782.

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

Olivetti G., Abbi R., Quaini F., Kajstura J., Cheng W., Nitahara J.A., Quaini E., Di Loreto C., Beltrami C.A., Krajewski S., Reed J.C. and Anversa P. (1997) Apoptosis in the failing human heart. *New Eng. J of Med.* **336:** 1131-1141.

Olson M., Hood L., Cantor C. and Botstein D. (1989) A common language for physical mapping of the human genome. *Science*. **245:** 1434-1435.

Park J.R. and Hochenberry D.M. (1996) BCL-2, a novel regulator of apoptosis. *Journal Cell Biochem.* **60:**12-17.

Pietu G., Albert O., Guichard V., Lamy B., Bois F., Leroy E., Mariage-Samson F., Houlgatte R., Soularue P and Auffray C. (1996) Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridisation of a high density cDNA array. *Genome Res.* **6:** 492-503.

Pietu G., Mariage-Samson R., Fayein N-A., Matingou C., Eveno E., Houlgatte R., Decraene C., Vandenbrouck Y., Tahi F., Devignes M-D., Wirkner U., Ansorge W., Cox D., Nagase T., Nomura N. and Auffray C. (1999) The genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Res.* **9:** 195-209.

Poulos J.E., Gower W.R Jr., Sullebarger J.T., Fontanet H.L. and Vesely D.L. (1996) Congestive heart failure: increased cardiac and extracardiac atrial natriuretic peptide gene expression. *Cardiovasc Res.* **32:** 909-919.

Prange C., Folta P., Harsch T., Johnson G., Kale P., Kuczmarski T., Lato B., Mila L., Nelson D. and Carrano A. (2000) The I.M.A.G.E consortium: expanding a publicly available cDNA resource. Genome Sequencing and Biology. Cold Spring Harbor Laboratories, Cold Spring Harbor. New York. 10-14 May.

Ptitsyn A. (2001) New algorithms for EST clustering. PhD thesis, South African National Bioinformatics Institute, University of Western Cape, Cape Town, South Africa.

Putnam E.A., Zhang H., Ramirez F. and Milewicz D.M. (1995) Fibrillin-2 (FBN2) mutations result in the Marfan-like disorder, congenital contractural arachnodactyly. *Nat. Genet*. **11:** 456-458.

Quackenbush J., Cho J., Lee D., Liang F., Holt I., karamycheva S., Parvizi B., Pertea G., Sultana R. and and White J. (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29:** 159-164.

Rawn J. D (1989) RNA processing. *In:* Biochemistry. Neil Patterson Publishers. USA. 781-820.

Rezvani M. and Liew C-C. (2000) Role of the Adenomatous Polyposis Coli gene product in human cardiac development and disease. *Journal of Biological Chemistry* **275:** 18470-18475.

Rezvani M., Barrans JD., Dai K-S and Liew CC. (2000) Apoptosis-related genes expressed in cardiovascular development and disease: an EST approach. *Cardiovascular Res.* **45:** 621-629.

Richard I., Broux O., Allamand V., Fougerousse F., Chiannikulchai N., Bourg N., Brengier L., Devaud C., Pasturaud P., Roudaut C. et al. (1995) Mutations in the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A. *Cell* **81:** 27-40.

Rochefort H., Capony F., Garcia M., Cavailles V., Freiss G., Chambon M., Morisset M. and Vignon. (1987) Estrogen-induced lysosomal proteases secreted by breast cancer cells: a role in carcinogenesis? *J Cell. Biochem.* **35:** 17-29.

Rogaev E.I., Sherrington R., Rogaeva E.A., Leveseque G., Dseda M., Liang Y., Chi H., Lin C., Holman K., Tsunda T. et al. (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature.* **376:** 775-778.

Saikumar P., Dong Z., Weinberg JM., Venkatachalam MA. (1998) Mechanisms of cell death in hypoxia/reoxygenation injury. *Oncogene.* **17:** 3341-3349.

Sarraf P., Mueller E., Smith W. M., Wright H. M., Kum J. B., Aaltonen L. A., de la Chapelle A., Spiegelman B. M. and Eng C. (1999) Loss-of-function mutations in PPAR-gamma associated with human colon cancer. *Molec. Cell* **3:** 799-804.

Schellenberg G.D., Bird T.D., Wijsman E.M., Orr H.T., Anderson L., Nemens E., White J.A., Bonnycastle L., Weber J.L., Alonso M.E., Potter H., Heston L.L. and Martin G.M. (1992) Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science* **258:** 668.

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 tumor 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., et al. (1996) A gene map of the human genome. *Science* **274:** 540-546.

Schuler G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Mol. Med.* **75:** 688-694

Semple C.A., Morris S.W., Porteous D.J. and Evans K.L. (2000) In silico identification of transcripts and SNPs from a region of 4p linked with bipolar affective disorder. *Bioinformatics* **16:** 735-738.

Sherrington R., Rogaev E.I., Liang Y., Rogaeva E.A., Levesque G., Ikeda M., Chi H., Lin C., Li G., Holman K., et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375:** 754-760.

Shibahara K., Asano M., Ishida Y., Aoki T., Koike T., Honjo T. (1995) Isolation of a novel mouse gene MA-3 that is induced upon programmed cell death. *Gene* **166:** 297-301.

Singer M.F. (1982a). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28:** 433-434.

Singer M.F. (1982b). Highly repeated sequences in mammalian chromosomes. Int. Rev. Cytol **76:** 67-112.

Smit A, F, A. and Green, P. (1999) http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl.

Strachan T. and Read A.P. (1998) *In:* Human Molecular Genetics. Bios Scientific Publishers. Oxford. United Kingdom.

Stoeckert C., Pizarro A., Manduchi E., Gibson M., Brunk B., Crabtree J., Schug J., Shen-Orr S. and Overton G.C. (2001) A relational schema for both array-based and SAGE gene expression experiments. *Bioinformatics* **17:** 300-308.

Sutton G., White O., Adams M.D. and Kerlavage A.R. (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Science and technology* **1:** 9-18.

Tassabehji M., Newton V.E. and Read A.P. (1994) Waardenburg syndrome type 2 caused by mutations in the microphthalmia (MTTF) gene. *Nat Genet.* **8:** 251-255.

Tisch R.and McDevitt H. (1996) Insulin-dependent diabetes mellitus. *Cell* **85:** 291-97.

Todd J. (1996) Genetic analysis of type 1 diabetes using whole genome approaches. *Proc. Natl. Acad. Sci. USA* **92:** 8560-8565.

Todd J. and Farrall M. (1996) Panning for gold: genome-wide scanning in type 1 diabetes. *Hum. Mol. Genet.* **5:** 1443-1448.

Tóth G., Gáspári Z. and Jurka J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.***10:** 967-981.

Torney D., Burks C., Davison D. and Sirotkin K. (1990a) A simple measure of sequence divergence. *Los Alamos National Laboratory technical report LAUR* 89-946. LANL, Los Alamos, NM.

Torney D. (1990b) Computation of d2. A measure of sequence dissimilarity. In Computers and DNA (ed. G. Bell and T. Marr), Vol. Santa Fe Institute studies in the sciences of complexity, vol. VII. Addison-Wesley, New York, NY.

Usui H., Falk J.D., Dopazo A., de Lecea L., Erlander M.G. and Sutcliffe J.G. (1994) Isolation of clones of rat striatum-specific mRNAs by directional tag PCR subtraction. *J. Neurosci.* **14:** 4915-4926.

Vansant G. and Reynolds W.F. (1995) The consensus sequence of a major Alu subfamily containsa functional retinoic acid response element. *Proc Natl Acad Sci (USA)* **92:** 8229-8233.

Vasmatzis G., Essand M., Brinkmann U., Lee B. and Pastan I. (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci. USA* **95:** 300-304.

Vingron M. and Hoheisel J. (1999) Computational aspects of expression data. *J. Mol. Med.* **77:** 3-7.

von Harsdorf R., Li PF. and Dietz R. (1999) Signaling pathways in reactive oxygen species-induced cardiomyocyte apoptosis. *Circulation* **99:** 21934-21941

Vyse T. and Todd J. (1996) Genetic analysis of autoimmune disease. *Cell* **85:** 311-318.

Wagner L., Schriml L., Pontius J., Maglott D. and Schuler G. (2000) Comparative genomics: The challenge of ESTs. Genome Sequencing and Biology. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. May 10-14 2000.

Wang L., Miura M., Bergeron L., Zhu H. and Yuan J. (1994) Ich-1, an Ice/ced-3-related gene, encodes both positive and negative regulators of programmed cell death. *Cell* **78:** 739-750.

Wang K., Yin X.M., Copeland N.G., Gilbert D.J., Jenkins N.A., Keck C.L., Zimonjic D.B., Popescu N.C. and Korsmeyer S.J. (1998) BID, a proapoptotic BCL-2 family member, is localised to mouse chromosome 6 and human chromosome 22q11. *Genomics.* **53:** 235-238.

Watson P.A., Hannan R., Carl L.L. and Giger K.E. (1996) Desmin gene expression in cardiac myocytes is responsive to contractile activity and stretch. *Am J Physiol.* **270:** C1228-C1235.

Welle S., Bhart K. and Thornton C.A. (1999) Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Res.* **9:** 506-513.

Welsh J., Chada K., Dalal S.S., Cheng R., Ralph D. and McClelland M. (1992) Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res.* **20:** 4965-4970.

Wheeler D.L., Church D.M., Lash A.E., Leipe D.D., Madden T.L., Pontius J.U., Schuler G.D., Schriml L.M., Tatusova T.A., Wagner L. and Rapp B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleids Acids Res.* **29:** 11-16.

Wilcox A.S., Khan A.S., Hopkins J.A. and Sikela J.M. (1991) Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion of STSs: implications for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837-1843.

Wildrick D. M. and Boman B. M. (1988) Chromosome 5 allele loss at the glucocorticoid receptor locus in human colorectal carcinomas. *Biochem. Biophys. Res. Commun.* **150:** 591-598.

Williamson A.R. (1999) Merck Gene Index Project (MGIP) *Bioinformatics.* **4:** 115-122.

Wittenburger T., Schaller H.C. and Helebrand S. (2001) An expresse sequence tag (EST) data mining strategy succeeding in the discovery of new G-protein coupled receptors. *J.Mol.Biol.* **307:** 799-813.

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

Wu G., Su S. and Bird B.C. (1994) Optimisation of subtractive hybridisation in construction of subtractive cDNA libraries. *Genet Analysis, Techniques & Applications* **11:** 29-33.

Wu T.J., Burke J.P. and Davison D.B. (1997) A Measure of DNA Sequence Dissimilarity Based on Mahalanobis Distance Between Frequencies of Words. *Biometrics* **53:** 1431-1439.

Zhang Z., Schwartz S., Wagner L. and Miller W. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*.**7:** 203-214.

CHAPTER 2

Optimisation of D2_CLUSTER for clustering of large EST data sets

Optimisation of D2_CLUSTER for clustering of large EST data sets

Contents **Page**

List of Figures

List of Tables

Summary:

Maximum value is extracted from fragmented EST resources by constructing gene indices, where all transcripts are partitioned into index classes such that transcripts are put into the same index class if they represent the same gene or gene isoform. EST projects implement different EST clustering methodologies to partition ESTs into index classes. The use of a non-alignment based algorithm, such as D2_CLUSTER is ideal for clustering ESTs that are known to contain sequence errors. However, the millions of ESTs in the EST databases present a formidable task for EST clustering tools such as D2_CLUSTER, because the computational cost of clustering a set of sequences is quadratic in the number of sequences. Therefore, the ability to cluster all human ESTs, in view of generating a STACK human gene index, required new algorithms or high performance machines. In the absence of new EST algorithms to cope with the deluge of EST data (at the time), we embarked on an approach to utilise high performance multiple processor machines to accelerate D2_CLUSTER processing. In this chapter, I report on the optimisation for porting D2_CLUSTER to the Origin2000 architecture and the modifications made to the code in order to accelerate EST clustering on multiple processors. Test data sets ranging from 4000-100000 sequences were used to bench-mark D2_CLUSTER performance on a multi-processor Origin2000. A restart capability was added to D2_CLUSTER that allowed the clustering procedure to be restarted at the same point at which it was interrupted. The implementation of restart functionality was enhanced by the ability to break the work into a number of pieces such that each piece processes more sequences and each successive piece uses less time. The ability to cluster large data sets was enhanced by replicating the database onto each of 128 processors.

A test data set of 15876 sequences demonstrated a reduction in time from 1000 CPU seconds to 800 CPU seconds on a 16 CPU Origin2000. D2_CLUSTER was used to cluster 490293 sequences on 128 CPU R12000 300 MHZ Origin2000 in 31 hours. The successful porting of D2_CLUSTER to the Origin2000 put the generation of a STACK human gene index into the realm of feasibility.

2.1 Introduction

D2_CLUSTER uses a word multiplicity comparison method that does not rely on alignment to derive a distance measure (referred to as the d^2 score in this chapter) between two sequences (Torney et al., 1990; Burke et al., 1999). Instead, D2_CLUSTER utilises an agglomerative algorithm where each sequence begins in its own cluster and the final clustering is constructed through a series of mergers that have been described in terms of minimal linkage by Burke et al (1999)(see chapter1 section 1.4.1.3b). Briefly minimal linkage (also referred to as transitive closure) refers to the property that two sequences (A and B) are in the same cluster even if they share no similarity but there exists a sequence C with enough similarity to both A and B. The d^2 score (i.e., measurement of similarity) between two sequences is calculated based on word matching within a window, together with a measure of the multiplicity of that word within a window. Therefore, the results of comparison are derived directly from the comparison of word composition (word identity and multiplicity) of sequence windows where very high similarities are detected (i.e., above 96% identity within a window as defined in Miller et al. (1999)).

The use of a non-alignment based algorithm, such as D2_CLUSTER is ideal for clustering ESTs that are known to contain sequence errors. However, the millions of ESTs in the EST databases present a formidable task for EST clustering tools such as D2_CLUSTER, because the computational cost of clustering a set of sequences is quadratic in the number of sequences and this quadratic cost arises from the need to compute the d^2 score for all pairs of sequences in the database and then merge the results into a growing set of clusters.

In order to reduce the computational cost of clustering large EST datasets in the initial versions of STACK production, input sequences for D2_CLUSTER processing were reduced in size based on arbitrary tissue partitioning (Miller et al., 1999). These “tissue bins” represented managable data sizes (maximum of 60000 EST sequences) for clustering on the MasPar MP 22-16 SIMD architecture. However, the MasPar implementation of D2_CLUSTER failed to meet the demands of an exponentially increasing EST database (Benson et al., 1999), as 100000 ESTs required six days of processing time on 16000 SIMD processors. The human division of dbEST (GenBank 110 release) was approaching 1.3 million ESTs, at that time, and the ability to cluster all human ESTs in view of generating a STACK human gene index required new algorithms or high performance machines.

In the absence of new EST algorithms to cope with the deluge of EST data, we embarked on an approach to utilise high performance multiple processor machines to accelerate D2_CLUSTER processing. A United States Department of Energy grant (DE-FCO3-95ER62062) was successfully obtained to develop clustering on multiple processor machines at the National Center for Super Computing Applications (NCSA). D2_CLUSTER had undergone initial parallelisation (Yael Weinbach, SGI, pers comm.) but had not been exhaustively tested on an Origin2000. Therefore, we set out to cluster sequence sets ranging from 4,000 to 100,000 sequences using D2_CLUSTER on multiple processors on an Origin2000 in order to benchmark D2_CLUSTER performance with a view of clustering sequences in excess of 500000 sequences on 128 processors. The development of a high performance shared memory parallel (SMP) model for D2_CLUSTER required:

1. **hardware configuration tuning**

The high performance NCSA Origin2000 machines are modified weekly with respect to hardware configuration including total memory, CPU availability and programming language-specific compilers. These factors need to be taken into account while optimising D2_CLUSTER's parallel performance.

2. **isolating parts of the D2_CLUSTER code for parallelisation**

Sections of the D2_CLUSTER code that require little processing time will not benefit from parallel processing and therefore only parts of the code that are computationally expensive needs to be parallelised.

3. **memory management**

ESTs have been the main focus for the application of D2_CLUSTER. However, mRNA sequences are becoming readily available and their increased length comapred to ESTs requires additional computer memory for clustering purposes. Therefore, the parallel version of D2_CLUSTER needs to cope with additional memory demands placed on it.

4. **Verification (i.e., verifying consistent results between serial and parallel clustering using D2_CLUSTER)**

The parallel version of D2_CLUSTER has to produce consistent results for clustering when compared to EST clusters generated by a serial run of D2_CLUSTER. Any conflicting results need to be analysed in order to understand how the code is behaving.

2.2 D2_CLUSTER program

The three steps involved in clustering sequence data using D2_CLUSTER are:

(1) Preprocessing of the input data set with a program called enc_db that generates a file of compressed sequence data and an index file giving the start position of every sequence in the compressed file.

(2) Clustering the input data using D2_CLUSTER. D2_CLUSTER saves all relationships between clustered sequences in a five-column matrix that is stored in a text file called "CLUSTER_TABLE" (see Appendix 1.6 for a description)

(3) interpretation of the CLUSTER_TABLE by a program, post_proc, in which each cluster of ESTs is transformed into a FASTA file of sequences.

2.2.1 D2_CLUSTER algorithm description

A description of D2_CLUSTER has been given in the context of its implementation in the STACK database as an EST clustering tool (chapter one section 1.4.1.3b). However, a description of the D2_CLUSTER algorithm is provided below to place the D2_CLUSTER optimisation in its proper context.

D2_CLUSTER comprises different blocks of code or routines that perform a specific function. The “bin2” routine converts each sequence to an array of “word scores” where “word scores” refer to the occurrence of all words (length = 6) in a sequence. The “compare2” routine then proceeds to compare all sequences to each other by comparing the array of word scores for each pair of sequences to calculate a d^2 score utilising a window of WINDOW_SIZE length (WINDOW_SIZE = 150bp) (Figure 2.2 line5). For example, if the database contains 5 sequences then for I=1,the iterations occur over J=2 through J=5, so that each of the four sequences (2-5) are compared with the first sequence (I=1) (Figure 2.2 lines 4-11). When I=2, then the iterations occur over J=3 through J=5, and each of the three sequences are compared with the second sequence (I=2) etc. The loop is decremented each time (J> I) so that a sequence is never compared to itself. This comparison results in sequence combinations: 1-2,1-3,1-4,1-5,2-3,2-4,2-5,3-4,3-5,4-5. The sequence windows of identical length can be visualised as a square matrix where an "x" represents the comparisons (Figure 2.1). All sequence pairs with a d^2 score above a theshold value are passed to the MERGE function (Figure 2.2, line6-8). Therefore the decision as to which sequences are identical is made prior to arriving at the MERGE function. The MERGE function sees sequences of (I,J) pairs and assigns sequence J to the cluster that represents sequence I or vice versa (figure 2.2 line7-8). The order of these (I,J) pairs varies depending on the number of processors being used. The MERGE function updates the variables that hold the sequence relationships such as the variables captured in the CLUSTER_TABLE (ie., LINK, MEMB and ORIENT variables, see Appendix I).

	1	2	3	4	5
1		x	x	x	x
2			x	x	x
3				x	x
4					x
5					

Figure 2.1 Square matrix indicating sequence windows of identical length. The "x" represets the comparison between two sequences. A sequence is never compared to itself.

2.3 Glossary of terms

A list of definitions is provided below that describes the Origin2000 hardware and software environment at NCSA. All terms are defined by the free online dictionary of computing (http://foldoc.doc.ic.ac.uk/foldoc/index.html) unless otherwise stated.

2.3.1 Mass storage system (mss) (http://archive.ncsa.uiuc.edu/SCD)

The UniTree system is a data storage system for NCSA users and comprises a single node Origin2000 server. The server contains eight 250MHz R10000 processors with 8Gb of memory. In addition, 18 tape drives provides 2 terabytes of disk space. Access to the UniTree system is only via ftp.

2.3.2 Compiler

A program that converts another program from some programming language (source) to machine language (i.e., representation of a computer program which is actually read and interpreted by the computer.

2.3.3 MIPSpro Compilers

(http://www.sgi.com/developers/devtools/languages/mipspro.html)

The MIPSpro compilers represent parallelised compilers from SGI and support C, C++ and Fortran 77/90 prorgamming languages. The MIPSpro compilers carry out high-level and architecture-specific optimisations to automatically improve the performance of a wide range of applications. Optimisation is achieved through the performance-orientated features in the MIPS microprocessors such as high-speed calling conventions and 32 -bit and 64-bit floating-point registers.

2.3.4 Modules (http://www.mcsr.olemiss.edu/cgi-bin/man-cgi?mmci+5)

Modules control the environment used to access software on the Origin2000 machines.The use of modules allows the user greater control over which programs and what versions of those programs are available for use. Modules works by bundling all the setup routines for a program into a single "modulefile". These modulefiles can be loaded or unloaded through a "module" command. For example, the MIPSpro compilers are loaded by the MIPSpro modules. Since different modules exist for various releases of the compilers including MIPSpro 7.2.1.3 and MIPSpro 7.3, a user can change the version of the MIPSpro compiler by loading the specific modulefile. Modulefiles are usually loaded by specifying the modulefile in the ".cshrc" file: a file that controls a user's unix environment.

2.3.5 Policy modules

The ability of applications to control memory management becomes an essential feature in multiprocessor system in order to maximise code performance. A policy module contains methods used to handle operations pertaining to memory management. For example, "initial memory allocation" is an operation that is handled by the "placement policy module" that determines what memory to use when memory is being allocated. Policy modules are created using a built-in function, policy_set_t, and a reference to the newly created policy module is returned using the "pm_create” routine. The "pm_create" routine returns a negative number when an error has occured in the policy module creation step.

2.3.6 Routine

A sequence of instructions for performing a particular task.

2.3.7 batch processing

A system that takes a set of commands or jobs, executes them and returns the results, all without human intervention. This contrasts with an interactive system where the user's commands and the computer's responses are interleaved during a single run.

2.3.8 LSBATCH

A load sharing system that provides distributed batch job scheduling services.

2.3.9 stack

A data structure for storing items which are to be accessed in “last-in first-out” order. The operations on a stack are to create a new stack, to add a new item onto the top of a stack and to remove the top item off. Errors occur when an attempt is made to remove items from an empty stack or add items to a stack that has no more room. A stack is used to store subroutine arguments and return calls at the machine code level. The user defines an area of memory for use as a stack.

2.4 D2_CLUSTER and the hardware environment optimisation

The human division of dbEST (GenBank release 110) was extracted and partitioned into "tissue-bins" using the protocol described in chapter one section 1.4.1.3a. All the sequence data was stored on a mass storage system (mss) (see section 2.3.1) at NCSA and transferred to a computer named modi4 (64 CPU Origin2000, 195Mhz) when needed. The Origin2000 machines at NCSA operate on a batch submission process (see section 2.3.7) that had undergone modification during our usage of modi4. This new environment provided initial problems for our development work as we had to benchmark the parameters needed to run our clustering on multiple processors. D2_CLUSTER was used successfully to benchmark clustering of sequences ranging from 15,000-56,000 on 3 processors (Table 2.1). However, errors were encountered when D2_CLUSTER was used to cluster data sets larger than 56,000 sequences. The errors were related to (a) the hardware configuration and (b) the D2_CLUSTER code.

```
Line1   Start program
line2           Read in database of sequences into memory
line3           loop over sequence I
line4                   loop over sequence J > I
line5                           compute d2(I, J)
line6                           if d2(I,J) < threshold
line7                                   merge J into I's cluster.tar
line8                                   or vice versa
line9                           end if
line10                  end loop
line11          end loop
line12          write cluster membership to disk
line13  end program
```

Figure 2.2 Pseudo code for D2_CLUSTER. The line numbers are indicated at the far left.

2.4.1 Hardware configuration errors

2.4.1.1 Insufficient memory to grow stack

On the assumption that a linear relationship exists between the size of the data and number of processors required, data sets ranging from 60,000 to 200,000 sequences would be clustered within 48 hours on 32 processors (extrapolated from Table 2.1). Attempts to cluster data sets in excess of 60,000 sequences using 32 CPUs failed due to insufficient memory to grow the stack. The "stack" refers to the portion of memory that is used for procedure calls and storage of temporary variables (see section 2.3.9). A detailed description of the errors causing the core dumps were located in "/var/adm/SYSLOG/". The use of 32 CPU exceeded the maximum memory (15GB) if no change was made to the stack parameter (i.e., 4GB x 32 processors =128GB). The stack limit was reduced to 200MB (see appendix 1 for parameter usage) which translated to about 6.4GB for 32 CPU (200 x 32), well within the boundary of the 15GB limit. Two large sequence sets namely reproductive and gland were successfully clustered and the clustered data was transferred, using the file transfer protocol (ftp), to SANBI for further processing.

2.4.1.2 Error message: "error:pm_create: Invalid argument" on one CPU

The clustering code was transferred from the 48 CPU Origin2000 machine to 64CPU and 128 CPU Origin2000 machines, named arctic and flurry respectively, in order to derive D2_CLUSTER benchmarks for a broaded range of CPUs. The first attempt at running D2_CLUSTER on arctic failed with an error message, "error:pm_create: Invalid argument". Attempts to reduce the stack limit (see section 2.3.9) on arctic failed to resolve the problem. Initially it was thought that there were problems with the default modules on the Origin2000 machines but addition of the latest MIPSpro module (section 2.3.4) did not correct the problem. The MIPSpro module ensures the loading of the correct MIPSpro compiler (see section 2.3.3) but since the latest MIPSpro compiler was installed on arctic, there was no need to load any MIPSpro module. Therefore, the module specification was removed from the user defined environment (.cshrc file) and the "pm_create" problem disappeared. Batch submission (see section 2.3.7 and 2.3.8) parameters are processed via a queueing system and once these parameters were optimised (Appendix I), it resulted in the successful clustering of a test data set of 24000 sequences using 22MB of memory on four processors within 26 hours.

2.4.1.3 SIGTTOU signal

A signal is a way of telling a process that “something” has happened and this “something” needs to be delt with. A process can be defined as a UNIX abstraction that manages the memory, CPU and input/output resources for a specific program that has been executed. When a signal is delivered, one of two things happen: (1) a routine that handles this specific signal is called with the information about the context in which the signal was delivered or (2) a default action is taken on behalf of the process such as terminating the process (Nemeth E et al., 1995). SIGTTOU, represents one of more than 30 signals that are defined for a UNIX system. SIGTTOU signals are sent to a process that attempts to write information to disk without the necessary permissions.

SIGTTOU errors were generated if the time limit for a batch submission was exceeded. The choice of queues for batch submission can be made by examining the output of the "qstat" command (Appendix 1). A 32 processor queue with 500MB memory constraint (cpu32_unl_500Mb) was used for sequence sets that had less than 80,000 sequences. The clustering of data sets containing greater than 80,000 sequences required clustering with 64 processors in order to reduce the processing time.

2.4.2 Restart module development

Availability of 64 CPUs for clustering was restricted to a 48 hour session over a weekend. Clustering jobs were checkpointed (i.e., temporarily interrupted) if they were not completed within 48 hours. D2_CLUSTER processing could not be resumed after checkpointing. The loss of clustering information due to the interruption of a D2_CLUSTER run led to the development of a version of D2_CLUSTER that was restartable , ie., the information for a D2_CLUSTER run would be saved if the code was interrupted before completion.

Information relating to a cluster's membership, after comparing two sequences, is saved in the MEMB, LINK, ORIENT and SEQ variables (Appendix 1.6). These variables only store data and all four variables are used by iterations over a specific loop (initialization loop; Figure 2.2, line3-11), within compare.c (virt_start_pos). The capture of information within the initialisation loop would be sufficient to recover an interrupted clustering job. The "loop over sequence I" (Figure 2.2, line 5) was split up into a number of pieces so that each piece, once completed, would be written to the disk. This rationale was used to produce a version of D2_CLUSTER that was restartable. Two arguments were added to the program input namely, "number of pieces" and "restart flag" where "number of pieces" is the number of pieces to break the outer loop into, and "restart flag" that refers to the restart file that is written at the end of each loop (Figure 2.2). The use of "pieces" refers to the change introduced in the loop structure (illustrated in bold text) from:

```
for (virt_start_pos=0;virt_start_pos<num_seq;
     virt_start_pos++){
    for (l=virt_start_pos+1; l < num_seq: l++){
                compare sequences
                        merge scores in cluster arrays
    }
}
```

to

```
for (ipiece=0;ipiece < npieces; ipiece++) {
   for (virt_start_pos=vstart;virt_start_pos<vend;
        virt_start_pos++){
       for (l=virt_start_pos+1; l < num_seq: l++){
                  compare sequences
                        merge scores in cluster arrays
       }
   }
        write restart file
}
```

The contents of a "restart" file are the number of pieces, current piece, number of sequences, MEMB array, LINK array, ORIENT array and NEW array. Improvements to this code could include storage of the whole input state so that inconsistent restarts would not be possible. The ipiece loop starts at 0 and writes the information for that loop to RESTART.1. Therefore ipiece=1 writes RESTART.2 and ipiece=2 writes RESTART.1, with the result that if the last ipiece done is odd, then RESTART.2 is the most recent restart file whereas if the last ipiece done is even, then RESTART.1 is the most recent restart file.

2.4.3 D2_CLUSTER errors

2.4.3.1 CLUSTER_TABLE inconsistencies

Differences in the content of each CLUSTER_TABLE were observed when D2_CLUSTER was executed on different number of processors, i.e., cluster assignments varied with a change in the number of processors. A test data set of 2826 sequences was used to try and debug the code on more than 32 CPUs. The test data set was successfully clustered on 1, 2, 4, 8, 16, 32, 44, 64, 80, 96, and 120 processors where the 120 processor clustering ran to completion within five minutes. The CLUSTER_TABLES for the use of more than 32 CPUs consistently showed differences from the 1 CPU job for the 2826 sequence set. For example, four ESTs were assigned to a cluster using 32 CPU, whereas the same four ESTs were partitioned into a two member cluster and two singletons. Another example shows that 15 ESTs were assigned to one cluster when run on 32 CPU but three of the 15 ESTs were placed in a separate cluster when using 2 CPU.

In order to determine the basis for this bug, the MERGE operation was isolated from the d^2 computation so that all the $d^2(I,J)$ scores were stored in an array before being presented to the MERGE operation. This rationale meant that the order of presentation of $d^2(I,J)$ pairs to the MERGE operation was independent of the order in which $d^2(I,J)$ was calculated. The results showed a consistent CLUSTER_TABLE regardless of the number of processors. The MERGE operation could be restricted to a single processor without affecting the parallel performance because the time needed to perform the MERGE operation was 0.1 microseconds on a single processor compared to 99 microseconds for the d^2 computation. The storage of d^2 scores in an array meant that 500GB would be required for 1 million sequences. The memory cost was reduced by splitting the clustering into a number of "pieces" (Figure 2.2, line 4) as outlined in section 2.4.2.

2.4.3.2 Segmentation faults in compare.c

Segmentation faults and bus errors occurred because the compare routine was being executed beyond the length of a predefined hash query sequence array (query_arr) that indexes the word count array. In compare.c, there is a loop toward the end of the routine that is labeled as an initialization loop that ranges from Q_start to Q_start+Q_windowsize+n+1 (Figure 2.3).

The loop reads in an index from the set of words in the query sequence and then uses that index as an offset into the count2 and VISIT arrays. The original loop runs over the end of the query sequence words and then tries to initialize some random location. The segmentation faults would occur when executing the count2[pos] = MIN_FREQ statement (Figure 2.3) because "pos" was out of bounds, i.e., the summation was extending the prefined query_arr by 1. The actual error was very dependent on the size of the stackspace, the number of CPUs and the parameters in the header2.h as was expected from an "array out of bounds" error. The maximum limit for the initialisation loop was changed to "Q_start + Q_windowsize" (figure 2.3) to compensate for violating the predefined hash query array (i.e., query_arr). The code was tested on 2,4,16,32 and 48 CPU and each CLUSTER_TABLE gave the same results. The consistent CLUSTER_TABLE on multiple CPU prompted the generation of the whole-body index1.0, that included clustering of 330000 sequences in 400 pieces on 128 CPU (Table 2.2).

2.4.3.3 Memory allocation management.

A mRNA data set of 15000 sequences were downloaded from Baylor College of Medicine (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/HTDB/HTDB_unique) and used as a test data set for clustering sequences in excess of 700 bases with D2_CLUSTER (average EST length of 700 bases defined by Miller et al., 1999). The mRNA clustering failed due to insufficient memory and resulted in a stack_ptr is NULL”error.

Memory is dynamically allocated in two places in D2_CLUSTER. The first is for a large array to hold the database in memory. Insufficient memory will lead to failure to print the message “nnnnn packed words read ...". Memory is also allocated for arrays used by each parallel thread. Each parallel thread uses three character arrays and two integer arrays to hold unpacked sequence data and word count data. In order to do this allocation, one needs to know how long the input sequences are. This was simplified by calculating the length of each sequence in the database and using the value for the longest sequence for the memory allocation in each parallel thread. Clearly, that is inefficient for a data set where the data includes a few long sequences and many short sequences. A data set was used where the longest sequence was 83,000 nucleotides in length and the majority of sequences were on average 2172 nucleotides long. Given the longest sequence length of 83000, each parallel thread is trying to allocate just under 1MB of memory for each sequence.

The mRNA data set required a lot of memory because of the 83000 nucleotide sequence. The memory estimate on 16 CPU for this case was 213 MB when a copy of the database was accessible from each CPU, but only 60 MB when one copy of the database was accessed by 16 CPU at the same time. On 32 processors, the memory estimate essentially doubled for each case. D2_CLUSTER allocates six scratch arrays based on the maximum sequence length. Additional modifications (section 2.3.3) were added to the code that (a) prints the estimated memory usage at the start of a D2_CLUSTER run and (b) allows the decision of replicating the database on all processors, to be made on the command line.

A "stack_ptr is NULL" error was corrected when the MAX_LIB_SEQ, MAX_QUERY_SIZE and TABLESIZE parameters were made irrelevant by making the variables dynamic and dependent on the real maximum sequence size in the database. One significant parameter that was left was MAX_DIM which is the maximum number of sequences in the database. This parameter could be made irrelevant too, but it would require reading the ".ind" file twice.

Clustering the mRNA on 32 CPU exceeded the 500MB memory limit for the specific queue with "Unable to allocate db_private" error message. The mRNA data set was 10MB in size but the NQE batch submission system had miscalcuated the amount of memory needed by using the memory usage values (RSS value) from the "ps" command. The required memory was calculated to be 12GB (377MB x32 processors) whereas the true memory being used was 377MB in total. This problem was overcome by setting the memory requirement for each processor to 1GB and used a queue of unlimited memory (32_CPU_unlimited).

```
  for (k=Q_start;k<Q_start+Q_windowsize+n+1;k++) {
   pos = query_arr[k];
   count2[pos] = MIN_FREQ;
   VISIT[pos] = 0;
  }

changed to:

  for (k=Q_start;k<Q_start+Q_windowsize;k++) {
   pos = query_arr[k];
   count2[pos] = MIN_FREQ;
   VISIT[pos] = 0;
  }
```

Figure 2.3 Change made to the compare.c (initialization loop) to avoid having to over run the query_arr by 1 (bold text).

2.4.4 Speed Improvements in view of clustering 500,000 sequences

As outlined in section 2.2, the sequence database is uncompressed and word values are calculated for each sequence every time the sequence is required for a d^2 computation. While this is most efficient memory wise, redundant work is being done since the word values of a given sequence do not change. D2_CLUSTER was modified in order to remove the redundant work at the expense of using more memory. Four programs, namely version 1.1, 1.2, 1.25 and 1.3 were generated while trying to change the way the program stores its sequences and the optimisation for each program was tested with the 15,876 sequence set on 16 CPU (Table 2.3).

Versions 1.25 and 1.3 should not have taken longer to complete its processing because theoretically they were doing fewer operations. For example, version 1.25 and 1.3 did require repeated unpacking of a compressed database each time a sequence was required. However, this observation could be attributed to a cache effect: The compressed database of sequences for version 1.2 (1.2MB) fits in the 4MB secondary cache of the Origin2000 whereas the 5.0MB and 19MB databases of version 1.25 and 1.3 respectively do not fit into the secondary cache. The larger sizes of the databases for version 1.25 and 1.3 is due to the storage of uncompressed bases as characters (char). Assuming, that 120 CPU will be 7.2 times faster than 16 CPUs, version1.2 will need 1.1 microseconds per d2 calculation. An estimated 38 hours on a 128 CPU Origin (R12K, 300 Mhz) would be required to cluster 500,000 sequences using the formula t = (1.1 usecs)*n*(n-1)/2 where n is 500,000.

Additional optimisations were incorporated into D2_CLUSTER (version 1.21) prior to carrying out the clustering of 470,293 sequences. The optimisations included;

(a) Changes to the algorithm for breaking the work into pieces.

(b) Printing a histogram of cluster sizes versus number of clusters

(c) Replicating the database onto all the processors.

(a) Changes to the algorithm for breaking the work into pieces.

Since the j loop runs from i to N (figure 2.2 line 4; figure 2.4 lines 4-8), later pieces take less time if the pieces have a constant number of sequences. The algorithm was changed so that each successive piece does more sequences and at the same time each piece takes roughly the same amount of time (see Table 2.5 for mathematical formula). A consequence of this was that the program set the number of pieces.

(b) Printing a histogram of the clustering output
The code outputs a distribution of number of clusters versus cluster size.

(c) Replicating the database onto multiple processors.
To improve scalability, a copy of the packed database was made available to each CPU. On 16 CPU the database was accessed every 80 microseconds whereas on 128 CPU, the database was accessed every 8 microseconds. The increased memory usage had a negligible effect since the packed database was only 1.6 MB for the test data set of 15,876 sequences. However, for 1 million sequences with an average length of 600 bases, the packed database is 150 MB in size. A routine was embedded in the code that will not replicate the database if the computer's memory was limiting (see Table 2.5 for memory calculation).

Benchmarks were generated using the 15876 sequence data set on 128 CPU, R12000 300 Mhz (Table 2.4). A reduction in time from 1000 seconds to 800 seconds on 16 CPU was observed when compared to the previous benchmark (Table 3 and 4). This increase in scalability was due to a copy of the database being accessed on all the processors. The above-mentioned improvements marked the possibility of clustering 500,000 sequences. The modified D2_CLUSTER was used to cluster the STACKv2.3 tissue data sets on 126CPU (Table 2) to generate the whole-body index2.0. The D2_CLUSTER run was completed over 74 pieces in 31 hours on 128 R12000 300Mhz CPU. The postclustering processing of the whole-body index is discussed in chapter 3.

2.5 Future directions

D2_CLUSTER reads in a database of sequences into memory as a compressed file that is created by "enc_db". Each sequence has to be uncompressed before the wordsizes are calculated. Once the d2 comparisons are completed, the sequences are compressed again with the result that each sequence is uncompressed multiple times. An improvement could be to calculate the wordsizes on the compressed database. Alternately, the step of uncompressing the database can be circumvented by reading in the sequences directly from the FASTA file.

Profiling tests run on D2_CLUSTER has demonstrated that the bulk of the computational time is spent in the compare routine (90%, Cofer H. pers. comm). Any significant speed improvements have to be focused on modifications to the code that calculates the d^2 scores for all sequence pairs.

```
     Read database of sequences
     LOOP OVER PIECE K
#pragma parallel
           LOOP OVER SEQUENCE I in PIECE K
                 LOOP OVER SEQUENCE J > I
                       COMPUTE $d^2$
                 END LOOP OVER J
           END LOOP OVER I
#pragma end parallel
           LOOP OVER SEQUENCE I in PIECE K
                 LOOP OVER SEQUENCE J > I
                       IF ($d^2$ > THRESHOLD) MERGE(J,I)
                 END LOOP OVER J
           END LOOP OVER I
           Write restart file
     END LOOP OVER K
     Write cluster membership
```

Figure 2.4 The final structure of the shared memory parallel version of D2_CLUSTER

Table 2.1 Memory usage to cluster sequences ranging from 15712 to 56141 sequences.

Tissue type	**Number of sequences**	**CPU time (hours)**	**Max. Memory (megabytes)**	**Max. Swap (megabytes)**	**Processors**
Connective	15712	11	8	33	3
Eye	24878	31.9	8	33	2
Lung	25555	28.97	8	33	3
Genomic	38424	63.12	8	33	3
Gland	50710	295.5	8	33	2
Heart	56141	131.88	8	33	3

Table 2.2 Benchmarks for clustering two large EST data sets

Tissue	Number of sequences	Number of processors	Stacksize (MB)	TABLE_SIZE	Time (hours)
Whole-body index1.0	330000	128 (R10000, 195Mhz)	32	200000	35
Whole-body index2.0*	470293	128 (R12000, 300Mhz)			31

*Average length of each EST was 379 bases. The longest sequence was 6741 bases and the shortest sequence was 50 bases in length.

Table 2.3 Four versions of D2_CLUSTER tested with a lung data set (15,876 sequences) on 16 CPU (Origin2000 R12000, 300Mhz processors).

Version	Description of d2 modification	Time (seconds)	Time per d2 comparison (microseconds)
1	Current program	1351.5	10.7
1.1	Optimised compare.c	1211.9	9.6
1.2	Optimised compare.c, unpack.c and bin.c	981.8	7.8
1.25	Optimised compare.c, unpack.c and bin.c; store uncompressed bases as char	1145	9.1
1.3	Optimised compare.c; store uncompressed word values as int; remove unpack() and bin() from loops; rewrite revcomp word value operation	1025	8.1

Table 2.4 Additional optimisations of D2_CLUSTER tested on 128 CPU, R12000, 300Mhz Origin2000.

Number of processors	Time (seconds)	Speedup
4	3106.42	4
8	1586.13	7.84
16	809.44	15.36
32	422.1	29.44
48	303.67	40.92
64	230.45	53.92
96	154.66	80.36
126	123.5	100.61

Table 2.5 Formulas used in the optimisation of D2_CLUSTER

Implementation	**Formula**

Calculate the number of pieces needed to complete a clustering run such that more sequences are processed with successive pieces and the time taken remains constant	Ak = N(1-sqrt(P-k)/P)
	where P = number of pieces
	N= number of sequences
	The integral runs from Ak to Ak+1.

Memory estimate at the start of D2_CLUSTER: (i) if the database is replicated then the memory estimate is the sum of (a) and (c). (ii) if the database is not replicated then the memory estimate is the sum of (b) and (c) .	(a) memory_estimate = ncpus*(1.2+((float)(len_db+20*maxlen)/(1024.0*1024.0))) (b) memory_estimate = ncpus*(1.2+((float)(20*maxlen)/(1024.0*1024.0))) (c) memory_estimate += 10.0 + 20.0*num_seq/ (1024.0*1024.0)

2.6 References

Benson D.A., Boguski M.S., Lipman D.J., Ostell J., Ouetelle B.F., Rapp B.A., and Wheeler D.L. (1999) Release Notes for NCBI-GenBank Flat File Release 112.0. *Nucleic Acids Res.* **27:** 12-17.

Burke J., Davidson D. and Hide W. (1999) D2_CLUSTER: A validated method for clustering EST and full-length cDNA. *Genome Res.* **9:** 1135-1142.

Hide W., Burke J. and Davidson D. (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol*. **1:** 199-215.

Nemeth E., Snyder G., Seebass S. and Hein T.R. (1995) Controlling processes. In: Unix system adminstration handbook. Prentice-Hall Inc. Upper Saddle River, New Jersey 07458.

Torney D.C., Burks C., Davison D., and Sirotkin K.M. In *Computer and DNA*, Bell, G. I., Marr, T. Eds, pp109-125, Addison-Wesley: New York, 1990.

CHAPTER 3

Generation of a STACK Human Gene Index

Generation of a STACK Human Gene Index

Contents Page

List of Figures

List of Tables

Summary:

Expressed sequence tags (ESTs) remain an important resource for gene discovery, mapping and genome annotation. Given that EST data is fragmented and error-prone, a number of groups have attempted to add value to EST data by generating indices where ESTs are placed into categories such that each category represents a gene. Gene index formation through EST clustering is hindered by transcript variation, sequence and annotation errors, paralogous expressed genes and artifactual matches. The South African National Bioinformatics Institute has initiated the sequence tag alignment and consensus knowledgebase (STACK) aimed at detecting and visualising expressed transcript diversity in the context of developmental and pathological context. The initial implementation of STACK focused on tissue partitioned EST data arising from the limitation of EST clustering algorithms, at that time, to process large EST data sets. However, a comprehensive view of sequence variation in its proper context requires EST assemblies sampling genes expressed in multiple tissues and expression states (i.e., developmental and pathology).

An hierarchical approach for generating the STACK human gene index (i.e., whole-body index) was undertaken where 1,198,607 sequences from the human EST division of GenBank were partitioned into tissue categories and processed through the pipeline of tissue partitioning, masking, clustering, assembly, assembly analysis, clone linking and radiation hybrid mapping. The resulting consensus sequences for each tissue category were clustered followed by assembly of all constituent ESTs to construct the whole-body index sequences. A non-redundant data set of whole-body index sequences was made blast searchable (http://ziggy.sanbi.ac.za/alan/researchINDEX.html).

An update schema (STACK_ADD) is described for the addition of sequences to pre-existing EST clusters. The STACK_ADD protocol was implemented by adding mRNA and EST sequences extracted from UniGene (build #106) to an existing STACK whole-body index database.

3.1 Introduction

Expressed sequence tags (ESTs) remain an important resource for gene discovery (Matsubara and Okubo 1993; Vasmatzis et al., 1998), mapping (Schuler et al., 1996; Deloukas et al., 1998) and genome annotation (http://www.ensembl.org). Given that EST data is fragmented and error-prone, a number of groups have attempted to add value to EST data by generating indices where ESTs are placed into categories such that each category represents a gene (Schuler et al., 1996; Cariaso et al., 1999; Quackenbush et al., 2001; see chapter1). Gene index formation through EST clustering is hindered by transcript variation, sequence and annotation errors, paralogous expressed genes and artifactual matches (Jongeneel 2000). Most EST clustering methods employed in gene index projects rely on alignment-based algorithms to assign ESTs to specific clusters (Sutton et al., 1995), and are often intolerant of sequence errors (Liang et al., 2000). However, use of a non-alignment based methodology such as D2_CLUSTER has been shown to be tolerant of sequencing errors (Hide et al., 1994; Burke et al., 1999; Miller et al., 1999; Christoffels et al., 2001).

A non-alignment based approach tends to capture gene variants and contaminating sequences that could represent chimeric clones (Hide et a., 1997; Burke et al., 1999; Miller et al., 1999). The use of a loose clustering approach such as D2_CLUSTER also allows for the incorporation of sequences that would otherwise be discarded as error-laden. An accurate assembly of this error-prone sequence data require the use of additional error checking tools to extract high quality bases and the ability to partition different isoforms corresponding to the same gene. Shotgun assembly tools such as MSA_CONTIG and PHRAP were available at the start of this project and were considered for integration into our gene indexing system even though these tools were not designed initially for EST processing. For example, shotgun sequences usually have a high degree of identity and they are derived from a single clone source (Liang et al., 2000). ESTs, on the other hand, have numerous sequence irregularities and are derived from a variety of DNA sources and therefore contain more sequence variation than shotgun sequences. The variation in EST data cannot necessarily be assessed by shotgun assembly tools. STACK clustering was initially performed on the MasPar architecture, as outlined in chapter 2, followed by assembly using the MasPar implementation of MSA_CONTIG, a multiple sequence alignment program. The size of the EST clusters soon exceeded 60,000 sequences and placed memory constraints on MSA_CONTIG that affected the alignment quality. The MasPar machines were decommissioned a year after the initiation of the STACK project and the STACK development shifted to the SGI high performance architecture where MSA_CONTIG was replaced with PHRAP as the assembly tool. Over the past year there has been reports documenting the advantages of tools such as CAP3 over PHRAP (Liang et al., 2000). However, in the absence of other assembly tools, at that time, PHRAP was chosen as a replacement for MSA_CONTIG. STACK technology was developed with the ability to replace any external software as publicly available assembly tools undergo improvements (Christoffels et al., 2001). The optimisation of D2_CLUSTER for use on high performance architecture has allowed for the processing of large quantities of EST data needed to generate the STACK human gene index (see Chapter 2). In addition to consolidating all ESTs from dbEST human division, the STACK human gene index can provide added value (i.e., accelerate disease gene discovery) if the reconstructed transcripts can be positioned into context of the genetic mapping information.

Radiation hybrid mapping information

Integration of chromosomal mapping information with EST assemblies provide an enriched resource for disease gene discovery (Deloukas et al., 1998). Historically, mapping methodologies have centered around the use of sequence tag sites (STSs) as unique landmarks across the genome (Olson et al., 1989). EST-based landmarks entered the realm of feasibility when it was demonstrated that single-pass sequences provide suitable templates for the design of gene-based STSs (Wilcox et al., 1991). An international consortium was established to develop STSs from expressed sequence tags for mapping studies using primarily radiation hybrid (RH) techniques (Schuler 1997a). Recently, about 45,000 markers were placed onto radiation hybrid panels and formed the basis of Genemap'98 (Deloukas et al., 1998). The *in-silico* assignment of radiation hybrid markers to transcripts was achieved through the development of the electronic polymerase chain reaction tool (ePCR; Schuler 1997b) which was used in the integration of mapping information with the STACK gene indices.

3.2 Methods

All methods summarised below have been semi-automated as detailed in Appendix III.

3.2.1 Generation of the STACK whole-body index

3.2.1.1 Subpartitioning

The EST division of GenBank (Release 110) was downloaded from the National Center for Biotechnolgy and Information (http://www.ncbi.nlm.nih.gov). All human ESTs were extracted from GenBank formatted EST files and partitioned into tissue bins (Table 3.1). The tissue sets were organised arbitrarily according to organ system relationships. The "tissue_type" subkey of the "FEATURES" key is only provided sometimes with nonstandardised terms in the data field. As a result, the assignment of an output file name for each sequence is based on (1) FEATURES/tissue_type, (2) FEATURES/cell_type, (3) FEATURES/clone_lib or SOURCE/library, (4) FEATURES/chromosome or (5) FEATURES/map. These rules were incorporated into a script that provides automation of the tissue paritioning step (see Appendix II). The resulting sequence files were placed directly into a hierachy (Table 3.1). All sequences that were annotated as derived from a disease-related tissue were duplicated and placed in a single set to facilitate the exploration of differentially expressed genes in the context of disease.

3.2.1.2 Masking

The clustering procedure is intended to group together those sequences that share identical regions. It is therefore necessary to ensure that ESTs submitted for clustering are free of artifactual sequence identical to the expressed transcript under study. All input sequences were subjected to masking against human repeat sequences using RepBase (Jurka 1998), common vector sequences (ftp://ncbi.nlm.nih.gov/repository/vector), and potentially contaminant species such as rodent, human mitochondrial and ribosomal DNA. Sequences were masked using CROSS_MATCH (P.Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/swat.htm) and later replaced with RepeatMasker (Smit and Green 1999).

3.2.1.3 Clustering

Clustering of the GenBank human EST division (Release 110) was achieved through a hierachical approach whereby ESTs for each tissue data set were clustered and assembled and followed by clustering of all tissue consensus sequences. The tissue data were transferred using ftp to NCSA (64 CPU Origin2000, 195Mhz) and SGI (64 CPU Origin2000, R12000 300 Mhz) for clustering using D2_CLUSTER (Torney et al., 1990; Burke et al., 1999). The clustering of all tissue-level data was used to optimise the D2_CLUSTER code (see chapter 2; Carpenter et al. in prep). The successful clustering and assembly of all tissue data sets was followed by the clustering of all tissue consensus sequences on a 126 CPU, R12000 300MHz Origin2000. Two sequences or their reverse complement fall into the same cluster if they share word multiplicities (where word length=6) of at least 96% identity in a 150-base window (see Appendix III for command line usage).

3.2.1.4 Alignment

Initial STACK development relied on the MasPar implementation of MSA_CONTIG for its cluster alignments. However, memory constraints on data sets in excess of 60000 sequences prompted the use of PHRAP for cluster alignments. At the level of the whole-body index PHRAP assembly, the tissue consensus sequences within a cluster are decomposed into their constituent ESTs. Clusters generated by D2_CLUSTER that were fragmented into subclusters during the PHRAP assembly, can be identified by their clusterID . For example, clusters 3_1 and 3_2 refer to “cluster 3” generated by D2_CLUSTER that was subsequently fragmented by PHRAP into two subclusters namely “3_1” and “3_2. The accuracy of the EST orientation description (as captured in the GenBank record), the cluster assembly and the alignment cannot be guaranteed. PHRAP generates sequence alignments but does not provide any subclusters to distinguish alternative splice or other scientifically interesting data from alignment problems induced by low sequence quality or experimental artifacts. To leverage the availability of loose clusters, the alignments have to undergo additional processing. CRAW and CONTIGPROC were developed to address postclustering and assembly artifacts (Miller et al., 1999; Christoffels et al., 2001).

3.2.1.5 Assembly analysis

CRAW is used to maximize consensus length, partition subassemblies and provide a simple means to view clusters (Burke et al., 1998). CRAW checks the agreement along the columns of a multiple sequence alignment and uses this information to sort related sequences within each cluster and generates a consensus sequence for each subcluster. A subcluster is generated if 50% or more of a 100-base window differs from the remaining sequences of a cluster, excluding the initial 100 bases of any read. The approach depends fundamentally on the alignment quality of each assembly generated by the assembly tool. For example, a poor alignment will yield erroneous sub-clusters, and too low a gap penality may yield too many columns in agreement and thus not create subclusters where they would be appropriate.

3.2.1.6 Consensus Partitioning

CONTIGPROC independently partitions the aligned sequences generated from the CRAW consensus sequences then ranks the consensus sequences according to the number of assigned sequences and number of called bases. The best ranking consensus sequence is taken as the primary representative of a cluster, whereas the remaining consensus sequences are logged with the best consensus sequence in the Genetic data environment (GDE, Smith et al., 1994) file format. The 5' or 3' orientation of each cluster is determined by a vote of the individual EST annotations and all output consensus sequences are arranged to read 5' to 3'. Low-quality regions defined as 2 N's followed by at least thirteen IUPAC codes with four or less clear A, T, C or G calls are replaced by a single run of 10 N's. A high-confidence subset called SANIGENE, consisting of only those consensus regions representing at least two reads, is also generated from the multi-sequence clusters.

3.2.1.7 Clonelinking

Each EST from GenBank is searched for clone information to trace the transcripts corresponding to the same gene. The clone information is used to extend the length of the cluster consensus sequences by joining clusters containing ESTs with shared cloneIDs. The presence of inaccurate cloneID names in EST records can cause false clone links between clusters (Aaronson et al., 1996). A stringent clone-linking criterium was used for the whole-body index to avoid the joining of false links between clusters (Figure 3.3). Clone links were accepted if two EST pairs joined two clusters where each EST pair had a different cloneID (Miller et al., 1999).

3.2.1.8 Clone-library information

The organ categories used to describe the STACK tissues included a range of different tissues. These tissue descriptions could not accurately describe the whole-body index because there was an integration of ESTs originating from different clone libraries. A library field was added to the FASTA header line for each whole-body entry and the information was extracted from the original EST record via a lookup table (Figure 3.6).

3.2.1.9 Incorporating radiation hybrid mapping data

The whole-body index consensus sequences were assigned radiation hybrid map positions using the e-PCR program developed by Schuler (1997b) (ftp://ncbi.nlm.nih.gov/pub/schuler/e-PCR) which uses published primer sequences and PCR product size (ftp://ncbi.nlm.nih.gov/repository/genemap/) to electronically map markers onto the consensus sequences.

3.2.2 Development of an updating schema

The STACK_ADD phase (Figure 3.1, blue arrows) is a protocol for the addition of sequences to an existing gene index without the need to re-cluster the data present in the existing database. The process involves an initial pairwise comparison between the exisiting database and new sequence data, reprocessing of clusters that have expanded due to incorporation of new data and the clustering of all sequences that do not match any existing indices. The STACK_ADD schema was implemented with the addition of UniGene's consensus sequences (mRNA and ESTs) to the STACK whole-body index. The UniGene consensus sequences refer to the representative sequence for each cluster that usually reflects the longest sequence in the cluster.

All representative sequences for UniGene (build #106) were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/repository/unigene/Hs.seq.uniq.Z) and cleaned to remove untrimmed vector, ribosomal, mitochondrial and low quality sequences using CROSS_MATCH. Cleaned sequences were compared to the whole-body index using CROSS_MATCH. All matching entries were traced using a method called transitive closure where a sequence A can be in the same cluster as sequence C if both sequence A and C match sequence B. The expanded clusters were collapsed such that the matching gene indices were replaced with their constituent ESTs. Inspection of these clusters led to the detection of erroneous cluster memberships due to inclusion of repetitive sequences (see section 3.2.3). UniGene sequences that did not find matching gene indices were clustered using D2_CLUSTER. The clusters generated by CROSS_MATCH and those generated by D2_CLUSTER were assigned new clusterIDs so that each clusterID would remain unique in the database. These clusters were assembled using PHRAP and subjected to error analysis using CONTIGPROC. The expanded clusters were represented at different levels in the database including initial PHRAP alignments, CRAW reports, consensus sequences and final cluster alignments. These clusters were removed from the existing database at all levels of representation to avoid the duplication of cluster records. A total of 145053 whole-body index sequences did not match any UniGene data and were compared all-versus-all to identify any redundancy in the existing database. As few as 21097 sequences found matching entries and were reassembled into 10,063 clusters.

The clusters derived from merging UniGene and STACK sequences and the unique UniGene clusters were appended to the existing whole-body index clusters and all existing clonelinking data was removed. The memory requirements for clonelinking during the processing of the whole-body index exceeded the 1gigabyte memory limit on the Origin2000 (the machine on which this processing was executed) because all clone links and EST orientations were stored in memory. The code was modified so that all data relating to the clonelinking phase could be accessed from the disk. Clone library information was appended to a consensus sequence record to reflect the origin of all ESTs contributing to the STACK multi-sequence cluster or singleton. Chromosomal locations were electronically mapped onto the consensus sequences using e-PCR. The entire process outlined above has been semi-automated.

3.2.2.1 Sequence contamination

The identification of a cluster containing 18000 sequences after the pairwise comparison between the whole-body index and UniGene data suggested the presence of contaminating sequences. The 18000 sequences were masked using RepeatMasker and then reassembled. Contaminants identified by RepeatMasker included ALU repeats and low complexity regions such as consecutive copies of A's and T's. The effect of masking with RepeatMasker was seen when the 18000 sequences were partitioned into 934 clusters. The removal of repetitive sequences and low complexity regions using RepeatMasker led to the screening of all sequences comprising the whole-body index consensus sequences and the UniGene sequences. The cleaned data (whole-body index and UniGene sequences) were then passed to the start of the STACK_ADD pipeline where the whole-body index sequences were compared against the UniGene sequences and the resulting clusters assembled.

3.2.3. Annotation

A nonredundant protein data set was downloaded from NCBI (ftp://ncbi.nlm.nih.gov/blast/db/nr.Z) on the 7th July 2000. All singletons and multi-sequence cluster consensus sequences were searched locally against the non-redundant protein data set using BLASTX. Sequences that matched with an E-value of $1x10^{-10}$ or lower were retained for putative functional assignment. The protein annotations were incorporated into the STACK whole-body index web search engine (http://ziggy.sanbi.ac.za/alan/researchINDEX.html).

3.2.4 Web-based searching

The whole-body index data is represented by multi-sequence clusters, singletons and clonelinked sequences. All clusters that did not contribute to the clonelinked entries were concatenated to the clonelinked records to provide a non-redundant data set for BLAST searching (http://ziggy.sanbi.ac.za/alan/researchINDEX.html). The UniGene cluster descriptions (ftp://ftp.ncbi.nlm.nih.gov/repository/unigene/Hs.data) were parsed to create a lookup table of UniGene clusterIDs and EST assigments so that all whole-body index BLAST hits could be cross-referenced to UniGene. A perl extraction tool was added to the BLAST search engine so that detailed information could be retrieved across the internet for each matching STACK cluster. All whole-body index clusters that are retrieved from the BLAST search engine are searched on-the-fly for any protein matches using the BLASTX results that were described in section 3.2.3.

3.3 Results

Tissue-level and whole-body index clustering

A total of 1,198,607 sequences were downloaded from the EST division of GenBank and cleaned as described in the section 3.2.1. The 50 base limit for the number of informative bases were not met by 32,240 (2.7%) sequences after masking, and these were removed from the input data. A total of 1,166,367 sequences were partitioned into 334,822 singletons and 143,885 multi-sequence clusters (Table 3.2). CloneID tracking led to the creation of 68,701 linked sets which represents 50% of the total cluster consensus sequences and 30% of the total singletons. Complete results are given in Tables 2 (clustering), 3 (linking) and 4 (errors).

The tissue-level consensus sequences were clustered on a 126 CPU 12000 300Mhz Origin2000 in 31 hours. A total of 470,293 consensus sequences were partitioned into 162,439 singletons and 69,158 multi-sequence clusters (Table 3.5). A fraction (5%) of the multi-sequence clusters (whole-body index2.30) generated by the clustering step were fragmented during assembly by PHRAP such that D2_CLUSTER-generated clusters were subdivided into multiple subclusters. Clone linked entries reduced dramatically from 30,665 to 8638 when the clone link criteria were increased to two EST pairs sharing two independent cloneIDs (Table 3.5). The effect of false clone links between two clusters due to incorrect annotation is illustrated in Figure 3.3.

mRNA incorporation into the whole-body index

The STACK_ADD protocol was implemented to add mRNA and EST sequences extracted from UniGene (build #106) to an existing STACK whole-body index. A total of 92,182 sequences were downloaded from UniGene and compared with the whole-body index (Figure 3.2). The pairwise comparison between the whole-body index and UniGene data merged a total of 96,618 STACK sequences (multisequence cluster and singletons) and 34,779 UniGene sequences into 50,201 clusters, also termed UniSTACK clusters of which 17853 (36%) were contaminated with repetitive elements. The presence of repetitive elements in the UniSTACK clusters was exposed through the capture of one cluster containing 18,000 sequences. The whole-body index clusters that did not match any UniGene sequences were found to contain 21,097 clusters (14.5%) that were redundant. The redundant clusters were collapsed into 10,063 clusters. UniGene sequences that did not match any whole-body index clusters were clustered using D2_CLUSTER after masking for repeats. A total of 6589 sequences were removed by the masking step and the remainder 51,085 sequences were partitioned into 50352 singletons and 310 multi-sequence clusters.

Sequence contamination

The UniGene data was added to the whole-body index by a pairwise comparison and resulted in the generation of 50,201 clusters that represent merged whole-body index and UniGene sequences. All matching whole-body index sequences and UniGene sequences were screened for additional repeats which resulted in the trimming of 28,513/96,618 (29.5%) of the whole-body index sequences. These contaminating sequences contributed to 17,853 clusters (14.5%) of the UniSTACK clusters. These clusters were collapsed to their original ESTs and mRNA and masked using RepeatMasker. The additional masking identified a range of repeat sequences (Figure 3.5). The cleaned data was clustered using D2_CLUSTER.

Capture of alternate gene expression forms

Alternatively spliced transcripts represent important biological information that has to be handled appropriately within a cluster assembly. Two or more alternatively spliced transcript isoforms may contain regions of identity as well as disparate regions and so require specialised tools to capture the regions of dissimilarity. STACK incorporates CRAW as a post assembly step to clustering and alignment in order to facilitate discrimination between distinct gene isoforms (Burke et al., 1998). Transcript variants are partitioned into sub-clusters that allow for simultaneous viewing of inconsistencies within a cluster. An example is fibulin (expressed in brain, parathyroid tumor, placenta, fibroblast, pancreas, heart, lung, testis, skin tumor) which exists as four or more isoforms(A-D) and each is clearly partitioned within the STACK whole-body index cluster 133232_3 (Figure 3.4). Fibulin's B isoform (X53742) and its corresponding ESTs are displayed as a stretch of 1's in the ASCII representation of each sequence (blue box Figure 3.4). Sequences corresponding to isoform C have been partitioned into a sub-cluster displayed as a string of 2's (red box Figure 3.4).

Assessing the STACK whole-body index consensus sequence fidelity by e-PCR

Of the 52,825 EST-based markers placed on the radiation hybrid maps, 25793 markers were assigned to one or more consensus sequences. STACK clusters are defined as redundant if there are multiple clusters that potentially represent the same gene. In total, 26,944 map assignments were made, suggesting a redundancy in the whole-body data set of 1.04-fold (26944/25793), i.e., multiple assignments to the same marker could reflect fragments of the same gene that could not be clustered because of insufficient overlapping sequences. Mapping inconsistencies, i.e., one sequence with different chromosome locations, accounted for 135 (0.5%) clusters.

Sequence Annotation

The construction of a consensus sequence from a set of clustered ESTs has the advantage that the sequence is longer than its component ESTs. This increase in sequence length facilitates functional assignment, transcript mapping and genomic annotation. The consensus sequences were assigned a function based on significant (E value = 1e-10) similarity to known SWISSPROT records and served as an indirect measure for assessing the consensus sequences. An in depth assessment of the consensus sequences was carried out on known genomic sequences (see chapter 4).

A total of 54354/66188 (82.1%) whole-body index sequences did not show any significant matches with database entries. A putative function could be assigned to 11834 (17.8%) whole-body index sequences based on a significant match to a SWISSPROT entry. The functional assignments provided evidence to assess the accuracy of 272 clone linked entries. A total of 232/272 (85%) linked clusters showed the correct SWISSPROT annotation for all constituent multi-sequence clusters. False clone links were generated for 21/272 (8%) linked clusters. A total of 19/272 (7%) linked clusters demonstrated clone links between SWISSPROT annotated clusters and clusters with hypothetical annotations.

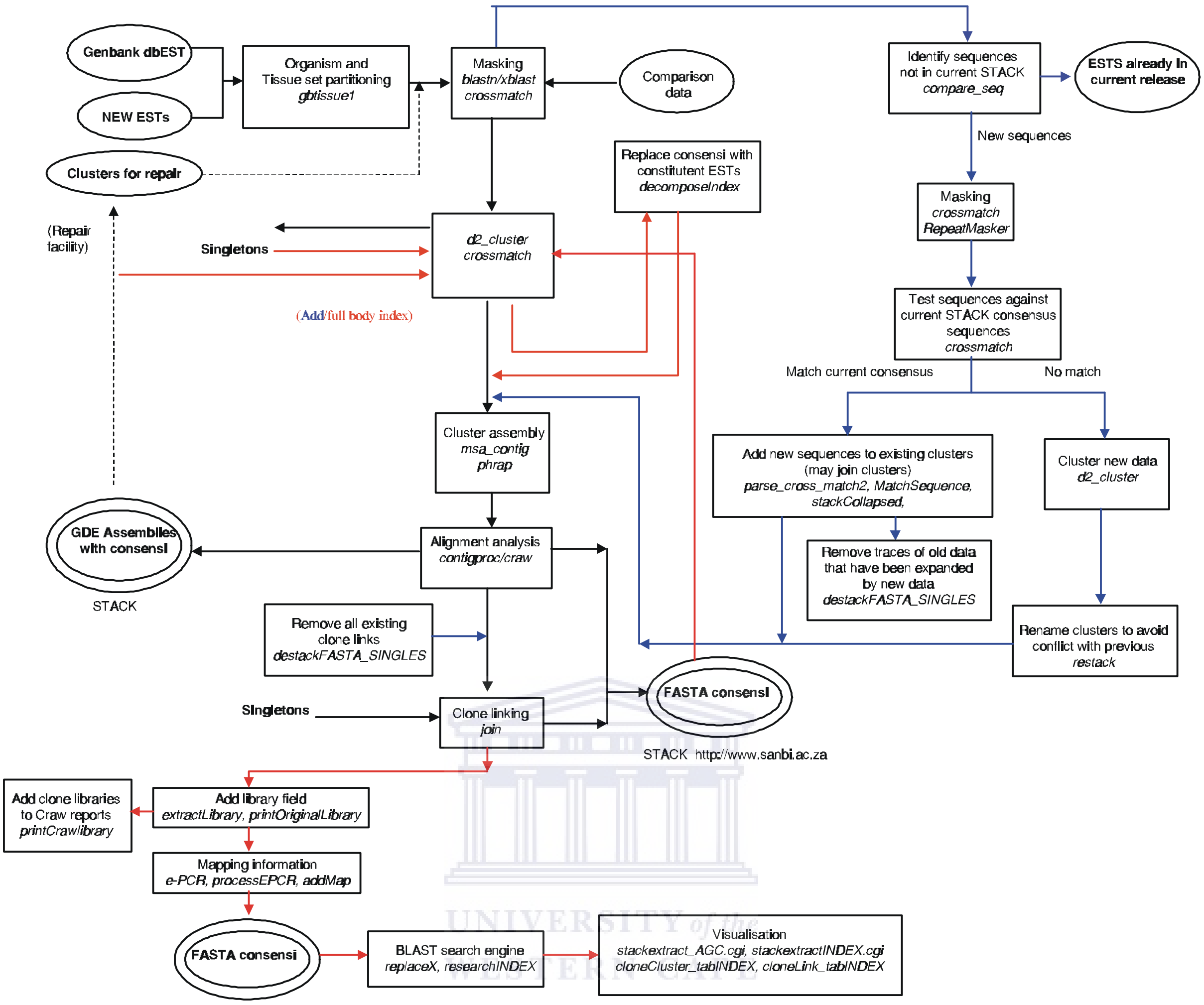

Genbank dbEST
NEW ESTs
Organism and Tissue set partitioning gbtissue1
Masking blastn/xblast crossmatch
Comparison data
Identify sequences not in current STACK compare_seq
ESTS already in current release
New sequences
Clusters for repair
Replace consensi with constitutent ESTs decomposeIndex
Masking crossmatch RepeatMasker
(Repair facility)
Singletons
d2_cluster crossmatch
Test sequences against current STACK consensus sequences crossmatch
(Add/full body index)
Match current consensus
No match
Cluster assembly msa_contig phrap
Add new sequences to existing clusters (may join clusters) parse_cross_match2, MatchSequence, stackCollapsed,
Cluster new data d2_cluster
GDE Assemblies with consensi
STACK
Alignment analysis contigproc/craw
Remove traces of old data that have been expanded by new data destackFASTA_SINGLES
Remove all existing clone links destackFASTA_SINGLES
Rename clusters to avoid conflict with previous restack
FASTA consensi
Singletons
Clone linking join
STACK http://www.sanbi.ac.za
Add clone libraries to Craw reports printCrawlibrary
Add library field extractLibrary, printOriginalLibrary
Mapping information e-PCR, processEPCR, addMap
FASTA consensi
BLAST search engine replaceX, researchINDEX
Visualisation stackextract_AGC.cgi, stackextractINDEX.cgi cloneCluster_tablNDEX, cloneLink_tablNDEX

Figure 3.1 STACK Overview

STACK processing overview. Inputs are shown in single-line ellipses, outputs in double-line ellipses. STACK first iteration, ADD, INDEX phases, and the repair facility are indicated by black, blue, red and black-dotted arrows respectively. The red and blue arrows indicate the contribution made by this thesis. In the first iteration (black arrows), human sequences from GenBank dbEST are partitioned into manageable, tissue-related sets. Common vector and repeat sequences are masked, and the resulting entries are subjected to loose clustering by d2_cluster. Clusters of related sequences are assembled by PHRAP, and their alignments are analyzed by CRAW. GDE format assembly data are output, and CONTIGPROC selects appropriate consensus and subconsensus sequences. Available clone-ID information is used to identify clone-linked clusters, after which full-length, joined consensus sequences are output in FASTA (Pearson) format. ADD (blue arrows) incorporates new sequence data by comparison to existing STACK consensus sequences. Existing clusters that are identified as members of the same group are reassembled and submitted as a single set to d2_cluster during the whole-body index phase (red arrows). The resulting index clusters are then expanded prior to assembly by replacing each consensus with the sequences that contribute to it. A library field is added to the header line of the sequence to reflect the origins of all the constituent ESTs. The EST accessions within the CRAW report are appended with their clone library names. Radiation hybrid markers are added to the consensus sequences using ePCR. The final consensus sequences are added to the blast search engine. Visualisation of the BLAST results include cluster consensus sequences, constituent EST sequencse and updated hyperlinks to UniGene.

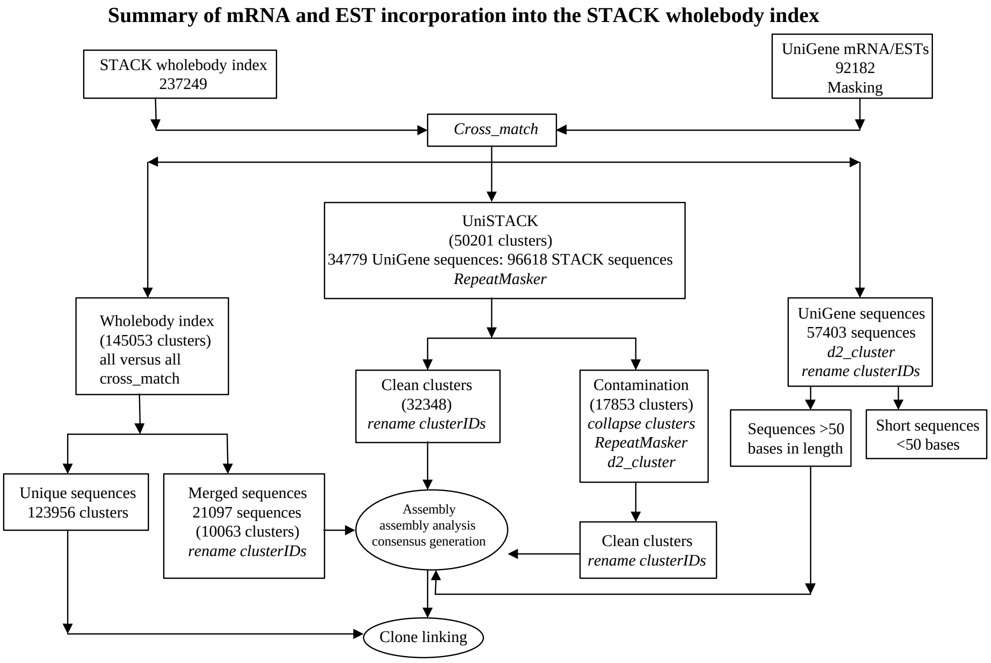


Figure 3.2 Summary of mRNA sequence incorporation into the whole-body index
UniGene's representative sequences were masked and compared to the STACK whole-body index sequences using cross_match. STACK sequences that did not overlap with UniGene sequences were searched for redundancy using cross_match. A total of 21,097 consensus sequences collapsed to 10,063 clusters that had to be assembled prior to clone linking. 17,853 UniSTACK (UniGene + STACK sequences) clusters contained contaminating sequences (i.e., repeats). ESTs corresponding to the contaminant clusters were masked for repeats using RepeatMasker prior to re-clustering. All clusterIDs were renamed to ensure their uniqueness. UniSTACK clusters that were free of contamination were assembled using PHRAP prior to clone linking. UniGene clusters that do not match any STACK sequences were clustered using d2_cluster prior to assembly and clonelinking.

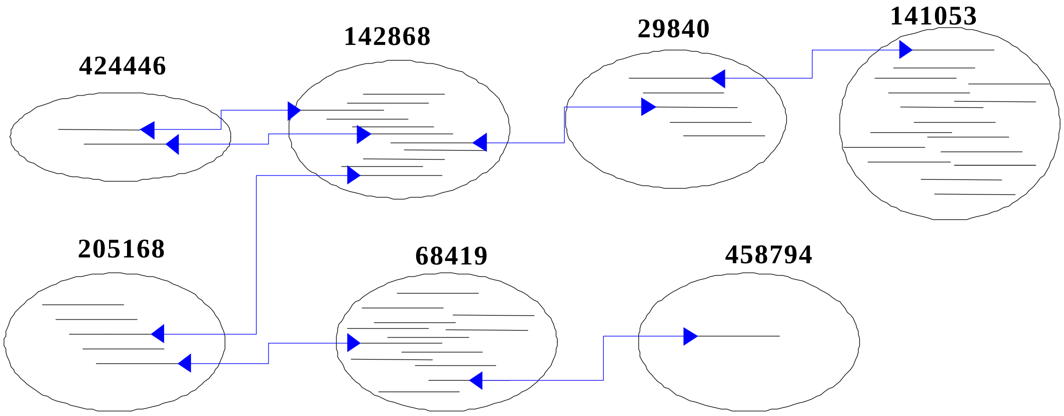
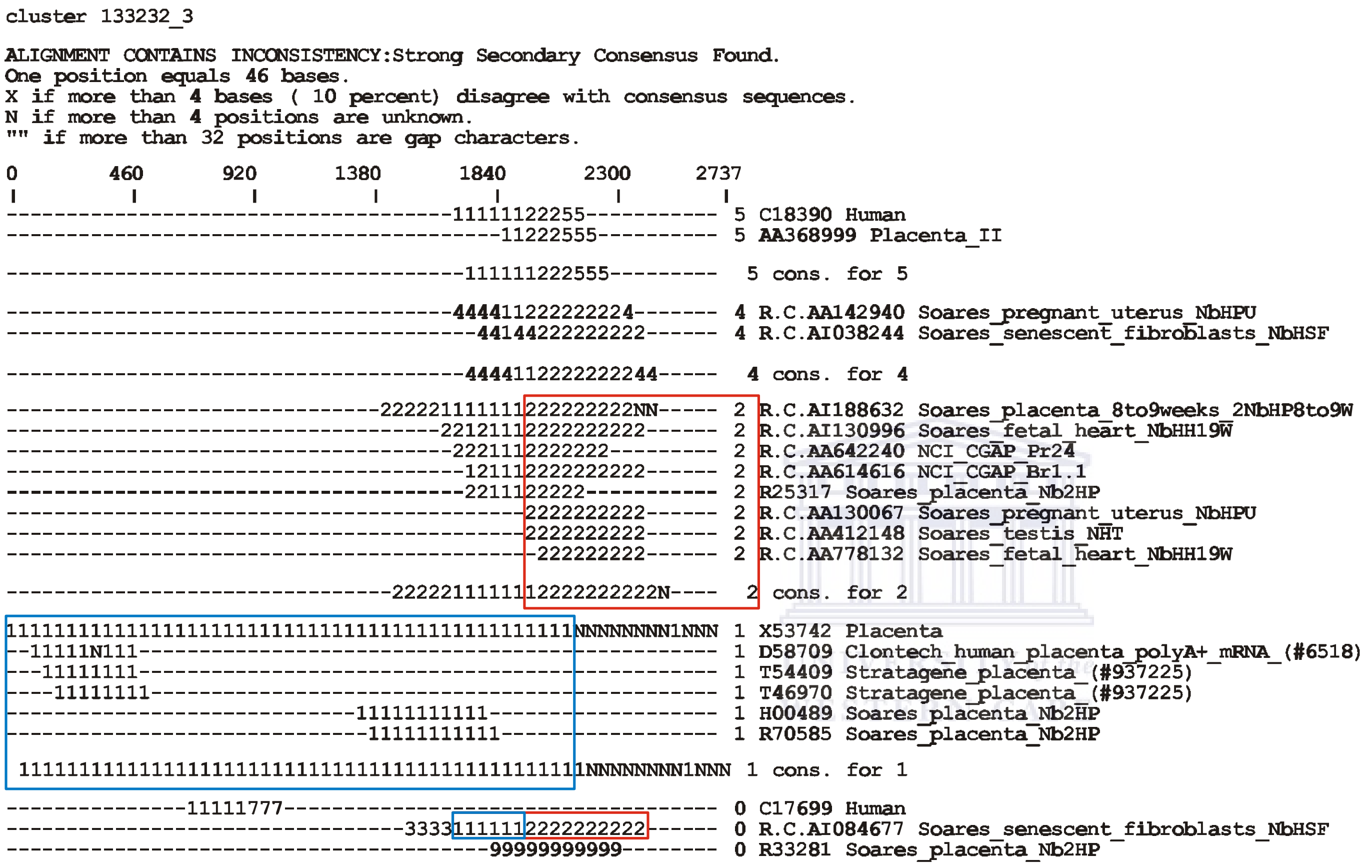


Figure 3.3 Diagram illustrating the erroneous clonelink relationships between clusters
Clusters that share at least one cloneID were joined during the clonelinking step to create extended entries as indicated by STACK whole-body entry index2607 (release 2.31). STACK clusterIDs (bold) are indicated above each cluster of ESTs (oval shapes). The ESTs are depicted as horizontal lines and each clonelink between two ESTs is shown by a blue double-headed arrow. Clusters **424446** and **142868** represent portions of the myosin-binding protein (NM_004997.1 e-value=0.0). Cluster **29840** represent an anti-oncogene on chromosome 8p21.3-p22 (AK001608 e-value=0.0) and cluster **141053** represent a fragment of the REC mRNA (NM_016353.1, e-value=0.0). Clusters 205168, 68419 and 458794 have no identity to any sequence in the non-redundant database (GenBank 27July2000). Accurate clone-links have been generated between clusters 424446 and 142868 that share at least two EST pairs with different cloneIDs.



```
cluster 133232_3

ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.
One position equals 46 bases.
X if more than 4 bases ( 10 percent) disagree with consensus sequences.
N if more than 4 positions are unknown.
"" if more than 32 positions are gap characters.

0        460       920       1380      1840      2300      2737
|         |         |         |         |         |         |
--------------------------------------11111122255----------- 5 C18390 Human
------------------------------------------11222555---------- 5 AA368999 Placenta_II

---------------------------------------111111222555---------  5 cons. for 5

--------------------------------------44441122222224-------- 4 R.C.AA142940 Soares_pregnant_uterus_NbHPU
----------------------------------------44144222222222------ 4 R.C.AI038244 Soares_senescent_fibroblasts_NbHSF

---------------------------------------444411222222244------  4 cons. for 4

--------------------------------222221111111222222222NN----- 2 R.C.AI188632 Soares_placenta_8to9weeks_2NbHP8to9W
-------------------------------------22121112222222222------ 2 R.C.AI130996 Soares_fetal_heart_NbHH19W
--------------------------------------2221112222222--------- 2 R.C.AA642240 NCI_CGAP_Pr24
---------------------------------------121112222222222------ 2 R.C.AA614616 NCI_CGAP_Br1.1
---------------------------------------2211122222----------- 2 R25317 Soares_placenta_Nb2HP
--------------------------------------------2222222222------ 2 R.C.AA130067 Soares_pregnant_uterus_NbHPU
--------------------------------------------2222222222------ 2 R.C.AA412148 Soares_testis_NHT
---------------------------------------------222222222------ 2 R.C.AA778132 Soares_fetal_heart_NbHH19W

---------------------------------2222211111112222222222N----  2 cons. for 2

111111111111111111111111111111111111111111111111NNNNNNNN1NNN 1 X53742 Placenta
--11111N111------------------------------------------------- 1 D58709 Clontech_human_placenta_polyA+_mRNA_(#6518)
---11111111------------------------------------------------- 1 T54409 Stratagene_placenta_(#937225)
----11111111------------------------------------------------ 1 T46970 Stratagene_placenta_(#937225)
------------------------------11111111111------------------- 1 H00489 Soares_placenta_Nb2HP
-------------------------------11111111111------------------ 1 R70585 Soares_placenta_Nb2HP

 111111111111111111111111111111111111111111111111NNNNNNNN1NNN 1 cons. for 1

---------------11111777------------------------------------- 0 C17699 Human
----------------------------------33331111112222222222------ 0 R.C.AI084677 Soares_senescent_fibroblasts_NbHSF
-----------------------------------------99999999999-------- 0 R33281 Soares_placenta_Nb2HP
```

Figure 3.4 CRAW output for a whole-body index cluster displaying alternate gene isoforms of the fibulin gene. The blue box indicates the region capturing the fibulin-1B isoform whereas sequences capturing fibulin-1C are surrounded by a red box.

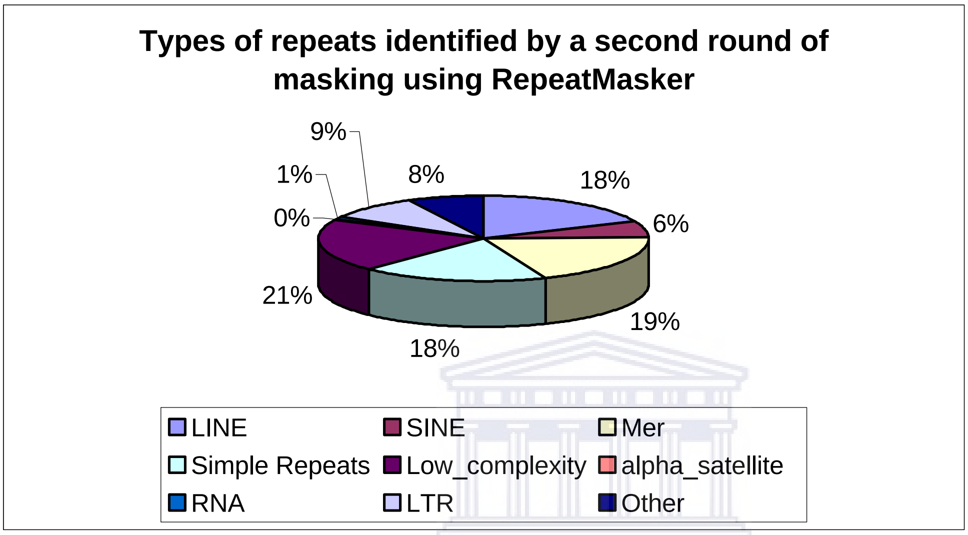


Figure 3.5 Range of repeats found with RepeatMasker.
A number of repeats were not found when sequences were masked using cross_match. A second round of masking using RepeatMasker identified additional repeat regions that caused incorrect cluster assignments.

```
>100037-0-index-001-2000-2.35 COVERAGE: 0.9924 OTHER_CONSENSI: 0    ASSIGNED: AA348809 R77952 R85973 AA973910 LIBRARY: Right_hemisphere Soares_placenta_Nb2HP Soares_retina_N2b4HR NCI_CGAP_Lu5 MAP: 14
CCCTACAAGnGGGGCAACCTTAnGATTACCTATTATTGGGGCCTTAAGAGATTTGAGAAGTTGGGGAGCTAACCA
AATTAAGACCCCTAAnGGTATATGTCCCTTAACCCTGTAAACAGGAGTTTATTTTGTCAAGCCTATTTTTCCCCA
TCCCTATTCATTACCTACCTAAAAAATATTGCCTTAAAAACATTTTCTTCCTTCGTGAGGCTTCTTAGAATGTTA
AATTTACCTTCTAAAAATTATACACTAAGTTATTTTACAGGAAAACAGCTTCTATAGGATTAATGTAATATATAT
ATGCAAAGGTCAAATGAAATATTTTTGTGGATGGTAAGAAAAATTTCAACTTACATTTTTGCAACTTCTTTAACG
ATATCACGGCAGGTCATTTCTTTCATCTATAGAAAATAAAATGTATACTGTTCATCAGAAGAACATTTTCCACTT
GTGTAATAACTATTTTCACTTTTATACTCAGATATAAAACCAAGGAAAATAACCTAAAGTCTGAAAAAGACCAGA
ATCGAAGTTTCCTGATTCATATTTTAATGTTTTGGAAATTTATAGACCGGGGTGGGTGCAGTGGCTTGTGC
```

Figure 3.6 FASTA formatted multi-sequence cluster consensus sequence.
The headerline documents the unique stackID, the ESTs used to generate the cluster, the clone libraries and the chromosomal position (radiation hybrid mapped) if it is known.

Table 3.1 List of arbitrary tissue divisions used by STACK

dbEST 101598 Homo Sapiens tissue partitioning		
Abitrary tissue partitions	Substituent tissues types	Total ESTs
Adipose	Brown, white	2376
Brain	Frontal lobe, cerebrum, cerebellum, cortex,	177719
Cochlea	Fetal cochlea	4304
Connective	Bone, skin, synovial membrane	40753
Digestive	Stomach, colon, gall bladder	51032
Disease	Duplicates of ESTs annotated as tumors	114496
Eye	Retina, cornea, ocular	28514
Genomic	Specified chromosomes	101986
Glands	Breast, endocrine	112346
Heart	Fetal heart, aorta	69830
Hemato-lymphatic	Blood, kidney, liver-spleen	255565
Lung	Trachea, larynx, lung	70259
Muscle	Leg, pectoral	16237
Olfactory	Olfactory epithelium	2600
Other	Monocytes, mononuclear cells	25925
Reproductive	Ovary, testis, uterus	239161

Sequences were partitioned over an arbitrarily defined tissue hierarchy designed to group physically related tissues and remain within constraints of computational resources. Genomic tissue is a set of ESTs labeled only as having a genomic region of hybridisation without a tissue source. The set of duplicate copies of disease-related sequences is loosely referred to as a tissue.

Table 3.2 STACK tissue-level clustering and alignment analysis results

Tissue	Singletons		Multi-sequence clusters			Small sequences		Total sequences
	Total	% total sequences	Total	Sequence in MSC	%Total sequences (clustering efficiency)	Total	%Total sequences	
Adipose	1693	71	181	572	24	111	5	2376
Brain	42245	24	22848	130573	73	4458	3	177719
Cochlea	1973	46	710	2213	51	118	3	4304
Connective	12652	31	4646	26210	64	876	2	40753
Digestive	17398	34	6734	32124	63	1481	3	51032
Disease	29139	25	12513	79433	69	4056	4	114496
Eye	13867	49	3448	12933	45	1388	5	28514
Genomic	38481	38	16314	72066	71	4457	4	101986
Gland	25836	23	12307	62176	55	1672	1	112346
Heart	20782	30	8341	45795	66	217	0.3	69830
Hemato-lymph	51654	20	17378	113147	44	2582	1	255565
Lung	20129	29	8554	47151	67	2726	4	70259
Muscle	4534	28	1183	8792	54	1037	6	16237
Olfactory	1478	56	248	830	32	283	11	2600
Other	9392	36	4315	15663	60	575	2	25925
Reproductive	43569	18	24165	188088	79	6321	3	239161
Totals	334822	26	143885	837766	64	32240	2	1313103

The total sequences in each tissue set are partitioned by d2_cluster into unique sequences (singletons) and clusters containing multiple related sequences, whereas sequences of <50 bases are excluded from clustering (small sequences).

Table 3.3 STACK tissue-level clone linking results

Tissue	**STACK**						**SANIGENE**			
	Total linked sets	Linked consensi and singles	Non-linked consensi	% total consensi	Non-linked singles	% total singles	Total linked sets	Linked consensi	Non-linked consensi	% total consensi
Adipose	0	0	181	100	1693	100	0	0	181	100
Brain	13157	35123	4282	19	25688	61	52909	11490	11358	50
Connective	1561	3433	3266	71	10599	84	10	20	690	97
Cochlea	323	666	601	85	1416	72	86	183	4462	96
Digestive	2165	4915	4761	71	14456	83	188	384	6350	94
Disease	6106	14103	6623	53	20926	72	725	1477	11036	89
Eye	3988	8616	1027	30	7672	55	699	1424	2024	59
Genomic	4168	9131	9997	74	29221	84	665	1358	10949	89
Gland	5056	11275	7242	59	19624	76	323	655	14070	96
Heart	3630	7937	5462	65	15724	76	295	594	7747	93
Hemato-lymph	10952	25388	9648	56	33996	66	1432	2958	14419	83
Lung	4222	9640	5142	60	13901	69	339	694	7860	92
Muscle	1164	2694	622	53	2400	53	52	112	1071	91
Olfactory	458	944	138	56	644	44	21	42	206	83
Other	3700	8901	929	22	3877	41	656	1346	2969	69
Reproductive	8051	26475	11268	47	29991	69	1854	3916	20249	84
Totals:	68701	169241	71189	50	231828	70	12635	26653	115641	81

Clone-ID annotations are grouped for all ESTs in a cluster, after which clusters or singletons containing matching cloneIDs are added to a linked set. The process is continued until no additional cloneID partners can be found. Each linked set may therefore contain singleton sequences and a cluster consensus; hence, the linking success rate is expressed in terms of the fraction of consensus and singletons that remain non-linked

Table 3.4 STACK tissue-level error analysis.

Tissue	Single cons.clusters	% of total clusters	Multi-cons clusters	% of the total clusters	Total only singletons	% total clusters	3’/5’ disagreement	% of total clusters
Adipose	173	96	5	3	3	2	23	13
Brain	19933	87	1850	8	296	1	2552	11
Cochlea	689	97	13	2	4	6	18	3
Connective	4098	88	316	7	93	2	358	8
Digestive	6089	90	370	6	82	1	493	7
Disease	10845	87	989	8	198	1	2589	21
Eye	2799	81	288	8	229	7	303	9
Genomic	14924	91	792	5	177	1	2550	16
Gland	10843	88	820	7	237	0.2	1096	9
Heart	7341	88	622	7	104	1	699	8
Hemato-lymph	14639	84	1774	10	271	2	2731	16
Lung	7483	87	667	8	137	2	1828	21
Muscle	1084	92	64	5	12	1	67	6
Olfactory	238	96	7	3	2	1	4	2
Other	3675	85	285	67	184	4	172	4
Reproductive	19178	79	3196	13	533	2	3373	14
Totals:	124031	86	12058	8	2562	2	18856	13

CRAW analyzes cluster alignments generated by PHRAP and partitions consistent ESTs into subclusters based on agreement with other sequences. The ideal result is a single consensus cluster, accounting for 86% of the STACK output, while the remaining clusters may contain multiple sequence subclusters (resulting a multiconsensus cluster), a primary consensus with one or more singleton sequences (data not shown), singleton ESTs according to the CRAW parameters. STACK clusters are generated by word identity counts and their read direction determined by majority vote of the annotations of constituent ESTs; clusters for which this vote is not unanimous are noted in the right-most two columns.

Table 3.5 Cluster analysis and clone-link information for two releases of the whole-body index.
The steep drop in clone-linked entries are due to stringent criteria that were implemented in the recent release of the whole-body index. Two EST pairs and two different cloneIDs were required to create a clone-link.

Whole-body index version	Singletons	Multi-sequence clusters (MSC)	Number of clusters that were fragmented during the PHRAP assembly. (%)	STACK					SANIGENE			
				Linked sequences	Average length of linked sequences	Non-linked MSC	Non-linked singletons	Average length of non-linked MSC	Linked sequences	Average length of linked sequences	Non-linked MSC	Average length of non-linked MSC
Index2.31[a]	162439	69158	3455 (5)	30665	1276.1	26889	118561	457.3	8638	1044.4	50219	344.9
Index2.35[b]	159840	66188	5634 (8)	7133	1948.4	50591	159840	700.0	7108	1346.7	50591	437.6

[a]Whole-body index2.31 was generated by clustering 470293 tissue-level consensus sequences
[b]Whole-body index2.35 was generated by adding UniGene(build #106) to an existing whole-body index2.30.

Table 3.6 Orientation for all consensus sequences in the whole-body index2.35

Orientation	Singletons*	Multi-sequence clusters		Clone-linked sequences	Totals
	Clusters not contributing to the clonelinked entries	Clusters contributing to the clonelinked entries	Clusters that do not contribute to the clonelinked entries		
3’	67451	6260	24423	5460	97334**
5’	57026	7016	16765	1410	75201
End-not-specified	35361	2311	9406	263	45030
Totals	159838	15587	50594	7133	

*No singleton records contributed to the clonelinked entries

** The total number of 3’ end sequences do not include the number of multi-sequence clusters that were clonelinked. The latter sequences have been accounted for in the clonelinked entries.

3.4 Discussion

The STACK gene index represents an attempt to add value to the EST data by generating unique indices and providing a resource to capture sequence variation. A hierarchical approach was implemented with the clustering of EST data captured in “tissue bins” followed by clustering of all tissue datasets.

The process of tissue partitioning highlighted the use of non-standardised terms in the EST GenBank records and the need to parse multiple features keys. The recent development of a controlled vocabulary at SANBI has used the cDNA library (clone_lib feature) names to partition sequences into different organ and tissue categories. As of February 2001, there have been approximately 5700 clone libraries represented in the EST database. In retrospect, tissue partitioning of EST data in STACK can be more easily automated by subdividing a finite set of clone libraries followed by the importing of EST records based on the clone library field.

The correct assignment of ESTs to specific clusters is hampered by inadequate masking of repetitive elements (Jongeneel 2000). The removal of 2.7% of ESTs during the STACK masking step ensured minimal generation of chimeric clusters. RepeatMasker was shown to be a more sensitive technique in comparison to command-line execution of cross_match for the removal of repetitive elements and low complexity sequences from an additional 28,513 consensus sequences. UniGene on the other hand, uses the DUST program at NCBI to remove low complexity regions from the EST data.

A total of 1,166,367 ESTs were partitioned into 162,439 singletons and 69,158 multi-sequence clusters by d2_cluster. A subset of 21,097 clusters (8.8%) represented sequences that should have been merged by d2_cluster. Type I errors (i.e., sequences that should be merged by d2_cluster) for d2_cluster were reported by Burke et al (1999) in 4.4% of clusters. The increased proportion of type I errors reported in the STACK gene index reflects the fragmentation of the assembly process where PHRAP was reported to partition 5% of the gene indices. The redundancy in the STACK gene index was resolved by a final round of pairwise comparison between all consensus sequences using cross_match (Figure 3.2). Indices such as TIGR and UniGene employ additional alignment-based processing to merge redundant clusters that are generated during the initial clustering stage (Liang et al., 2000).

Assembly analysis

The most extensively used assembly programs in genomic sequencing projects include PHRAP (http://www.genome.washington.edu/uwgc/analysistools/phrap.htm), TIGR assembler (Sutton et al., 1995) and CAP3 (Huang and Madan 1999). Recent reports by Liang et al (2000) suggested that PHRAP and CAP3, in comparison to TIGR assembler-EST, are more tolerant of sequencing discrepencies. The initial STACK development incorporated the use of PHRAP as an assembly tool due to the inability of MSA_CONTIG to handle EST datasets in excess of 60,000 sequences (see chapter 2). TIGR's gene indices have adopted CAP3 as their assembly tool because it provided higher fidelity consensus sequences and generated few assemblies for each gene (Liang et al., 2000; Quackenbush et al., 2001). STACK, however, does not rely on PHRAP to generate its consensus sequences because of the inherent problems associated with using PHRAP in the absence of trace files for EST data. For example, quality values indicate how accurate the base call is (i.e., values > 20 (99% confidence) represent high confidence calls) (Ewing and Green 1998). In the absence of quality values, PHRAP assigns a default quality of 15 to each base. During the consensus generation, when several sequences disagree, PHRAP resolves the problem by inserting two different bases in the final consensus sequences, producing insertion errors (Ewing and Green 1998; Liang et al., 2000). Despite recent evidence that CAP3 produces fewer assemblies than PHRAP (Liang et al., 2000), only 5% of the STACK gene indices were split into multiple assemblies during the assembly stage.

Radiation hybrid mapping

EST sequencing is intrinsically inadequate for identifying truly rare transcripts (Bortoluzzi et al., 2000). Therefore, the use of EST-based STSs will tend not to capture rare transcripts and as a result, STSs would provide optimistic estimates of cluster accuracy. However, we have used radiation hybrid markers to assess the fidelity of the STACK consensus sequences. An analysis of our mapping information suggest a 1.08-fold redundancy in the STACK clusters and approximately 0.5% of clusters represented inconsistent mapping information. The level of error reported in the stack gene index mapping data is supported by the 1% error reported in mapping laboratories. On the other hand, the possibility exists that the assignment of more than one cluster to a STS could represent the capture of paralogous genes in different clusters. Map locations were assigned to 40% (26944/66188) of the STACK gene indices and provide a resource for positional candidate gene selection relevant to both physical location and source of gene expression.

Capture of alternate gene expression forms

Databases such as TIGR and UniGene have focused on reconstructing the gene complement of the human genome and their technological developments have been directed towards achieving that goal. STACK, however, focuses on the detection and visualisation of transcript variation in the context of developmental and pathological states. Alternatively spliced transcripts that capture important biological information within the same cluster assembly are handled by specialised tools, CRAW and CONTIGPROC, within the STACK process (Miller et al., 1999; Christoffels et al., 2001). The ability to discriminate between distinct gene isoforms was illustrated by the partitioning of isoforms of fibulin within the STACK whole-body index cluster 133232_3 (Figure 3.4; see chapter 4 for additional illustrations of detecting transcript diversity).

Ongoing STACK development

STACK will make increasing use of the relational database architecture to enhance data access. The maintenance of clusterIDs or links to new IDs from release to release, are being planned. Inclusion of genomic information will be used to map clusters and expression states to genome location.

3.5 References

Benson D.A., Boguski M.S., Lipman D.J., Ostell J., Ouetelle B.F., Rapp B.A., Wheeler D.L. (1999) Release Notes for NCBI-GenBank Flat File Release 112.0. *Nucleic Acids Res.* **27:** 12-17.

Bortoluzzi S., d’Alessi F., Romualdi C. and Danieli G.A. (2000) The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.* **10:** 344-349.

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

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.

Deloukas P., Schuler G.D., Gyapay G., Beasley E.M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T.C., McKusick K.B., Beckmann J.S. et al. (1998) A physical map of 30,000 human genes. *Science* **282:** 744-746.

Ewing B. and Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 868-877.

Hide W., Burke J. and Davidson D. (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol*. **1:** 199-215.

Hide W., Burke J., Christoffels A. and Miller R. (1997) A novel approach towards a comprehensive consensus representation of the expressed human genome. *Genome Informatics. Universal Academy Press Inc. Tokyo Japan. Ed. S. Muiyano and T. Takagi.* 187-195.

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

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

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

Matsubara K. and Okubo K. (1993) Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotechnol.* **4:** 672-677.

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

Olson M., Hood L., Cantor C. and Botstein D. (1989) A common language for physical mapping of the human genome. *Science* **245:** 1434-1435.

Quackenbush J., Cho J., Lee D., Liang F., Holt I., karamycheva S., Parvizi B., Pertea G., Sultana R. and and White J. (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29:** 159-164.

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., et al. (1996) A gene map of the human genome. *Science* **274:** 540-546.

Schuler G.D. (1997a) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Mol.Med* **75:** 688-694.

Schuler G.D. (1997b) Sequence mapping by electronic PCR. *Genome Res*. **7:** 541-550.

Smit A. and Green P. (1999) http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl.

Smith S.W., Overbeek R., Woese C.R., Gilbert W. and Gillevet P.M. (1994) The genetic data environment: An expandable GUI for multiple sequence analysis. *Comp. Appl. Biosci.* **10:** 671-675.

Torney D.C., Burks C., Davison D. and Sirotkin K.M., In *Computer and DNA*, Bell, G. I., Marr, T. Eds.; Addison-Wesley: New York, 1990, pp. 109-125.

Vasmatzis G., Essand M., Brinkmann U., Lee B. and Pastan I. (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci.* **95:** 300-304.

Wilcox A.S., Khan A.S., Hopkins J.A. and Sikela J.M. (1991) Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion of STSs: implications for an expression map of the genome. *Nucleic Acids Res*. **19:** 1837-184.

Chapter 4

STACK whole-body index validation

STACK whole-body index validation

Contents **Page**

List of Figures

List of Table

Summary

This chapter describes the use of the completed chromosome 22 sequence to assess the accuracy of the whole-body index consensus sequences, the identification of the range of transcript structures identified in the whole-body index, and the application and context of singletons and EST assemblies for identifying alternative splicing events.

Overall 63.3% of the annotated chromosome 22 genes had significant hits to whole-body index sequences. Twenty-five whole-body indices matched regions of chromosome 22 that had not been described in the past. Alignment of whole-body indices to chromosome 22 genes demonstrated a 0.96 fold redundancy in STACK, similar to the radiation hybrid mapping data. A total of 84,387 genes in the human genome were estimated from the chromosome 22 verified STACK sequences.

Two novel splice variants were identified in the whole-body index data corresponding to *neurofibromatosis2* gene and *fibulin1* gene. A detailed report for the characterised events in the 25 known alternatively spliced genes can be viewed at http://www.sanbi.ac.za/~alan/twentyfive_splicegenes.htm. Sim4 alignments for 493 indices mapped onto the chromosome 22 genes were analysed and classified as exon sequence (349/493; 5 exon-skips), intron sequence (3/493), gapped exons (8/493; 3 exon skips) and combined intron-exon transcripts (133/493; 8 exon skips).

4.1 Introduction

Gene identification in eukaryotic genomes is more difficult than in prokaryotes because of multiple exons separated by large intronic sequence. Current gene finding programs identify exons based on multiple properties such as identification of potential coding regions combined with matches to the consensus splice sites (Burge and Karlin, 1997; Haussler 1998). These computational methods have been shown to be inadequate for the identification of genes in eukaryotes. For example, almost 40% of GENSCAN-predicted genes on chromosome 22 did not form part of any gene confirmed by other means and included an unknown proportion of false positives (Durham et al., 1999). The failure rate of computational tools designed for gene finding can be attributed in part to our inability to understand the rules governing the choice of splice sites. However, recent studies on the spliceosome machinery in eukaroytes have revealed new insights into the mechanism of RNA splicing and mRNA stability (Mitchell and Tollervey 2000).

The absence of adequate computational tools for gene finding and the ever -increasing human genomic sequence data (Durham et al., 1999; Hattori et al., 2000) have spawned protocols that align genomic sequence to transcribed sequences in order to define exon boundaries and ultimately mRNA structure (Ewing and Green 2000; Liang et al., 2000; de Souza et al., 2000). Variation in mRNA structure occurs through alternative splicing and is very common in vertebrates. For example, a minimum estimate that 35% of human genes show variably spliced products has been derived from aligned ESTs that were mapped to the human genome sequence (Croft et al., 2000). The true estimate is probably much higher, considering that ESTs sample a portion of each mRNA and the origins of each EST is biased towards certain tissue types and developmental states. A variety of mRNA structures are produced through splicing (Lopez, 1998; review by Black 2000). Exons can be spliced into the mRNA or skipped; introns that are normally excised can be retained in the mRNA; the positions of either 5’ or 3’ splice sites can shift to make exons longer or shorter and alterations in transcriptional start sites or polyadenylation sites also allow production of multiple mRNAs from a single gene (Gautheret et al., 1998; Beaudoing et al., 2000).

In this chapter, I report on the use of the completed chromosome 22 sequence to assess the accuracy of the whole-body index consensus sequences, the identification of the range of transcript structures identified in the whole-body index, and the application and context of singletons and EST assemblies for identifying alternative splicing events.

4.2 Method

Comparison to chromosome 22 genomic data

EMBL records for the published 311 mRNAs and 234 predicted genes on chromosome 22 were downloaded from the Sanger ftp site on the 7 November 2000 (ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/). The 134 pseudogenes were extracted from the chromosome 22 genomic sequence (ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/chr_22_analysis_version_22-10-1999.fa.gz) using the start and end positions of each pseudogene as reported in the gene table on the Sanger website (http://www.sanger.ac.uk/cgi-bin/c22_genes_table.pl). The 679 chromosome 22 genes were masked for contaminating sequences as outlined in chapter 3 using RepeatMasker (Smit and Green 1999). The 226,028 whole-body index sequences were searched against the database of chromosome 22 genes using BLASTN. A whole-body index sequence was accepted as a significant match to the chromosome 22 genes if it had at least 94% identity over more than 80% of the whole-body index sequence length. The whole-body index sequences that did not match to the documented chromosome 22 genes were searched against the masked version of the chromosome 22 genomic sequence (1999 release: ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_analysis_version_22-10-1999) to identify unreported transcribed genes on chromosome 22. The results were compared to the gene table on the Sanger website (http://www.sanger.ac.uk/cgi-bin/c22_genes_table.pl). This process was repeated for the updated chromosome 22 genomic sequence (ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_19-05-2000.masked.fa.gz) and the updated list of genes published on the Sanger website on the 15th November 2000.

Identification of splice events in the whole-body index data

Twenty-five experimentally identified alternatively spliced genes on chromosome 22 were extracted in FASTA and EMBL format. The whole-body index sequences that capture the 25 alternatively spliced genes were identified by BLASTN and aligned to the alternatively spliced genes using Sim4 (Florea et al., 1998). The alignment coordinates were cross-referenced against the exon boundaries documented in the EMBL records for each of the alternatively spliced genes. The known exon boundaries for all the splice variants allowed for the identification of novel splice variants, exon skipping and alternate donor and acceptor sites. This protocol was semi-automated and applied to 226 chromosome 22 genes that encoded more than 2 exons and for which there were documented intron-exon boundaries.

4.3 Results

Comparison to chromosome 22 genomic data

Six hundred and seventy seven of the whole-body index sequences matched 345/545 chromosome 22 genes (excluding the pseudogenes) and demonstrated a 1.9-fold redundancy in the whole-body index sequences that occurred as a result of non-overlapping fragments. The remainder of the whole-body index sequences (225,351) were searched against the masked chromosome 22 genomic sequence (release 1999). A total of 1691 consensus sequences matched regions of the chromosome 22 genomic sequence that did not overlap with the 545 known genes and predicted genes.

All the whole-body index sequences were searched against the most recent release of the masked chromosome 22 genomic sequence (release 2000) using BLASTN. Twenty-five consensus sequences matched regions of chromosome 22 (release 2000) that did not match any of the annotated genes on chromosome 22. The twenty-five unique consensus sequences represent a more accurate data set compared to the 1691 sequences that were searched against an older version of chromosome 22 genomic sequence. The twenty-five unique consensus transcripts were not identified in the list of genes published for chromosome 22 on the Sanger website (15th November 2000). No putative function could be assigned to the twenty-five novel genes identified on chromosome 22 after searching the protein non-redundant database (release January 2001). However, 318/677 STACK sequences that did match the chromosome 22 genes were assigned putative functions based on the outcome of a search against the non-redundant protein database (January 2001) (http://www.sanbi.ac.za/~alan/677chr22putativefunction.htm).

Identification of splice events

A splice variant for each of 11/25 (44%) alternate splice genes on chromosome 22 was captured by 1 or more whole-body index multi-sequence cluster or singleton. Of the remaining 14 alternate splice genes, all were identified in the whole-body index data but no distinction could be made between the various isoforms. Three multi-sequence clusters showed partitioning of the splice variants for 3/11 alternatively spliced genes namely; BK1191B2.3.1 (similar to Malonyl coA-acyl carrier, Figure 4.4), DJ1042K10.1.2 (adenylosuccinate lyase gene, Figure 4.3) and BK941F9.1 (fibulin, Figure 4.2). Two novel splice variants were identified in the whole-body index data corresponding to *neurofibromatosis2* gene and *fibulin1* gene (Figure 4.5a and b). Nine exon skipping events, twenty one alternate acceptors and eleven alternate donors were observed by comparing the

25 known alternatively spliced genes with the whole-body index sequences (Table 4.1). A summary of the captured events in the alternatively spliced genes using processed ESTs is presented in Table 4.1. A detailed report for the characterised events in the 25 known alternatively spliced genes can be viewed at http://www.sanbi.ac.za/~alan/twentyfive_splicegenes.htm.

A total of 493 whole-body index transcripts matched 226 chromosome 22 genes for which there was evidence of intron-exon boundaries and the presence of at least two exons in the chromosome 22 genes. Sim4 alignments for 493 indices mapped onto the chromosome 22 genes were analysed and classified as exon sequence (349/493; 5 exon-skips), intron sequence (3/493), gapped exons (8/493; 3 exon skips) and combined intron-exon transcripts (133/493; 8 exon skips).

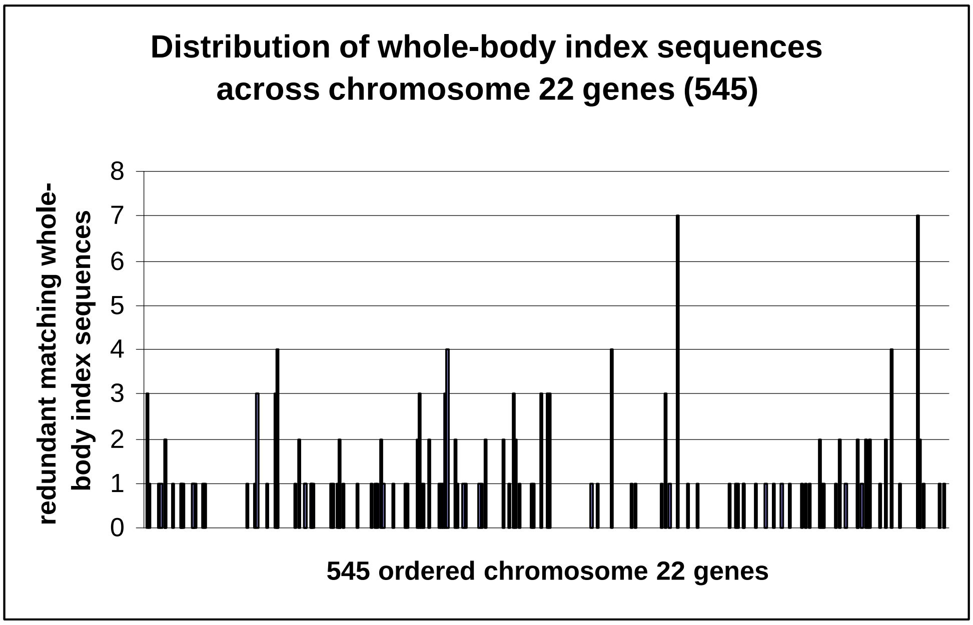
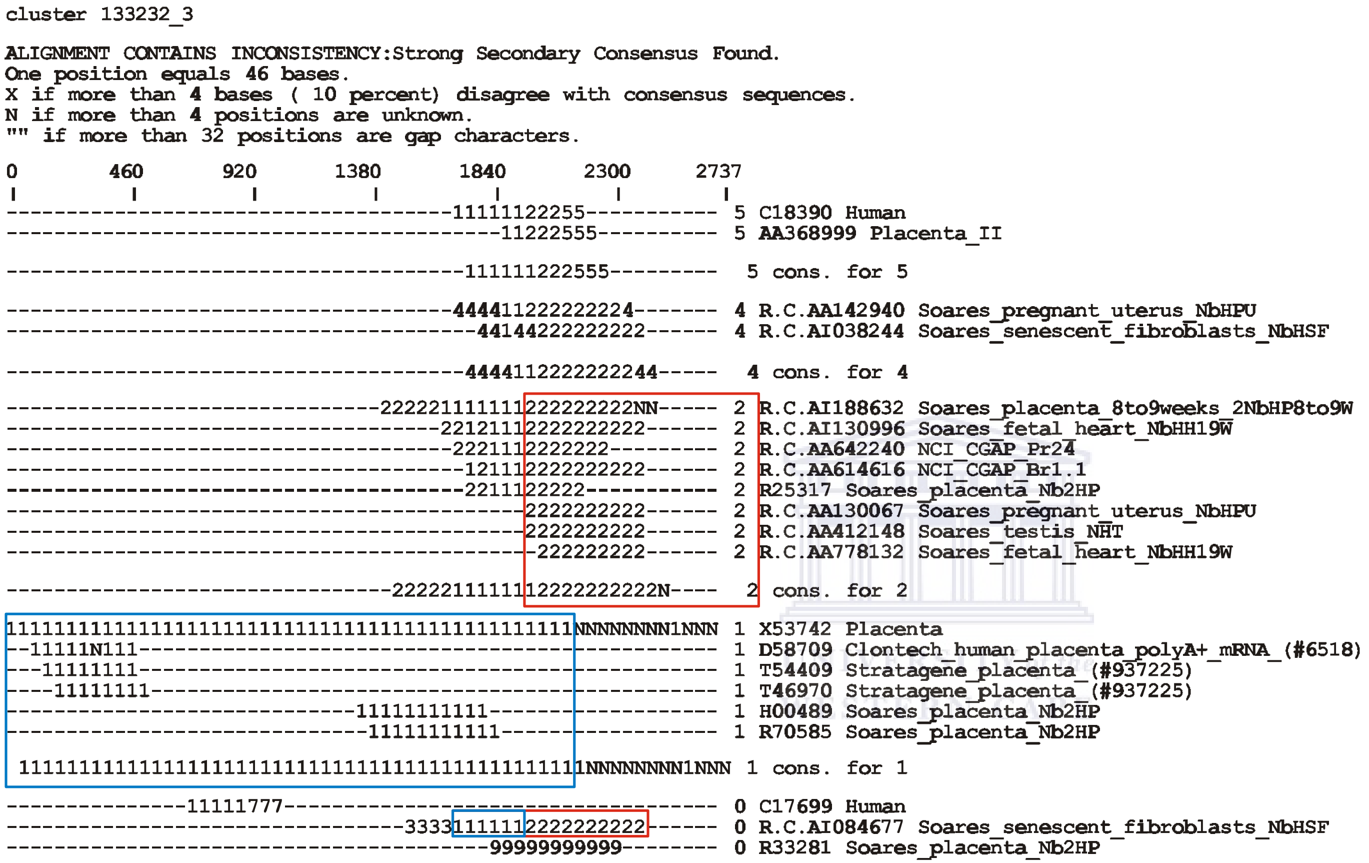
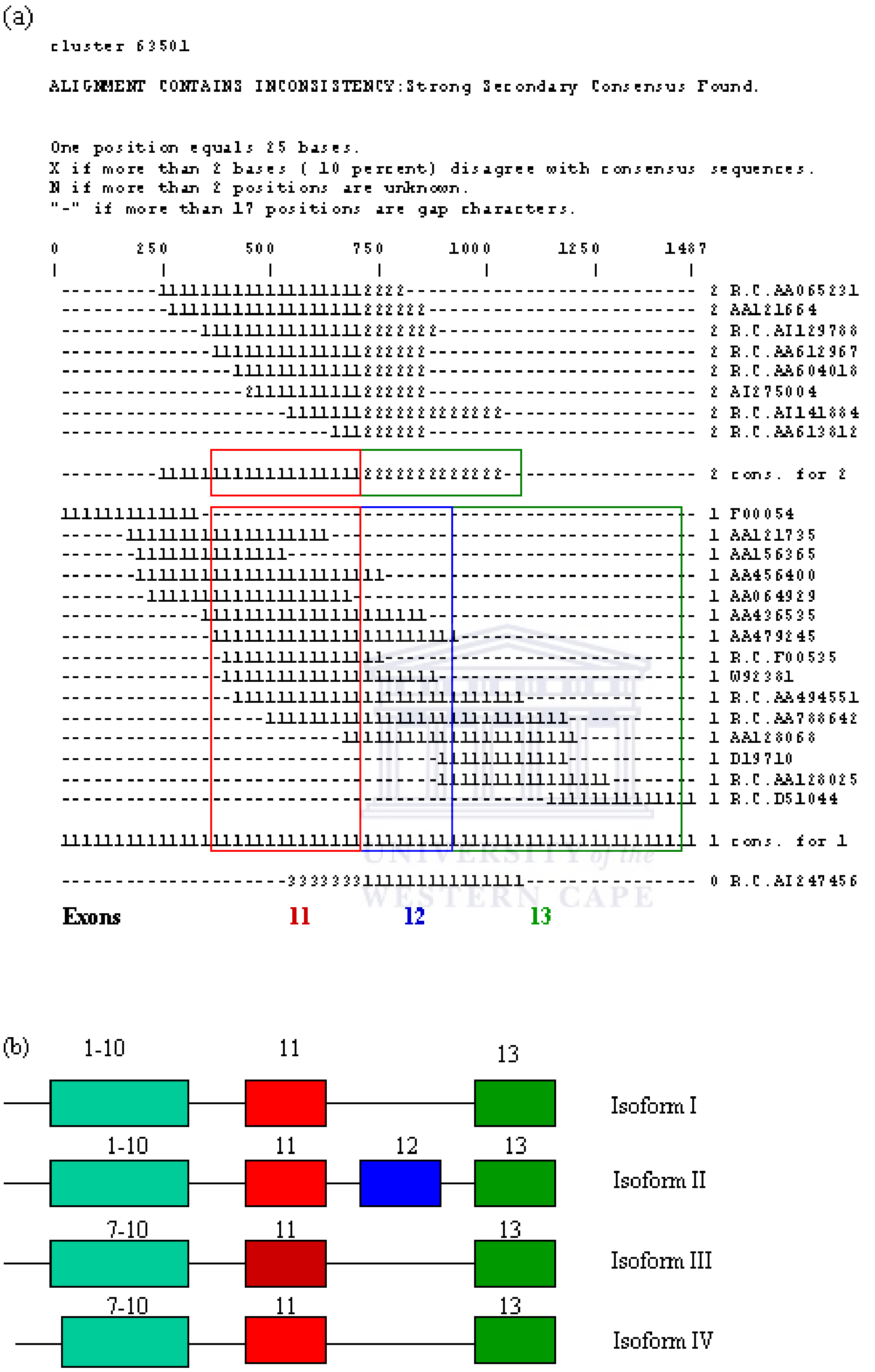


Figure 4.1 Distribution of 677 whole-body index sequences across the 545 chromosome 22 genes (mRNA sequences) (excluding the pseudogenes). The whole-body index sequences that are plotted on the x-axis represent redundant sequences because more than one sequence match the same mRNA. In these multiple hits to the same mRNA there are no overlapping regions between the whole-body index sequences.



```
cluster 133232_3

ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.
One position equals 46 bases.
X if more than 4 bases ( 10 percent) disagree with consensus sequences.
N if more than 4 positions are unknown.
"" if more than 32 positions are gap characters.

0        460      920       1380      1840       2300     2737
I          I         I         I         I         I         I
--------------------------------------11111122255----------- 5 C18390 Human
------------------------------------------11222555---------- 5 AA368999 Placenta_II

--------------------------------------111111222555---------  5 cons. for 5

--------------------------------------444411222222224------- 4 R.C.AA142940 Soares_pregnant_uterus_NbHPU
----------------------------------------44144222222222------ 4 R.C.AI038244 Soares_senescent_fibroblasts_NbHSF

--------------------------------------4444112222222244-----  4 cons. for 4

--------------------------------222221111111222222222NN----- 2 R.C.AI188632 Soares_placenta_8to9weeks_2NbHP8to9W
-------------------------------------22121112222222222------ 2 R.C.AI130996 Soares_fetal_heart_NbHH19W
--------------------------------------2221112222222--------- 2 R.C.AA642240 NCI_CGAP_Pr24
---------------------------------------121112222222222------ 2 R.C.AA614616 NCI_CGAP_Br1.1
---------------------------------------2211122222----------- 2 R25317 Soares_placenta_Nb2HP
--------------------------------------------2222222222------ 2 R.C.AA130067 Soares_pregnant_uterus_NbHPU
--------------------------------------------2222222222------ 2 R.C.AA412148 Soares_testis_NHT
---------------------------------------------222222222------ 2 R.C.AA778132 Soares_fetal_heart_NbHH19W

--------------------------------2222211111112222222222N----  2 cons. for 2

111111111111111111111111111111111111111111111111NNNNNNNN1NNN 1 X53742 Placenta
--11111N111------------------------------------------------- 1 D58709 Clontech_human_placenta_polyA+_mRNA_(#6518)
---11111111------------------------------------------------- 1 T54409 Stratagene_placenta_(#937225)
----11111111------------------------------------------------ 1 T46970 Stratagene_placenta_(#937225)
------------------------------11111111111------------------- 1 H00489 Soares_placenta_Nb2HP
-------------------------------11111111111------------------ 1 R70585 Soares_placenta_Nb2HP

 111111111111111111111111111111111111111111111111NNNNNNNN1NNN 1 cons. for 1

---------------11111777------------------------------------- 0 C17699 Human
----------------------------------33331111112222222222------ 0 R.C.AI084677 Soares_senescent_fibroblasts_NbHSF
-----------------------------------------99999999999-------- 0 R33281 Soares_placenta_Nb2HP
```

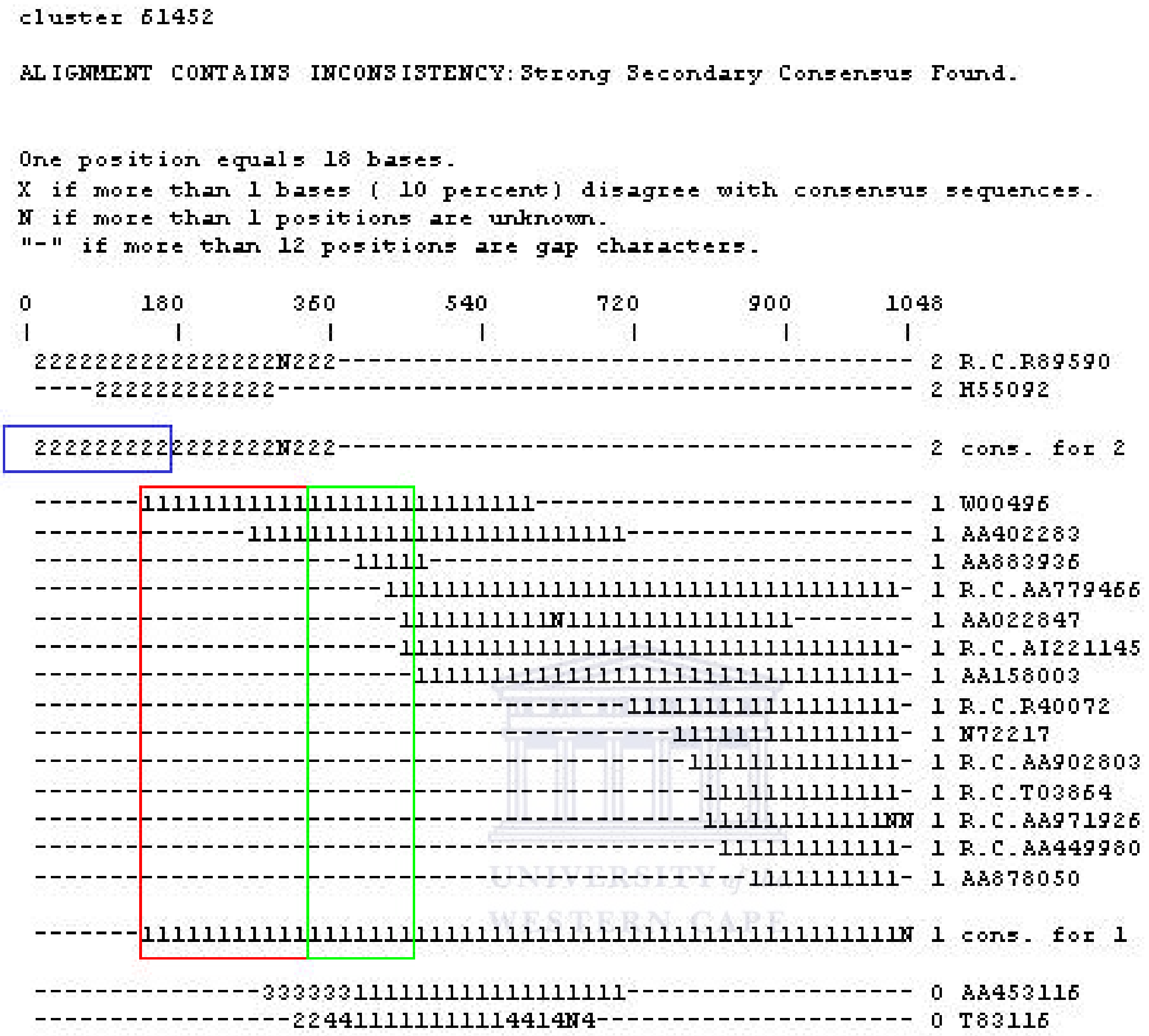
Figure 4.2 CRAW output for a whole-body index cluster displaying alternate gene isoforms of the *fibulin* gene. The blue box indicates the region capturing the fibulin-1B isoform whereas sequences capturing fibulin-1C are surrounded by a red box.



```
cluster 63501

ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.


One position equals 25 bases.
X if more than 2 bases ( 10 percent) disagree with consensus sequences.
N if more than 2 positions are unknown.
"-" if more than 17 positions are gap characters.

0         250        500        750        1000       1250       1487
|          |          |          |          |          |          |
 ---------11111111111111111112222---------------------------- 2 R.C.AA065231
 ----------111111111111111111222222-------------------------- 2 AA121664
 -------------1111111111111112222222------------------------- 2 R.C.AI129788
 --------------11111111111111222222-------------------------- 2 R.C.AA612967
 ----------------111111111111222222-------------------------- 2 R.C.AA604018
 -----------------21111111111222222-------------------------- 2 AI275004
 ---------------------11111112222222222222------------------- 2 R.C.AI141884
 -------------------------111222222-------------------------- 2 R.C.AA613812

 ---------11111111111111111112222222222222------------------- 2 cons. for 2

 1111111111111----------------------------------------------- 1 F00054
 ------1111111111111111111----------------------------------- 1 AA121735
 -------11111111111111--------------------------------------- 1 AA156365
 -------111111111111111111111111----------------------------- 1 AA456400
 --------1111111111111111111--------------------------------- 1 AA064929
 -------------111111111111111111111-------------------------- 1 AA436535
 --------------11111111111111111111111----------------------- 1 AA479245
 ---------------1111111111111111----------------------------- 1 R.C.F00535
 ---------------111111111111111111111------------------------ 1 W92381
 ----------------1111111111111111111111111111---------------- 1 R.C.AA494551
 -------------------11111111111111111111111111111------------ 1 R.C.AA788642
 --------------------------11111111111111111111111----------- 1 AA128068
 -----------------------------------1111111111111------------ 1 D19710
 -----------------------------------11111111111111111-------- 1 R.C.AA128025
 ----------------------------------------------11111111111111 1 R.C.D51044

 111111111111111111111111111111111111111111111111111111111111 1 cons. for 1

 ---------------------33333331111111111111111---------------- 0 R.C.AI247456
```

Figure 4.3 CRAW output for a whole-body index cluster displaying alternate gene isoforms of the *adenylosuccinate lyase* gene. (a) The sequences depicted by a string of 1's represent exons 11, 12 and 13 (red, blue and green boxes respectively; isoform II). The sequences depicted by a string of 1's followed by 2's represent exons 11 and 13 (red and green boxes respectively; isoform I, III or IV). There is not enough sequence to distinguish isoform I, III and IV). (b) A diagram illustrating the variety of isoforms identified for the ADSL gene.

(a)



```
cluster 61452

ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.


One position equals 18 bases.
X if more than 1 bases ( 10 percent) disagree with consensus sequences.
N if more than 1 positions are unknown.
"-" if more than 12 positions are gap characters.

0       180       360       540       720       900      1048
|         |         |         |         |         |       |
 2222222222222222N222-------------------------------------- 2 R.C.R89590
 ----222222222222------------------------------------------ 2 H55092

 2222222222222222N222-------------------------------------- 2 cons. for 2

 -------11111111111111111111111111------------------------- 1 W00496
 --------------1111111111111111111111111------------------- 1 AA402283
 ---------------------11111-------------------------------- 1 AA883936
 -----------------------1111111111111111111111111111111111- 1 R.C.AA779466
 ------------------------1111111111N111111111111111-------- 1 AA022847
 ------------------------111111111111111111111111111111111- 1 R.C.AI221145
 -------------------------11111111111111111111111111111111- 1 AA158003
 ---------------------------------------111111111111111111- 1 R.C.R40072
 ------------------------------------------111111111111111- 1 N72217
 -------------------------------------------11111111111111- 1 R.C.AA902803
 --------------------------------------------1111111111111- 1 R.C.T03864
 --------------------------------------------111111111111NN 1 R.C.AA971926
 ---------------------------------------------111111111111- 1 R.C.AA449980
 -----------------------------------------------1111111111- 1 AA878050

 -------11111111111111111111111111111111111111111111111111N 1 cons. for 1

 ---------------333333111111111111111111------------------- 0 AA453116
 -----------------224411111111114414N4--------------------- 0 T83116
```

(b)

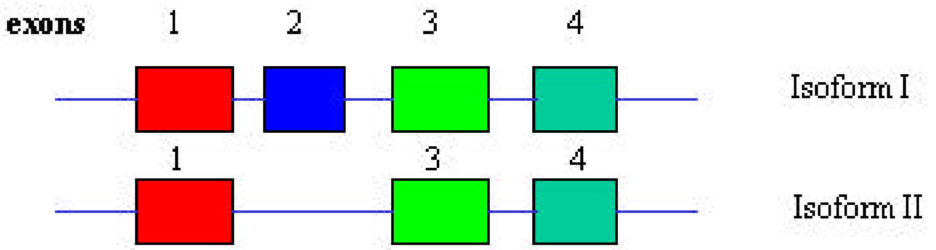


Figure 4.4 CRAW output for a whole-body index cluster displaying alternate gene isoforms for the gene similar to *Malonyl coA-acyl carrier* gene. (a) Sequences depicted by a string of 1’s represent isoform II capturing exon 1 followed by exon3 (red and bright geen boxes respectively). Sequences depicted by a string of 2’s represent isoform I capturing exon 2 and three (blue box). (b) Illustration of all the exons for the two isoforms.

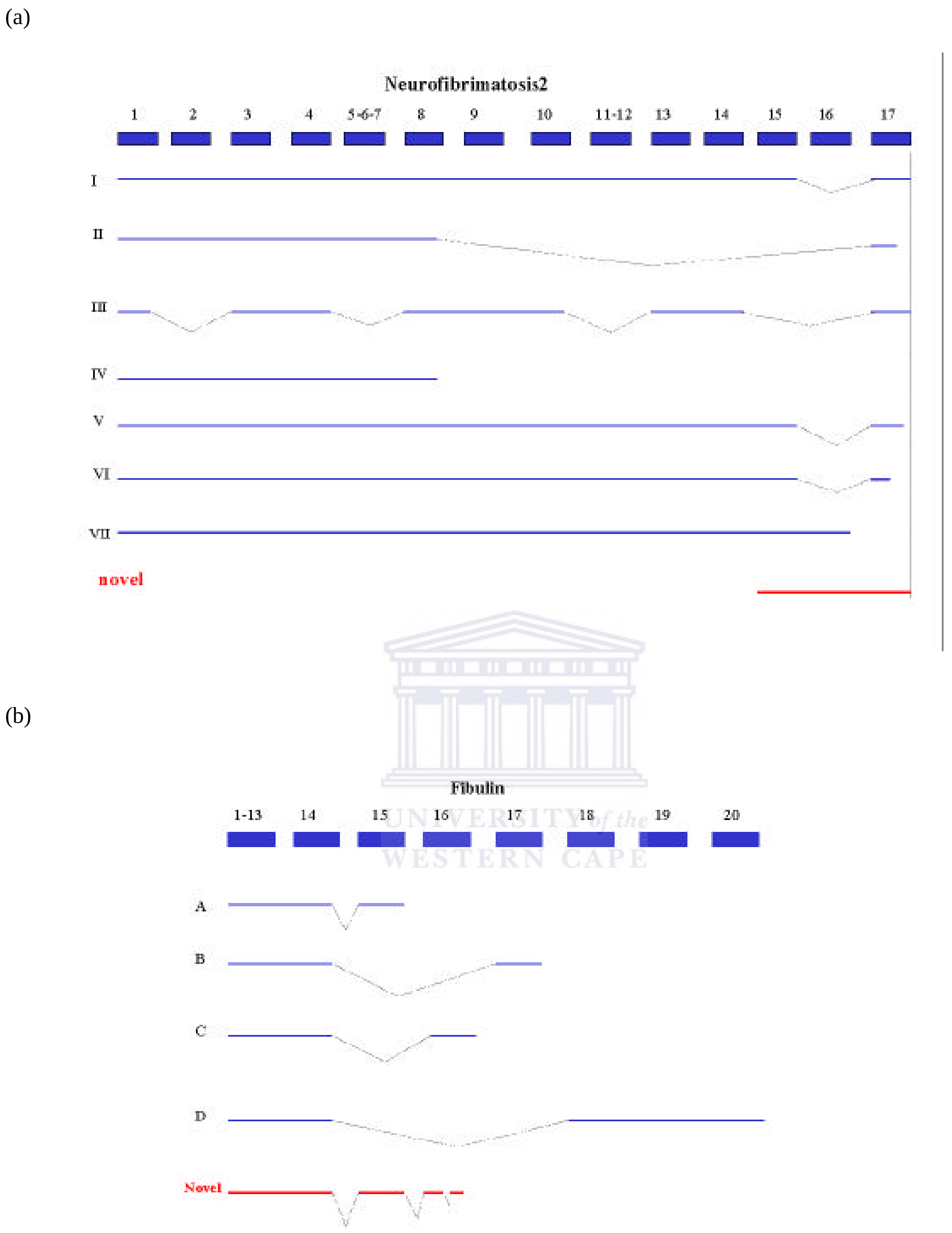


Figure 4.5 Novel alternative splice events captured in the whole-body index data.

(a) Seven documented isoforms for *neurfibromatosis2* gene (blue lines; exons depicted as blue rectangles). A novel isoform identified in an EST representing exons 15, 16 and 17(red line).

(b) Four documented isoforms of *fibulin* (blue lines; exons depicted as blue rectangles). A novel isoform captured in a EST representing exons1-16 where exon 16 has a deleted middle portion

Table 4.1. List of all the captured events when comparing processed ESTs to 25 chromosome 22 alternatively spliced genes.

Captured events	**Number of processed ESTs capturing alternate splice genes on chromosome 22**	
	Singletons (65)	**Multi-sequence clusters (50)**
exon sequence	39	23
cryptic intron	1	1
splice variants*	7	17
exon skipping++	8 [2,3,4 and 9 exons]	2 [1 exon]
Alternate donor	8 [intron]	3 [intron]
Alternate acceptor	12 [3 intron]	9 [7 intron]
novel splice variant	0	2

*The splice variants also captured alternate donor and acceptor sites and therefore the numbers in each column do not equal the total singletons and multi-sequence clusters.

++All the singletons capturing exon skip events have been found in multi-sequence clusters in the recent release of UniGene (30Th November 2000)

4.4 Discussion

Comparison between the whole-body index2.35 and chromosome 22

The whole-body index2.35 was searched against the 679 chromosome 22 genes and captured 677 (including 349 singletons) index sequences with more than 93% identity to 345/659 (52%) chromosome 22 genes where each gene spanned at least 80% of the matching whole-body index sequence. This finding is consistent with previous observations that approximately half of the identified genes have EST support (Liang et al., 2000). Overall 63.3% of the annotated chromosome 22 genes (excluding pseudogenes) had significant hits to the whole-body index2.35. If we only consider multi-sequence clusters that match chromosome 22 genes, then despite the difference in assembly methodologies, consistent reports are published for significant hits of annotated genes to cluster consensus sequences for STACK, TIGR's THCs and ORESTES namely 60.2% (extrapolated from 677 minus 349 singletons), 60.7% (Fiang et al., 2000 and 50.2% (de Souza et al., 2000) respectively. The increased number of identified annotated genes (10%) using STACK and TIGR’s THCs as compared to the ORESTES data suggests that there is added value provided by the assembled ESTs.

The whole-body index2.35 sequences that match to chromosome 22 genes also provide an independent assessment of the redundancy in the whole-body index sequences. Singletons have been reported to represent low quality and intron-containing sequences (Liang et al., 2000, de Souza et al., 2000). However, the 349 singletons captured as part of the 677 whole-body index sequences were incorporated into the assessment because of the high degree of similarity to the chromosome 22 genes. Burke et al. (1999) characterised type I error rate (i.e., inability to join to sequences that belong in the same cluster) in d2_cluster, (the algorithm used to cluster the whole-body index sequences) and found an upper limit for this error rate to be 0.4%. Applying the type I error rate in d2_cluster, reduces the 677 matching whole-body index sequences to 676 sequences that actually match 345 chromosome 22 genes. This accounts for a 1.95 fold redundancy that is higher than our radiation hybrid mapping estimate of 1.04 fold. The discrepancy in the whole-body index2.35 redundancy values is a result of the absence of singletons in the mapping of RH marker data. Excluding the singletons from the indices that match chromosome 22 genes produces a 0.96 fold redundancy (677-349/345) that similar to the results obtained with the RH mapping data (Chapter 3 section 3.3).

The whole-body index2.35 was searched against the chromosome 22 genomic sequence. A total of 2368 whole-body index sequences were identified with at least 93% identity and a minimum of 80% coverage across the length of the whole-body index consensus sequences. A subset of these sequences (1691 sequences) represented unique matches to the chromosome 22 genomic sequence that did not match any annotated regions of chromosome 22 (Durham et al., 2000; http://www.sanger.ac.uk/cgibin/c_22genes.pl). The unique whole-body index2.35 sequence matches were reduced from 1691 to 25 when compared to the updated chromosome 22 genomic sequence (ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_19-05-2000.masked.fa.gz). The high number of whole-body index sequence matches are consistent with previous reports of high gene density on chromosome 22 (Deloukas et al., 1998; Saccone et al., 1996).

The 702 (677+25) whole-body index sequences that mapped to chromosome 22 allows for an estimation of the number of genes in the genome. Using an approach reported by Fiang et al. (2000); if we combine the measurement of redundancy for the whole-body index2.35 using RH markers and the multi-sequence cluster hits to the annotated genes on chromosome 22 (average 1.0-fold), then there is EST support for approximately 702 genes on chromosome 22. Fiang et al. (2000) reported that approximately 54.8% of genes are represented by ESTs. Our data suggests that chromosome 22 contains as many as 1281 genes (100/54.8x702). Chromosome 22 is reported to represent 1.1% of the genome (Dunham et al., 1999) and is 1.38-fold gene rich (Deloukas et al., 1998) that suggests that the genome could contain approximately 84387 genes (1281x(100/1.1)/1.38). A number of reports have been published where the total genes in the human genome has been as low as 35000 (Ewing and Green, 2000) and as high as 140000 (Fiang et al., 2000). The discrepancy in gene numbers arises from the different approaches used by independent laboratories. Genes numbers derived from ESTs should be accepted with caution because the EST data represent a resource that is riddled with errors including unprocessed mRNA, low quality sequence and alternatively spliced genes.

The international human genome consortium predicted a total of 30000-40000 genes after generating an international gene index (International Genome Consortium 2001). The international gene index was generated using ENSEMBL, RefSeq, SWISSPROT and Trembl and each of these genes are either protein computer predictions. The approach adopted taken by the international consortium does not take into account genes that are expressed at low levels or in rare tissues. These genes will be missing or under represented in mRNA databases and hard to detect by protein homology. Single-exon genes encoding small proteins may also be missed as it is difficult to distinguish them from genome contamination. These considerations could inflate the total number of genes predicted by the international consortium. On the other hand, the total number of genes predicted from the STACK gene index could be an overestimate based on the fragmentation of sequence data in the absence of mRNA and other full length sequences.

Capturing alternate splicing events

Alternative splicing is seen as a means of producing functionally diverse polypeptides from a single gene especially in vertebrates (Lopez et al., 1998). A range of reports have documented occurrence of alternative splicing. Prevalence figures range from 35-38% (Mironov et al., 1999; Brett et al., 2000; Croft et al., 2000). Using the whole-body index 2.35, 11/25 (44%) of the known alternatively spliced genes on chromosome 22 were detected. Approximately 46% of the processed EST data matching the 25 alternatively spliced genes capture some form of alternate splicing namely, alternate donor/acceptor sites, exon skipping and intron retention. This data represent a comparison against the known exon-intron boundaries documented in EMBL. These alternate splicing events may represent artifacts of incomplete mRNA processing as suggested by Wolfsberg and Landsman (1997) who reported that approximately 1/5th of the EST database contains aberrant or incomplete mRNA.

Alternative splicing is a key problem in clustering as it results in fragmentation of EST clusters especially in strict alignment-based algorithms (Fiang et al., 2000). The use of a non-alignment based clustering system in STACK results in tolerance of variation in EST data and subsequent capture of alternative splicing events (Burke et al., 1999; Miller et al., 1999; Christoffels et al., 2001). This has been clearly demonstrated in the capture of three alternatively spliced genes that map to chromosome 22 (Figures 4.2-4.4). The use of error-checking tools during STACK processing ensures that the gene variants are partitioned within a cluster of ESTs. It is interesting to note that each of the identified alternatively spliced genes was detected using assembled EST data. The advantage of an EST assembly is demonstrated by the wealth of information that can be extracted from the STACK cluster analysis records (Figures 4.2-4.4).

4.5 References

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.

Black D. L. (2000) Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* **103:** 367-370.

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

Burge C. and Karlin S. (1997) Prediction of complete gene structures in huamn genomic DNA. *J. Mol. Biol.* **270:** 2411-2414.

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

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.

Croft L., Schandroff 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.

Deloukas P., Schuler G.D., Gyapay G., Beasley E.M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T.C., McKusick K.B., Beckmann J.S. et al. (1998) A physical map of 30,000 human genes. *Science* **282:** 744-746.

de Souza S. J., Camargo A. A., Briones R.S., Costa F. F. et al. (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci.USA* **9(23):** 12690-12693.

Durham I., Shimizu N., Roe B. A., Chissoe S. et al. (1999) The DNA sequence of human chromosome 22. *Nature* **402:** 489-496.

Ewing B.and Green P. (2000) Analysis of expressed sequence tags indicates 35000 human genes. *Nat. Genet.* **25:** 232-234.

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.

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.

Hattori M., Fujiyama A., Taylor T. D., Watanabe H., Yada T et al. (2000) The DNA sequence of human chromosome 21. *Nature.* **405:** 311-319.

Haussler D. (1998) Computational genefinding. *Trends Biochem. Sci. Supplementary guide to bioinformatics*. 12-15.

Liang F., Holt Ingeborg., Perea G., Karamycheva S., Salzberg S. L. and Quackenbush J. (2000) Gene index analysis of the human genome estimates approximately 120000 genes. *Nat. Genet*. **25:** 239-240.

Lopez A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32:** 279-305.

Miller R., Christoffels A., Gopalakrishnan C., Burke J., Ptitsyn A.A., Broveak T.R., and Hide W. (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. (1999) Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288-1293.

Mitchell P. and Tollervey D. (2000) MRNA stability in eukaryotes. *Current Opinion in Genetics and Development*. **10:** 193-198.

Saccone S., Caccio S., Kusuda J., Andreozzi L. and Bernardi G. (1996) Identification of the gene-richest bands in human chromosomes. *Gene* **174:** 85-94.

Smit A. and Green P. (1999) http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl.

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

Chapter 5

The identification of potential novel microsatellite markers and the assembly of a 4 megabase (Mb) region of chromosome 19 draft sequence to accelerate *PFHB1* candidate gene discovery.

The identification of potential novel microsatellite markers and the assembly of a 4 megabase (Mb) region of chromosome 19 draft sequence to accelerate *PFHB1* candidate gene discovery.

Contents **Page**

List of Figures

List of Tables

Summary:

Progressive familial heart block1 (PFHB1) is a cardiac conduction disorder that has been mapped to chromosome 19q13.3. The release of chromosome 19 draft sequence (June 2000) included 36 BAC and cosmid clones (1184 fragments) spanning the *PFHB1* locus for which there were no assemblies available. In an attempt to reduce the size of the *PFHB1* disease locus and accelerate the identification of candidate genes, we (i) performed *in-silico* screening for microsatellites, (ii) assembled the 1184 genomic fragments and (iii) mapped expressed transcripts including STACK and BodyMap onto the assembled data. This work presents the integration of the genetic and physical maps for the *PFHB1* locus, STACK and BodyMap transcripts, mouse developmental ESTs and RefSeq contigs. Potential novel microsatellites were identified in 29 out of 36 BAC and cosmid clones. PHRAP assembly reduced the 1184 chromosome 19 genomic fragments to 370 contigs and 874 singletons. The assemblies were annotated by mapping 119 STACK transcripts, 24 BodyMap transcripts, mouse ESTs and six RefSeq contigs. Seven positional candidates, previously shown to be expressed in heart tissue, have been identified including GLTSRC2, DKF2P761A179, Kaptin, T- elongation factor 4, nucleobindin, CGI-123 protein and CD37-antigen.

5.1 Introduction

During the last fifteen months we have witnessed accelerated sequencing of the human genome which has culminated in the release of draft sequence for 94% of the 24 individual human chromosomes (Genome Consortium 2001). The draft sequence provides a readily available resource that needs to be exploited, in tandem with wet-bench techniques, in order to accelerate the discovery of disease genes. The first steps in disease gene identification usually include genetic linkage analysis and fine mapping, which rely on the availability of an abundance of highly polymorphic markers (see section 5.1.2) spaced at relatively short intervals along the genome (reviewed in Keating 1992).

5.1.1 Genetic linkage analysis

Genetic linkage analysis facilitates the identification of the chromosomal location of a gene without any prior knowledge of its function. In order to map a disease gene, a search is undertaken for the co-inheritance of alleles at a specific genetic marker (see section 5.1.2 for an example) with the clinical phenotype of the disease within a family. The alleles at two loci that are situated physically very close to each other tend to be co-inherited but as the distance between them increases, the creation of new combinations of alleles by recombination becomes more likely (Terwilliger and Ott 1994). Recombination occurs during meiosis when two homologous chromosomes line up on the spindle and exchange DNA segments by a process termed “crossing over”. Crossing over event occurs between precisely corresponding sequences, so that no base pairs are added to or lost from the recombinant chromosome. The probability of a recombination occurring between two loci is termed the “recombination fraction” (denoted θ) and ranges in value from 0.00 (for loci next to each other) to 0.5 (for unlinked loci situated either far apart or on different chromosomes). In linkage analysis a calculation is made for the probability of an observed association between the inheritance of a specific DNA marker allele and the presence of a phenotypic trait. A comparison is made between the probability that the observed distribution of alleles would arise under the hypothesis of linkage (i.e., $0.000 < \theta < 0.5$) to the probability that this distribution would occur randomly (i.e., $\theta = 0.5$). The ratio of these two possibilities is the odds ratio (L). For convenience, L is converted to a decimal logarithm termed a lod score (“log of the odds”).

The formula of the lod score (Z) is as follows:

$$Z(\theta) = \log_{10} \frac{L(0.00 < \theta < 0.5)}{L\,(\theta = 0.5)}$$

Odds of more than 1000 to 1 (lod score of > 3) are necessary to prove significant evidence for linkage and odds of less than 1 to 100 (lod score of < -2.00) are sufficent to reject linkage. Lod scores between –2 and +3 are inconclusive (Terwillinger and Ott 1994).

5.1.2 Microsatellite DNA markers

In the past, human genetic markers included: (i) blood groups (Emery et al., 1969; Ghosh 1977), (ii) electrophoretic mobility variants of serum proteins (Hill et al., 1975; Johnson et al., 1981), (iii) human leukocyte antigen (HLA) tissue types (Wastiaux et al., 1978), (iv) DNA restriction fragment length polymorphisms(RFLPs) (Donis-Keller et al., 1987), (v) DNA mini satellites (Jeffreys et al.,1985) and variable number tandem repeats (VNTRs) (Nakamura et al., 1987). However, the discovery of microsatellite DNA families (Weber and May 1989) together with the development of the polymerase chain reaction (PCR) (Saiki et al., 1985; Mullis and Faloona 1987) has made genetic linkage analysis more powerful and informative than was previously possible with RFLPs.

Microsatellite DNA families comprise tandem repeats that have repeating units of length 1-6 base pairs, which are interspersed throughout the genome (Tόrth et al., 2000). A class of dinucleotide repeats, designated (CA)n.(GT)n (hereafter referred to as CA) constitute one of the most abundant families of human repetitive DNA elements, accounting for 50,000-100,000 stretches of repeats (Hamada 1982; Jeang 1983; Tautz 1984). CA repeats represent the most frequent simple sequence repeats found in the human genome and are interspersed every 30-60kb (Weber and May 1989; Beckman et al., 1992). CA repeats have been shown to be spaced at 5kb and 18 kb intervals on chromosome 21 and 22 respectively (Durham et al., 1999; Hattori et al., 2000). Tautz (1989) demonstrated hypervariability in the length of CA repeats when he observed length polymorphisms for two loci in three generations of a family, where corresponding alleles showed a Mendelian pattern of inheritance. This length polymorphism was later exploited as a general source of polymorphic markers for genome mapping and linkage studies (Gyapay et al., 1994).

Trinucleotide repeats, depending on the repeat class, are one to two orders of magnitude less frequent than CA repeats (Gastier et al., 1995). Database searches to estimate the distribution of trinucleotide repeats in the human genome have revealed that (AAT)n, (AAC)n and (AGC)n repeats are the most frequent in the human genome (Stallings 1994). Gastier et al. (1995) confirmed the abundance of (AAT)n and (AAC)n in the human genome but could not replicate the result for (AGC)n. Recently, (AAC)n repeats were shown to be the most frequent triplet repeat in mammalian introns (Tόrth et al., 2000). Expansions of trinucleotide repeat length in the coding regions of genes are known to cause neurodegenerative diseases such as fragile X syndrome, Huntington's disease, myotonic dystrophy and spinocerebellar ataxia) (reviewed in Warren and Nelson 1993; Bates and Lehrach 1994; Reddy and Housman 1997) and human cancers (Wooster et al., 1994; Arzimanoglou et al., 1998).

Tetranucleotide repeats have been reported to be advantageous because they produce cleaner PCR amplification products than dinucleotide repeats and are more readily co-amplified (Gastier et al., 1995). Exons seldom contain tetranucleotide repeats and the intronic and intergenic regions of vertebrate genomes have been shown to contain more tetranucleotide repeats than trinucleotide repeats (Tόrth et al., 2000).

Currently, numerous positional cloning projects are hampered by a lack of known polymorphic genetic markers situated within the disease gene critical interval. The detection of such markers could be used to (i) improve lod scores and (ii) identify recombination events, thereby narrowing the search area, and, ultimately, reducing the number of candidate genes to be screened. Traditional methods for the isolation and characterisation of microsatellites using molecular biology techniques have included southern blotting, in order to detect the genomic clone containing a di-, tri- or tetra-nucleotide repeat, followed by sub-cloning and sequencing of the desired fragment. Primers are then, designed to the regions flanking the stretch of repeats and PCR amplification carried out on a group of unrelated individuals to test the marker's polymorphic information content (Christoffels A, thesis 1997; GenBank submissions: U89020, U89021, U89022, AF003935, U88960, G31336). Recently, an *in-silico* method for identifying tandem repeats has been described which paved the way for accelerating microsatellite detection (Benson et al., 1999).

The generation of chromosome maps has included transcript maps for defined regions of human chromosomes (Wang et al., 1999; Hamshere et al., 2000; Lee et al., 2000) and mapping cDNA sequences for particular cell types, eg., skeletal muscle, onto a physical map (Pallavicini et al., 1997). These maps rely on radiation hybrid panels for their mapping, which allows for assignment of sequences to a region of about 1Mb. Recently, Hamshere et al. (2000) reported a kilobase resolution transcript map of a 10Mb region of the chromosome 19 cosmid library together with an *in-silico* northern analysis of these transcripts. The region covered by this refined map covers at least nine human diseases, of which five are still unidentified namely; nonsyndromic deafness (Chen et al., 1995), retinitis pigmentosa locus (*RP11*) (McGee et al., 1997), isolated cardiac conduction disease (de Meeus et al., 1997), progressive familial heart block type 1 (PFHB1) (Brink 1997), nonsyndromic orofacial cleft malformation (Martinelli et al., 1998) and asthma susceptibility loci (Ober et al., 1998).

PFHB1 is a cardiac conduction disorder that has been mapped to chromosome 19q13.3 in South African families (Brink et al., 1995). Recently, *PFHB1* has been fine mapped to a region flanked by *D19S606* and *D19S866* that spans a genetic map distance of 4 centimorgans (cM) (Arieff et al., 1999) (Figure 5.1). The limited number of known polymorphic markers in this region has hindered the fine mapping efforts to reduce the *PFHB1* locus. Refining the region harboring the *PFHB1* gene would reduce the selection of plausible candidate genes from a chromosome that has been reported to be particularly gene rich (Ashworth et al., 1995). Identification of novel genes on chromosome 19 through annotation has been made possible through the availability of the completed first draft sequence for chromosome 19 (June 2000). An assembly of the draft sequence in the *PFHB1* region would provide the scaffold for annotation using processed ESTs, such as the STACK whole-body index (Miller et al., 1999; Christoffels et al., 2001).

I report on the identification of potential novel microsatellite markers and the assembly of chromosome 19 draft sequences (release November 2000) across the *PFHB1* disease-gene region. In addition, the chromosome 19 contigs were annotated using the STACK whole-body index in order to identify candidate genes.

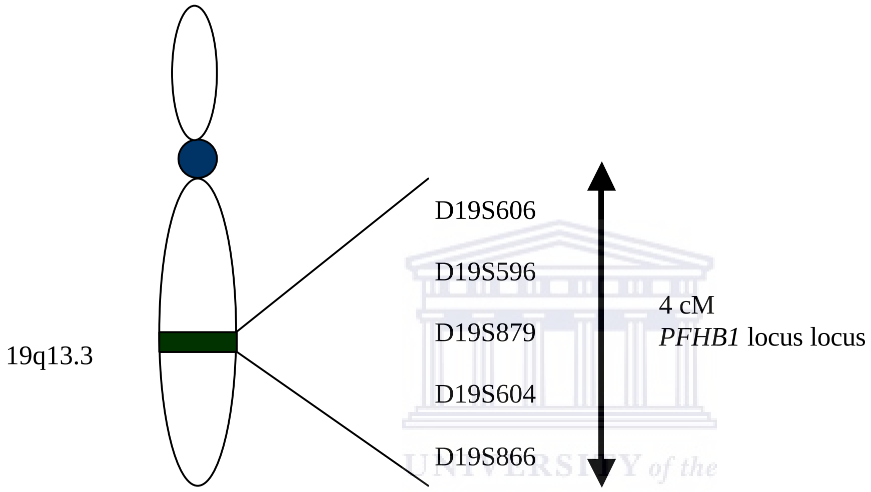


Figure 5.1 An ideogram of chromosome 19 depicting the position of the *PFHB1* locus relative to microsatellite markers on 19q13.3. The *PFHB1* locus is indicated by a double-headed arrow.

5.2 Methods

All methods summarised below have been semi-automated as detailed in Appendix IV.

5.2.1 Data acquisition

The first completed draft sequence of chromosome 19 has been made available to the public through the Joint Genome Initiative (JGI) in the form of sequenced cosmid and BAC clones. A total of 1184 sequences, representing 23 BAC and 13 cosmid clones, spanning a 4Mb region harbouring the *PFHB1* disease gene, on chromosome 19q13.3, were downloaded from the JGI ftp site (ftp://sawdoff.llnl.gov/pub/JGI_data/Human/Ch19). The header line for each FASTA record was transformed so that the BAC and cosmid names were captured. Standardisation of each record was essential for semi-automating the analysis of the *PFHB1* locus

Sequences were masked for *E. coli* (ftp://ncbi.nlm.nih.gov/repository/genomes/ecoli) using cross_match (P.Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/swat.htm) prior to screening for tandem repeats. The *E. coli* masked sequences were then screened for other contaminating sequences such as vector, simple repeats, mitochondrial and ribosomal regions using RepeatMasker (Smit and Green 1999), prior to the PHRAP assembly.

One hundred and sixteen chromosome 19 Refseq sequences (i.e., NCBI reference sequences) were retrieved from NCBI using the batch entrez system (http://www.ncbi.nlm.nih.gov/Entrez/batch.html). A total of six of the 116 Refseq sequences were mapped to the sequences that span the *PFHB1* region. The six records were processed to remove sequences that do not lie within the *PFHB1* region.

5.2.2 Screening the BAC and cosmid fragments for tandem repeats

Tandem Repeat Finder (TRF) (trf.irix.exe), a program designed to detect long stretches of tandem repeats (Benson 1999), was downloaded from http://c3.biomatch.mssm.edu/trf.html. All chromosome 19 sequences, masked for *E.coli* sequences, were screened for tandem repeats with a maximum period size of four.

5.2.3 Detection of known microsatellites

Four genetic markers are known to rivet onto five of the clones (*D19S604* onto R29295; *D19S596* onto R31763; *D19S879* onto BC52309 and *D19S866* onto BC61330/R28901) (S. Bardien-Kruger pers. comm.). These were used as positive controls to verify our *in-silico* screening for tandem repeats. The fragments containing the above-mentioned microsatellites were extracted from Entrez (http://www.ncbi.nlm.nih.gov/Entrez) and searched against the 1184 sequences using est2genome (Mott 1997) to confirm that they were present only in the five clones described. The clones harbouring the above-mentioned microsatellites were compared with the *in-silico* identified tandem repeats to confirm the accuracy of TRF.

5.2.4 Genomic sequence assembly of the PFHB1 disease-gene region

The masked chromosome 19 sequences were assembled using PHRAP (P.Green, unpublished, http://www.genome.washington.edu/uwgc/analysistools/phrap.htm) with the implementation of the default PHRAP parameters. The data represented in the “ace” (phrap output file) file was parsed in order to determine the sequence membership of each contig.

5.2.5. Annotating the PHRAP contigs and singletons

The contigs and singletons generated by PHRAP (hereafter referred to as phrap contigs and singleton contigs, respectively) were searched against the stack whole-body index2.35 using BLASTN with a score of $1e^{-40}$. The matching stack consensus sequences were searched against a protein non-redundant database (27 April 2000) and a DNA non-redundant database (7 July 2000).

BodyMap (http://bodymap.ims.u-tokyo.ac.jp) represents a database of expression profiles of human and mouse genes, known and novel, in various tissues. Approximately 18,000 transcripts were downloaded (release 27 November 2000; http://bodymap.ims.u-tokyo.ac.jp/datasets/gs_seq.all), together with gene expression information for each of the Bodymap transcripts (http://bodymap.ims.u-tokyo.ac.jp/datasets/gene_tissue_matrix). The Bodymap transcripts were searched against the PHRAP contigs using BLASTN to identify gene expression information pertaining to the matching coding portions of the PHRAP contigs.

A total of 110,000 mouse ESTs were obtained from the National Institute of Aging of the National Institutes of Health, and used to identify mouse genes that show similarity to the phrap contigs. The mouse ESTs were used as a blast database and each phrap contig was searched against the mouse database using BLASTN.

5.3 Results

5.3.1 Identification of tandem repeat regions using Tandem Repeat Finder

Tandem repeats (di, tri and tetra nucleotides) were identified in 29 out of the 36 BAC and cosmid clones (Figure 5.2). A detailed list of the *in-silico* identified tandem repeats can be viewed at http://ziggy.sanbi.ac.za/alan/TandemRepeats.html.

5.3.2 PHRAP assembly of chromosome 19 data

The screening for contaminating sequences resulted in the removal of 150 out of the 1188 chromosome 19 fragments that were masked across most of its length barring 10 nucleotides. A total of 118 sequences of the 1188 sequences contained *E.coli* contamination and represented 23 BAC/cosmid clones. Twenty six out of 118 sequences were masked across the entire length of the sequence for *E.coli* contamination. Eleven of the 23 *E. coli*-containing

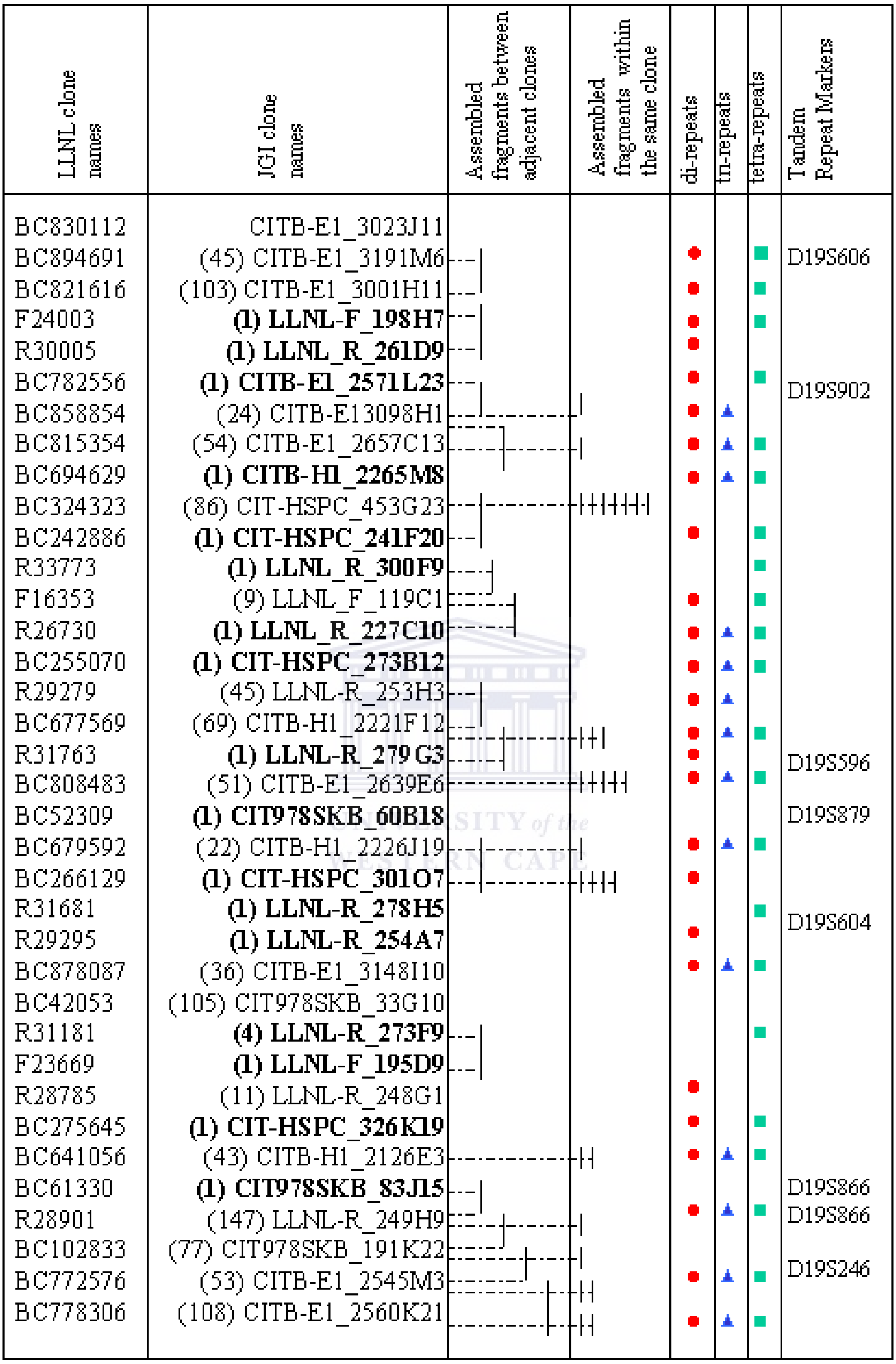
BAC/cosmid clones were “finished” phase sequences and contained a stretch of *E.coli* sequences ranging from 24-119 bases. The presence of *E. coli* contamination was not unexpected, given the disclaimer by genome centers that their draft sequences have not been cleaned from contaminating sequences (FTP site at http://www.jgi.doe.gov/JGI_home.html).

The phrap assembly of 1038 fragments was condensed into 313 contigs and 528 singletons. The contigs represent (i) overlapping fragments from adjacent clones (14), (ii) overlapping fragments from clones that do not map adjacent to each other (errors) (62), (iii) overlapping sequences within the same clone (35) and (iv) PHRAP generated singletons that represent high quality bases (202) (Figure 5.2).

5.3.3 Annotation of the PHRAP assemblies

One hundred and ninety one stack whole-body index sequences matched 243 phrap contigs. A total of 34/191 stack consensus sequences showed heart tissue expression. A total of 113 stack sequences showed significant similarity to genes in the non-redundant DNA database (Table 5.1, Figure 5.3). A search of the protein database identified significant hits for 53/191 stack sequences (Table 5.2).

Twenty-four BodyMap transcripts were identified with significant similarity to 13 *PFHB* contigs. A total of 20 out of the 24 BodyMap transcripts showed gene expression in portions of the heart (Table 5.3). Fifty-two mouse EST matched nine PFHB1 contigs (Table 5.4).

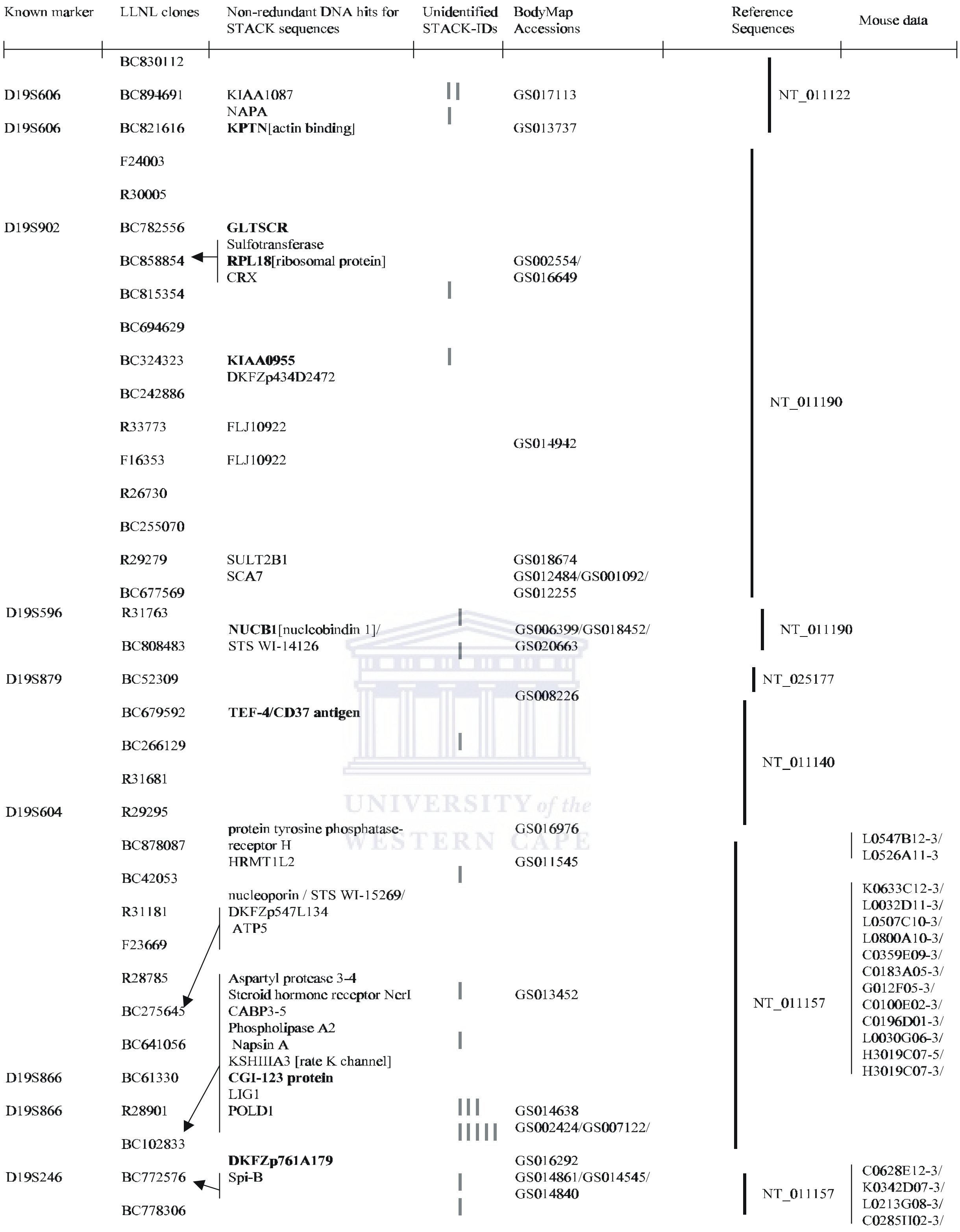


LLNL clone names	JGI clone names	Assembled fragments between adjacent clones	Assembled fragments within the same clone	di-repeats	tri-repeats	tetra-repeats	Tandem Repeat Markers
BC830112	CITB-E1_3023J11						
BC894691	(45) CITB-E1_3191M6			●		■	D19S606
BC821616	(103) CITB-E1_3001H11			●		■	
F24003	**(1) LLNL-F_198H7**			●		■	
R30005	**(1) LLNL_R_261D9**			●			
BC782556	**(1) CITB-E1_2571L23**			●		■	D19S902
BC858854	(24) CITB-E13098H1			●	▲		
BC815354	(54) CITB-E1_2657C13			●	▲	■	
BC694629	**(1) CITB-H1_2265M8**			●	▲	■	
BC324323	(86) CIT-HSPC_453G23						
BC242886	**(1) CIT-HSPC_241F20**			●		■	
R33773	**(1) LLNL_R_300F9**					■	
F16353	(9) LLNL_F_119C1			●		■	
R26730	**(1) LLNL_R_227C10**			●	▲	■	
BC255070	**(1) CIT-HSPC_273B12**			●	▲	■	
R29279	(45) LLNL-R_253H3			●	▲		
BC677569	(69) CITB-H1_2221F12			●	▲	■	
R31763	**(1) LLNL-R_279G3**			●			D19S596
BC808483	(51) CITB-E1_2639E6			●	▲	■	
BC52309	**(1) CIT978SKB_60B18**						D19S879
BC679592	(22) CITB-H1_2226J19			●	▲	■	
BC266129	**(1) CIT-HSPC_301O7**			●			
R31681	**(1) LLNL-R_278H5**					■	D19S604
R29295	**(1) LLNL-R_254A7**			●			
BC878087	(36) CITB-E1_3148I10			●	▲	■	
BC42053	(105) CIT978SKB_33G10						
R31181	**(4) LLNL-R_273F9**					■	
F23669	**(1) LLNL-F_195D9**						
R28785	(11) LLNL-R_248G1			●			
BC275645	**(1) CIT-HSPC_326K19**			●		■	
BC641056	(43) CITB-H1_2126E3			●	▲	■	
BC61330	**(1) CIT978SKB_83J15**						D19S866
R28901	(147) LLNL-R_249H9			●	▲	■	D19S866
BC102833	(77) CIT978SKB_191K22						
BC772576	(53) CITB-E1_2545M3			●	▲	■	D19S246
BC778306	(108) CITB-E1_2560K21			●	▲	■	

Figure 5.2

Figure 5.2 Diagram illustrating the position of identified tandem repeats and PHRAP assembled chromosome 19 fragments relative to the ordered chromosome19 BAC/cosmid clones.

From top to bottom, LLNL clone names are prefixed with “BC”, “R” or “F”. The number of fragments corresponding to each clone are indicated in brackets and precedes the JGI clone names. The JGI clones that represent finished phase sequence are indicated by bold text. All the “finished” clones are represented by one sequence. Vertical dotted lines indicate the clones that were used to produce the assembled data (horizontal bars). PHRAP assembly produced overlapping fragments between adjacent clones (long horizontal bars). The multiple sequences corresponding to the draft clones were assembled into non-overlapping contigs (short horizotal bars). Di-, tri- and tetra- repeats were found in most of the clones and indicated by circles, triangles and squares, respectively.



Known marker	LLNL clones	Non-redundant DNA hits for STACK sequences	Unidentified STACK-IDs	BodyMap Accessions	Reference Sequences	Mouse data
	BC830112					
D19S606	BC894691	KIAA1087 NAPA		GS017113	NT_011122	
D19S606	BC821616	**KPTN**[actin binding]		GS013737		
	F24003					
	R30005					
D19S902	BC782556	**GLTSCR**				
		Sulfotransferase				
	BC858854	**RPL18**[ribosomal protein]		GS002554/		
		CRX		GS016649		
	BC815354					
	BC694629					
	BC324323	**KIAA0955**				
		DKFZp434D2472				
	BC242886				NT_011190	
	R33773	FLJ10922				
				GS014942		
	F16353	FLJ10922				
	R26730					
	BC255070					
	R29279	SULT2B1		GS018674		
		SCA7		GS012484/GS001092/		
	BC677569			GS012255		
D19S596	R31763					
		NUCB1[nucleobindin 1]/		GS006399/GS018452/	NT_011190	
	BC808483	STS WI-14126		GS020663		
D19S879	BC52309				NT_025177	
				GS008226		
	BC679592	**TEF-4/CD37 antigen**				
	BC266129					
	R31681				NT_011140	
D19S604	R29295					
		protein tyrosine phosphatase-		GS016976		
	BC878087	receptor H				L0547B12-3/ L0526A11-3
		HRMT1L2		GS011545		
	BC42053					
		nucleoporin / STS WI-15269/				K0633C12-3/ L0032D11-3/ L0507C10-3/ L0800A10-3/ C0359E09-3/ C0183A05-3/ G012F05-3/ C0100E02-3/ C0196D01-3/ L0030G06-3/ H3019C07-5/ H3019C07-3/
	R31181	DKFZp547L134				
		ATP5				
	F23669					
	R28785	Aspartyl protease 3-4				
		Steroid hormone receptor Ncr1		GS013452		
	BC275645	CABP3-5			NT_011157	
		Phospholipase A2				
	BC641056	Napsin A				
		KSHIIIA3 [rate K channel]				
D19S866	BC61330	**CGI-123 protein**				
		LIG1				
D19S866	R28901	POLD1		GS014638		
				GS002424/GS007122/		
	BC102833					
		DKFZp761A179		GS016292		
D19S246	BC772576	Spi-B		GS014861/GS014545/		C0628E12-3/ K0342D07-3/ L0213G08-3/ C0285H02-3/
				GS014840	NT_011157	
	BC778306					

Figure 5.3 Diagram integrating the positions of genes, STACK sequences (grey vertical bars), BodyMap transcripts, RefSeq sequences (black vertical bars) and mouse ESTs relative to the chromosome 19q13.3 BAC/cosmid mapped clones and known microsatellites. Seven genes that demonstrate heart expression are indicated by bold text. The arrows point a collection of genes to one clone

Table 5.1 Summary table for stack consensus sequence hits to PHRAP contigs and hits for stack sequences to the DNA non-redundant database. BLASTN scores for the PHRAP contigs in square brackets. STACK sequences with heart expression are indicated in bold font.

Phrap Contigs [BLASTN exp score for best hit to stack consensus sequences]	**STACK ID (stack sequences matching the phrap contigs)**	**BLASTN results for STACK sequences matching a DNA non-redundant database (DNA non-redundant database 7 July 2000)**	**Exp score (1e-40 cutoff)**
Contig175[1e-68]	463864	Homo sapiens chromosome 19 clone LLNLR-254A7, complete sequence	6e-69
Contig137[1e-157]	3274	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	1e-158
Contig229[7e-56]	13325	Homo sapiens CD37 antigen (CD37) mRNA	3e-66
Contig229[3e-67]	13323	Homo sapiens CD37 antigen (CD37) mRNA	1e-122
Contig229[5e-66]	13322	Homo sapiens CD37 antigen (CD37) mRNA	0.0
Contig68[3e-80]	122914_1	Human dehydroepiandrosterone sulfotransferase (STD) gene, exon 6 and complete cds	5e-80
Contig66[1e-99] Contig68[1e-139] Contig69[5e-67]	122914_2	Human sulfotransferase-related mRNA sequence	0.0
Contig290[1e-128] Contig137[1e-127]	115913_1	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	1e-128
Contig137[1e-106]	**122927**	Homo sapiens sulfotransferase family 2B, member 1 (SULT2B1) mRNA	0.0
Contig68[1e-133]	122921	Human dehydroepiandrosterone sulfotransferase (STD) gene, exon 6 and complete cds	1e-133
Contig307[0.0]	141012_2	Homo sapiens mRNA; cDNA DKFZp564O0463 (from clone DKFZp564O0463); partial cds	0.0
Contig290[1e-108] Contig137[1e-107]	22725	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	1e-108
Contig290[0.0] Contig137[0.0]	22723	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	0.0
Contig137[1e-47]	12023	Homo sapiens ribosomal protein L18 (RPL18) mRNA	1e-71
Contig290[0.0] Contig137[0.0]	2036	Homo sapiens ataxin-7 (SCA7) gene, partial cds	0.0
Contig137[6e-50]	12003	Homo sapiens ribosomal protein L18 (RPL18) mRNA	1e-83
Contig137[1e-78]	12014	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	2e-79
Contig137[1e-41]	12019	Homo sapiens ribosomal protein L18 (RPL18) mRNA	7e-77
Contig180[1e-148]	**22110**	Homo sapiens mRNA; cDNA DKFZp547L134 (from clone DKFZp547L134); complete cds	0.0
Contig293[0.0]	52299	Expression vector Ad5CMV-p53 tumor suppressor protein p53 (p53) expression cassette, complete sequence	0.0
Contig288[0.0]	32320	Homo sapiens inward rectifier potassium channel (KIR2.4), mRNA	0.0
Contig74[3e-72]	162300	Homo sapiens CaBP3 (CABP3) mRNA, complete cds	1e-163
Contig288[1e-123]	32319	Homo sapiens inward rectifier potassium channel (KIR2.4), mRNA	1e-123
Contig229[1e-130]	13320_1	Homo sapiens CD37 antigen (CD37) mRNA	1e-94
Contig229[1e-134]	**13320_2**	Homo sapiens CD37 antigen (CD37) mRNA	7e-97
Contig229[7e-96]	**121013_3**	H.sapiens mRNA for TEF-4 protein	1e-135
Contig137[7e-65]	**12002**	Homo sapiens ribosomal protein L18 (RPL18) mRNA	0.0
Contig270[7e-89]	113526_2	Homo sapiens protein tyrosine phosphatase, receptor type, H (PTPRH) mRNA	0.0
Contig265[3e-56]	51811	Homo sapiens protein arginine N-methyltransferase 1 (HRMT1L2) gene, complete cds, alternatively spliced	6e-51
Contig180[1e-151]	139723_1	Homo sapiens activating transcription factor 5 (ATF5), mRNA	1e-168
Contig180[0.0]	139723_3	Homo sapiens activating transcription factor 5 (ATF5), mRNA	0.0
Contig227[1e-134]	71906	Homo sapiens mRNA; cDNA DKFZp434D2472 (from clone DKFZp434D2472); partial cds	1e-128
Contig265[1e-	51797	Homo sapiens protein arginine N-methyltransferase 1 (HRMT1L2) gene,	1e-111

111]		complete cds, alternatively spliced	
Contig290[2e-78] Contig137[5e-78]	211147	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	6e-79
Contig212[1e-92]	**140634**	Homo sapiens glioma tumor suppressor candidate region protein 2 (GLTSCR2) mRNA, complete cds	0.0
Contig250[0.0] Contig173[1e-146]	140137_5	Homo sapiens mRNA for KIAA0955 protein, complete cds	0.0
Contig173[3e-91]	140137_1	Homo sapiens mRNA for KIAA0955 protein, complete cds	1e-173
Contig250[1e-179]	140137_2	Homo sapiens mRNA for KIAA0955 protein, complete cds	1e-179
Contig229[4e-51]	150584	Homo sapiens CD37 antigen (CD37) mRNA	1e-117
Contig75[2e-49]	100115	Human steroid hormone receptor Ner-I mRNA, complete cds	1e-103
Contig137[0.0]	30163	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	0.0
Contig175[3e-82]	410148	Homo sapiens chromosome 19 clone LLNLR-254A7, complete sequence	9e-83
Contig273[1e-111] Contig98[1e-103]	130208	H.sapiens CpG island DNA genomic Mse1 fragment, clone 54d4, reverse read cpg54d4.rt1a	1e-65
Contig180[3e-57]	50265	Homo sapiens VRK3 mRNA for vaccinia related kinase 3, complete cds	0.0
Contig254[1e-133]	**140137_6**	Homo sapiens mRNA for KIAA0955 protein, complete cds	1e-105
Contig99[0.0] Contig103[8e-70] Contig141[1e-101]	**160574**	Homo sapiens kaptin (actin-binding protein) (KPTN), mRNA	0.0
Contig267[1e-77]	**133446_3**	Homo sapiens nucleobindin 1 (NUCB1), mRNA	0.0
Contig137[4e-57]	277407	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	5e-58
Contig288[0.0]	47458	Homo sapiens inward rectifier potassium channel (KIR2.4), mRNA	0.0
Contig137[1e-90]	467531	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	5e-90
Contig79[1e-179]	27762	Homo sapiens mRNA; cDNA DKFZp761A179 (from clone DKFZp761A179); partial cds	1e-179
Contig9[6e-56]	107099	Homo sapiens aspartyl protease 3 mRNA, partial cds	4e-56
Contig75[9e-55]	107097	Homo sapiens aspartyl protease 4 mRNA, complete cds	9e-56
Contig75[6e-59]	107103	Homo sapiens aspartyl protease 4 mRNA, complete cds	2e-64
Contig267[6e-86]	**133446_4**	Homo sapiens nucleobindin 1 (NUCB1), mRNA	0.0
Contig75[6e-96]	100056_1	Human steroid hormone receptor Ner-I mRNA, complete cds	1e-126
Contig75[0.0]	**100056_2**	Human steroid hormone receptor Ner-I mRNA, complete cds	0.0
Contig62[1e-131]	156637	Homo sapiens cone rod homeobox protein (CRX) gene, complete cds	1e-131
Contig56[1e-102]	166737	human STS SHGC-30732	1e-154
Contig79[1e-156]	**276003**	Homo sapiens mRNA; cDNA DKFZp761A179 (from clone DKFZp761A179); partial cds	1e-156
Contig180[0.0]	26212	human STS WI-15269	0.0
Contig290[0.0] Contig137[0.0]	16371	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	0.0
Contig44[1e-53] Contig90[1e-44]	**56389**	Homo sapiens CGI-123 protein mRNA, complete cds	0.0
Contig180[0.0]	128364_1	Homo sapiens nucleoporin 62kD (NUP62), mRNA	0.0
Contig180[0.0]	128364_3	Homo sapiens mRNA; cDNA DKFZp547L134 (from clone DKFZp547L134); complete cds	0.0
Contig180[0.0]	128364_2	Homo sapiens mRNA; cDNA DKFZp547L134 (from clone DKFZp547L134); complete cds	0.0
Contig298[0.0]	95925	Homo sapiens mRNA for KIAA1141 protein, partial cds	0.0
Contig29[0.0]	5955	human STS WI-14126	0.0
Contig41[1e-48] Contig56[3e-61]	145671	Homo sapiens CaBP5 (CABP5) mRNA, complete cds	1e-140
Contig251[4e-55] Contig75[3e-54]	85686	Homo sapiens polymerase (DNA directed), delta 1, catalytic subunit (125kD) (POLD1) mRNA	0.0
Contig62[1e-167]	35740	Homo sapiens cone rod homeobox protein (CRX) gene, complete cds	1e-166
Contig265[1e-135]	5024	Homo sapiens protein arginine N-methyltransferase 1 (HRMT1L2) gene, complete cds, alternatively spliced	3e-67
Contig211[1e-163]	315281	Homo sapiens cDNA FLJ10922 fis, clone OVARC1000420	1e-160
Contig266[0.0] Contig274[8e-66] Contig278[0.0] Contig123[2e-65] Contig282[7e-44] Contig34[7e-52]	137874_2	Homo sapiens mRNA for KIAA1087 protein, partial cds	0.0
Contig214[1e-105]	134410	H.sapiens Spi-B mRNA	0.0
Contig29[7e-46]	84599	Homo sapiens branched chain aminotransferase 2, mitochondrial (BCAT2) mRNA	1e-101
Contig180[3e-66]	164712	Homo sapiens sialic acid binding Ig-like lectin 5 (SIGLEC5), mRNA	0.0

Contig59[4e-55]	64091	Rattus rattus K+ channel protein (KSHIIIA3) mRNA, complete cds	7e-87
Contig59[1e-73]	64099	Human Chromosome 11p14.3 PAC clone pDJ1082L12 containing KNCN1 and MyoD, complete sequence [Homo sapiens]	0.0
Contig59[1e-58]	64105	Homo sapiens potassium voltage-gated channel, Shaw-related subfamily, member 4 (KCNC4) mRNA	1e-142
Contig59[6e-42]	64103	Homo sapiens potassium voltage-gated channel, Shaw-related subfamily, member 4 (KCNC4) mRNA	1e-152
Contig137[0.0]	174327	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	0.0
Contig242[2e-83]	**97440_7**	Homo sapiens monocyte/macrophage Ig-related receptor MIR-7 (MIR cl-7) mRNA, complete cds	0.0
Contig242[1e-147]	97440_6	Homo sapiens monocyte/macrophage Ig-related receptor MIR-10 (MIR cl-10) mRNA, complete cds	0.0
Contig9[6e-56] Contig75[1e-173]	107089_1	Homo sapiens napsin A mRNA, complete cds	1e-172
Contig9[2e-71] Contig75[1e-172]	107089_2	Homo sapiens napsin A mRNA, complete cds	0.0
Contig9[1e-159] Contig75[2e-61]	107089_3	Homo sapiens aspartyl protease 3 mRNA, partial cds	0.0
Contig288[2e-82]	**113290_2**	Human cytohesin-2 mRNA, complete cds	0.0
Contig175[0.0]	149850	Homo sapiens chromosome 19 clone LLNLR-254A7, complete sequence	0.0
Contig298[0.0]	95099_1	Homo sapiens mRNA; cDNA DKFZp434N043 (from clone DKFZp434N043); partial cds	0.0
Contig298[0.0]	95099_2	Homo sapiens mRNA; cDNA DKFZp434N043 (from clone DKFZp434N043); partial cds	0.0
Contig173[1e-56]	129590	Homo sapiens mRNA for KIAA0955 protein, complete cds	1e-131
Contig290[1e-163] Contig137[4e-54]	79654	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	1e-163
Contig137[2e-43]	**149711**	Homo sapiens chromosome 19 clone LLNLR-221E8, complete sequence	7e-96
Contig76[2e-53]	49728	Homo sapiens cDNA FLJ10922 fis, clone OVARC1000420	1e-115
Contig62[2e-43]	159738	human STS SHGC-31437	5e-44
Contig180[4e-53]	139726	Homo sapiens activating transcription factor 5 (ATF5), mRNA	2e-54
Contig81[3e-65]	69719	Homo sapiens ligase I, DNA, ATP-dependent (LIG1) mRNA	2e-69
Contig53[0.0] Contig302[0.0]	69718	Homo sapiens ligase I, DNA, ATP-dependent (LIG1) mRNA	0.0
Contig180[2e-61]	139722	Homo sapiens activating transcription factor 5 (ATF5), mRNA	9e-63
Contig229[9e-74]	**279757**	Homo sapiens CD37 antigen (CD37) mRNA	8e-74
Contig265[0.0]	51794_3	Homo sapiens protein arginine N-methyltransferase 1 (HRMT1L2) gene, complete cds, alternatively spliced	0.0
Contig265[1e-80]	51794_2	Homo sapiens HMT1 (hnRNP methyltransferase, S. cerevisiae)-like 2 (HRMT1L2) mRNA	0.0
Contig85[5e-60] Contig89[8e-74] Contig18[2e-53] Contig91[0.0] Contig60[1e-159] Contig70[3e-50]	**159569**	Homo sapiens phospholipase A2, group IVC (cytosolic, calcium-independent) (PLA2G4C) mRNA, and translated products	0.0
Contig211[0.0]	38481	Homo sapiens cDNA FLJ10922 fis, clone OVARC1000420	0.0
Contig211[0.0]	38482	Homo sapiens cDNA FLJ10922 fis, clone OVARC1000420	0.0
Contig265[1e-115]	48756	Homo sapiens protein arginine N-methyltransferase 1 (HRMT1L2) gene, complete cds, alternatively spliced	1e-115
Contig288[0.0]	148393	Homo sapiens inward rectifier potassium channel (KIR2.4), mRNA	0.0
Contig180[2e-55]	128384	Homo sapiens mRNA; cDNA DKFZp547L134 (from clone DKFZp547L134); complete cds	5e-57
Contig273[3e-73]	77580_1	Homo sapiens N-ethylmaleimide-sensitive factor attachment protein, alpha (NAPA), mRNA	2e-71
Contig239[0.0]	77580_3	Homo sapiens N-ethylmaleimide-sensitive factor attachment protein, alpha (NAPA), mRNA	9e-81
Contig273[9e-64]	77580_2	Homo sapiens N-ethylmaleimide-sensitive factor attachment protein, alpha (NAPA), mRNA	1e-117

Table 5.2 Summary table for stack consensus sequence hits to phrap contigs and hits for stack sequences to the protein non-redundant database. BLASTX scores for the phrap contigs in square brackets. STACK sequences with heart expression are indicated in bold font.

Phrap Contigs [BLASTX exp score for best hit to stack consensus sequences]	**STACK ID (stack sequences matching the phrap contigs)**	**BLASTX results (Protein non-redundant database 27April2000)**	**Exp score (1e-20 cutoff)**
Contig137[1e-157]	3274	(AK000207) unnamed protein product [Homo sapiens]	2e-36
Contig229[5e-66]	13322	CD37 antigen sp|P11049|CD37_HUMAN LEUKOCYTE ANTIGEN CD37	1e-22
Contig66[1e-99] Contig68[1e-139] Contig69[5e-67]	122914_2	ALCOHOL SULFOTRANSFERASE (HYDROXYSTEROID SULFOTRANSFERASE) (HST)	1e-159
Contig290[1e-128] Contig137[1e-127]	115913_1	DNA-binding protein TAXREB302 - human T-cell lymphotropic virus type	1e-27
Contig137[1e-106]	**122927**	(U92322) hydroxysteroid sulfotransferase SULT2B1a [Homo sapiens]	1e-165
Contig307[0.0]	141012_2	transposase - Escherichia coli insertion sequence IS10 gb|AAB28848.1| (S67119)	0.0
Contig137[6e-50]	12003	ribosomal protein L18 sp|Q07020|RL18_HUMAN 60S	6e-26
Contig137[1e-41]	12019	ribosomal protein L18 sp|Q07020|RL18_HUMAN 60S	7e-22
Contig293[0.0]	52299	(X01405) p53 [Homo sapiens]	2e-72
Contig288[0.0]	32320	inward rectifier potassium channel gb|AAD51376.1|AF081466_1 (AF081466)	2e-50
Contig74[3e-72]	162300	(AF224511) Ca2+-binding protein CaBP3	8e-34
Contig288[1e-123]	32319	(AJ003065) Kir2.4 protein [Rattus norvegicus]	1e-42
Contig229[7e-96]	**121013_3**	TEA domain family member 2 sp|P48301|TEF4_MOUSE TRANSCRIPTIONAL ENHANCER FACTOR TEF-4	4e-54
Contig137[7e-65]	12002	ribosomal protein L18 sp|Q07020|RL18_HUMAN 60S	1e-104
Contig270[7e-89]	113526_2	protein tyrosine phosphatase, receptor type, H pir||A49724 protein-tyrosine-phosphatase	0.0
Contig212[1e-92]	**140634**	(AF182076) glioma tumor suppressor candidate region protein 2	2e-65
Contig250[0.0] Contig173[1e-146]	140137_5	(AB023172) KIAA0955 protein	0.0
Contig173[3e-91]	140137_1	(AB023172) KIAA0955 protein	8e-59
Contig180[3e-57]	50265	(AB031052) vaccinia related kinase 3	2e-67
Contig99[0.0] Contig103[8e-70] Contig141[1e-101]	**160574**	kaptin (actin-binding protein) gb|AAD39358.1|AF105369_1 (AF105369)	0.0
Contig267[1e-77]	**133446_3**	nucleobindin 1 sp|Q02818|NUBN_HUMAN NUCLEOBINDIN PRECURSOR (NUCB1) gb|AAB60431.1| (U31342	1e-107
Contig79[1e-179]	27762	(AL137451) hypothetical protein	7e-63
Contig9[6e-56]	107099	(AF200344) aspartyl protease 3	1e-24
Contig75[6e-59]	107103	pronapsin A precursor sp|O96009|NAP1_HUMAN NAPSIN 1 PRECURSOR	6e-35
Contig267[6e-86]	133446_4	nucleobindin 1 sp|Q02818|NUBN_HUMAN NUCLEOBINDIN PRECURSOR (NUCB1) gb|AAB60431.1|	0.0
Contig75[0.0]	100056_2	OXYSTEROLS RECEPTOR LXR-BETA (LIVER X RECEPTOR BETA) gb|AAA61783.1| (U07132)	1e-157
Contig307[1e-100]	106641	(AB022023) nonmuscle myosin heavy chain B [Bos taurus]	2e-60
Contig79[1e-156]	**276003**	(AL137451) hypothetical protein	9e-27
Contig44[1e-53] Contig90[1e-44]	56389	ring finger protein 11 (AB024427) Sid1669p [Mus musculus]	1e-71
Contig180[0.0]	128364_1	nucleoporin 62kD sp|P37198|NU62_HUMAN NUCLEAR PORE GLYCOPROTEIN P62 (X58521	1e-152
Contig180[0.0]	128364_3	nucleoporin p62 - human emb|CAB82399.1| (AL162061) hypothetical protein	3e-68
Contig180[0.0]	128364_2	nucleoporin p62 - human emb|CAB82399.1| (AL162061) hypothetical protein	3e-22
Contig251[4e-55] Contig75[3e-54]	85686	DNA-directed DNA polymerase delta - human gb|AAA35768.1| (M81735)	2e-81
Contig266[0.0] Contig274[8e-66] Contig278[0.0] Contig123[2e-65] Contig282[7e-44] Contig34[7e-52]	137874_2	(AB029010) KIAA1087 protein [Homo sapiens]	0.0
Contig214[1e-105]	134410	TRANSCRIPTION FACTOR SPI-B pir||S25655 Spi-B protein - human emb|CAA46878.1| (X66079)	1e-128
Contig29[7e-46]	84599	branched chain aminotransferase 2, mitochondrial sp|O15382|BCAM_HUMAN	4e-31
Contig180[3e-66]	164712	sialic acid binding Ig-like lectin 5 gb|AAB70703.1| (U71383)	0.0
Contig59[4e-55]	64091	VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV3.2 (KSHIIIA)	4e-32

		gb|AAA41819.1| (M59211) potassium channel Kv3.2b [Rattus norvegicus] gb|AAA42143.1| (M84203)	
Contig59[1e-73]	64099	potassium channel gene 1 (alternative splicng product described in Luneau et al 1991) sp|P25122|CIKD_RAT VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV3.1 (KV4) (NGK2) (RAW2) gb|AAA41501.1| (M68880)	5e-82
Contig59[1e-58]	64105	potassium voltage-gated channel, Shaw-related subfamily, member 4 sp|Q03721|CIKG_HUMAN VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV3.4 (KSHIIIC) gb|AAA57263.1| (M64676)	3e-23
Contig242[2e-83]	97440_7	(AF004230) MIR-7 [Homo sapiens]	0.0
Contig242[1e-147]	97440_6	(AF009637) immunoglobulin-like transcript 5 protein [Homo sapiens]	1e-130
Contig9[6e-56] Contig75[1e-173]	107089_1	pronapsin A precursor sp|O96009|NAP1_HUMAN NAPSIN 1 PRECURSOR (NAPSIN A) (NAPA) (TA01/TA02) gb|AAD04917.1| (AF090386)	6e-58
Contig9[2e-71] Contig75[1e-172]	107089_2	pronapsin A precursor sp|O96009|NAP1_HUMAN NAPSIN 1 PRECURSOR (NAPSIN A) (NAPA) (TA01/TA02) gb|AAD04917.1| (AF090386) napsin A	0.0
Contig9[1e-159] Contig75[2e-61]	107089_3	pronapsin A precursor sp|O96009|NAP1_HUMAN NAPSIN 1 PRECURSOR (NAPSIN A) (NAPA) (TA01/TA02) gb|AAD04917.1| (AF090386)	0.0
Contig288[2e-82]	**113290_2**	(U70728) cytohesin-2 [Homo sapiens]	1e-167
Contig298[0.0]	95099_2	hypothetical protein DKFZp434N043.1 - human (fragment) emb|CAB45736.1| (AL080143) hypothetical protein	0.0
Contig173[1e-56]	129590	(AB023172) KIAA0955 protein	8e-41
Contig53[0.0] Contig302[0.0]	69718	DNA ligase I sp|P18858|DNL1_HUMAN DNA LIGASE I gb|AAA59518.1| (M36067)	4e-66
Contig265[1e-80]	51794_2	(AF232716) protein arginine N-methyltransferase 1 [Mus musculus]	3e-64
Contig85[5e-60] Contig89[8e-74] Contig18[2e-53] Contig91[0.0] Contig60[1e-159] Contig70[3e-50]	159569	phospholipase A2, group IVC (cytosolic) gb|AAC32823.1| (AF058921) cytosolic phospholipase A2-gamma	0.0

Table 5.3 Tissue expression information for all BodyMap accessions that match the *PFHB1* contigs

BodyMap accessions	Phrap Contigs	Expression information associated with each BodyMap accession																				
		Adipose	astrocyte	CD8_Tcell	colon	colon_epithel	conjunctiva	cornea	cortex	fetal_brain	fibro	GenBank	ileum_end	intest_metaplasia	iris	liver	papilla_pilli	pmn	reaming_bone	skeletal_muscle	substantia_nigra	ventrical_muscle
GS006399	CITB-E1_2639E6#46										X	X	X								X	
GS002424	pfhb_seq.Contig75		X			X			X		X	X	X		X	X					X	
GS011545	pfhb_seq.Contig265		X			X			X		X	X	X		X	X				X	X	
GS008226	CITB-H1_2226J19#22		X			X		X	X		X	X	X		X	X				X	X	
GS014942	pfhb_seq.Contig211		X			X		X	X		X	X	X			X				X	X	X
GS001092	pfhb_seq.Contig290		X			X		X	X	X	X	X	X		X	X		X		X	X	X
GS007122	pfhb_seq.Contig91		X			X		X	X	X	X	X	X		X	X		X		X	X	X
GS013452	pfhb_seq.Contig180					X		X	X	X	X	X	X		X	X	X	X		X		X
GS014638	pfhb_seq.Contig301		X			X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
GS014840	CITB-E1_2545M3#49		X			X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
GS002554	pfhb_seq.Contig68		X			X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
GS004097	pfhb_seq.Contig288		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
GS012255	pfhb_seq.Contig290		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
GS012484	pfhb_seq.Contig290		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS013737	pfhb_seq.Contig99	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS014545	CITB-E1_2545M3#43	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS014861	CITB-E1_2545M3#43	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS016292	Pfhb_seq.Contig75	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS016649	Pfhb_seq.Contig68	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS016976	CITB-E1_3148I10#31	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS017113	Pfhb_seq.Contig239	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS018542	CITB-E1_2639E6#46	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS018674	Pfhb_seq.Contig137	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GS020663	CITB-E1_2639E6#46	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 5.4 Significant similarity hits to a non-redundant DNA database for mouse ESTs that match to the *PFHB1* contigs.

Mouse ID	**PFHB contig**	**BLASTN non-redundant DNA search (27th April 2000)**	**Exp. score**
H3057G10-5	pfhb_seq.Contig288	gi|3885502|gb|AF079971.1|AF079971 Mus musculus cytohesin-2 mRNA,	0.0
H3057G10-3	pfhb_seq.Contig288	gi|3885502|gb|AF079971.1|AF079971 Mus musculus cytohesin-2 mRNA,	0.0
H3057H11-5	pfhb_seq.Contig75	gi|6678506|ref|NM_009473.1|| Mus musculus nuclear receptor	0.0
H3057H11-3	pfhb_seq.Contig75	gi|6678506|ref|NM_009473.1|| Mus musculus nuclear receptor subfamily	0.0
L0032D11-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0628E12-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
K0633C12-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
L0547B12-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
L0507C10-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	e-167
L0526A11-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
L0800A10-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
K0342D07-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
L0213G08-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
L0292F04-3	pfhb_seq.Contig180	gi|558040|gb|S71575.1|S71575 ADS39 [mice, DDS, androgen-dependent	0.0
C0285H02-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
C0254F01-3	pfhb_seq.Contig180	gi|558040|gb|S71575.1|S71575 ADS39 [mice, DDS, androgen-dependent	0.0
C0359E09-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0183A05-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
G0112F05-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0100E02-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0114F03-3	pfhb_seq.Contig180	gi|558040|gb|S71575.1|S71575 ADS39 [mice, DDS, androgen-dependent	0.0
C0196D01-3	CIT978SKB_33G10#59	gi|6754695|ref|NM_010798.1|| Mus musculus macrophage migration	0.0
H3116G01-5	pfhb_seq.Contig180	gi|236260|gb|S59342.1|S59342 nuclear pore complex glycoprotein p62	0.0
H3004H06-5	pfhb_seq.Contig265	gi|7141327|gb|AF232717.1|AF232717 Mus musculus protein arginine	0.0
H3004H06-3	pfhb_seq.Contig265	gi|7141327|gb|AF232717.1|AF232717 Mus musculus protein arginine	0.0
L0030G06-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
H3098E07-5	pfhb_seq.Contig307	gi|6752235|emb|AL133224.2|CNS01DU9 Human	0.0

		chromosome 14 DNA sequence ***	
H3019C07-5	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
H3019C07-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0029D04-5	pfhb_seq.Contig265	gi|7141327|gb|AF232717.1|AF232717 Mus musculus protein arginine	0.0
H3075C01-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
C0674E08-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
K0647D06-3	pfhb_seq.Contig180	gi|558040|gb|S71575.1|S71575 ADS39 [mice, DDS, androgen-dependent	0.0
C0653F07-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0602C04-3	pfhb_seq.Contig75	gi|6678506|ref|NM_009473.1|| Mus musculus nuclear receptor subfamily	0.0
C0666D05-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
C0674C05-3	CITB-E1_2545M3#47	gi|438133|emb|Z21848.1|MMDPDCS M.musculus mRNA for DNA-polymerase	0.0
C0651E01-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
L0604G01-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
L0508F07-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
L0531E04-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	e-152
L0506A05-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
L0509D02-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
L0549B03-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
J0543C01-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	e-177
L0506A01-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
H0520A01-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	e-121
L0538C01-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0
C0459F06-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	0.0
L0405E03-3	CIT978SKB_33G10#59	gi|193641|gb|L02913.1|MUSGROEARA Mouse growth factor-induced	e-169
C0932C01-3	CITB-E1_3148I10#30	gi|6678392|ref|NM_009406.1|| Mus musculus troponin I, cardiac	0.0

Table 5.5 Summary of candidate genes identified through annotation of STACK transcripts that were mapped to the *PFHB1* disease locus.

Candidate genes classified by functional classes	
Enzymes Napsin A (proteinase) Sulfotransferase SULT2B1 LIG1 (ligase) phospholipase A2 Aspartyl protease 3-4 HRMT1L2 (methyl transferase) POLD1 (DNA polymerase	**Unknown mRNA** KIAA1087 **GLTSCR2** KIAA0955 FLJ10922 DKF2P547434 **DKF2P761A179** DKF2P434D2492
Transcription Factors Activating transcription factor 5 (ATF-5) **T- elongation factor 4 (TEF-4)** Cone rod homeobox gene (CRX) Spi-B (expressed in lymphoid cells)	**Binding proteins** Calcium binding protein (CABP3-5) **KAPTIN (KPTN, actin-binding)** **nucleobindin (NUCB1)** **CGI-123 protein**
Channels Rat potassium channel Nucleoporin	**Receptors** Steroid receptor (Ner1) Phospho-tyrosine phospho-tase receptor H2 (PTPRH2)
Immune system **CD37-antigen**	

5.4 Discussion

Tandem Repeats

The informativeness of the tandem repeats increases with increasing average number of repeats (Weber 1990). The polymorphic information content (PIC) is used as a measure of the variability/informativeness for a microsatellite and is reported between 0 and 1.0 (Strachan and Read 1997). Weber (1990) reported that tandem repeats as short as 10 repeating units and lower had PIC values of zero whereas tandem repeating units of 24 and greater had PIC values as high as 0.8. Therefore, in the present study, a minimum of 12 repeating units was used for the screening of the chromosome 19 BAC/cosmid fragments. Tandem repeats were identified in 29 out of 36 BAC/cosmid clones. The length of the identified tandem repeats ranged between 12 and 58 for di-nucleotides, 16-76 for tri-nucleotides and 8-171 for tetra-nucleotides. Draft phase sequence accounts for 19 out of the 29 clones for which tandem repeats were identified. The remainder ten clones represent finished phase sequence therefore represent an accurate, valuable source of potential novel microsatellites for saturating the *PFHB1* disease-gene region.

PHRAP assembly

The draft and finish phase sequence for chromosome 19 has not been free of *E. coli* contamination and users are warned about the quality of the sequence data when accessing the JGI ftp site (http://www.jgi.doe.gov). One hundred and eighteen out of the 1184 BAC/cosmid fragments contained some form of *E. coli* contamination. These *E. coli*-containing sequences accounted for 11 BAC/cosmid clones that were classified as "finished phase" clones by the sequence center. Masking across the entire sequence length of a chromosome 19 fragment occurred for 12.6% of the sequences.

The remainder of the sequences were assembled into 310 contigs and 874 singletons. The 36 BAC/cosmid clones across the *PFHB1* locus represent overlapping clones (Ashworth pers. comm). The absence of overlapping contigs across the entire *PFHB1* locus together with the large amount of singletons suggest that the draft sequence for the *PFHB1* locus is incomplete.

Four reference sequences have been generated for the *PFHB1* region but on closer examination it is clear that a number of gaps exists in these reference sequence data (Figure 3). The combined PHRAP assembly and reference sequences (NCBI 14th January 2001) represent the most complete status on the sequence coverage for the *PFHB1* region.

Chromosome 19 has been reported to be a rich source of repeats (Ashworth et al., 1995; Puttagunta et al., 2000; Ashworth pers. comm). The prevalence of repeats was observed with the erroneous assembly of 57 contigs. The reference sequences obtained from NCBI had to be further processed to remove any genomic segments that did not follow the same order as that which appears on the chromosome 19 map at LLNL. For example, NT_011157 shows clone CTB-33G10 (AC011495) followed by a portion of clone CTD-2560K21 (AC008743). This chromosome map order is incorrect as there at least 9 other clones placed between CTb-33G10 and CTD-2560K21 (Figure 5.3).

The recent publication of the draft sequence for the human genome has documented the assembly process using the GigAssembler (Genome Consortium 2001). GigAssembler has been used at the Santa Cruz Genome Center to assemble each of the human chromosome genomic sequence fragments (http://genome.ucsc.edu/). An assembly of chromosome 19 using the GigAssembler has generated a reference sequence interspersed with regions of unknown nucleotides (i.e., stretches of N's). This illustrates the gaps that have yet to be filled in the chromosome 19 sequencing effort at JGI. However, the chromosome 19 scaffold generated at the Santa Cruz Genome Center has been integrated with EST and mRNA data available for chromosome 19 and represent a framework on which to validate the assembly across the *PFHB1* region.

Candidate genes

PFHB1 is a cardiac conduction disorder probably of the bundle of His and the bundle branches (Brink and Torrington 1977; Brink et al., 1995). Clinical features of PFHB1 include right bundle branch block, and/or left anterior hemiblock and complete heart block. Evidence suggests that familial bundle branch diseases similar to PFHB1 do occur, although identified under different names (Mosetti 1954; Trivella et al., 1960; Steenkamp et al., 1973; Vallianos et al., 1974; Stephan 1979 and van der Merwe 1988). The identification of genes that are expressed in the conduction system would provide plausible candidates for *PFHB1*. Recently, Nguyên-Tran et al. (2000) elucidated a novel genetic pathway for sudden cardiac death via defects in the transition between ventricular and conduction system cell lineages. Using a knockin of *lacZ* into the endogenous *HF1-b* locus, Nguyên-Tran et al. (2000) demonstrated that *HF1-b* displayed a restricted pattern of expression within the ventricular chamber and was preferentially expressed in conduction system lineages, including the atrioventricular (AV) node, atrioventricular ring, branching bundles in the interventricular node and the distal His-Purkinje fibres. *HF1-b* deficient mice displayed an increased incidence of postnatal mortality. The sudden death of *HF1-b* mutant mice was examined using implanted radio telemetry to monitor electrocardiographical data from wild type and mutant mice. No arrhythmias or conduction abnormalities were observed in the wild type mice. However, conduction defects were observed in the mutant mice that indicated physiological dysfunction at all levels of the conduction system including, ie., sino-atrial node and AV node, His bundle and the distal Purkinje fibres. *HF1-b* and genes involved in the HF1-b-associated pathway therefore serves as plausible candidate genes for conduction disorders based on the anatomical localisation of HF1-b and its association with malfunctioning of the conduction system.

A search for the human homolog of *HF-1b* was performed in the SWISSPROT protein database, where the map position of the HF1-b encoding gene is recorded as chromosome 7. No genes similar to the *HF-1b* were identified in the BAC/cosmid fragments that map to the *PFHB1* region. The absence of overlapping fragments covering the *PFHB1* region suggests that a thorough screen of chromosome 19 BAC and cosmid libraries at the wetbench is essential for determining the presence or absence of an *HF-1b*-like transcription factor within the *PFHB1* locus. Additional evidence that warrants the investigation of transcription factors as causative agents for PFHB1 has been provided recently by Jimenez-Sanchez et al. (2001). In this study, Jimenez-Sanchez et al (2001) compared phenotypes of known diseases to their corresponding disease genes and found that, for a subset of 1000 genes, autosomal dominant diseases were associated with genes that encode transcription factors.

The connexin family represent candidate genes that have not been examined in this study. Increasing evidence support the role for altered connexin distribution in arrhythmogenesis (Kirchoff et al., 1998; Simon et al., 1998; van der Velden et al., 1998). Connexin 40 (Cx-40) has been shown to be a sensitive marker for central and peripheral conduction system in the murine heart (Gourdie et al., 1993; Delorme et al., 1995). Nguyên-Tran et al. (2000) demonstrated distinct differences in the cellular distribution of Cx-40-containing gap junctional plaques between wild type mice compared to *HF1-b* mutant mice, particularly in the distal Purkinje cells. In addition, the distribution of Cx-40 within the cells themselves was altered in the *HF1-b* mutant hearts. Normally Cx-40 is redistributed from the cytosol to the cell membrane during early postnatal development representing formation of functional gap junction plaques at the cell membrane (Litchenberg et al., 2000). *HF1-b* mutant mice showed significantly fewer Cx40-positive staining at the cell borders and more random distribution of Cx-40 compared to wildtype mice. Future studies in search for the *PFHB1* causative gene should include the screening for connexin genes.

The zinc finger protein family (ZFP) contains many of the currently known transcription factors (Dai and Liew 1999). Seven types of *ZFPs* have been described in a human heart EST database (Dai and Liew 1998). The function of most ZFPs have not been completely characterised but some have been implicated in cardiac developmental or pathological processes (Molkentin et al., 1994, 1997, 1998; Hasegawa et al., 1997; Dai and Liew 1998; Margolin et al., 1994; Witzgall et al., 1994; Vissing et al., 1995; Mendelsohn et al., 1994; Arber et al., 1997). Dai and Liew (1999) reported on the enrichment of chromosome 19 for ZFP-encoding genes and mapped a total of 6 out of 126 cardiovascular-based *ZFPs* (cvbZFPs) to chromosome 19p (Dai and Liew 1999; GenBank: R98367, X82125, U52096, M99593, U37263 and M63625). However, none of the six chromosome 19 *ZFP* genes showed identity to the chromosome 19 BAC/cosmid fragments when screened with SIM4 (data not shown). The absence of similar *ZFP* genes from the *PFHB1* region does not exclude them as candidates because the region covered by the BAC/cosmid sequenced clones is incomplete. Stronger supporting evidence for *ZFPs* as candidate genes by position as well as function could be saught after by probing BAC/cosmid libraries using the six chromosome 19 *ZFPs* as probes.

Kaptin (KPTN), glioma tumor suppressor candidate region protein 2 (GLTSCR2), KIAA0955, transcriptional enhancer factor (TEF-4), CD37, nucleobindin1, Steroid hormone receptor Ner1, cyclic AMP-dependent transcription factor (ATP5), aspartyl protease 3 and 4 and protein tyrosine phosphatase receptor type H (PTPRH) were identified through stack consensus sequence BLAST searches of the DNA non-redundant database. Seven of the 11 genes mentioned above show identity to heart expressed sequences in STACK (bold text Figure 5.3). Kaptin, an actin-binding protein, was isolated originally from platelets and recently localised to the tips of elongating stereocilium found in the embryonic inner ear and correspond to sites of actin polymeristaion (Brearer and Abraham 1999). Additional evidence is required to exclude this gene as a plausible candidate, including absence of expression in cardiac tissue.

Twenty-four BodyMap transcripts had significant identity with the *PFHB1* genome fragments and twenty of the twenty-four BodyMap transcripts code for genes that are expressed in the heart (Table 5.4).

In summary, this work-in-progress presents a list of STACK consensus sequences, BodyMap transcripts, mouse ESTs and known genes that have been extracted and placed relative to the BAC/cosmid clones and the genetic markers. The integration of information, in this study, from a range of resources provide the platform for further study in order to reduce the *PFHB1* candidate gene list. Functional assignments of the identified candidate transcripts provide an immediate way forward in reducing the *PFHB1* candidate genes.

5.5 References

Arber S., Hunter J.J., Ross J., Hongo M., Sansig G., Borg J., Perriard J.C., Chien K.R. and Caroni P. (1997) MLP-deficient mice exhibit a disruption of cardiac cyto-architectural organisation, dilated cardiomyopathy and heart failure. *Cell* **88:** 393-403.

Arief Z., Christofffels, A.G., Hide W and Corfield V.A (1999) Using software trapping to identify PFHB1 candidate genes.

Arzimanoglou I.I., Gilbert F. and Barber H.R.K. (1998) Microsatellite instability in human solid tumors. *Cancer* **82:** 1808-1820.

Ashworth L.K., Batzer M.A., Brandriff B., Branscomb E., de Jong P., Garcia E., Garnes J.A., Gordon L.A., Lamerdin J.E., Lennon G., Mohrenweiser H., Olsen A.S., Slezak T. and Carrano A.V. (1995) An integrated metric physical map of human chromosome 19. *Nat. Genet.* **11:** 422-427.

Bates G. and Lehrach H. (1994) Trinucleotide repeat expansions and human genetic disease. *BioEssays.* **16:** 277-284.

Beckman J.S and Weber J.L. (1992) Survey of human and rat microsatellites. *Genomics* 12: 627-631.

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573-580.

Brearer E.L. and Abraham M.T. (1999) 2E4 (kaptin): a novel actin-associated protein from human blood platelets found in lamellipodia and the tips of the stereocilia of the inner ear. *Eur J Cell Biol*. **78:** 117-126.

Brink A.J. and Torrington M. (1977) Progressive familial heart block-Two types. *S. Afr. Med. J*. **52:** 53-59.

Brink P. A., Ferreirra A., Moolman J.C., Weymar H.W., van der Merwe P. and Corfield V. (1995) Gene for progressive familial heart block type-1 maps to chromosome 19q13. *Circulation* **91:** 1633-1640.

Chen A.H., Ni L., Fukushima K., Marietta J., O'Neill M., Couke P., Willems P. and Smith R.J.H. (1995) Linkage of a gene for dominant non-syndromic deafness to chromosome 19. *Hum. Mol. Genet.* **4:** 1073-1076.

Dai K-S.and Liew C-C. (1998) Characterisation of a novel gene encoding zinc finger domains identified from expressed sequence tags of a human heart cDNA database. *J Mol Cell Cardiol*. **30:** 2365-2375.

Dai K-S. and Liew C-C. (1999) Chromosomal, *in silico* and *in vitro* expressin analysis of cardiovascular-based genes encoding zinc finger proteins. *J Mol Cell Cardiol.* **31:** 1749-1769.

Delorme B., Dahl E., Jarry-Guichard T., Marics I., Briand J-P., Willecke K., Gros D. and Theveniau-Ruissy M. (1995) Developmental regulation of connexin 40 gene expression in mouse heart correlates with the differentiation of the conduction system. *Dev. Dynam.* **204:** 358-371.

De Meeus A., Stephan E., Debrus S., Jean M-K., Loiselet J., Weissenbach J., Demaille J and Bouvagnet P. (1995) An isolated cardiac conduction disease maps to chromosome 19q. *Circulation Res.* **77:** 735--140**.**

Donis-Keller H., Green P., Helms C., Cartinhour S., Weiffenbach B., Stephens K., Keith T.P., Bowden D.W., Smith D.R., Lander ES. et al. (1987) A genetic linkage map of the human genome. *Cell* **51:** 319-337.

Emery A.E., Smith C.A. and Sanger R. (1969) The linkage relations of the loci for benign (Becker type) X-borne muscular dystrophy, colour blindness and the Xg blood groups. *Ann Hum Genet.* **32:** 261-269.

Gastier J.M., Pulido J.C., Sunden S., Brody T., Buetow K.H., Murray J.C., Weber J.L., Hudson T.J., Sheffield V.C. and Duyk G. (1995) Survey of trinucleotide repeats in the human genome: assessment of their utility as genetic markers. *Hum. Mol. Genet.* **4:** 1829-1836.

Gayapay G., Morrissette J., Vignal A., Dib C., Fizames C., Millasseau P., Marc S., Bernadi G., Lathrop M and Weissenbach J. (1994) The 1993-1994 Genethon human genetic linkage map. *Molecular Cell Biology.* 6: 3010-3013.

Ghosh T. (1977) Blood group genotypes in Zambia and linkage to sickle cell disease. *Med J Zambia* **11:** 87-89.

Gourdie R.G., Stevens N.J., Green C.R., Rothery S., Germroth P and Thompson R.P. (1993) The spatial distribution and relative abundance of gap-junctional connexin40 and connexin43 correlate to functional properties of components of the cardiac atrioventricular conduction system. *J. Cell Sci.* **105:** 985-991.

Hamada H., Petrino M.G and Takunga T. (1982) A novel repeated element with Z-DNA forming potential is widely found in evolutionary diverse eukaryotic genomes. *Proc. Natl. Acad. Sci. USA.* **79:** 6465-6469.

Hamshere M., Cross ., Daniels M., Lennon G. and Brook J.D. (2000) A transcript map of a 10-Mb region of chromosome 19: A source of genes for human disorders, including candidates for genes involved in asthma, heart defects and eye disorders. *Genomics* **63:** 425-429.

Hasegawa K., Lee S.J., Jobe S.M., Markham B.E. and Kitsis R.N. (1997) Cis-acting sequences that mediate induction of beta-myosin heavy chain gene expression during left ventricular hypertrophy due to aortic constriction. *Circulation* **96:** 3943-3953.

Hill S.Y., Goodwin D.W., Cadoret R., Osterland C.K. and Doner S.M. (1975) Association and linkage between alcoholism and eleven serological markers. *J Stud Alcohol*. **36:** 981-982.

Jeang K-T. and Hayward G.S. (1983) A cytomegalovirus DNA sequence containing tracts of tandemly repeated CA dinucleotide hybridizes to highly dispersed elements in mammalian cell genomes. *Molecular Cell Biology* 3: 1389-1402.

Jeffreys A.J., Wilson V. and Thein S.L. (1985) Hypervariability 'minisatellite' regions in human DNA. *Nature* **314:** 67-73.

Jimmenez-Sanchez G., Childs B. and Valle D. (2001) Human disease genes. *Nature* **409:** 853-855.

Johnson G.F., Hunt G.E., Robertson S. and Doran T.J. (1981) A linkage study of manic-depressive disorder with HLA antigens, blood groups, serum proteins and red cell enzymes. *J Affect Disord.* **3:** 43-58.

Jurka J. (1998) Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* 8:333-337. http://charon.girinst.org/~server/repbase.html.

Keating M.(1992) Linkage analysis and long QT syndrome. Using genetics to study cardiovascular disease. *Circulation* 85: 1973-1986.

Kirchoff S., Nelles E., Hagendorff A., Kruger L., Traub O. and Willecke K. (1998) Reduced cardiac conduction velocity and predisposition to arrhythmias in connexin 40-deficient mice. *Curr. Biol.* **8:** 299-302.

Lee H., Choi E., Seomun Y., Montgomery K., Huebner A., Lee E., Lau S., Joo C-K., Kucherlapati R. and Yoon S-J. (2000) High-resolution transcript map of the region spanning

D12S1629 and D12S312 at chromosome 12q13: Triple A syndrome-linked region. *Genome Res*. **10:** 1561-1567.

Litchenberg W.H., Norman L.W., Holwell A.K., Martin K.L., Hewett K.W. and Gourdie R.G.(2000) The rate and anisotropy of impulse propagation in the postnatal terminal crest are correlated with remodeling of Cx43 gap junction pattern. *Cardiovasc. Res.* **45:** 379-387.

Margolin J.F., Friedman J.R., Meyer K.H., Vissing H., Thiesen H.J. and Rauscher F.R. (1994) Kruppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci USA*. **91:** 4509-4513.

Martinelli M., Scapoli L., Pezzetti F., Carcini F., Carinci P., Baciliero U., Padula E. and Tognon M (1998) Suggestive linkage between markers on chromosome 19q13.2 and nonsyndromic orofacial cleft malformation. *Genomics* **51:** 177-181.

McGee T.L., Devoto M., Ott J., Berson E.L., Dryja T.P. (1997) Evidence that the penetrance of mutations at the *RP11* locus causing dominant retinitis pigmentosa is influenced by a gene linked to the homologous *RP11* allele. *Am. J. Hum. Genet.* **61:** 1059-1066.

Mendelsohn C., Lohnes D., Decimo D., Lufkin T., LeMeur M., Chambon P. and Mark M. (1994) Function of the retinoic acid receptors (RARs) during development(II). Mltiple abnormalities at various stages of organogenesis in RAR double mutants. *Development* **120:** 2749-2771.

Molkentin J.D., Kalvakolanu D.V. and Markham B.E. (1994) Transcription factor GATA-4 regulates cardiac muscle-specific expression of the alpha-myosin heavy-chain gene. *Mol Cell Biol*. **14:** 4947-4957.

Molkentin J.D., Lu J.R., Antos C.L., Markham B., Richardson J., Robbins J., Grant S.R. and Olson E.N. (1998) A calcineurin dependent transcriptional pathway for cardiac hypertrophy. *Cell* **93:** 215-228.

Mosetti A. (1954) Blocco di branca familiare. *Folia cardiol.* **13:** 527-534.

Mott R. (1997) EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS* **13:** 477-478.

Mullis K.B. and Faloona F.A. (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155:** 335-350.

Nakamura Y., Leppert M., O’Connell P., Wolff R., Holm T., Culver M., Martin C., Fujimoto E., Hoff M., Kumlim E. and White R. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235:** 1616-1622.

Nguyên-Tran Van T.B., Kubalak S.W., Minamisawa S., Fiset C., Wollert K.C., Brown A.B., Ruiz-Lozano P., Barrere-Lemaire S., Kondo R., Norman L.W., Gourdie R.G., Rahme M.M., Feld G.K., Clark R.B., Giles W.R. and Chien K.R. (2000) A novel pathway for sudden cardiac death via defects in the transition between ventricular and conduction system cell lineages. *Cell* **102:** 671-682.

Ober C., Cox N.J., Abney M., Rienzo A., Lander E.S., Changyaleket B., Gidley H., Kurtz B., Lee J., Nance M., Pettersson A et al. (1998) Genome-wide search for asthma susceptibility loci in a founder population. *Hum. Mol. Genet.* **7:** 1393-1398.

Pallavicini A., ZimbelloR., Tiso N., Muraro T., Rampoldi L., Bortoluzzi S., Valle G., Lanfranchi G. and Danieli G.A. (1997) The preliminary transcript map of a human skeletal muscle. *Hum. Mol. Genet.* **6:** 1445-1450.

Puttagunta R., Gordon L.A., Meyer G.E., Kapfhamer D., Lamerdin J.E., Kantheti P., Portman K.M., Chung W.K., Jenne D.E., Olsen A.S. and Burmeister M. (2000) Comparative maps of human 19q13.3 and mouse chromosome 10 allows identification of sequences at evolutionary breakpoints. *Genome Res.* **10:** 1369-1380.

Reddy P.S. And Housman D.E. (1997) The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.* **9:** 364-372.

Saiki R.K., Scharf ., Faloona F., Mullis K.B., Horn G.T., Erlich H.A. and Arnheim N. (1985) Enzymatic amplification of β-globin genomic sequences and rstriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350-1354.

Simon A.M., Goodenough D.A. and Paul D.L. (1998) Mice lacking connexin 40 have cardiac conduction abnormalities characteristic of atrioventricular block and bundle branch block. *Curr. Biol.* **8:** 295-298.

Smit A. and Green P. (1999) http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl.

Stallings R.L. (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implication for human genetic diseases. *Genomics* **21:** 116-121.

Steenkamp W.F.J. (1972) Familial trifascicular block. *Am. Heart J.* **84:** 758-760.

Stephan E. (1979) Hereditary bundle branch system defect. A new genetic entity? *Am. Heart J.* **97:** 708-718.

Tautz D. and Renz M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res*. 12: 4127-4138.

Tautz D. (1989) Hypervariability of simple sequences as a general source of polymorphic DNA markers. *Nucleic Acids Res.* 17(16): 6463-6471.

Terwillinger J.D. and Ott J. (1994) *In* “Handbook of Human Genetic Linkage”. Pp.224-225, Johns Hopkins Univ. Press Baltimore/London.

Tórth G., Gáspári Z. and Jurka J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.***10:** 967-981.

Trivella P. and Chella E.L. (1960) Blocco di branca destra familiare e congenito. *Minerva cadioangiol.* **8:** 186-188.

Vallianos G. and Sideris D.A. (1974) Familial conduction defects. *Cardiology* **59:** 190-197.

van der Merwe P.L., Weymar H.W., Torrington and Brink A.J. (1986) Progressive familial heart block. Part II clinical and ECG conformation of progression-report on 4 cases. *S. Afr. Med. J.* **70:** 356-357.

van der Merwe P-L. (1988) Die identifikasie can hoe-risiko pasiente met latente siekte in families met tipe I progressiewe familiele hartblok. Ph.D. Thesis, University of Stellenbosch, South Africa.

van der Merwe P.L., Weymar H.W., Torrington and Brink A.J. (1988) Progressive familial heart block (type 1) A follow-up study after 10 years, *S. Afr. Med. J.* **73:** 275-276.

Van der Velden H.M., van Kempen M.J., Wijffels M.C., can Zijverden M., Groenewegen W.A., Allessie M.A. and Jongsma H.J. (1998) Altered pattern of connexin40 distribution in persistent atrial fibrillation in the goat. *J. Cardiovas. Electrophysiol.* **9:** 596-607.

Vissing H., Meyer W.K., Aagaard L., Tommerup N. and Thiesen H.J. (1995) Repression of transcriptional activity by heterologous KRAB domains present in zinc finger proteins. *FEBS Lett*. **369:** 153-157.

Wang C-Y., Shi J-D., Huang Y-Q., Cruz P.E., Ochoa B., Hawkins-Lee B., Davoodi-Semiromi A. and She J-X. (1999) Construction of a physical and transcript map for a 1-Mb genomic region containing the urofacial (Ochoa) syndrome gene on 10q23-q24 and localisation of the disease gene within two overlapping BAC clones (<360kb). *Genomics* **60:** 12-19.

Warren S.T. and Nelson D.L. (1993) Trinucleotide repeat expansions in neurological disease. *Curr. Opin. Neurobiol.* **3:** 757-759.

Wastiaux J.P. Lamoureux G., Bouchard J.P., Durivage A., Barbeau C. and Barbeau A. (1978) HLA and complement typing in olivo-ponto-cerebellar atrophy. *Can J Neurol Sci.* **5:** 75-81.

Weber J.L. (1990) Informativeness of human (dC-dA)$_n$.(dG-dT)$_n$ polymorphisms. *Genomics.* **7:** 524-530.

Witzgall R., O' Leary E., Leaf A., Onaldi D. and Bonventre J.V. (1994) The Kruppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proc Natl Acad Sci.USA* **91:** 4514-4518.

Wooster R., Cleton-Jansen A-M., Collins N., Mangion J., Cornelis R.S., Cooper C.S., Gusterson B.A., Ponder B.A.J., von Deimling A., Wiestler O.D. et al. (1994) Instability of short tandem repeats (microsatellites) in human cancer. *Nat. Genet.* **6:** 152-156.

Chapter 6

CONCLUDING REMARKS

CONCLUDING REMARKS

The primary goals of this dissertation were the generation of a human gene index and its application to disease candidate gene discovery. Preliminary work on the STACK database circumvented the problem of large sequence data sets and limited algorithms by producing a tissue partitioned database of processed ESTs. However, the generation of a STACK human gene index required the ability to cluster as many as 500,000sequences, within a limited time frame. To this end, chapter 2 reports the successful porting of D2_CLUSTER to the Origin2000 architecture and the modifications made to the code in order to accelerate EST clustering on multiple processors. Modifications included (1) a restart capability that allowed the clustering procedure to be restarted at the same point at which it was interrupted, (2) the ability to break the work into a number of pieces such that each piece processes more sequences and each successive piece uses less time, and (3) the ability to make the database available to each of the processors in order to enhance the speed performance. The modified version of D2_CLUSTER was used successfully to cluster 490,293 sequences on 128 CPU R12000 300 MHZ Origin2000 in 31 hours. D2_CLUSTER performs redundant tasks that need to be addressed for future implementations of the code. For example, a database of sequences is read into memory as a compressed file and each sequence has to be uncompressed before the wordsizes are calculated. Once the d2 comparisons are completed, the sequences are compressed again with the result that each sequence is uncompressed multiple times. Future improvements to D2_CLUSTER should attempt to calculate the wordsizes on the compressed database. Alternately, the step of uncompressing the database can be circumvented by reading in the sequences directly from the FASTA file. The addition of ESTs to an existing STACK database requires minimal processing time in order to cope with demands of an exponential increase in the release of EST data. The time taken to generate a new release of STACK can be reduced by not clustering sequences that have been processed in a previous release. Functionality could be built into D2_CLUSTER so that each sequence comparison is only done on sequences that have not processed before (i.e., some mechanism needs to be put in place that tracks an "old" cluster membership).

A hierarchical approach was used to generate the STACK gene index. ESTs were partitioned into arbitrary tissue categories and clustered using high performance computing resources. Consensus sequences were generated using inhouse tools and these sequences were used to generate a whole-body index. The tools used to generate the STACK reconstructed transcripts capture alternative splicing events (Miller et al., 1999; Christoffels et al., 2001). The exponential increase of EST data in GenBank requires rapid processing in order to ensure the STACK database remains current. To this end, I have implemented a STACK_ADD protocol that incorporates new sequences to an existing database and ensures that sequences are not processed more than once. This STACK_ADD implementation was tested on UniGene build #106.

STACK development represents a work-in-progress. Future development of STACK focuses on linking the underlying data more firmly to biological processes and making the resultant information accessible to a widening range of users. Protein predictions from transcript isoforms and cross-references to known protein records will allow for association with standardised anntations such as gene ontology (http://www.geneontology.org).

The implementation of STACK, as described in this thesis does not keep track of clusterIDs from one release to the next i.e., new clusters are generated for each release of STACK. Organisationally, STACK will make increasing use of the relational database architecture to enhance data access. This will pave the way for maintenance of clusterIDs, or links to new clusterIDs from release to release. Entrez-styled querying capabilities are needed to allow for (i) access to specialised subsets of the STACK database, (ii) identification of isoforms based on phyiscal or developmental expression states and (iii) locating entries based on physical location within the genome. The above-mentioned functionality should accelerate gene candidate discovery and provide an enhancement on the methodolgy described in chapter five.

The STACK technology (STACK_PACK) represent a distributable clustering system and management tool set that can be applied to any genome project. The ongoing development of a visualisation tool for STACK-processed transcripts will be distributed with the STACK_PACK tools.

Progressive familial heart block1

Chapter five details an approach for reducing the *PFHB1* disease locus and accelerating the identification of candidate genes. A three-fold approach was undertaken namely, (a) *in-silico* screening for microsatellites, (b) assembly of the 1184 chromosome 19 genomic fragments mapping to the *PFHB1* locus and (c) mapping expressed transcripts including STACK and BodyMap onto the assembled data. The successful *in-silico* detection of tandem repeats, which are potential microsatellite markers, in genome data represent a means of integrating the genetic and physical maps for an unidentified disease gene. The *in-silico* identified tandem repeats are in excess of 12 repeating units and therefore potential polymorphic and justify further analyses in the laboratory for their informativeness as microsatellites. A plausible approach includes the design of primers that will allow for the PCR amplification of these loci in a panel of unrelated individuals to test the variability of the amplified alleles. Expression information and sequence coverage were integrated from resources such as STACK and BodyMap transcripts, mouse developmental ESTs and RefSeq contigs. Seven positional candidates, previously shown to be expressed in heart tissue, have been identified: GLTSRC2, DKF2P761A179, Kaptin, T- elongation factor 4, nucleobindin, CGI-123 protein and CD37-antigen. In addition, a list of unidentified transcripts were mapped to the *PFHB1* locus. Functional assignment of the mapped transcripts is required to prioritise the disease candidate gene list. For example, tools such as Interpro (http://www.ebi.ac.uk/interpro) provide automation for an ongoing assessment of the functional classes represented by the *PFHB1* candidate transcripts.

The Santa Cruz Genome Center has generated assemblies for each of the human chromosome genome sequence fragments (http://genome.ucsc.edu/, Genome Consortium 2001). The chromosome 19 scaffold generated at the Santa Cruz Genome Center has been integrated with EST and mRNA data available for chromosome 19 and represents a framework on which to validate the assembly across the *PFHB1* region. For example, *PFHB1* flanking markers could be mapped on the Santa Cruz assembled chromosome 19 genome sequence using ePCR. The ePCR mapped segment should be extracted and assembled with the JGI assemblies generated in chapter five. Memory constraints for PHRAP will require the generation of overlapping fragments of the 4Mb region extracted from the Santa Cruz data.

Appendix I (Chapter 2)

Protocol for using the supercomputers at the National Center for Supercompting Applications (NCSA)

1.1 Commands on the NCSA machine (modi4)

The history of a job “*bhist -l jobnumber*”

set the number (##) of processors on a machine “*setenv MP_SET_NUMTHREADS ##*”

1.2 batch script for modi4

```
#!/bin/sh
#BSUB -M 6g           '164MBx32'
#BSUB -n 32                   'processors'
#BSUB -c 1536:00              'time 48hours x32'
#BSUB -o brain.out            'outfile'
#BSUB -N                      ''
#BSUB -J brain                'job name'
limit
limit stacksize 200000
cd /scratch/$USER
/bin/rm -r brain
mkdir brain
cd brain
msscmd cd d2_cluster, get brain.out.gz
gunzip brain.out.gz
mkdir clusters
touch fastafiles
echo "brain.out" >> fastafiles
echo " clustering started at " >& brain.log
date >>  brain.log
enc_db brain.out
setenv MP_SET_NUMTHREADS 32
d2_cluster 6 ./brain.out 0.96 50 150 1 0 1
post_proc
```

```
tar cvf brain_cluster.tar clusters
gzip brain_cluster.tar
mv brain_cluster.tar.gz /scratch-modi4/n8644
tail \20 /var/adm/SYSLOG >> brain.log
```

1.3 batch script for CRAY/SGI machine (eg., arctic and flurry)

```
#QSUB -r name_of_job
#QSUB -s /bin/csh
#QSUB -o
#QSUB - eye.log
#QSUB -l mpp_p =4
#QSUB -lT 7200 -lt 7200
#QSUB -j eye.joblog
#QSUB -lM 800Mb -lm 800Mb
#QSUB -mb
#QSUB -me
#QSUB -mu alan@sanbi.ac.za
cd $QSUB_WORKDIR
setenv MP_SET_NUMTHREADS 4
```

1.4 commands on arctic

```
qstat -f cpu1_unl | egrep -i "stack|mem"
qstat -u  name  "list the queues used by 'name'"
ps -flu         "status of jobs in the queue"
```

1.5 Queues on arctic

NQE queues on arctic are listed using the qstat command.

Queue Name	CPU_limit	Time_limit	Memory_limit	Off/On
cpu1_unl_1.5Gb	1	Unlimited	1.5Gb	
cpu1_6hr	1	6hr	unlimited	Off
cpu1_unl	1	Unlimited	Unlimited	Off
cpu4_unl_1Gb	4	Unlimited	1Gb	
cpu4_6hr	4	6hr	Umlimited	
cpu4_unl	4	Unlimited	unlimited	Off
cpu16_6hr	16	6hr	Unlimited	
cpu16_24hr_1Gb	16	24hr	1Gb	
cpu16_24hr	16	24hr	unlimited	
cpu16_unl_1Gb	16	Unlimited	1Gb	
cpu16_unl	16	Unlimited	Unlimited	Off
cpu32_6hr	32	6hr	unlimited	
cpu32_24h_500Mb	32	24hr	500Mb	
cpu32_24hr	32	24hr	unlimited	
cpu32_unl_500Mb	32	Unlimited	500Mb	
cpu32_unl	32	Unlimited	Unlimited	Off
cpu64_24hr	64	24hr	Unlimited	Off
cpu64_unl	64	Unlimited	Unlimited	

A list of queues are provided with information such as "STS=on", "TOT =0" and "LIMIT=1". STS indicates whether a queue is open for use. "TOT" indicates the number of jobs present in a specific queue and "LIMIT" indicates the maximum jobs that can be submitted for a queue.

1.6 Description of a D2_CLUSTER CLUSTER_TABLE

The syntax of the CLUSTER TABLE is one line per sequence in the database. Each line contains five columns namely; SEQ, MEMB, LINK, ORIENT and NEW.

(i) SEQ refers to the sequence number ranging from 0 to N.

(ii) MEMB is an integer indicating cluster membership. If it is equal to -1, then the sequence was too short to be evaluated.

(iii)LINK is an integer that gives identity of another sequence in that cluster by having a low score with the current sequence. If LINK equal -1 (and MEMB is not equal to -1), then this sequence is the last in the cluster. The entire membership of a given cluster can be found by following LINK numbers.

(iv) ORIENT is a product of three integers. The orientation of the current i j pair (1 for i j; -1 for i rev j where rev j denotes the complement of sequence j), the orientation of i with its current LINK sequence and the orientation of j with its current LINK sequence). ORIENT can be either 1 or -1 and i and j denote sequence numbers.

(v) NEW means nothing. It is always equal to zero.

For example,

SEQ	***MEM***	***LINK***	***ORIENT***	***NEW***
0	-1	-1	1	0
1	-1	-1	1	0
2	2	8	1	0
3	3	9	-1	0
4	4	-1	1	0
5	3	-1	-1	0
6	2	-1	1	0
7	3	5	1	0
8	2	6	-1	0
9	3	7	-1	0

There are three clusters originating with three sequences: 2, 3, and 4. The clusters are (i) 2, 8 and 6 (ii) 3, 9, 7 and 5, and (iii) 4. Sequence 0 and 1 were too short.

Appendix II (Chapter3)

Protocols developed for generating a STACK human gene index and its visualisation on the world wide web

Scripts written for this project are indicated in bold.

Running D2_CLUSTER

make a "clusters directory for d2 output
#mkdir clusters

set num of processors
#setenv MP_SET_NUMTHREADS ##

convert input file to binary
#enc_db inputfile

d2_cluster [wordsize] [inputseq] [%similarity] [minseqlen] [windowsize] [revcomp] [restart] [replicate database]
#d2_cluster 6 ./inputseqs 0.96 100 150 1 0 1
=>output at this stage is a CLUSTER_TABLE (5 column matrix showing the relationships between sequences

extract FASTA files using the CLUSTER_TABLE
#post_proc

ADD_schema

<u>Remove sequences that are shorter than 50bases</u>
#**removeShortseq.pl** newsequence_file > newseq

<u>Comparison of old and new data</u>
compare the present database consensus sequences against the
new data set using crossmatch.
#cross_match database.seq newseq -masklevel 101 > crossmatch.log

extract all the matching entry names from the crossmatch output
#**parse_cross_match2.pl** [crossmatch.log] [newseq] [tissue] [cutoff]
=>output files are
crossed -> list of stackIDs vs newsequence IDs
cross.parsed-> list of newsequence IDs vs all the matching stackIDs

All newsequence IDs that share stackIDs are merged with the stackIDs
through a process of transitive closure.
#**MatchSequence** cross.parsed > newclusterlist

Remove traces of the old data that is now expanded by the new
Copy the old database directory to a working directory (eg., /work/)
change directory to /work/tissue/
use the stackIDs inthe crossmatch output to remove all affected clusters

from the tissue.contiglist, all.* files, gde directory and the crawlog

```
#destackFASTA_SINGLES.pl [crossed] [tissue] [oldcontiglist]
```

=>output

destack.contiglist -> contains only those clusters not affected
fasta and singles directories of all unaffected clusters.
new gde directory of all unaffected clusters
new crawfile called crawfile.new

Expanded clusters are collapsed

Collapse the consensus sequences within the expanded clusters to their constituent ESTs so that the assembly step can begin. At the same timewe make a lookup table of the mRNA sequences to that we can add them.

Each stack consensus sequence is replaced with the ESTs in the gde file

```
#stackCollapsed.pl -u unigenefasta singletonfasta outdir tissue
```

=>output

the FASTA formatted cluster files are saved in the outdir.

PHRAP assembly

assemble the sequences in the 'outdir'.

```
#runphrap.pl outdir outdir.contig_set
```

=>alignments are saved in 'outdir.contig_set'. These assemblies have to be analysed by the stack_pack system and appended to the old data.

copy the above assemblies to the working directory

```
#cp outdir.contig_set /work/tissue/
#../stack.bin/contiglist outdir.contig_set > outdir.contiglist
```

STACK_PACK

run contigproc4.pl

```
#../stack.bin/contigproc4.pl tissue outdir.contiglist > tissue.cp_log
```

append this contiglist to the contiglist of all the clusters that remained unchange after the addition of new sequences

```
#cat outdir.contiglist >> destack.contiglist
#mv destack.contiglist tissue.contiglist
```

run contigclone,clustlink,join and finish.pl. These are run from the shell script (cpj.sh).

```
#cd /work/
#./stack.bin/indexcpj.sh tissue tissue > tissue.log
```

Need to add the LIBRARY field.

<u>UniGene library information</u>

extract the library information from the unigene datafile (Hs.data)

```
#extractUGexpression.pl Hs.data > unigentissue
```

=>output

file containing Hs IDs and tissue names

GenBank EST library information
print a list of ESTs with its clone libraries using the raw est file

```
#extractLibrary.pl [estfile] > tissuelist
```

Integrating library info onto headerline

```
#printOriginalLibrary.pl -l tissuelist -t unigentissue -u unigeneACC -s seqfile
```

where -> -l will create a hashtable of the stack EST accessions
-> -L will create a hashtable of ETSs vs library info
-> -t create a hashtable of UniGene clusterIDs and tissues
-> -u create a hashtable of UniGene accessions (Hs### or GenBank)
-> -s FASTA formatted file of stack consensus sequences
=>output
seqfile.originTissue ->contain a LIBRARY field with original library
Once all the hashtables have been written to disk you can just run the script on each of the all.* files.
eg., **printOriginalLibrary.pl** -s all.fasta

```
:mv all.fasta.originTissue all.fasta
:mv all.fasta_link_duplicates.originTissue all.fasta_link_duplicates
:mv all.singles.originTissue all.singles
:mv all.singles_link_duplicates.originTissue all.singles_link_duplicates
```

Addition of Radiation hybrid mapping information

```
#cat all.* > totalseq
#mv totalseq /data4/alan/EPCR/
#cd /data4/alan/EPCR/
#/data9/alan/EPCR/bin/sgi/e-PCR /data9/alan/EPCR/db/genemap99.sts totalseq >
totalseq.genemap
```

The ePCR output represents a list of stackIDs vs RH markers.

```
#processEPCR.pl totalseq.genemap genemap
```

=>output
stackMarkerlist.genemap ->list of each stackID and its corresponding RH markers
bogusmap.genemap -> list of stackIDs with different RH markers
markerERRORS.genemap -> list of RH markers that map to two clusters (these stack clusters are mostly phrap-fragments)
statistics.genemap -> gives the summary of the mapping information (this includes; total hits, unique RHmarkers, unique clusters, mapping errors

The mapping info gets incorporated into the FASTA records

```
#addMap.pl [stackMarkerlist] [FASTA seq] ....
```

=>output
FASTAfile.new -> contains a MAP field after the library field.

```
#mv FASTAfile.new FASTAfile
```

Determine the orientation of all clusters(3/5 prime or end-not-spec)
cluster data (eg., all.* files)

```
#countOrientation.pl -l clonelist -s stack_fasta_file(eg., all.fasta)
```

where -> -l creates a list of accession vs orientation
-s seq file such as all.fasta
=>output

stack_fasta_file.stat containing a list of each clusterID and its orientation (3-PRIME, 5-PRIME or end-not-spec)

```
#grep "3-PRIME" stack_fasta_file.stat | wc -l
#grep "5-PRIME" stack_fasta_file.stat | wc -l
#grep "end-not-spec" stack_fasta_file.stat | wc -l
```

linked fasta file
use the all.fasta_link_duplicate.stat and all.singles_link_duplicate.stat
files to find out what the orientation is of the linked clusters

```
# linkdirection.pl [file.stat] link_tissue.fasta > outfile
=>output
```

outfile contains a list of linked entries with their orientation

Make the data ready for the BLAST web-interface
make file blastable for blast1 (only blast1 webinterface ready)

```
#cat all.fasta all.singles and link_tissue.fasta > researchINDEX.seq
```

The X's in the sequence causes problems for blast1 so I replace then with N's

```
#replaceX.pl researchINDEX.seq > newfile
#mv newfile researchINDEX.seq
```

```
#pressdb researchINDEX.seq
=>output
researchINDEX.seq.csq
researchINDEX.seq.nhd
researchINDEX.seq.ntb
```

Web interface
http://ziggy.sanbi.ac.za/alan/researchINDEX.html is stored in
/var/www/htdocs/alan/**researchINDEX.html**

This html page calls the cgi-script (/var/www/cgi-bin/researchINDEX.pl)

researchINDEX.pl uses a configuration file called (/usr/local/lib/blast_search/research-config.pl)
This script sets the blastdatabase location to "/data4/blast"
It also sets the lookup for the blastable file to "'researchINDEX' =>
'researchINDEX.seq'

researchINDEX.pl calls an extract script (**stackextract_AGC.cgi**) to pull out the stack consensus sequences. The extraction process needs two files (index.clonecluster.tab and index.clonelink.tab)

```
#cloneCluster_tabINDEX.pl all.fasta* > index.clonecluster.tab
#cloneLink_tabINDEX.pl index.link.table > index.clonelink.tab
```

create an indexing file for researchINDEX.seq

```
#index researchINDEX.seq
=>output
researchINDEX.seq.ndx
```

stackextractINDEX.cgi is called from within the stackextract_AGC.cgi.

researchINDEX.pl returns a webpage that has created links on_the_fly to the **stackextract_AGC.cgi**. Each stackID in the blast output can be clicked on and the stackextract_AGC.cgi sets up a webpage with links to each EST and the consensus sequence and UniGene links. The consensus link is executed by **stackextractINDEX.cgi**.

Appendix III (Chapter4)

Protocols developed for mapping STACK transcripts to the human genome
Scripts written for this project are indicated in bold.

<u>Find novel genes on chr22 using STACK gene index</u>

download the chr22 gene table from http://www.sanger.ac.uk/cgi-bin/c22_genes_table.pl
save the table as text
extract the start and stop positions and then accessions and
save it in three files (genes, predicted and pseudogenes)

```
foreach table {
        run printPseudogene.pl [seqfile] [genetable]
        sequences are saved in a directory for each table
}
```

cat all sequences in a dir using :
#/data4/alan/bin/MRNA/**cat_long_list.pl** [dir] [outfile]

formatdb -i above_outfile -p F
search the index against this database.
#blastall -p BLASTN -d aboveoutfile -i stackindex -e 1e-40 -o outfile

all blast results in the file that meet the criteria are printed to a directory. #parse_blastFile.pl blastfile [outdir]
Then run **parse_STACK_blast.pl** on the above blast dir.the results show the stackids that found matches.

print the stackids that do not find matches.use these sequences to search the chr22 data.
#blastall -p BLASTN -d chr22genomeseq -i remainder stackindexseq -e 1e-40 -o uniquematches

print each blast record that meet criterium of 1e-40 to a directory
#**parse_blastFile.pl** uniqueINDEX.blast_2000chr22 novelhits_2000 > & novelog2000 &

parse all stack sequences that match 94% and over 80% of its length.
#**parse_swissprot_identities.pl** novelhits_2000 > novelhits_2000_blast
awk '{if ($4 > 80 && $5 >=94) print $0 }' novelhits_2000_blast > novelchr22_2000.hits

#mkdir old
leave only the accurate matches in the novelhits_2000 directory
#move.pl novelhits_2000_blast novelhits_2000
#this will look at the IDs in the first file and then move all other sequence IDs to another directory. NOw the novelhits_2000 directory contain the accurate matches. These hits has to be parsed so that we print the start and end positions of the matching bases
#**printSubjectStartStop.pl** novelhits_2000 > novelhits_2000.positions

compare the above matching bases to the genetable
#**compare_chr22_genePos.pl** /deepsea/alan/validity/Chr22/oldsanger/Sanger_genetable
novelhits_1999.positions

The updated gene table at the sanger site (15th Nov2000) reported 802 genes as oppose to the 545 genes in 1999. There does not appear to be any standardisation of gene names so I searched the entire stack gene index against the new chr22 genome sequence. These results were then compared to the sanger gene table to find the unique hits.

Appendix IV (Chapter 5)

Protocol for candidate gene discovery on chromosome 19q13.3

Scripts written for this project are indicated in bold.

<u>download sequences</u>

download sequences from ftp://sawedoff.llnl.gov/pub/JGI_data/Human/Draft/

Each sequence file contain multiple sequences.Concatenate all the sequences in one file and rename the sequences so that the BAC clone name is reflected in the headerline

#**renameCHR19clonesFILE.pl** [directory of sequences] > newfile

<u>produce stack header:</u> standardise header for automation

#**printPFHB_header.pl** [newfile] > nefile.stackheader

<u>masking</u>

Ecoli- masking prior to tandem repeat screen

downloaded the ecoli genome and split it into fragments

#cross_match seqfile ecoli.seq -minmatch 12 -minscore 20 -screen

<u>RepeatMasking prior to PHRAP assembly</u>

#RepeatMasker -x seqfile

<u>Tandem repeat finder</u>

downloaded the Tandem RepeatFinder for sgi platform (trf.irix.exe)
from http://c3.biomatch.mssm.edu/trf.html
processed all the chromosome 19 sequences for tandem repeats using
the following parameters:
trf.irix.exe seqfile 2 7 7 80 10 40 500 -f
where 2=match weight; 7=mismatch; 7=indel; 80=matching prob;
10=indel prob; 40=min align_score; 500 max periodsize; -f print flank
sequence

#**runRepeatfinder.pl** [seqdir]
output: TANDEM_REPEATS directory

<u>parse the data in TANDEM_REPEATS</u>

```
#runTandemRepeatParser.pl [tandem_outputdir] [max period size]
```

search STACK whole-body index

```
#blastall -p blastn -d /data4/blast/researchINDEX.seq -i PFHB.phrapseq -e 1e-20 -o PFHB.phrapseq.stackblastn
```

split blastfile into multiple files

```
#parse_blastFile.pl blastfile [outdir]
```

parse stackdata

```
#print_blastDetails.pl [outdir]
output: pfhb.StackLibrary   # print stack library field
               pfhb.StackDescription # print stack header line
               pfhb.StackExpectScore # print expection scores
               pfhb.stackseq                # print the matching stackseq
#finddupstackSeq.pl [pfhb.StackExpectScore] [../phrap.multicontigs] [../phrap.mergedclones]
output: UniqueStackIDs - stackids that match one contig
                       duplicateStackIDs - stackids that match to more than one contig
```

annotate stack seqs with nonredundant blast search

```
#blastall -p blastn -d /data10/nr/nt -i stackseq -e 1e-20 -o pfhb.stackseq.blastnr
#blastall -p blastx -d /angis/dbases/nr/nr -i stackseq -e 1e-20 -o pfhb.stackseq.blastx
```

```
#printStackblastn.pl [pfhb.StackExpectScore] [pfhb.stackseq.blastnr]
output: phrapcontig stackID blastDescription ExpectScore
```

blast phrapcontigs against nonredundant database

```
#blastall -p blastn -d /data10/nr/nt -i phrap.mergedclones -e 1e-20 -o mergedclones.blastn
#blastall -p blastn -d /data10/nr/nt -i phrap.multicontigs -e 1e-20 -o multicontigs.blastn
```

download the bodymap matrix from bodymap and save it as text with tab delimiters.

download the bodymap sequences.

search it against the PFHB contigs

```
#blastall -p -d PFHB -i bodymapseq -o bodymap.blast
```

parse blastdata

```
#parse_blastFile.pl bodymap.blast outdir
#parse_blastMatches.pl outdir > bodymap.out
```

output is a list of query|hit|exp|match%|matchlen|

find the expression for each matching bodymap seq

```
#bodymap.pl bodymapMatrix bodymap.out > bodymap.chr19
```

output is a list of bodymapIDs and tissues

Mouse data obtained from Minoru Ko

search these against PFHB sequences

```
# blastall -p blastn -d PFHB -i mouseseq -o mouse.blast
```

parse mousedata

```
# parse_blastFile.pl mouse.blast mousedir
# parse_blastMatches.pl mousedir > mouse.ch19hits
```

fetch the mouse sequences that match

```
#fetchmouse.pl mouse.chr19hits mouse.seq > mousechr19.seq
```

annotate the mouse sequences

```
# blastall -p blastn -d /data10/nr/nt -i mousechr19.seq -o mousechr19.blast
# parse_blastFile.pl mousechr19.blast outdir
# parse_blastMatches.pl outdir > mousechr19.dnanr
```