High Performance Computing and Algorithm Development: Application of Dataset Development to Algorithm Parameterization

NAME : Mario Ricardo Edward Jonas

SUPERVISOR: Professor Winston A. Hide

A minithesis submitted in partial fulfillment of the requirements for the degree of *Magister Scientiae* at the South African National Bioinformatics Institute, University of the Western Cape

2006

Keywords

Algorithm testing

Biological sequence fidelity

Sensitivity

Specificity

Expressed Sequence Tags (EST)

Dataset creation

EST Clustering

EST Assembly

Alternative splicing

Sequence Error

Abstract

A number of technologies exist that capture data from biological systems. In addition, several computational tools which aim to organize the data resulting from these technologies have been created. The ability of these tools to organize the information into biologically meaningful results, however, needs to be stringently tested.

The research contained herein focus on data produced by technology that records short Expressed Sequence Tags (ESTs). An EST reference dataset was generated that can be used to test the set of tools which use ESTs to reconstruct expression events. The EST reference dataset contains well-characterized biological phenomena (exon-skipping, paralogy) and quantified sequence error.

A subset of computational tools (*d2_cluster*, WCD, *phrap*, CAP3) were tested using the reference dataset and it was found that CAP3 produces higher integrity sequences at the cost of losing alternative splicing information. *Phrap*, the looser clustering algorithms implemented in *d2_cluster* and the novel tool WCD, produce results which capture the alternatively expressed sequence information.

Future related research should focus on elucidating the internal gene structure of the results produced by the computational tools evaluated in order to determine the biological validity of the results beyond the level of sequence similarity.

Availability of dataset: www.sanbi.ac.za/~mario/dataset.tgz

Declaration:

I declare that "*High Performance Computing and Algorithm Development: Application of Dataset Development to Algorithm Parameterization*" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full Name: Mario R. E. Jonas **Date:** 21 May 2006

Acknowledgements

Done! Now, onto the best part, for me at least: thanking you!

To the students who were there at the start, thanks for the moral support and encouragement (Anelda, Buki, Estienne, Faisel, Nothemba, Sumir, Vicky). To the ones who came after (Adele, Farahnaz) and helped make SANBI an enjoyable place to work in, thanks. AlanP, your humor brightened many a day, and your level-headedness helped balance my views. AllanK, your generous spirit has lifted me out of many a ditch. To the old Honors students Fezi and Thurayah, you have given me an opportunity to extend myself into an area that I love; education. Portia, thanks for always supplying the stationery needed. Judith, thanks for being sweet! ☺

Ferial, thanks for telling me to "trek vinger!" You have become a friend for whom I will "killa-de-bull!" Cathal, thanks for trying to teach me the one thing I am VERY slowly making part of me, "Don't take any nonsense from myself!" Vlad, thanks for invaluable assistance in such a short period of time.

Last, but definitely not least, to my supervisor, Winston Hide, thank you. You are full of nonsense (the more colloquial term escapes me now), but thanks for believing in me and giving me a chance. Thanks for the attempts, hopefully not in vain, to teach me to be delivery-driven.

Now on to the individuals who have more permanence and therefore more significance for me: my family, especially my mom. It has been an infuriating period for you, I know. However you handled it with the grace and the quiet spirit that I have come to love and respect. My sisters and dad, I love you guys more than you will ever know. Elize, Hein, Reg, and Rod: who always believed that there is more to me than what circumstances have shown; you may be right. To my girlfriend, Chriszaan: Love you lots, princess!

Saving the best for last: I would like to thank God. Somehow, somewhere in a secular environment, I started to believe in His existence again. Strange that!

Table of Contents

List of Figures

List of Tables

Abbreviations

AceDB	A Ceanorhabditis Elegans Database
AS	Alternative Splicing
BAP	Bacterial alkaline phosphatase
BLAST	Basic Local Alignment Search Tool
CAGE	Cap Analysis of Gene Expression
CAP	Contig Assembly Program
cDNA	Complementary DNA
DNA	De-oxyribonucleic Acid
EMBL	European Molecular Biology Laboratory
ESTs	Expressed Sequence Tags
FN	False Negative
FP	False Positive
GI	Gene Index
HUGO	Human Genome Organisation
ISO	Insufficient Sequence Overlap
MGC	Mammalian Gene Collection consortium
MPSS	Massively Parallel Signature Sequencing
mRNA	messenger RNA
msbar	Mutate Sequences Beyond All Recognition
NCBI	National Centre for Biotechnology Information
NEDO	New Energy and Industrial Technology Development Organization
ORESTES	Open Reading frame Expressed Sequence Tags
phrap	Phragment Assembly Program
RefSeq	Reference Sequence
RI	Rand Index
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
Sn	Sensitivity
SNP	Single Nucleotide Polymorphism

Sp	Specificity
TA	TIGR Assembler
TAP	Tobacco acid pyrophosphatase
TGI	TIGR Gene Index
TGICL	TIGR Gene Indices Clustering Tool
TIGR	The Institute for Genomic Research
TN	True Negative
TP	True Positive
Tremble	translated EMBL
UCSC	University of California, Santa Cruz
UTR	Untranslated Region

Chapter 1 Introduction

With the increasing number of sequenced genomes (188 as on 15 May 2006[*]), one of the next challenges for researchers is to characterize the transcriptome, which is defined as the complete transcribed complement of the genome. Characterization includes transcript cataloging (including determination of all possible gene transcripts and expression events), transcript profiling (the spatio-temporal expression patterns of gene transcripts), and understanding the transcription regulatory networks[1]. Transcript cataloging is defined as the recording and description of all expressed genomic sequences, including alternative transcripts, anti-sense transcripts non-protein coding RNA. Several technologies exist, each with inherent sampling bias which attempt to characterize and catalogue expression products,.

Examples of these technologies include Serial Analysis of Gene Expression (SAGE[2,3]), Cap Analysis Gene Expression (CAGE[4-6]) and Massively Parallel Signature Sequencing (MPSS[7-9]), Expressed Sequence Tags (ESTs) and Microarrays.

The main tools used for cataloging and characterizing gene expression products are based on the use of cDNA's, whether that be partial cDNA fragments (as used by SAGE, CAGE, MPSS, ESTs and Microarrays) or full-length cDNA sequences (as used by the Mammalian Gene Collection consortium (MGC)[10] and the NEDO project[11]).

All of the above-mentioned methods will be discussed in more detail in the following sections. Several genome-based computational tools also exist which aim to catalog gene transcripts through gene prediction. These tools try to infer the gene structure from the intrinsic genome sequence properties, and as such, fall outside the scope of this discussion.

[*] http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_index.shtml

1.1 Transcriptome characterization technologies

1.1.1 Serial Analysis of Gene Expression (SAGE)

SAGE allows the researcher to determine the number and relative abundance of a gene transcript in a biological sample. cDNA is labeled at the 3'-end with biotin and immobilized on streptavidin-coated magnetic beads. The immobilized cDNA fragments are restricted with a 4-base restriction enzyme (NlaIII or Sau3A) which generates a 'sticky' CATG or GATC-end. An adaptor containing a recognition site for class II restriction enzymes (BsmFI or MmeI) is then ligated to the 'sticky' end[1]. Both of these restriction enzymes cut upstream of their recognition sites; BsmFi cuts 14bp upstream, and MmeI cuts 18-20bp upstream. Restriction with these enzymes then produce 14bp SAGE[3] tags or 21bp LongSAGE[2] tags. Tags are concatenated into longer sequences which are then sequenced. Quantifying the number of unique markers gives an estimate of the expression of a gene under a specific set of conditions.

One of the disadvantages of SAGE is that the short tag size introduces ambiguities in the identification of gene transcripts since the fragments may not necessarily be unique. The ambiguity problem has been alleviated somewhat by LongSAGE which produces longer 21bp tags. An additional disadvantage is that a large number of clones need to be purified and sequenced, leading to increased cost and limited throughput[9]. In addition, the fact that there may not be a cut-site for the enzymes (NlaIII and Sau3A) acting as anchoring enzymes[12] means that some transcripts may not be represented at all.

1.1.2 Cap Analysis Gene Expression (CAGE)

CAGE is similar to SAGE in that short nucleotide fragments (typically 20 bp) are generated via class II restriction enzymes. These generated nucleotide fragments are concatenated, cloned and sequenced. The major difference between CAGE and SAGE is that CAGE tags are generated from the 5' end of the capped mRNA, as opposed to

the 3' end for SAGE. CAGE relies on the CAP trapper method developed by Carninci et al[4,6,13] which selectively captures 5' capped mRNAs, leading to the use of CAGE tags in characterizing transcription start sites.

The process (**Figure 1, p4**) starts with first cDNA-strand synthesis, followed by biotinylation of the diol moieties unique to the cap structure and polyA tail. Subsequent degradation by RNase I removes single-stranded RNA, as well as the polyA tail (most of which will be unprotected by the polyT primer), leaving a full-length mRNA-cDNA hybrid which is biotinylated only at the 5' cap structure[4,13-15]. These full-length hybrids are then isolated on streptavidin beads, subjected to RNA hydrolysis to remove the mRNA, and subsequent second-strand cDNA synthesis.

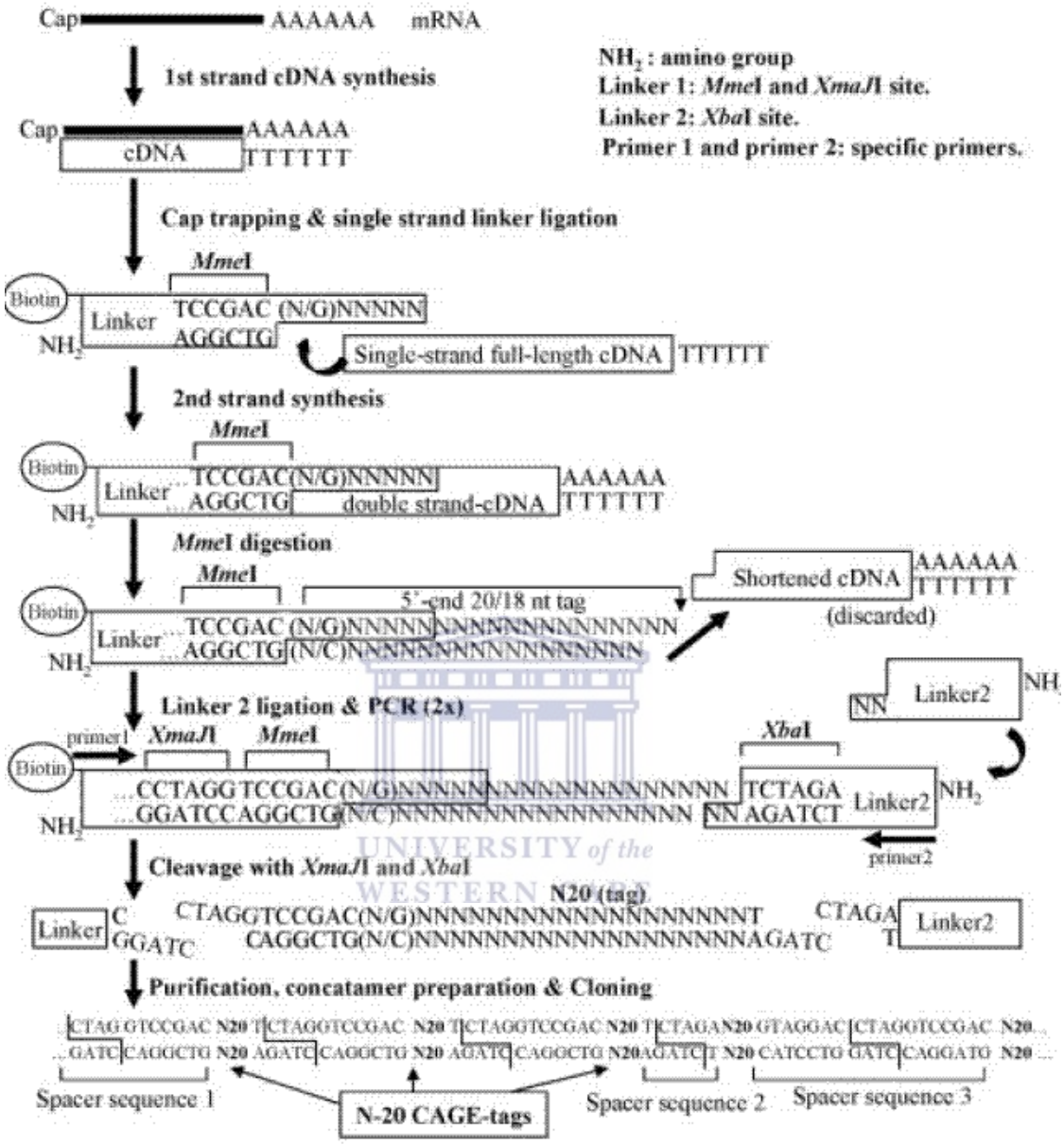

Figure 1: Diagrammatical overview of CAGE Technology[6]. See text (Section 1.1.2, p2) for further details.

1.1.3 Oligo-capping

Sugano *et al*[16] developed the oligo-capping method (**Figure 2, p5**) in which the cap structure of an mRNA molecule is replaced with a synthetic oligonucleotide. The synthetic oligonucleotide serves to label the capped end of the mRNA, thereby ensuring that only full-length mRNAs are captured for library construction. Bacterial

alkaline phosphatase (BAP) hydrolyses the phosphate of truncated mRNA 5' ends whose cap structures have been broken down, and leaves a hydroxyl group at the 5' position. Tobacco acid pyrophosphatase (TAP) removes any intact cap structure, leaving the phosphate at the 5' end. T4 RNA ligase then selectively ligates the synthetic oligo to the 5' phosphate, ignoring the mRNA molecules containing the 5' hydroxyl moiety.

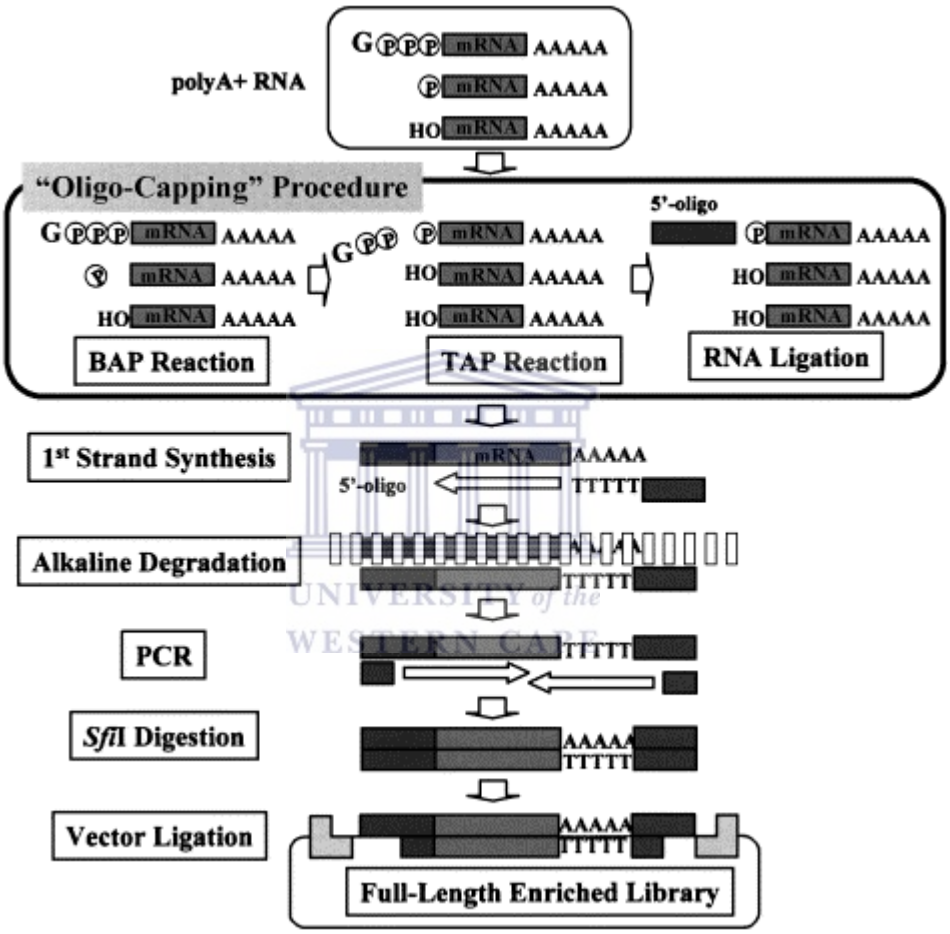

Figure 2: Diagram of steps involved in oligo-capping (obtained from Sugano et al[16]). Explanation in text (Section1.1.3 "Oligo-capping", p4).

1.1.4 Massively Parallel Signature Sequencing (MPSS)

Like SAGE, MPSS generates a 17-20bp tag (called a signature sequence) extending from the 3'-most Sau3A restriction site. These unique signature sequences are then

attached to micro-beads via proprietary technology called Megaclone[*]. The signature sequences are sequenced in a parallel fashion, resulting in massive reduction in time and effort[9]. No prior knowledge of any of the sequences is needed, and characterising differential expression allows for counting transcript numbers as low as 5 transcripts per million (tpm)[17†], making it the most sensitive of all the technologies reviewed here. The higher sensitivity of MPSS is an advantage when considering that certain transcripts are present at levels as low as 0.001 copies per cell[18].

Unfortunately, due to the complexity of this method, specialized equipment is needed which, for most laboratories, is not financially viable[1]. The proprietary nature of the technology also limits potential users to a single supplier, Lynx Therapeutics.

1.1.5 Microarrays

Microarrays consist of a grid of sequence probes attached to either a glass slide or nylon membrane as a support medium. Based on the type of probe used, two types of microarrays exist: the probes on the support medium can either be cDNA or oligonucleotides (High-Density Oligonucleotide Arrays - HDOAs). As much as 30,000 probes can be placed on a slide. The sequence for the probe does not need to be known. The targets are either cDNA synthesized from the transcript mRNA, or total RNA from the cell or tissue under investigation. Microarrays allow thousands of genes to be assayed.

1.1.6 Expressed Sequence Tags (ESTs)

ESTs are single-pass reads of the cDNA obtained from reverse-transcribing mRNA which is present as consequence of gene expression[19,20]. ESTs do not require a known template, and is therefore a good method for finding novel genes. Although ESTs can be used to quantify the level of transcription, the technology is not as sensitive as SAGE, CAGE or MPSS in detecting low-abundance transcripts (see for example Sun

[*] http://www.lynxgen.com
[†] http://www.takarabioeurope.com/news/mpss_faq.html#q7

et al[21]), leading to an under-representation of these low abundance transcripts in EST databases.

ESTs have been extensively used for novel gene discovery, gene mapping, generating gene indices and gene annotation. For a more extensive discussion on ESTs, see the following **section 1.2**, "***Rationale for Using ESTs***" (**p7**).

1.2 Rationale for using EST data

SAGE, MPSS, Microarrays and ESTs give quantitative information with regards to expression levels of gene products. In addition, these technologies can be used to compare expression levels and products across various biological conditions.

Oligo-capping and CAGE allow the generation of full-length cDNA and the subsequent characterization of the gene product. Full-length cDNA sequences (FL-cDNA's) are generally accepted as the best sources for transcript cataloging. In the MGC[22] project pipeline, which aim to generate full-length cDNA's, 5' and 3' ESTs are generated first. Therefore ESTs for a transcript is available before the FL-cDNA's are[23]. In addition, although an FL-cDNA may be present in the database, it may not necessarily reflect all the alternative transcript isoforms which exist[24] for a particular gene. Thus, ESTs represent an inexpensive and fast way of generating quantitative expression data, as well as for characterization of gene transcripts.

It needs to be stressed that an approach which uses complementary methods of transcript cataloging is more sensible and provides more solid results than a single approach. The caveat with regard to the use of EST data for transcript cataloging is that it needs to be well organized and characterized. This is done in the context of a Gene Index, which aims to group together all ESTs emanating from the same gene locus[25,26].

1.2.1 Characteristics of ESTs

ESTs represent one of the most useful means of reconstructing virtual transcripts because they have broad expression state (e.g. species, anatomical location, disease state) and coverage (e.g. if humans are excluded, ESTs exist for 768 species, 519 with more than 100 ESTs per organism[*]). When considering that only about 188 genomes have been sequenced, it means that for most organisms, only EST data exist.

The high number of ESTs in EST databases is another reason for their usefulness. Human ESTs account for 21.4% (7,741,240) of dbEST (which contained 36,241,897 ESTs as on 12 May 2006)*.

The initial bias towards 3' ESTs in EST database has been met by the increase in the number of 5' ESTs, as well as the presence of Open Reading frame Expressed Sequence Tags (ORESTES[27-29]). ORESTES have been shown to be distributed throughout the transcript length, but preferentially generate ESTs from the central regions of gene transcripts. The presence of ORESTES in EST databases mean that there is distributed transcript localization, i.e. a more representative view of the complete transcript, which adds to the 5' and 3' ESTs already present in the database.

A Gene Index attempts to cluster ESTs such that ESTs belonging to a specific gene is assigned to a single class. ESTs have been used to generate Gene Indices such as STACKdb[30,31], Unigene[32,33] and TIGR Gene Indices (TGI[34-36]). These indices also attempt to reflect alternative splicing of these genes.

ESTs have been used to assist in gene identification, i.e. the detection and characterization of novel genes, through the use of tissue-specific EST libraries, as well as in gene expression studies. In addition to this, ESTs can be used to identify genetic variations such as Single Nucleotide Polymorphisms (SNPs) and alternatively expressed genes[37].

[*] http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

1.2.2 EST Disadvantages

Although ESTs are extremely useful in reconstructing the virtual transcript, they are marred by several problems in the data. These include (1) low sequence quality, or sequencing error, (2) the presence of chimeras, (3) the existence of gene families, (4) the presence of repeats, (5) contamination with genomic sequence, (6) contamination with vector sequence, (7) alternatively expressed transcription, (8) the existence of database errors. These features complicate the organizing and usefulness of ESTs. In addition, the transcript sequence information may be incomplete, since it most frequently contains only incomplete fragments of gene transcripts. Wang *et al* conclude that most clustering errors occur because of Insufficient Sequence Overlap (*ISO*) errors[38]. *ISO* errors can, however, be countered by full-length cDNA cloning and sequencing.

1.2.2.1 Sequence Error

Sequence error refers to the random single-base errors that occur in biological sequence data. The causes for this may be biological or technical. Biological error may be due to polymerase decay (error probability increases with increasing sequence length), primer interference (primer interferes with the start of a sequencing read), or stuttering (a part of the DNA to be transcribed gets re-read; happens after repeated G's or T's). Technical errors occur during sequencing and include lane-tracking error. Depending on the level of error per sequence, related sequences may differ from each other to such an extent that they may be assessed to be unrelated. On the other hand, so much error may have been introduced that unrelated sequences may appear highly similar.

1.2.2.2 Chimeras

Chimeras are made up of sequence fragments from different sequence sources. These might be due to the artificial ligation of ESTs during EST production, or clones mistakenly ligated from different mRNA species. The presence of chimeras would

cause clustering and sequence assembly to associate totally disparate sequences with each other.

1.2.2.3 Gene Families (Paralogs)

Paralogs are accepted as having been derived from gene duplication, subsequent to which sequence divergence occurred[39]. Gene family members share similar nucleotide sequence motifs, as well as amino acid secondary and tertiary structure. According to Taylor and Brinkman[40], 10% of human genes have ancient paralogs. Depending on the level of sequence divergence i.e. sequence identity, between the family-derived ESTs, apparently homologous family members would tend to cluster together. Consequently, EST sequences from completely separate genes would then be merged into a single gene.

1.2.2.4 Repeats

Repeated DNA sequences (repeats) are ubiquitously dispersed throughout a particular genome. These repeats my vary both in length and copy number. Repeats are more prevalent than the coding regions of the genome. Most of these are found outside the coding regions, but are often found within the exonic parts of these genes. Sometimes these repeats even perform regulatory functions[41]. Repeats may cause false gene clustering and assemblies, since the common repeats would force unrelated sequences to group together based on assumed similarity; repeats should ideally be masked and not deleted from the sequences containing them.

1.2.2.5 Sequence Contamination

ESTs, like any other sequence data, may contain foreign sequence matter i.e. sequence derived from sources other than the intended source. The foreign sequence matter may comprise all, or part of the sequence. Sequence contamination common to sequences could lead to the erroneous clustering or grouping of unrelated sequences together[*].

[*] http://www.ncbi.nlm.nih.gov/VecScreen/contam.html

1.2.2.6 Vector Sequence

EST databases contain some EST sequences that have been found to contain vector sequences which were not removed prior to the sequence submission process. Although the quality control of sequence submission has increased, these contaminated sequences are still present and as such, vector sequences should be masked out. Vector masking is done by searching sequence against a known vector database and masking the appropriate vector fragment.

1.2.2.7 Alternative splicing

Alternative pre-mRNA splicing (*AS*) produces various gene products from the same gene template through the use of alternative transcription initiation and polyadenylation sites as well as alternative exon usage. *AS* appears to account for the large disparity between the number of genes found on a genome, and the expression products represented by the proteome. It is estimated that as much as 70% of all human genes are alternatively expressed. ESTs which represent a specific *AS* gene may be deemed to be so dissimilar to each other, that they are placed in different clusters or assemblies.

1.2.2.8 Database Errors

The deluge of biological data requires human intervention to create the relevant databases, as well as to capture the relevant data. In addition to this, complementary annotation data are added to characterize the relevant data points. Each step of this process presents possibilities of error introduction. Errors may include the format of the data files, syntactic, typographical and scientific error in the sequence, as well as the incorrect annotation of sequences[42].

All of the above-mentioned phenomena complicate the use of ESTs, and make the analysis of results obtained from EST data difficult. When using EST data, these phenomena need to be considered, and where possible, avoided or removed.

1.3 Some tools which use ESTs to reconstruct gene transcripts

This section gives an overview of some of the tools which use ESTs to reconstruct gene expression events. There are tools which do this reconstruction on the systems level by just classifying or clustering related sequences into a single class i.e. one class would ideally represent a single gene or transcript. These clustering tools include *d2_cluster*[43], and the clustering utilities of the TIGR Gene Indices Clustering Tool (TGICL). For these tools no further processing of these classified sequences are done.

The other set of tools attempt gene expression reconstruction on the assembly level i.e. if enough criteria are met, related sequences are assembled into linear contiguous sequences which are longer composites of the related sequences. Assembly tools include the Phragment Assembly Program (*phrap*), the Contig Assembly Program (CAP3), and TIGR Assembler (TA). Both CAP and *phrap* were designed to assemble fragments into a single linear sequence, and as such, the behavior of these programs in e.g. the presence of alternative splicing is uncertain.

These tools are commonly combined into a pipeline which clusters related sequences, upon which the clusters are then assembled. Examples of these pipelines are StackPack (clustering via *d2_cluster*, and assembly via *phrap*), TGICL (clustering via *tclust*, *nrcl* and *sclus*, and assembly via CAP3).

1.3.1 Assemblers

1.3.1.1 Phragment Assembly Program (*phrap*)

Phrap was originally designed for assembling shotgun genome DNA sequence. *Phrap* allows the usage of the complete sequence, not only the high quality sequence data. Instead of generating a consensus sequence, *phrap* uses the high quality data fragments to generate a ‘mosaic’ contig. Sequence similarity is based on “word-nucleated” local

alignment. The sequences to be compared are searched for identical subsequences or 'words' of a specified length. If there are multiple matching words between the input sequences, the diagonals in the Smith-Waterman alignment matrix representing these matches are extended. This process is done recursively, possibly resulting in multiple alignments with scores above the cut-off score. *Phrap* is able to generate its own quality scores if none are provided[*].

1.3.1.2 Contig Assembly Program (CAP3)

CAP is an assembly tool which has also been developed for genome assembly. It was designed by Huang in 1992[44], and has been improved several times[45,46]. The most recent version, CAP4, is a commercial product for which no algorithmic information is available.

CAP produces its assemblies in three phases:

Phase 1: 5' and 3' poor regions of each read are identified using local alignment and removed. Overlaps between reads are computed and false overlaps are identified and removed.

Phase 2: Reads are joined to form contigs in decreasing order of overlap scores. In CAP3, corrections are made to contigs via forward-reverse constraints. These constraints are obtained by sequencing both ends of a sub-clone and insist that "*the two reads should be on opposite strands … within a specified distance range*".

Phase 3: Multiple sequence alignment of reads is constructed and a consensus sequence with quality values is calculated for each base in the contig[46].

1.3.1.3 Partial Order Alignment (POA)

Lee *et al*[47] have suggested the use of partial order (PO) graphs as data structures to represent multiple sequence alignments (MSA). Dynamic programming is then used to align the PO-MSA. Dynamic programming starts of in the usual way with the

[*] http://www.phrap.org/phredphrap/phrap.html

alignment of two sequences. The resultant alignment is represented as a PO-MSA. All subsequent sequences to be added to the MSA, are aligned to the PO-MSA. The result is a graph representation of a MSA. Lee[48] then extended the work done on POA by developing the *heaviest_bundling* algorithm to use dynamic programming to construct consensus sequences.

1.3.2 Clustering tools

1.3.2.1 d2_cluster

d2_cluster[43] is based on the d2 distance function. Clustering of similar sequences is done in one of two ways: alignment-based (sequences are aligned to each other to determine the similarity) and non-alignment based. *d2_cluster* is a word-based method which falls into the second category. The d2 distance function is based on word count and the most similar sequences are the ones with the lowest d2 value. The d2 value for two sequences is calculated by determining the word frequencies of each sequence and then taking the sum of the square of the differences.

Mathematically:

$$\mathbf{d}_k^2(x, y) = \Sigma_{|w|=k}(\mathbf{c}_x(w) - \mathbf{c}_y(w))^2 ,$$

where x and y are sequences, w is a word which has length k.

Instead of calculating the d2 score over the complete sequence, it is calculated over a predefined contiguous length called a *window*. The d2 score for a pair of sequences is then the minimum score between all the pairs of windows for these sequences. The default window size for *d2_cluster* is 100 bp.

1.3.2.2 WCD

WCD is a novel extension of *d2_cluster* which, in addition to the d2 distance function, allows for the use of two additional distance functions; edit distance and a common

word heuristic. An added feature of WCD is the ability to do simple parallel processing.

Our selection of tools for evaluation were limited to the two clustering tools *d2_cluster* and the novel tool WCD, as well as the assembly tools *phrap* and CAP3.

1.3.3 Pipelines

1.3.3.1 stackPack

StackPack uses *d2_cluster* to do word-based clustering of EST sequences, *phrap* to assemble the clusters, and CRAW, which does additional sequence analysis to determine possible alternatively expressed transcripts.

1.3.3.2 TGICL

TGICL[49] is a pipeline of programs which first clusters ESTs using three clustering utilities *tclust* (a transitive closure clustering tool with overlap filtering options), *nrcl* (a containment clustering and layout utility which processes pairwise alignments) and *sclust* (a seeded clustering tool that processes pairwise alignments) and then assembles these clusters using CAP3[46].

1.4 Definition of fidelity of program reconstruction

The fidelity of reconstruction can be defined as the measure to which the virtual transcripts resemble the actual gene products. This would include the extent to which the tools assign ESTs to the correct (known) groupings, whether these groupings are clusters or assemblies. Fidelity is also affected by the ability of programs to reconstruct and record alternative splicing events.

To the best of my knowledge, a comprehensive fidelity assessment for reconstruction tools has not yet been performed. Bouck *et al*[50] did a cursory assessment of STACK and the HGI using one gene, whereas Liang *et al*[34] did a more extensive analysis using 73 genes and assessing CAP3, *phrap*, TIGR Assembler (TA) and their new EST-

specific implementation of TA, called TA-EST. Their analysis however, focused only on assembly level reconstruction, i.e. it only considered the assembled contigs. Determining the fidelity of these tools is dependent on knowledge of, not only the gene sequence boundaries, but also the structure internal to the gene extremities.

1.5 The importance of "stable gene structure"

There are common elements that define gene structure: (1) a transcription initiation site, (2) a 5' untranslated region (UTR) with transcription regulation signals, (3) an initiation site for the protein coding sequence, (4) exon-intron boundaries, with splice site signals at the termini, (5) a termination site for the protein coding sequence, and (6) a 3' UTR with signals for polyadenylation and regulation. The elucidation of gene structure is helped tremendously by the availability of full-length cDNA sequences[51].

In order to accurately measure the integrity of reconstruction obtained by the methods under investigation, the output generated by these methods needs to be compared against some form of consistently annotated gene structure. A minimal description of gene structure requires only the protein coding termini and the exon structure and is therefore sufficient to assess how well these programs use sequence data to reconstruct the underlying expression events.

1.6 Thesis Organization

Chapter 1 reviews the field of transcript characterization, the various challenges facing it, the means of characterizing, as well as the use of ESTs in transcript characterization. **Chapter 2** states the aims of this research. **Chapter 3** describes the generation of the dataset, and concludes with a summary of the dataset content. The performance of these programs in the presence of the various artifacts is described and discussed in **Chapter 4**. **Chapter 5** summarizes the findings of the research as captured in Chapters 3-4.

Chapter 2 Aims

The various transcriptome technologies capture various elements of gene expression. In an attempt to analyze and organize the vast amounts of data coming from the different transcriptome projects, computational tools are needed. How well the computational tools reconstruct the underlying biology as recorded by the specific technology needs to be accurately assessed.

Sensitivity (***Sn***) reflects the extent to which a tool detects, or fails to detect, the right object or a true positive (TP) as defined by a reference set. The specificity (***Sp***) is defined in terms of the success of the tool to NOT select a wrong object or a false positive (FP).

The biological efficiency or fidelity may be defined as the extent to which the results obtained from computational tools reflect actual biology in the presence of data containing biological artifacts and phenomena. In addition to the biological variability, the data upon which the various computational tools operate may also contain error introduced in the process of obtaining the biological data.

The aim of this research is to contribute to the assessment of computational tools by creating a reference dataset consisting of sequence data from a single transcriptome technology (ESTs). The reference dataset should be as reflective of a true biological system as possible. As such, the reference dataset should include well-characterized and quantified biological phenomena. Additionally, certain data-processing error should also be included. The clustering tools (*d2_cluster*, WCD), assembly tools (*phrap*, CAP3) will be assessed for ***Sn*** and ***Sp***, as well as for the biological fidelity of the results generated by these tools.

To accomplish the aim, the following approach will be followed:

1. A reference dataset will be produced using as basis the gene dataset created by Hide *et al*[52], in which they annotate 52 exon-skipped genes, as well as the ESTs which capture these exon-skips . In addition to the exon-skip data, paralogous EST data, as well as EST data simulating sequencing error, will be added to the reference dataset.
2. The behavior of the various programs will be assessed in the presence of known sequence error, gene paralogs and exon-skipping. The fidelity of reconstruction will be assessed for the programs in the presence of these artifacts.

Supervised clustering assumes the presence of a known homolog to the gene from which the EST transcripts are obtained, whereas such a homolog may not exist. In addition, a RefSeq sequence used for 'supervision' represents a single form of the gene transcript when several transcript isoforms may exist. Partly because of these limitations, this research follows an unsupervised approach, in which no parent mRNA is used to classify or order ESTs. However, since true biology is partially represented by expressed mRNAs, parent mRNAs will be used to assess the similarity of contigs generated by the assembly tools.

Chapter 3 Dataset generation

3.1 Introduction

In order to assess how well computational tools perform the tasks for which they were designed, there needs to be a standard reference dataset against which their results can be compared. The reference dataset can be compiled in one of two ways:

1. theoretically or synthetically,
2. based on specific empirical characteristics.

To ultimately assess the performance of the selected gene expression reconstruction systems, the EST test dataset to be generated will be of the second type, in which the unifying characteristics will be exon-skipping and paralogy. In order to derive this dataset, an existing gene dataset of 52 exon-skipped genes created by Hide *et al*[52] will be used. Hide *et al*[52] have manually curated the various transcript isoforms of these exon-skipped genes, as well as the exons which are skipped in each transcript isoform.

The characteristics around which the EST test dataset will be built will include:

1. genes with known alternative splicing,
2. genes for which paralogs exist, and
3. gene-specific ESTs with known sequence error-rates.

The last criterion will be met by generating gene-specific ESTs with known error-rates based on research done by Ewing *et al*[53,54] and Liang *et al*[34].

In order to be useful for this research, the generated dataset should:

1. be able to annotate i.e.
    a. The dataset should be able to relate ESTs to the gene from which they were derived,
    b. Provide a description of the genomic region from which the gene transcripts originate e.g. providing genomic coordinates for the genomic region spanned by the gene.

2. provide a measure of EST identity to the gene from which it originates.
3. provide a record of isoform presence, numbers and composition.
4. include paralogous genes.
5. include known errors with well defined properties and characteristics.

3.2 Methods

Unless otherwise stated, all sequence and sequence-related data used in this project were obtained from the UCSC Browser based on the May 2004 Assembly of the NCBI build 35.[*]

3.2.1 Gene selection

3.2.1.1 Confirming correct HUGO identifiers

It was decided to refer to the genes by the assigned HUGO name. The 52 genes in Hide *et al*[52] gene dataset have originally been annotated according to their ENSEMBL id's and therefore did not have consistent HUGO identifiers. MatchMiner[†] is a suite of tools that uses information from different sources to correlate disparate gene ID's with each other. MatchMiner succeeded in matching the ENSEMBL ID's for these genes to RefSeq accession numbers for 42 of the 52 genes. Using BLAST (v2.2.13), the remaining 11 genes for which MatchMiner could not find a RefSeq accession number were confirmed by using the nucleotide sequence for that gene as query sequence. The best BLAST hit using default search parameters and a significance cut-off of 10e-120 was selected as the gene accession number.

Where accession numbers were present, the most recent version of that sequence was found by searching NCBI and UCSC Browser (*May 2004 human assembly*). Where uncertainty existed about multiple Genbank accession numbers, the sequence data for the gene was used to BLAST-search for the most significant sequence match, and that

[*] http://www.genome.ucsc.edu
[†] http://discover.nci.nih.gov/matchminer/index.jsp

identifier was accepted. The identifier would be the HUGO name where it existed, the Refseq ID, or the DNA accession number.

From the 52 genes, 27 (**Table 1, p22**) were selected for which protein entries exist in the SwissProt* version 49.7 protein database (See **Table 15** in **Appendix 6.6, p54**). For some of these genes, actual PDB structures were found as well. An additional selection criterion was that the paralogs found for the Hide *et al*[52] dataset be as representative of real biology as possible. That would mean that 10% of the genes should have known paralogs[40]. Only 3 confirmed paralogs could be found for these 52 genes (See **Table 3, p26**), which limited the dataset to 27 genes.

The following information was obtained for each gene from the UCSC Genome Browser:

1. the HUGO gene name,
2. sequence information from the ”Known Gene” track (which excludes introns, as well as 5’ and 3’ UTR’s, but includes all exons)
3. the total number of exons as annotated in the longest “Known Gene”
4. gene-specific ESTs, which include spliced ESTs (ESTs that span intronic regions),
5. isoform number per gene (taken to be the number of mRNAs for each gene).

The data used is based on the May 2004 Assembly of the NCBI build 35.

* http://www.ebi.ac.uk/swissprot

Gene	Exon(s) skipped (*EST GenBank Accession Number*)
ARVCF	19 (*T79735, R08546*)
ATP6E	2 (*AA332132*), 5 (*BE735148, BE732718*), 5-7 (*AI929680*)
BCR	20 (*AW025032*)
CLTCL1	29 (*AA378884*)
DGCR2	2, 3 (*BE531182*)
ECGF1	5 (*AI347252*)
EWSR1	6 (*BE311429*)
G22P1	3 (*BE018656*)
GCAT	2 (*AA670436*), 2,3,5 (*AI198343*)
GGT1	3 (*AW903997, AU077341*), 7 (*AI222095, H27285, AA917932*)
GSTT1	2 (*AA280398, AA280360, AA689400, BE280663*), 2-3 (*R05684, AV650136*), 3 (*BF343733*), 3-4 (*AA298437*)
GTPBP1	2 (*AA418991, AW592929, AW510699, AW182864, AI652565, AI474631, AW015416*)
HMG2L1	2 (*AA053700, AA223380, AA192830, AA223568), 5 (AA595272, BE745167), 2,5 (BE793346, AW374294*)
LGALS1	3 (*BE738697, BE738430, BE738129, BE737824, AA095630, AW006485, AI922873*)
MFNG	2, (*BE254149, AU143259*), 7 (*AW170461, AW166072, AI762014*)
MIL1	2 (*BE741543*), 3 (*BE900458, BE798008, AW250153, AW580672*)
NF2	2, 3 (*BE265514*)
NPAP60L	4 (*H45683*)
PIK4CA	36,37,38,39,40,41,42 (*W04181*), 50 (*BE670661*)
PMM1	4 (*R36322*)
RBX1	2 (*AW163628, AW161957, AW161517, AA843156*), 4 (*AI140018*)
SEC14L2	10 (*H06489, AA147533*)
SLC25A17	2-4 (*AA326069*), 3 (*AU123445*), 3-4 (*BE298274*)
ST13	8 (*AI424473*)
TCF20	3 (*AW366548*)
UBE2L3	2 (*BE093601*)
UFD1L	2, 3 (*R08973*)

Table 1: Exon-skipped genes as annotated by Hide et al[52], with the EST support for the exon-skips recorded in column 2 e.g. in UFD1L, exons 2 and 3 are skipped, and this is confirmed by the EST with Accession Number RO8973**. Genes are ordered alphabetically according to HUGO gene symbol.**

3.2.1.2 Obtaining unambiguous genome coordinates

The coordinates for the RefSeq sequences were obtained from the UCSC Browser (*May 2004 human assembly: NCBI Build 35*). Where there was only one RefSeq gene/ sequence per gene, the location of the gene was taken to be the coordinates of the RefSeq gene. When multiple RefSeq transcripts existed per gene, the composite coordinates were taken as the location of the gene i.e. the maximum region which includes all RefSeq transcripts. Care was taken that all mRNA data used had genomic

coordinates within this maximum genomic region.

The genes identified by Hide et al[52] are distributed throughout chromosome 22. **Table 2 (p23)** summarizes the data gathered for these genes. It records the genomic location of each gene, as well as the total number of exons contained by these genes. In addition, it records the total number of UCSC-assigned ESTs and the number of spliced ESTs i.e. ESTs spanning intronic regions.

Gene	Exons	ESTs	Spliced ESTs	Genomic Position on Chr. 22
ARVCF	20	165	67	18331974-18378863
ATP6E	9	758	543	16449489-16486044
BCR	22	379	215	21847105-21982698
CLTCL1	33	106	46	17541541-17653751
DGCR2	10	595	177	17398353-17484458
ECGF1	9	328	271	49096589-49100664
EWSR1	17	1144	1010	27988824-28021059
G22P1	13	2296	1996	40260392-40303081
GCAT	9	125	112	36447010-36455942
GGT1	16	263	102	23323736-23349524
GSTT1	5	230	149	22700695-22708825
GTPBP1	12	197	69	37426468-37452744
HMG2L1	12	235	60	33978049-34016353
LGALS1	4	1048	953	36314681-36318846
MFNG	8	229	168	36108141-36125424
MIL1	4	360	96	16546303-16586545
NF2	14	49	31	28324118-28419137
NPAP60L	7	138	62	43840612-43857701
PIK4CA	54	584	312	19386544-19517555
PMM1	8	210	152	40215945-40228910
RBX1	5	315	245	39671884-39693168
SEC14L2	10	190	68	29117486-29144382
SLC25A17	9	203	121	39409130-39458363
ST13	12	1172	387	39545102-39577187
TCF20	4	175	16	40786901-40841468
UBE2L3	4	1147	554	20246572-20302877
UFD1L	12	406	344	17812394-17841280

Table 2: Summary of genes and the ESTs covering them and their splice sites. The exon count was confirmed by the existence of protein entries for the specific gene in the SwissProt database (See Table 14 in Appendix 6.6, p54). Genes are ordered alphabetically.

3.2.2 Data processing

3.2.2.1 EST pre-processing

3.2.2.1.1 Duplicate Accession number detection and removal

Each gene-specific EST set of sequences was searched for Accession number duplicates within and across sets. Where these sequences existed they were removed.

3.2.2.1.2 Sequence masking

All the EST sequences used were masked with RepeatMasker set to mask for human repeats (with the following options: **–mam**: mask repeats in non-primate, non-rodent animals, **-pa 4**: use 4 parallel processors, **-nocut**: do not excise masked bases, **-ace**: produce additional aceDB formatted output) and DUST (masks for Low Complexity Regions or simple repeats). **Figure 3** (**p25**) summarizes some of the information obtained through RepeatMasker (GC content, as well as the percentage of EST sequence masked). For the raw EST data for each of the genes, the longest, shortest and average EST length was determined. Subsequently, sequences shorter than 100 bp were removed. **Figure 3** (**p25**) is based on information recorded in **Table 9** in **Appendix 6.1 (p47)**.

3.2.2.1.3 Short sequence removal

Sequences shorter than 100 bp, as well as sequences having less than 100 unmasked bp were removed.

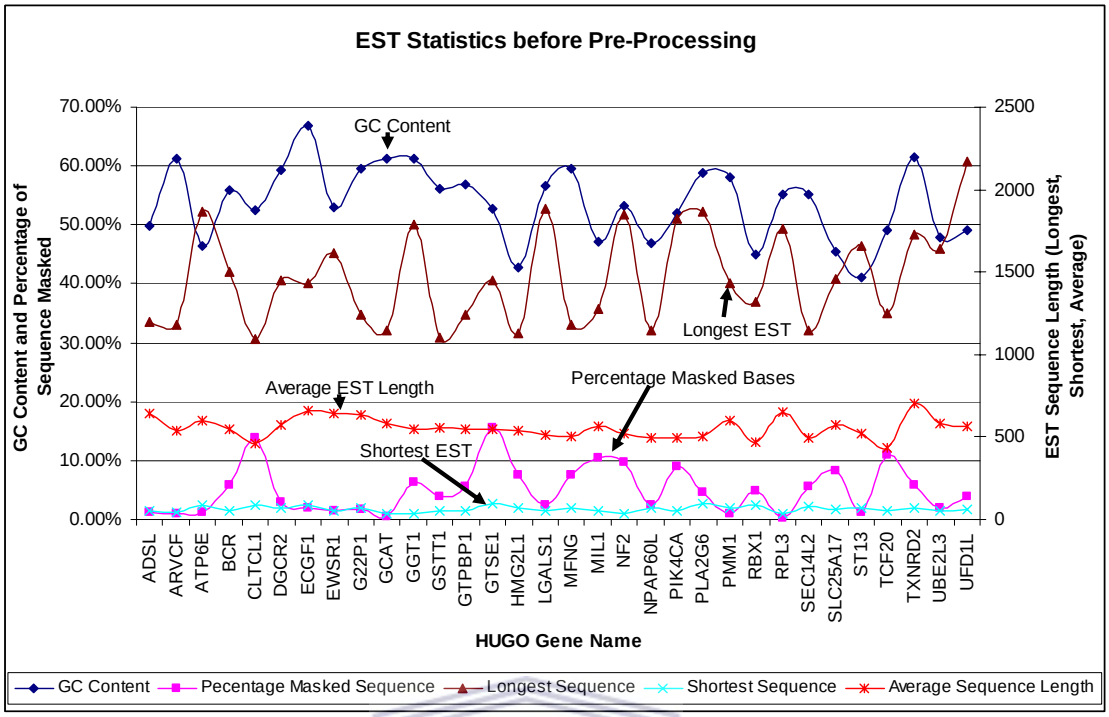

Figure 3: Raw EST data. GC content and percentage masked bases obtained from RepeatMasker. The data in Table 9 (p47) in Appendix 6.1 was used for this graph.

3.2.3 Artifactual data inclusion

3.2.3.1 Sequence error

It is commonly accepted that the most error-prone regions of ESTs are the sequence termini. Ewing and Green[53,54] empirically determined error rates for these. In order to introduce sequence error, the individual EST sequences were mutated using *msbar*, an application found in the EMBOSS suite of programs. *msbar* introduces random error within a specified sequence. These error rates varied from 1-11% as per Ewing and Green[53]. *msbar* was used as follows:

"**msbar -sequence** sequence_name **-count** number_of_mutations **-point** 1 **-block** 0 **-codon** 0 **-outseq** mutated_output_sequence"

msbar Parameters: **-sequence**: sequence to be mutated,

-count: number of mutations to introduce (*integer value*),

-point: whether to introduce point mutations (0=no, 1=yes),

-block: whether to introduce block mutations (0=no, 1=yes),

-codon: whether to introduce codon mutations (0=no, 1=yes),

-outseq: the name of the output sequence

The mutations that are introduced are (0=None, 1=Any of the following, 2=Insertions, 3=Deletions, 4=Changes, 5=Duplications, 6=Moves). Only point mutations were introduced (**-point** 1).

3.2.3.2 Paralogs

Searching GeneCards[55] and UCSC for paralogs of the Hide *et al*[52] dataset resulted in the genes summarized in **Table 3 (p26)**.

Gene	Paralog	Paralog Genomic location
*GGT1	GGT2	Chr 22: 19892266-19910582
*ST13	FAM10A3	Chr. 12: no known coordinates
	FAM10A4	Chr. 13: 49644155-49645750
	FAM10A5	Chr. 11: 18240031-18241622
	FAM10A6	Chr. 8: 134489324-134490574
	FAM10A7	Chr. 7: 132310019-132312611
†UBE2L3	UBE2L6	Chr. 11: 57075704-57091756
†GSTT1	GSTT2	Chr. 22: 22624162-22650652
†ATP6E	ATP6V1E2	Chr. 2: 46650638-46658747

Table 3: Paralogs found by searching GeneCards and UCSC for annotated paralogs. (*) GeneCards match, (†) UCSC match. Paralog Genomic Location: GeneCards/ UCSC coordinates for the genomic location of the paralog.

To have a dataset that is representative of real biological data with an estimated 10% paralog presence in gene data[40], only three of these paralogous genes were included: GGT2, UBE2L6 and ATP6V1E2.

The ST13 family members were excluded since 2 of them (FAM10A6 and FAM10A7) lacked mRNA sequence data, and one (FAM10A3) lacked genomic coordinates. GSTT2 was excluded because it has an ambiguous assignment to two separate genomic locations in alternative orientations.

3.2.3.3 Exon-skipping

The subset of genes used contains information about which exons are skipped as summarized by Hide et al[52] and recorded in **Table 1 (p22)**.

3.3 Discussion

For each of the genes, the mRNAs associated with each gene were downloaded from UCSC. As a measure of the integrity of the ESTs assigned to each gene, the ESTs were aligned to the each of the gene-specific mRNAs with BLAST. **Table 4 (p28)** shows the number of ESTs classified based on the sequence identity attained. If an EST aligns to an mRNA with a sequence identity of between 80 and 85% of the entire EST length, it is classified as a **Class D** EST, if 85 to 90% identity exists, it is a **Class C** EST, a **Class B** if 90-95% identical. A **Class A** EST is more than 95% identical to the target mRNA. **Table 4 (p28)** shows that most ESTs are more than 90% identical to the mRNAs assigned to each gene.

Gene	Class A	Class B	Class C	Class D
ARVCF	143	2	0	0
ATP6E	787	12	1	0
BCR	281	6	0	0
CLTCL1	71	0	0	0
DGCR2	638	7	0	0
ECGF1	301	10	3	0
EWSR1	1142	11	1	0
G22P1	0	0	0	0
GCAT	137	0	0	0
GSTT1	219	4	0	0
GTPBP1	178	1	0	0
HMG2L1	180	4	0	0
LGALS1	952	17	0	0
MFNG	185	2	1	0
MIL1	326	4	0	0
NF2	208	2	0	0
NPAP60L	103	0	0	0
PIK4CA	424	0	0	0
PMM1	208	4	0	0
RBX1	242	4	0	0
SEC14L2	173	1	0	0
SLC25A17	211	2	0	0
ST13	810	18	2	0
TCF20	69	0	0	0
UBE2L3	1063	10	1	0
UFD1L	367	3	0	0

Table 4: Indication of the sequence identity of the EST dataset to the parent mRNAs obtained from UCSC for the subset of Hide et al[52] geneset. Class A: 95-100% identity to mRNA, Class B: 90-95% identity to mRNA, Class C: 85-90% identity to mRNA, Class D: 80-85% identity to mRNA.

As an additional measure of the integrity of the UCSC EST assignment to a specific gene, two databases (TGI[*] and Unigene[32]) with their own methods for doing EST-to-gene assignments were selected, and their EST assignments were compared to those of UCSC.

[*] http://www.tigr.org/db.shtml

TGI identifies all sequence overlaps between EST sequences. It then uses TIGR Assembler to join, through transitive closure, sequences which are more than 95% identical over more than 40 base pairs. Unigene uses BLAST to compare the complete set of organism genes to itself. An initial cluster of highly similar genes is created and ESTs are aligned and added to these initial clusters. **Table 10 (p49)** in **Appendix 6.2** records the HUGO names for each gene, as well as the corresponding TGI and Unigene gene index ID's.

If an EST is assigned to a gene by all three databases, it would imply a high integrity sequence and as such, is assigned a ***Class I*** status. If only two out of three databases assign that EST to a gene, it is a ***Class II*** EST, else the EST in UCSC is a ***Class III*** EST (see **Table 11 (p51)** in **Appendix 6.3**). This class assignment is specific with respect to the ESTs contained in the UCSC data i.e. ***Class I*** + ***Class II*** + ***Class III*** = total number of UCSC ESTs.

Figure 4 (p30) summarizes the measure of confidence in the EST-to-gene assignment as annotated by UCSC. For all of the genes except for MIL1, both Unigene and TGI, or either of Unigene or TGI concurs with the UCSC EST-to-gene assignment.

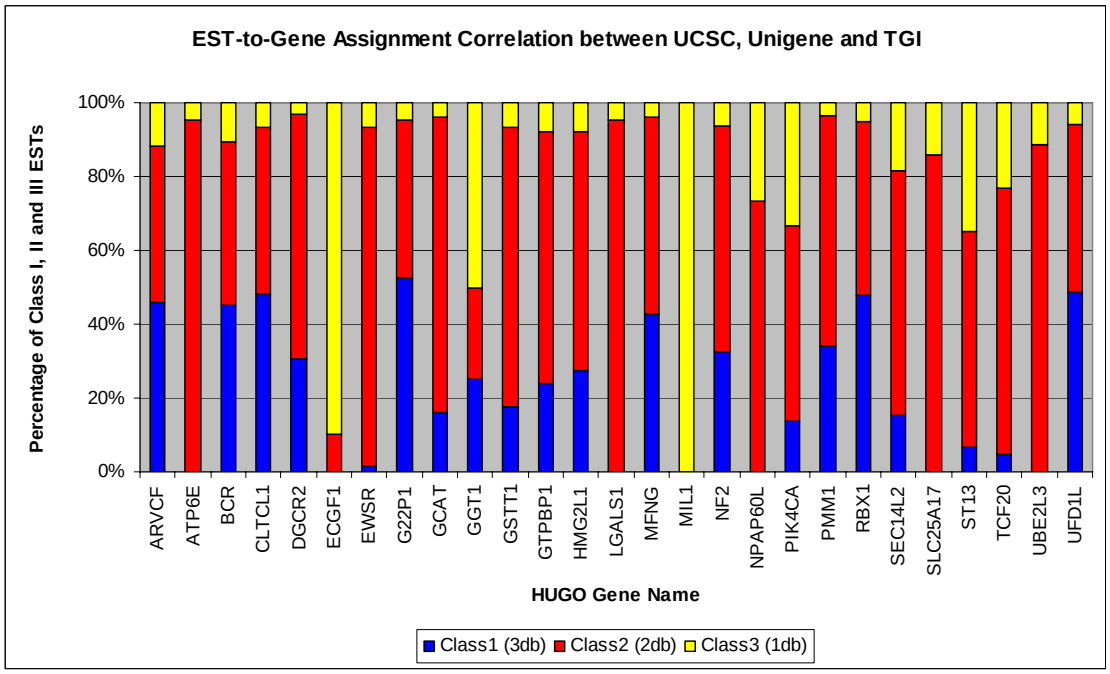

Figure 4: Classification of ESTs based on the concurrence in EST-to-gene assignment between UCSC, Unigene and TGI. Based on data contained in Table 11 (p51) in Appendix 6.3.

Chapter 4 Analysis of results produced by the evaluated programs

4.1 Introduction

4.1.1 Need for fidelity assessment (especially in the presence of artifacts)

The rationale for doing fidelity assessment of these programs is two-fold: firstly, there is the variability of biological systems and secondly, the errors present in the recording of biological sequence data.

Biology is unpredictable and does not always follow our distilled observations of phenomena nor our idealized hypotheses or theories thereof. Therefore, the tools which aim to discover biological features have to be assessed on their ability to do so in the presence of biological data variability, as well as on the ability to generate results which reflect real-life biology. In essence, given a dataset containing evidence for natural phenomena (e.g. ESTs which capture expression events), any program should be assessed on its ability to reconstruct that phenomenon (e.g. an expression event) as accurately as possible.

Although at present stricter quality control measures are being enforced with regard to biological sequence submission, there are already low-integrity sequences present in the existing databases. As a rule, when using EST databases, the first step should be standard cleaning procedures which include masking for contaminants (e.g. genomic, vector, bacterial, and mitochondrial sequences) as well as for the wide range of repeats present in human sequence data.

Once all cleaning measures have been implemented, certain sequence features still exist which can negatively impact the efforts of transcript reconstruction and therefore

of cataloging these transcripts. These features include, but are not limited to, possible chimeras, alternative transcripts, sequencing and database errors, as well as paralogs.

A truly successful program would provide the best possible reconstruction of an expression event amidst these additional sequence features. A consistent measure of the success of a reconstruction attempt is therefore crucial when assessing these programs.

4.2 Results

4.2.1 Rand Index for each program

The Rand Index (**RI**) is a measure of the similarity between two datasets. In this case the program results are compared to the reference EST dataset. The rand Index is calculated as follows:

$$RI = \frac{a+d}{a+b+c+d}$$

where *a* and *d* are the number of agreements between the two datasets, and *b* and *c* are the number of disagreements between the two datasets. Therefore, the lower the number of disagreements *b* and *c*, the more **RI** tends towards 1.

RI values range from 0 to 1, with higher values indicating higher similarity e.g. a value of 1 would mean the groupings are identical. The sets (clusters or contigs) produced by the various programs were compared to the reference dataset based on UCSC assignments. The RI values for each gene were calculated and averaged over the 27 selected genes (**Table 5 (p32)**).

	phrap	**CAP3**	***d2_cluster***	**WCD**
RI	0.8953	0.8750	0.9341	0.9329

Table 5: Average Rand Index (RI) results for *phrap*, CAP3, *d2_cluster* and WCD. RI gives an indication of the similarity between two groupings. The reference grouping is the known EST membership per gene, and the second grouping is the grouping obtained from a specific program.

4.2.2 Sensitivity (Sn) and Specificity (Sp)

Sensitivity (***Sn***) reflects the extent to which a tool detects, or fails to detect, the right object or a true positive (TP) as defined by a reference set. ***Sn*** is dependent on how many true positives are recognized out of the total number of reference set of true positives (TP+FN) where FN (false negatives) is the number of reference set objects a tool fails to detect. Therefore:

$$Sn = TP/ (TP+FN)$$

The specificity (***Sp***) is defined in terms of the success of the tool to NOT select a wrong object or a false positive (FP). ***Sp*** is affected by the number of objects rightly excluded from being selected i.e. true negatives (TN). Therefore:

$$Sp = TN/ (TN+FP)$$

For the analysis of ***Sn*** and ***Sp***, each program processed a composite EST dataset comprised of the reference dataset of 27 genes (all positive) from the Hide *et al*[52] set, as well as the ESTs belonging to the selected paralogous genes (all negatives). In this context, the true positives contained in a cluster or contig would be the ESTs which make up the majority of that cluster or contig. The rest of the members for this cluster or contig would be labeled false positives.

For example, if *cluster A* consists of 40% of ESTs from gene1, 30% of ESTs from gene2 and 30% of ESTs from gene3, *cluster A* is representative of gene1. For *cluster A* then, the 40% of ESTs for that cluster are counted as true positives, and the rest (60%) are false positives. False negatives are those ESTs belonging to gene 1 which have been assigned as singletons, or have been assigned to the cluster defined by another gene. True negatives are the paralog ESTs that have not been assigned to any of the clusters defined by the reference dataset. A summary of TP, FP, TN, FN, Sn and Sp is shown in **Table 6**, **p34**.

	TP	FP	TN	FN	Sn	Sp
CAP3	13055	49	570	879	0.94	0.92
Phrap	12372	627	564	637	0.95	0.47
D2	12471	478	552	561	0.96	0.54
WCD	12634	477	565	831	0.94	0.54

Table 6: Sensitivity and Specificity values for the composite set of 27 reference gene ESTs and 3 paralogous gene ESTs [Sn=TP/(TP+FN), Sp = TN/(TN+FP)].

4.2.3 Contigs generated by *phrap* and CAP3

Using the default parameter settings for *phrap* and CAP3, a number of contigs were produced. **Figure 5 (p34)** shows the contig-to-mRNA ratio for CAP3 and *phrap*. The data upon which this figure is based is shown in **Table 12** in **Appendix 6.4 (p53)**.

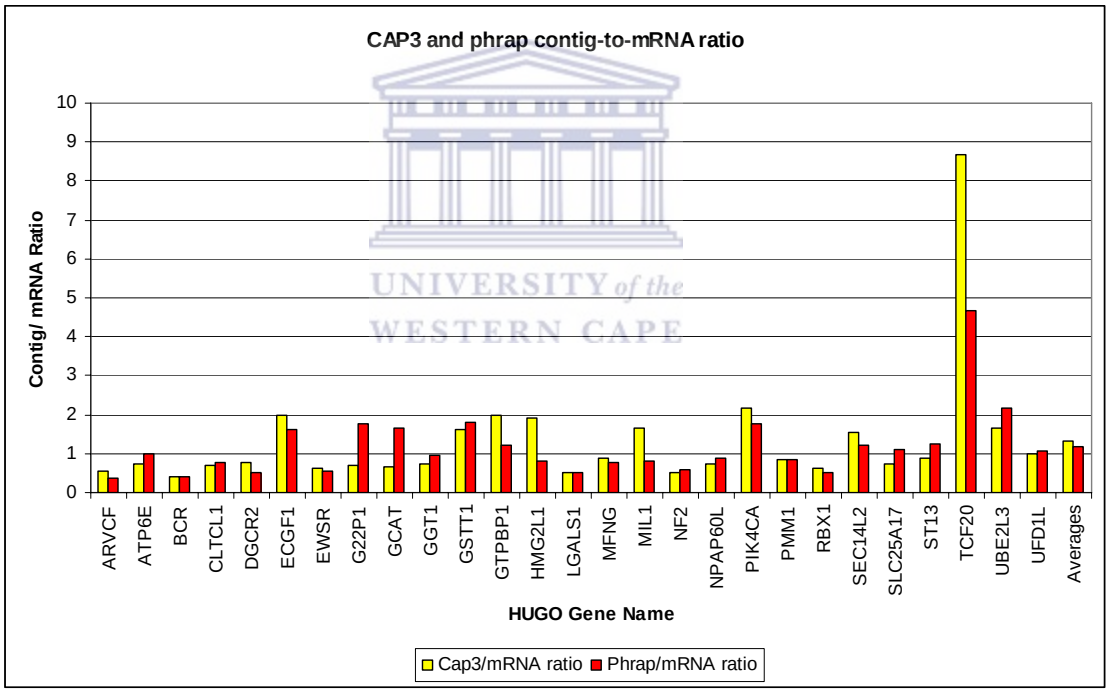

Figure 5: Ratio of contigs generated by CAP3 and *phrap* vs. the number of mRNAs assigned to each gene by UCSC. The data for this table is contained in Table 12 in Appendix 6.4, p53.

4.2.4 Skipped ESTs missed by CAP3

Skipped ESTs contain information about which exons are skipped. Their inclusion into any cluster and assembly therefore adds exon-skipping information to the resultant grouping. For all of the selected genes, *phrap* has incorporated the skipped ESTs in its analysis and in results. CAP3 fails to incorporate such exon-skipping information for 37% (10 of the 27) of genes analyzed. For those 10 genes, 12-100% of exon-skipping information is lost (See **Table 7**, **p35**). All of these missed ESTs have higher than 98% identity to the parent mRNAs.

HUGO Name	Skipped EST Missed	Total Skipped ESTs	Percentage of Skipped ESTs missed	EST Acc Number
ATP6E	1	4	25.00%	AI929680
GCAT	2	2	100.00%	AA670436, AI198343
GSTT1	1	8	12.50%	AA298437
LGALS1	1	7	14.29%	BE738129
NF2	1	1	100.00%	BE265514
NPAP60L	1	1	100.00%	H45683
RBX1	1	5	20.00%	AI140018
SLC25A17	1	3	33.33%	AU123445
ST13	1	1	100.00%	AI424473
UFD1L	1	1	100.00%	R08973

Table 7: CAP3 results with respect to exon-skipped ESTs. CAP3 assigns these skipped ESTs to the singlet class, thereby losing alternate transcript information. Phrap incorporates all of the exon-skipped ESTs.

4.2.5 Program output

The basic outputs obtained from the programs tested are contigs and clusters, in the instances where sequences could be grouped together. Sequences which could not be grouped together are labeled as singletons (in the case of clustering) or singlets (in the case of contig assembly). The results for the assemblers (*phrap*, CAP3) and the clusterers (WCD, *d2_cluster*) are shown in **Figure 6 (p36)** with information extracted from **Table 13** in **Appendix 6.5 (p54)**. The high contig-to-singlet ratio for *phrap*, and cluster-to-singleton ratio for *d2_cluster* and WCD are mostly due to the lower number of singlets/ singletons.

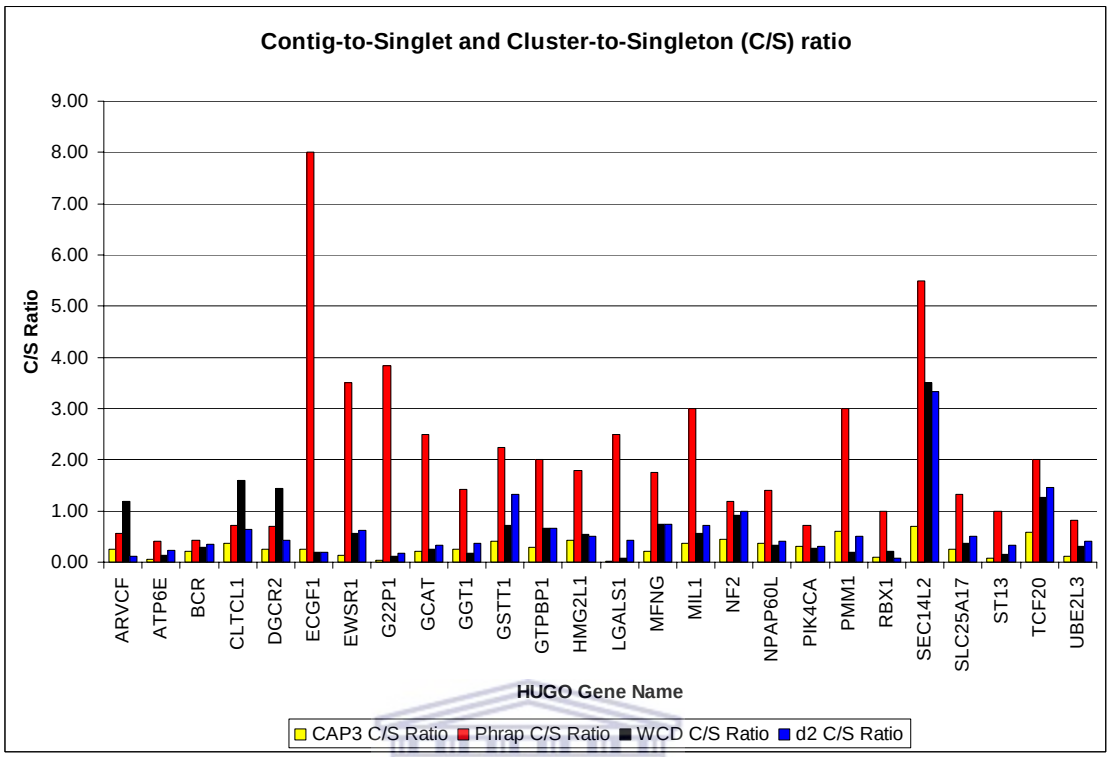

Figure 6: Summary of Contig-to-Singlet Ratio for assemblers and Cluster-to-Singleton ratio for clustering tools. Data for this table is recorded in Table 13 (p54) in Appendix 6.5.

4.2.6 BLAST matches to longest Assembler-generated contigs

In order to get some measure of how well the reconstructed genes resemble the known sequences, the contigs generated by *phrap* and CAP3 were searched against a database consisting of the longest representative mRNAs of the original genes. From the BLAST results, the extent to which the contig spans or covers the length of the representative mRNA, was recorded as 'coverage' (See **Table 8, p37**).

HUGO Name	Genbank Transcript Identifier	*phrap* Contig Length	*phrap* Best BLAST Match	*phrap* Coverage	CAP3 Contig Length	CAP3 Best BLAST Match	CAP3 Coverage	*phrap/* CAP3
ARVCF	U51269	3470	U51269	63.52%	4654	U51269	60.55%	0.75
ATP6E	BC004443	2133	BC004443	10.60%	2704	BC004443	47.12%	0.79
BCR	X02596	5392	X02596	68.01%	5339	X02596	78.74%	1.01
CLTCL1	X95486	3222	X95486	95.84%	3064	X95486	98.56%	1.05
DGCR2	D79985	4966	D79985	73.84%	3543	D79985	83.52%	1.40
ECGF1	BC052211	2221	BC052211	33.00%	2256	BC052211	69.90%	0.98
EWSR	X66899	3211	None	None	2629	X66899	88.66%	1.22
G22P1	BC008343	2833	BC008343	70.60%	3213	BC008343	64.43%	0.88
GCAT	AK123190	1588	AK123190	65.05%	1490	AK123190	87.38%	1.07
GSTT1	BC007065	1335	BC007065	17.98%	1676	BC007065	57.16%	0.80
GTPBP1	AF077204	4621	None	None	2554	None	None	1.81
HMG2L1	AL079310	4416	AL079310	86.30%	2578	AL079310	87.35%	1.71
LGALS1	BC020675	998	None	None	954	BC020675	55.14%	1.05
MFNG	U94352	2293	U94352	63.37%	2069	U94352	83.37%	1.11
MIL1	AF146568	4104	AF146568	39.47%	2041	None	None	2.01
NF2	AF369658	4631	AF369658	80.89%	2561	AF369658	99.41%	1.81
NPAP60L	BC028125	3326	BC028125	45.85%	3516	BC028125	45.11%	0.95
PIK4CA	AF012872	4821	AF012872	87.45%	3893	AF012872	99.67%	1.24
PMM1	BC016818	1605	BC016818	58.26%	2312	BC016818	52.94%	0.69
RBX1	BC017370	1818	BC017370	10.34%	2501	BC017370	7.36%	0.73
SEC14L2	AL096881	3321	AL096881, AB006630	80.75% 80.75%	3108	AL096881, AB006630	90.07% 90.07%	1.07
SLC25A17	BC005957	2435	BC005957	53.96%	2010	BC005957	76.67%	1.21
ST13	BC052982	4012	BC052982	27.54%	3533	BC052982	88.85%	1.14
TCF20	AB006630	3068	AB006630	91.75%	2793	AB006630	98.68%	1.10
UBE2L3	AJ000519	3320	AJ000519	27.62%	2674	AJ000519	84.37%	1.24
UFD1L	BC005087	1878	BC005087	9.27%	2280	BC005087	45.00%	0.82
Average		**3117**		**54.84%**	**2767**		**72.92%**	**1.14**

Table 8: Contig vs. mRNA BLAST results: This table summarises the results obtained after searching the longest contigs generated by *phrap* and CAP3. Column 1: HUGO name - the accepted HUGO identifier for the known gene. Column 2: Genbank Transcript Identifier - the identifier of the complete mRNA transcript. Columns 3 and 6: Contig Length - the longest contig generated each assembler. Columns 4 and 7: Best BLAST Match - the best BLAST match when the assembler-generated contig is searched against the database consisting of only the gene-specific mRNAs in column 2. Columns 5 and 8: Coverage - defined as the percentage of contigs that align to the total mRNA length. Column 9: *phrap*/CAP3 – the ratio of *phrap* (Column 3) and CAP3 (Column 6) contig length.

4.3 Discussion

The fidelity with which tools reconstruct an underlying expression event as captured by ESTs is determined by 1) the ability to correctly assign ESTs from a single gene to a single gene class, 2) the biological validity of the reconstructed event.

4.3.1 Correct assignment of member ESTs

It needs to be iterated that an unsupervised clustering and assembly method has been followed. The correct assignment of member ESTs were defined by the EST-to-gene assignments done by UCSC. UCSC uses BLAT[56] to align the ESTs to the genome, insisting that there be at least a 93% base identity over the entire alignment length. Therefore the reference set of ESTs has also been obtained by an unsupervised method.

4.3.1.1 The Rand Index values

The average Rand Index (**Table 5**, **p33**) appears to show that the clustering tools produce group assignments which correlate more highly with the reference dataset. The difference in method of finding related sequences is evident between the sequence similarity-based assemblers and the word-count-based clusterers. RI is a normalized count of the pairs of sequences that were treated alike by the different algorithms. Similarity-based methods would fail to detect sequence similarity between a pair of sequences where the word-count based method would determine a sequence relationship.

The similar RI values for *phrap* (0.89) and CAP3 (0.87) in **Table 5**, (**p32**) would imply similar results. However, the sensitivity and specificity values discussed in the following **section 4.3.1.2**, as well as the results in **Table 8**, **p37** (discussed in **section 4.3.2**) show that a high correlation in RI does not mean high integrity of reconstruction.

4.3.1.2 Sensitivity and Specificity

All the tested programs are fairly successful in determining the paralogous data e.g. all capture between 552 and 570 of the 574 paralogous ESTs (**Table 6**, **p34**). This would mean that at least 96.2% of all paralogous data would be distinguished from its family members. The failure of *phrap*, *d2_cluster* and WCD to be more discriminatory in the inclusion of paralogous data is evident from the very low ***Sp*** values (0.47-0.54). The programs are also relatively successful at determining which ESTs belong to a specific gene class with ***Sn*** values between 0.94 and 0.96. All the programs are also fairly consistent in assigning as false negatives those 160 true positives which are shorter than 50 bp.

4.3.2 Biological validity

From inspection of BLAST alignments (**Table 8**, **p37**), it can be seen that *phrap* generates contigs that have lower similarity to the representative mRNAs. The reduction in BLAST matches brought on by excluding matches with less than 95% identity could mean one of two things. Either, the low identity contigs are of such low integrity, that it generates spurious hits, or it does not map to contiguous regions. If the latter, that would mean that *phrap* is better able to capture exon-skips, as the inclusion of all of the skipped ESTs would suggest. *phrap* Generates longer contigs than CAP3, as can be seen by the average *phrap vs*. CAP3 length ratio of 1.14 (**Table 8**). This might be due to the higher number of included ESTs used by *phrap* for its assembly.

Contigs generated by both assemblers seem to differ from the parent mRNA to such an extent that no similarity can be found between contig query and target parent. This can be seen when looking at columns 4 and 7 of **Table 8** (**p37**). *phrap*-Generated contigs miss 3 out of 26 genes (11.54%) whereas CAP3-generated contigs miss 2 out of 26 genes (7.70%). Both assemblers generate contigs which fail to resemble GTPBP1. The reason for this is not clear. CAP3 also generates virtual transcripts which have better average width coverage (coverage over the length of the parent transcript) than *phrap* (72.92% vs. 54.84%). It would appear as if CAP3 is good at reconstructing a **single**

high–integrity sequence with longer coverage, whereas *phrap* has better ability to incorporate alternative splicing data (**Table 7**, **p35**), deemed to be low-integrity sequences.

Both the assemblers produce contigs which have high coverage statistics for the native SEC14L2 mRNA, as well as for the mRNA transcript associated with the gene TCF20. A search of the location of the genes (SEC14L2: 29,117,486 – 29,144,382 vs. TCF20: 40,880,516 – 40,935,078), shows that these genes have no overlap whatsoever. The paralog list which was determined for this dataset does not indicate that these genes are in any way paralogous.

CAP3 and *phrap* both use sequence identity to relate sequences to each other, making these programs more prone to insufficient sequence overlap (*ISO*)[38]. In the presence of *ISO*, sequences belonging to the same gene may be placed in a separate cluster or contig. This may explain the higher number of contigs than actual transcripts produced (**Figure 6**, **p36**). Whether these additional transcripts are novel expressions or just assembly artifacts has yet to be investigated. Non-alignment methods (WCD, *d2_cluster*) are not as sensitive to *ISO*, which may account for the higher correlation (**Table 5**, **p32**) to the known EST clustering/ grouping.

Chapter 5 Conclusion

In general, the highly variable nature of biological data requires diverse means of properly describing and mining this data. A single tool or utility is unlikely to do justice to the richness of biology and the data captured from biological systems. A suite of such tools, at best, would capture only a grainy snapshot of biological phenomena in time.

With regard to EST organization and ordering, it is no less true. In the face of alternative splicing, a looser grouping or clustering approach, as in *d2_cluster*, WCD and *phrap*, appears to be a better option for capturing that diversity. Unfortunately, this looser approach also allows the inclusion of lower integrity sequences under the guise of sequence variability. The low-integrity nature of ESTs makes the loose clustering approach appropriate.

The stricter approach used by CAP3 is more appropriate where the sequences are of higher integrity i.e. sequences with higher coverage than the single-pass nature of ESTs. It must be kept in mind that most of the assembly tools have been developed with high-quality sequences as source material.

With the vast amounts of biological data being generated, human analysis of said data becomes, at the very least, a daunting task and at most, impossible. Computational tools for analyzing data are becoming more ubiquitous. The success of these tools to extract the underlying biology that give rise to the data, needs to be measured consistently. A mere comparison of novel tools to existing tools only gives a relative, maybe erroneous measure of the success of the tool to reconstruct the underlying biology. The best means of assessment of computational tools remains biological data with well-characterized features.

This thesis has undertaken to generate a well-characterized dataset that can be used to test computational tools that use ESTs to reconstruct expression events, especially in the presence of sequence errors, the presence of gene families, and the presence of exon-skipping. The thesis has created the dataset aimed for:

- The reference dataset contains quality-characterized EST sequences which have been characterized according to there identity to the related gene (**Table 4, p28**), as well as to the fidelity of assignment by selected Gene Indices (**Figure 4**, **p30**; **Table 11**, **p51**).
- The reference dataset clearly relates each EST member to the gene from which it is estimated to originate.
- The reference dataset contains:
    1. quantified EST sequence error (1-11%)
    2. 10% annotated gene paralogs (GGT2, ATP6V1E2, UBE2L6) (**Table 3**, **p26**)
    3. EST's capturing the exon-skips recorded in **Table 1**, **p22**.
- The reference dataset unambiguously demarcates the genomic location of each gene.

The generated dataset can be found at http://www.sanbi.ac.za/~mario/dataset.tgz.

Extension of this research may focus on:

- Better annotation of the internal gene structure of each gene in order to elucidate the exon structure for each gene.
- Testing additional programs and tools on the paralogous, sequence error and the exon-skipping information contained in the generated test dataset.

References

1. Ruan,Y., Le Ber,P., Ng,H.H. & Liu,E.T. Interrogating the transcriptome. *Trends Biotechnol.* **22**, 23-30 (2004).

2. Saha,S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508-512 (2002).

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

4. Carninci,P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327-336 (1996).

5. Kodzius,R. *et al.* Absolute expression values for mouse transcripts: re-annotation of the READ expression database by the use of CAGE and EST sequence tags. *FEBS Letters* **559**, 22-26 (2004).

6. Shiraki,T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A* **100**, 15776-15781 (2003).

7. Brenner,S. *et al.* In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U. S. A* **97**, 1665-1670 (2000).

8. Brenner,S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630-634 (2000).

9. Pollock,J.D. Gene expression profiling: methodological challenges, results, and prospects for addiction research. *Chemistry and Physics of Lipids* **121**, 241-256 (2002).

10. Gerhard,D.S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14**, 2121-2127 (2004).

11. Yudate,H.T. *et al.* HUNT: launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res.* **29**, 185-188 (2001).

12. Anisimov,S.V. & Sharov,A.A. Incidence of "quasi-ditags" in catalogs generated by Serial Analysis of Gene Expression (SAGE). *BMC. Bioinformatics.* **5**, 152 (2004).

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

14. Bashiardes,S. & Lovett,M. cDNA detection and analysis. *Curr. Opin. Chem. Biol.* **5**, 15-20 (2001).

15. Hayashizaki,Y. & Kanamori,M. Dynamic transcriptome of mice. *Trends in Biotechnology* **22**, 161-167 (2004).

16. Sugano,S., Suzuki,Y., Yamashita,R. & Nakai,K. Oligo-capped cDNAs for promoter identification and annotation. *International Congress Series* **1246**, 233-239 (2002).

17. Brenner,S. *et al.* In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U. S. A* **97**, 1665-1670 (2000).

18. Czechowski,T., Bari,R.P., Stitt,M., Scheible,W.R. & Udvardi,M.K. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* **38**, 366-379 (2004).

19. Adams,M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991).

20. Venter,J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

21. Sun,M. *et al.* SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC. Genomics* **5**, 1 (2004).

22. Gerhard,D.S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14**, 2121-2127 (2004).

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

24. Shoemaker,D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922-927 (2001).

25. Aaronson,J.S. *et al.* 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 (1996).

26. Burke,J., Wang,H., Hide,W. & Davison,D.B. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Research* **8**, 276-290 (1998).

27. Camargo,A.A. *et al.* The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. U. S. A* **98**, 12103-12108 (2001).

28. de Souza,S.J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. U. S. A* **97**, 12690-12693 (2000).

29. Dias Neto,E. *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. U. S. A* **97**, 3491-3496 (2000).

30. Christoffels,A. *et al.* STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.* **29**, 234-238 (2001).

31. Miller,R.T. *et al.* A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9**, 1143-1155 (1999).

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

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

34. Liang,F. *et al.* An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* **28**, 3657-3665 (2000).

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

36. Quackenbush,J. *et al.* The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159-164 (2001).

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

38. Wang,J.P. *et al.* EST clustering error evaluation and correction. *Bioinformatics.* **20**, 2973-2984 (2004).

39. Fitch,W.M. Homology - a personal view on some of the problems. *Trends in Genetics* **16**, 227-231 (2000).

40. Taylor,J.S. & Brinkmann,H. 2R or not 2R? *Trends in Genetics* **17**, 488-489 (2001).

41. Jasinska,A. & Krzyzosiak,W.J. Repetitive sequences that shape the human transcriptome. *FEBS Lett.* **567**, 136-141 (2004).

42. Hadley,C. Righting the wrongs - DNA and protein sequence databases are increasingly useful research tools. But to maximize their potential, the errors in them need to be addressed. *Embo Reports* **4**, 829-831 (2003).

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

44. Huang,X. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics* **14**, 18-25 (1992).

45. Huang,X. An improved sequence assembly program. *Genomics* **33**, 21-31 (1996).

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

47. Lee,C., Grasso,C. & Sharlow,M.F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452-464 (2002).

48. Lee,C. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* **19**, 999-1008 (2003).

49. Pertea,G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics.* **19**, 651-652 (2003).

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

51. Strausberg,R.L. *et al.* An international database and integrated analysis tools for the study of cancer gene expression. *Pharmacogenomics J.* **2**, 156-164 (2002).

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

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

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

55. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. & Lancet,D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* **13**, 163 (1997).

56. Kent,W.J. BLAT---The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664 (2002).

Chapter 6 Appendices:

6.1 Summary of Raw EST data

All of the ESTs were masked with RepeatMasker and subsequently masked with DUST. Columns 2 and 3 in **Table 9** (**p47**) summarizes data reported by RepeatMasker. Columns 4-6 shows length statistics for the ESTs contained in the UCSC EST-to-gene assignment.

HUGO Gene Name	GC Content	Percentage Sequence Masked	Longest EST Sequence Length	Shortest EST Sequence Length	Average EST Sequence Length
ARVCF	61.15%	1.07%	1179	40	535
ATP6E	46.54%	1.17%	1868	90	596
BCR	55.85%	5.78%	1503	50	545
CLTCL1	52.53%	13.85%	1092	84	462
DGCR2	59.27%	2.93%	1447	68	576
ECGF1	66.96%	1.91%	1436	91	659
EWSR1	52.87%	1.36%	1616	50	646
G22P1	59.57%	1.82%	1244	72	635
GCAT	61.23%	0.52%	1143	31	581
GGT1	61.18%	6.26%	1789	37	545
GSTT1	56.25%	3.97%	1099	50	554
GTPBP1	56.93%	5.64%	1244	50	547
HMG2L1	42.81%	7.47%	1127	73	539
LGALS1	56.55%	2.32%	1884	50	515
MFNG	59.49%	7.46%	1178	68	504
MIL1	47.20%	10.57%	1277	50	567
NF2	53.34%	9.72%	1845	39	521
NPAP60L	46.94%	2.43%	1146	69	499
PIK4CA	52.03%	9.00%	1822	50	494
PMM1	58.00%	0.94%	1428	72	599
RBX1	44.87%	4.84%	1317	91	469
SEC14L2	55.14%	5.56%	1144	76	493
SLC25A17	45.34%	8.36%	1456	64	573
ST13	41.15%	1.27%	1660	68	518
TCF20	48.98%	11.00%	1250	50	430
UBE2L3	47.82%	1.84%	1640	50	578
UFD1L	49.11%	3.84%	2167	64	568

Table 9: Gene-specific EST statistics of raw EST data. Column 1 contains the HUGO name of gene, GC content: Total GC content of the ESTs for each gene, Percentage Sequence Masked: Percentage of bases masked by RepeatMasker, Longest, Shortest and Average Length of the ESTs for each gene.

6.2 Unigene clusters and TGI Tentative Human Consensi ID's corresponding to the UCSC gene

Both TGI and Unigene were searched for the cluster assignment correlating to the HUGO Gene name of the reference dataset. The results are shown in **Table 10** (**p49**). In order to do the analysis reported in **Table 11** (**p51**), multiple clusters for the each gene were combined into a single file e.g. for BCR, the Unigene files Hs.474328, Hs.517461, Hs.534451and Hs.551463 were combined into a single file. Similarly, the multiple TGI files for BCR (THC2243616, THC2256273, THC2430310, THC2434400, THC2264599, THC2445841) were combined. Where no cluster was found for a specific gene, this was indicated by "*No cluster found*".

Gene	Unigene ID	TGI Acc Num
ARVCF	Hs.326730	THC2261875
ATP6E	Hs.517338	*No cluster found*
BCR	Hs.474328, Hs.517461, Hs.534451, Hs.551463	THC2243616, THC2256273, THC2430310, THC2434400, THC2264599, THC2445841
CLTCL1	Hs.368266	THC2242120, THC2248235
DGCR2	Hs.517357	THC2244118, THC2401426
ECGF1	*No cluster found*	THC2246034, THC2256759, THC2256758, THC2256760
EWSR1	Hs.374477	THC2398091
G22P1	Hs.292493	THC2255256
GCAT	Hs.54609	THC2236271
GGT1	Hs.444164	THC2242753, THC2246917, THC2252944
GSTT1	Hs.268573	THC2235456, THC2240483, THC2335207, THC2346894
GTPBP1	Hs.276925	THC2233956, THC2257788, THC2371426
GTSE1	Hs.386189, Hs.475140	THC2256618, THC2264408, THC2257974, THC2361387
HMG2L1	Hs.197086, Hs.588815	THC2240131, THC2257233, THC2257234
LGALS1	Hs.445351	THC2233894,THC2272272, THC2254242, THC2398817
MFNG	Hs.517603	THC2234903, THC2409491
MIL1	Hs.118681	*No cluster found*
NF2	Hs.187898	THC2242050, THC2252009, THC2259070, THC2259071, THC2259072, THC2259073, THC2276539, THC2276541, THC2285460
NPAP60L	Hs.475103	THC2247389
PIK4CA	Hs.529438	THC2256070
PMM1	Hs.75835	THC2257433, THC2434093
RBX1	Hs.474949	THC2244889, THC2404816
SEC14L2	Hs.335614	THC2234283, THC2246132, THC2338679, THC2263909
SLC25A17	Hs.474938	THC2257286
ST13	Hs.546303, Hs.558698, Hs.567998	THC2254921, THC2262045
TCF20	Hs.475018	THC2246637, THC2264092
UBE2L3	Hs.108104	THC2250568, THC2309103

Table 10: For each gene, the corresponding matching Unigene and TIGR gene clusters were found that correspond to the ESTs assigned to a gene by UCSC. "*No cluster found*" means that no clusters were assigned to the specific HUGO gene name.

6.3 EST Classification Based on Database correlation

Table 11 (**p51**) shows the number of ESTs contained in each of the EST-to-gene assignments for UCSC, Unigene and TGI. The columns labeled "TGI Count", "Unigene Count" and "UCSC Count" record the number of ESTs assigned by each database. In most instances Unigene have larger EST datasets per gene than either UCSC or TGI. This is reflected in the lay-out of **Table 11**: EST assignment number increases across the table from left-to-right. The cells labeled "None" mean that no EST-to-gene assignment was found for that specific gene e.g. for ATP6E, MIL1, SLC25A17 and ST13, no TGI assignments were found. Since the basis for the reference dataset is data obtained from UCSC, the classification of ESTs is dependent on the data contained in the UCSC EST-to-gene assignments. Therefore, the values contained in columns 5-7 sum to the number of EST present in the UCSC assignment i.e.

ClassI + ClassII + ClassII = UCSC EST Count.

HUGO ID	TGI EST Count	UCSC EST Count	Unigene EST Count	Class I (3db)	Class II (2db)	Class III (1db)
ARVCF	80	163	164	75	69	19
ATP6E-ATP6V1E1	None	757	875	0	722	35
BCR	190	379	415	171	168	40
CLTCL1	55	106	130	51	48	7
DGCR2	248	594	793	182	394	18
ECGF1	35	328	None	0	33	295
EWSR	20	1141	1266	16	1050	75
G22P1	1602	2294	2715	1203	986	105
GCAT	21	125	155	20	100	5
GGT1	94	263	552	66	62	133
GSTT1	44	230	270	41	174	15
GTPBP1	56	195	300	47	133	15
HMG2L1	67	232	262	64	150	18
LGALS1	0	1048	1130	0	999	49
MFNG	115	229	257	98	122	9
MIL1-BCL2L13	None	357	423	0	0	357
NF2	105	263	308	86	160	17
NPAP60L-NUP50	78	138	308	40	74	24
PIK4CA	184	583	470	81	308	194
PMM1	76	210	237	72	131	7
RBX1	161	315	338	151	148	16
SEC14L2	30	190	198	29	126	35
SLC25A17	9	203	246	0	174	29
ST13	863	1165	88	78	682	405
TCF20	25	172	151	8	124	40
UBE2L3	3	1146	1149	2	1014	130
UFD1L	205	406	427	198	184	24

Table 11: Summary of ESTs assigned to each gene by each method (TIGR, UCSC, Unigene), as well as the number of ESTs common to the three methods. Class I ESTs are common to all 3 databases (3 db), Class II ESTs are only common to 2 out of the 3 databases (2db), and Class III ESTs are the remainder of the UCSC ESTs.

6.4 Contig-to-mRNA ratio

The number of mRNAs assigned by UCSC to belong to a specific gene has been downloaded and the numbers recorded. These numbers are reflected in Column 2 in **Table 12** (**p53**). The number of contigs generated by *phrap* and CAP3 are recorded in columns 3 and 5 of **Table 12**. As a crude measure of the success of gene transcript reconstruction from ESTs by CAP3 and *phrap*, the ratio of contigs generated *vs*. actual number of mRNAs recorded was calculated (columns 4 and 6 of **Table 12**). On average, CAP3 produces more contigs (1.33) than does *phrap* (1.17).

The number of mRNAs assigned to a gene does not necessarily reflect the alternative transcript count for that gene unless care has been taken to ensure that these mRNAs are non-redundant. No tests were done in this research to remove redundant mRNAs from the UCSC data, and therefore the contig-to-mRNA ratio remains a crude metric.

HUGO Name	Actual mRNAs	CAP3 Contigs	Cap3/mRNA ratio	*phrap* Contigs	*phrap*/mRNA ratio
ARVCF	11	6	0.55	4	0.36
ATP6E	4	3	0.75	4	1.00
BCR	36	14	0.39	15	0.42
CLTCL1	13	9	0.69	10	0.77
DGCR2	24	19	0.79	12	0.50
ECGF1	5	10	2.00	8	1.60
EWSR	26	16	0.62	14	0.54
G22P1	13	9	0.69	23	1.77
GCAT	3	2	0.67	5	1.67
GGT1	18	13	0.72	17	0.94
GSTT1	5	8	1.60	9	1.80
GTPBP1	5	10	2.00	6	1.20
HMG2L1	11	21	1.91	9	0.82
LGALS1	10	5	0.50	5	0.50
MFNG	9	8	0.89	7	0.78
MIL1	15	25	1.67	12	0.80
NF2	23	12	0.52	13	0.57
NPAP60L	8	6	0.75	7	0.88
PIK4CA	12	26	2.17	21	1.75
PMM1	7	6	0.86	6	0.86
RBX1	8	5	0.63	4	0.50
SEC14L2	9	14	1.56	11	1.22
SLC25A17	11	8	0.73	12	1.09
ST13	8	7	0.88	10	1.25
TCF20	3	26	8.67	14	4.67
UBE2L3	6	10	1.67	13	2.17
UFD1L	13	13	1.00	14	1.08
Averages			1.33		1.17

Table 12: Transcript isoform data: Relationship between the contigs generated by each assembler and the actual number of mRNAs (transcript isoforms). Actual mRNAs: Actual number of mRNAs are defined to be transcripts which fall well within the region defined by the RefSeq gene. The data therefore may not reflect unique transcripts, and contains a level of redundancy. CAP3 Contigs, *phrap* Contigs: The number of contigs generated by CAP3 and *phrap*. CAP3/mRNA, Phrap/mRNA: The ratio of CAP3/contigs vs. actual mRNAs.

6.5 Contig-to-Singlet and Cluster-to-Singleton (C/S) Ratio

Gene	CAP3		*phrap*		WCD		*d2_cluster*	
	Contig	Singlets	Contigs	Singlets	Clusters	Singletons	Clusters	Singletons
ARVCF	6	23	4	7	13	11	1	8
ATP6E	3	46	4	10	2	15	3	13
BCR	14	66	15	35	12	41	13	38
CLTCL1	9	24	10	14	24	15	9	14
DGCR2	19	75	12	17	26	18	8	19
ECGF1	10	40	8	1	1	5	1	5
EWSR1	16	114	14	4	5	9	5	8
G22P1	9	199	23	6	2	17	2	11
GCAT	2	9	5	2	1	4	1	3
GGT1	13	53	17	12	5	28	7	19
GSTT1	8	20	9	4	5	7	4	3
GTPBP1	10	34	6	3	6	9	6	9
HMG2L1	21	50	9	5	6	11	6	12
LGALS1	5	182	5	2	2	23	3	7
MFNG	8	37	7	4	3	4	3	4
MIL1	25	66	12	4	5	9	5	7
NF2	12	27	13	11	10	11	10	10
NPAP60L	6	16	7	5	2	6	2	5
PIK4CA	26	84	21	29	12	45	13	41
PMM1	6	10	6	2	1	5	1	2
RBX1	5	49	4	4	2	9	1	12
SEC14L2	14	20	11	2	14	4	10	3
SLC25A17	8	32	12	9	3	8	4	8
ST13	7	82	10	10	2	13	4	12
TCF20	26	44	14	7	14	11	16	11
UBE2L3	10	85	13	16	7	23	7	17
UFD1L	13	50	14	9	21	12	7	14
Total	311	1537	285	234	206	373	152	315

Table 13: Summary of assembler (CAP3, *phrap*) and clustering (WCD, *d2_cluster*) contig and singlet/ singleton results for individual genes. Genes are arranged in order of increasing number of ESTs.

Program	Contig/Cluster members	Contigs/ Clusters	Singlets/ Singletons	C/S Ratio
CAP3	12343	273	888	0.31
phrap	12771	335	639	0.52
WCD	12588	160	837	0.19
d2_cluster	12703	170	562	0.30

Table 14: Results of the composite dataset comprised of the reference set of 27 gene-specific ESTs and the 3 paralog ESTs

6.6 SwissProt information for the 27 selected genes

Genes were selected, as far as possible, if protein entries existed for them in the SwissProt protein database. For some of them, actual PDB structures were found as well. The only exception to this rule is TCF20, since the "Known Gene" track only supplies one representative mRNA (AB006630), which has a hypothetical trEMBL entry (Q9UGU0). These genes were selected in such a way that the gene/ mRNA which cover the most number of exons was used as the representative sequence for the gene. This approach gives us the total number of exons for the gene. This is an assumption that is valid only if account is kept of the transcripts which have not been included, since they only have hypothetical trEMBL proteins, or their sequences have been assigned "Provisional" status by NCBI annotators.

HUGO ID	Representative mRNA	Exons	Protein
arvcf	U51269	20	O00192
bcr1	X02596	23	P11274
cltcl1	X95486	33	P53675
dgcr2	D79985	10	P98153
ecgf1	BC052211	10	P19971
ewsr1	X66899	17	Q01844
g22p1	BC008343	12	P12956
gcat	BC014457	9	O75600
gstt1	BC007065	5	P30711
hmg2l1	AL079310	12	Q9UGU5
lgals1	BC020675	4	P09382
mfng	U94352	8	O00587
mil1	AF146568	4	Q9BXK5
nf2	AF369658	17	P35240
npap60l	AF116624	7	Q9UKX7
pik4ca	AF012872	54	P42356
pmm1	BC016818	8	Q92871
rbx1	BC017370	5	P62877
rpl3	BC012786	10	P39023
sec14l2	AL096881	12	O76054
slc25a17	BC005957	9	O43808
st13	BC052982	12	P50502
tcf20	AB006630	5	trEMBL: Q9UGU0
ube2l3	AJ000519	4	P68037
ufd1l	BC005087	12	Q92890

Table 15: SwissProt Proteins found for each of the reference dataset genes.

6.7 Default Program Parameter Settings

For each of the programs used, the default parameters were accepted. **Table 16** (**p56**) summarizes only the some of the parameters that have impacted this study.

Program	Variable Parameters
phrap	forcelevel=0, penalty=-2, gap_init=-4, gap_ext=-3, ins_gap_ext=-3, del_gap_ext=-3, maxgap=30
CAP3	-o N specify overlap length cutoff (40) -p N specify overlap percent identity cutoff (80) -r N specify reverse orientation value (1)
d2_cluster	window_size (100), word_size (6), sequence length cut-off (50), similarity cut-off (0.96), reverse_comparison (1)
WCD	window length (-l, 100), word size (-w, 6), sequence length cut-off (-T, 40), common word (-H, 5)

Table 16: The default parameters, which affect the performance of the various algorithms, have been applied for all the programs used.

6.8 Scripts used for data analysis

6.8.1 Python script for calculating the Rand Index (RI)

The script has been provided by Scott Hazelhurst of WITS University

```
import sys

from   string import split
import re

def compReader(inp,clustering):
    """ reads in cluster table from inp and produces a
        dictionary in clustering """
    cnum=0
    max = 0
    data = inp.readline()
    data = data.strip("\n.")
    while len(data) != 0:
        nums = split(data)
        rep  = nums[0]
        for n in nums:
            clustering[n]= rep
        data = inp.readline()
        data = data.strip("\n.")

def randIndex(clustering1, clustering2):
    # computes the rand index between clustering1 and clustering2
    # these are
    n=a=d=0
    for i in clustering1.keys():
        for j in clustering2.keys():
            if i != j:
                n=n+1
                if clustering1[i] == clustering1[j] and
clustering2[i]==clustering2[j]: a=a+1
                if clustering1[i] != clustering1[j] and
clustering2[i]!=clustering2[j]: d=d+1
    return float(a+d)/n

f1 = file(sys.argv[1])
f2 = file(sys.argv[2])

c1 = {}
c2 = {}
compReader(f1,c1)
compReader(f2,c2)

print randIndex(c1,c2)
```

6.8.2 Perl script to find duplicate accession numbers

This script finds duplicate EST accession numbers in a gene-specific EST fasta file. It uses system calls to *nix commands ***sort*** and ***diff*** to produce a file containing the duplicate accession number(s), if found.

Scriptname: *duplicate_finder.pl*

```
#!/usr/bin/perl -w
# Script uses some system calls to generate a file containing the
# duplicate EST's within a specific file. It does so as follows:
# 1. Extract the accession numbers from the FASTA headers and write
#     to a file
# 2. Create file from "1" above with the numbers ordered with "sort"
# 3. Create file from "1" above with the numbers uniquely ordered
#     with "sort -u"
# 4. Use "diff" to locate the differences between files created in
#     "2" and "3" above and write it to file. The differences would
#     be the duplicated accession numbers

foreach $file(@ARGV){
# Step 1: Extract Accession Numbers and write to file
      $gene = (split(/\./, $file))[0] ;
      $duplicate_file = $gene.".differences.txt" ;
      open(IN, $file) ;
      $unsorted = $gene.".unsorted.txt" ;
      open(OUT, ">>$unsorted") ;
      $sorted = $gene.".sorted.txt" ;
      $sorted_unique = $gene.".sorted_unique.txt" ;
      while(<IN>){
            if(/>.+\|.+\|.+\|(.+)\|/){# Accession number now in $1
                  $acc = $1 ;
                  $acc =~ s/\..+$// ;# Remove terminal version number
                  print OUT "$acc\n" ;
            }
      }
      close(IN) ; close(OUT) ;

# Step 2: Create sorted file from file created in Step 1 above
      system("sort $unsorted > $sorted") ;

# Step 3: Create uniquely sorted file from file created in Step 1
# above
      system("sort -u $unsorted > $sorted_unique") ;

# Step 4: Locate the differences/ duplicates between files created in
# Steps 2 and 3
      system("diff $sorted $sorted_unique > $duplicate_file") ;

# Create ouput file containing the duplicate accession numbers
      open(IN, $duplicate_file) ;
```

```
        open(OUT, ">$gene.duplicates.txt") ;
        while(<IN>){
                if(/^<\s(.+)\n/){
                        print OUT $1, "\n" ;
                }
        }
        close(IN) ; close(OUT) ;

# Cleaning up some files
        system("rm $duplicate_file $unsorted $sorted $sorted_unique") ;
}
```

6.8.3 Perl script to remove duplicate sequences from a FASTA file

Once duplicate accession numbers are found, this script uses the output file from **section 6.8.2** on **p58** above, to remove those sequence(s) from a FASTA file.

Scriptname: *duplicate_sequence_remover.pl*

```
#!/usr/bin/perl -w
# Given an input file with the duplicate accession numbers
# 1.  duplicate acc nums are read into an array
# 2.  a hash is created with duplicate acc nums as keys and the
#     values for each hash member is initialized to 0
# 3.  the accession number found in a fasta file is compared against
#     the duplicate acc num array
#     - if it is absent from the duplicate array, it is written to
#     the outfile
#     - if it is present in the duplicate array, it is written to
#     the outfile, and its value changed to "1" to indicate that the
#     entry has been written already, and to prevent it from being
#     added to the output file again
foreach my $file (@ARGV){
      %removal_hash = () ;
      my $out = $file."_minus_duplicates" ;
      my $gene = (split(/\./, $file))[0] ;
      my $duplicates = $gene.".duplicates.txt" ;
      my @remove = to_be_removed($duplicates) ;

      # Create a hash with acc nums as keys and value=0
      foreach my $acc_num(@remove){
            $removal_hash{$acc_num} = 0 ; # Means "not found"
      }
      $/ = "\n>" ;
      open(SEQUENCE, $file) ;
      open(CLEANED, ">>$out") ;
      while(<SEQUENCE>){
            $to_write = 1 ;
            if(/gb\|(.+)\.\d+\|/){
                  $to_write = searcher($1, $to_write) ;
            }
            if($to_write == 1){
                  print CLEANED ;
            }
      }
      close(SEQUENCE) ;
      close(CLEANED) ;
}

sub to_be_removed{
# Write the duplicate accession numbers into an array and returns the
array
      $/ = "\n" ;
      my ($duplicates) = $_[0] ;
      my @remove_acc_nums = () ;
```

```
	open(DUPLICATES, $duplicates) ;
	while(<DUPLICATES>){
		chomp ;
		push(@remove_acc_nums, $_) ;
	}
	close(DUPLICATES) ;
	return(@remove_acc_nums) ;
}

sub searcher{
# Determine whether a fasta entry has been added to the output file
# or not.
#	- If the accession number is not in the duplicate array,
#	"$to_write" remains unchanged
#	- If it is and the hash value is "0" it means it has not been
#	written yet
#	- If it is and the hash value is "1" it means it has already
#	been written and will not be written again
	my ($bait, $found) = @_ ;
	foreach my $match(keys %removal_hash){
		if($bait eq $match){
			if($removal_hash{$match} == 1){
				$found = 0 ;
			}elsif($removal_hash{$match} == 0){
				$removal_hash{$match} = 1 ;
				$found = 1 ;
			}
		}
	}
	return($found) ;
}
```

6.8.4 Perl script that calculates Sensitivity and Specificity values

Scriptname: *sn_sp.pl*

```
#!/usr/bin/perl -w
# Takes an algorithm file with cluster/assembly members on a single
# line, as well as singlets on individual lines and
# determines the gene each EST comes from.
# It needs the reference files which are files with gene-specific
# names, containing the EST accesion numbers for each gene

$results_file = shift(@ARGV) ;

foreach $file(@ARGV){
	my $gene = (split(/\./, $file))[0] ;

	# Get a list of all the genes from the filenames
	push(@genelist, $gene) ;
	gatherer($file, $gene) ;
}

$TP = $FP = $TN = $FN = 0 ;
$grand_total = 0 ;

open(IN, $results_file) ;
while(my $line = <IN>){
	$singleton = 0 ;
	# Initialize the counter for each gene in the genelist
	foreach(@genelist){
		$gene_counter{$_} = 0 ;
	}
	chomp($line) ;
	if($line =~ /\s/){
		@ests = split(/\s/, $line) ;
		@ests = sort(@ests) ;
	}else{
		@ests = $line ;	# Singletons: Single AccNum per line
		$singleton = 1 ;
	}
	$grand_total += @ests ;
	$total = @ests ;

	foreach my $est (@ests){
		foreach my $gene(keys %all_genes){
			if($all_genes{$gene} =~ /$est/){
				$gene_counter{$gene}++ ;
				if($singleton == 1){
					if($gene =~ /paralog/){ $TN++ ; }
					else{ $FN++ ; }
				}
			}
		}
	}
```

```
	# Obtain total number of ESTs in a grouping

	foreach my $gene(keys %gene_counter){
		if($total != 0){  # Ensures that a gene is represented:
				  "0" means no EST for that gene was
				  # found
		$gene_fraction{$gene} = sprintf("%0.2f",
$gene_counter{$gene}/$total) ;
		if($gene_fraction{$gene} >= 0.5){
			if($gene !~ /paralog/){
				$TP += $gene_counter{$gene} ;
			}elsif($gene =~ /paralog/){
				$TN += $gene_counter{$gene} ;
			}
		}elsif($gene_fraction{$gene} < 0.5){
			$FP += $gene_counter{$gene} ;
		}
		$members{$gene} = $gene_counter{$gene} ;
		}
	}

	foreach my $gene(sort{$gene_fraction{$b} cmp
$gene_fraction{$a}} keys %gene_fraction){
		if($gene_fraction{$gene} > 0){
		print "$gene $members{$gene} " ;
		}
	}
	print "\n"
}

print "TP: $TP\tFP: $FP\tTN: $TN\tFN: $FN --> Grand Total:
$grand_total\n" ;

close(IN) ;

sub gatherer{
	my ($filename, $gene) = @_ ;
	my @acc_nums = () ;
	open(IN, $filename) ;
	while(my $line = <IN>){
		chomp($line) ;
		push(@acc_nums, $line)
	}
	close(IN) ;
	@acc_nums = sort(@acc_nums) ;
	$all_genes{$gene} = join(",", @acc_nums) ;
}
```

6.8.5 Perl script that uses *msbar* to mutate sequences

This script uses *msbar* (Mutate Sequence Beyond All Recognition) to introduce random point mutations into the original EST sequence. For the sake of comparison, a range of error percentage values were selected (1%, 3%, 5%, 7%, 9% and 11%). The single UCSC FASTA file containing all the ESTs specific to a gene was fragmented such that the resultant files each contained a single EST fasta-formatted sequence. These single-sequence files were then used as input for *msbar* and the range of error percentages was introduced.

Scriptname: *msbar_mutator.pl*

```
#!/usr/bin/perl –w

# Script will add random error into the original EST dataset. In
# order to do this. msbar will be used. Msbar only operates on
# individual sequences, so the FASTA file containing all the ESTs for
# a specific gene has to be fragmented s.t. the sequences are all
# separated into individual files.
# Thereafter, msbar will mutate these individual sequence files by
# introducing 1, 3, 5, 7, 9 and 11% error.
# Program outline:
# 1. Take as input each gene-specific file and create individual
#     FASTA files consisting of a single FASTA sequence i.e. Gene1
#     contains 30 UCSC assigned ESTs. After this step, there will be
#     30 individual files for Gene1, each file containing a single
#     sequence from the original file
# 2. Use files created in 1 above as input for msbar and create one
#     file per sequence per percentage error i.e. after this step,
#     one of the 30 files produced in step 1 above would have
#     produced a file with 1% error introduced into the original
#     sequence, a file with 3% error, a file with 5% error, etc.

$/ = "\n>" ;
foreach $file(@ARGV){
# Step 1: Fragment original EST FASTA file into individual EST
sequences
      $dir = $file.".temp" ; $filenums = 1 ;
      system("mkdir $dir") ; system("chmod 777 $dir/") ;
      system("cp $file $dir/") ;

      open(IN, $file) ;
      while(<IN>){
            $out = $file.$filenums ;
            s/>\n$/\n/ ;
            open(OUT, ">$dir/$out") ;
            print OUT ">$_" ;
```

```
            close(OUT) ;
            $filenums++ ;
      }
      close(IN) ;

      $/ = "\n" ;
      for( $x = 1 ; $x <$filenums ; $x++){
            $single_seq = $file.$x ; $length = 0 ;
            open(IN, "$dir/$single_seq") ;
            while($line = <IN>){
                  chomp($line) ;
                  if($line !~ />/){
                        $length += length($line) ;
                  }
            }
            close(IN) ;

# Step 2: Introduce error into the individual EST sequences
            @percent = qw(1 3 5 7 9 11) ;
            foreach $perc (@percent){
                  $number = int(($perc/100)*$length) ;
                  $mute_file = $single_seq."_".$perc ;
                  system("msbar -sequence $single_seq -count $number
-point 1 -block 0 -codon 0 -outseq $mute_file") ;
            }
      }
}
```