BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 1. 6–21
Advance Access publication May 23, 2006

doi:10.1093/bib/bbl015

A hitchhiker's guide to expressed sequence tag (EST) analysis

Shivashankar H. Nagaraj, Robin B. Gasser and Shoba Ranganathan

Received 24th April 2006

Abstract

Expressed sequence tag (EST) sequencing projects are underway for numerous organisms, generating millions of short, single-pass nucleotide sequence reads, accumulating in EST databases. Extensive computational strategies have been developed to organize and analyse both small- and large-scale EST data for gene discovery, transcript and single nucleotide polymorphism analysis as well as functional annotation of putative gene products.

We provide an overview of the significance of ESTs in the genomic era, their properties and the applications of ESTs. Methods adopted for each step of EST analysis by various research groups have been compared. Challenges that lie ahead in organizing and analysing the ever increasing EST data have also been identified.

The most appropriate software tools for EST pre-processing, clustering and assembly, database matching and functional annotation have been compiled (available online from http://biolinfo.org/EST). We propose a road map for EST analysis to accelerate the effective analyses of EST data sets. An investigation of EST analysis platforms reveals that they all terminate prior to downstream functional annotation including gene ontologies, motif/pattern analysis and pathway mapping.

Keywords: *expressed sequence tags; sequence assembly and clustering; database similarity searches; functional annotation; conceptual translation; transcriptome analysis*

INTRODUCTION

To understand the behaviour of complex biological organization and processes in terms of their molecular constituents [1], we must not only identify, catalogue and assign the function of all of its genes and gene products, but also understand regulatory interconnections between DNA, RNA and proteins. Following on from significant advancement in high-throughput technologies (microarrays, automated sequencing and mass spectrometry), transcriptomics, the global study of transcription, together with genomics and proteomics, have undoubtedly contributed to a systems biology approach. These technologies have generated a deluge of data. Fortunately, efficient computational tools (intelligent data networks, query, retrieval, analysis and visualization tools) have now optimized data mining, accelerating the process of discovery.

Expressed sequence tag (EST) and complementary DNA (cDNA) sequences provide direct evidence for all the sampled transcripts and they are currently the most important resources for transcriptome exploration. ESTs are short (200–800 nucleotide bases in length), unedited, randomly selected single-pass sequence reads derived from cDNA libraries. High-throughput ESTs can be generated at a reasonably low cost from either the 5′ or 3′ end of a cDNA clone to get an insight into transcriptionally active regions in any organism. In 1991, ESTs were used as a primary resource for human gene discovery [2]. Thereafter, there has been an exponential growth in the generation and

Corresponding author. Prof. Shoba Ranganathan, Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, NSW 2109, Australia, Tel: +61-2-9850 6262; Fax: +61-2-9850 8313; E-mail: shoba@els.mq.edu.au

Shivashankar H. Nagaraj is an international Macquarie University scholarship graduate student at the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney. He is using comparative genomics to identify novel therapeutic gene targets.

Robin B. Gasser is a Professor and Reader in Veterinary Parasitology, Department of Veterinary Science, The University of Melbourne, working on generating gender-specific EST libraries of parasitic nematodes affecting livestock animals.

Shoba Ranganathan is a Chair Professor of Bioinformatics at Macquarie University and Adjunct Professor, National University of Singapore. Her research work focuses on computational structural biology and comparative genome sequence analysis.

accumulation of EST data in public databases for myriad organisms. At present, ESTs enable gene discovery, complement genome annotation, aid gene structure identification, establish the viability of alternative transcripts, guide single nucleotide polymorphism (SNP) characterization and facilitate proteome analysis [3–5].

Whole genome sequencing is currently impractical and expensive for organisms with large genome sizes. Such an approach is unlikely to be applied extensively, irrespective of the significance of such genome data in human and animal health, agriculture, ecology and evolution. In addition, genome expansion, as a result of retrotransposon repeats, makes whole genome sequencing less attractive for plants such as maize [6]. In this scenario, EST data sets have been utilized to complement genome sequencing or as an alternative to the genome sequencing of many organisms, earning the label, the 'poor man's genome' [5]. It must be noted that ESTs are subject to sampling bias resulting in under-representation of rare transcripts, often accounting for only 60% of an organism's genes [7]. However, ESTs in combination with reduced representation sequencing strategies, such as methylation filtration and high C_ot selection, have enabled the successful examination of the gene pool in plants like maize [8].

There are several steps in EST analysis and an overwhelming number of tools available for each step. These methods have different strengths and attempt to extract biological information systematically from ESTs, in spite of their error-prone nature. However, there exists confusion in choosing the right tool for each different step of EST analysis and the subsequent downstream annotation at DNA or protein level. The confusion is compounded by the ability of some tools to handle high-throughput EST data, while others cannot.

This review briefly describes how ESTs are generated, to get an idea of the possible sources of error in the sequences obtained; where they are deposited (EST data resources) and the bottlenecks in EST analysis. We also list their proven utility in different application areas, the general methods and protocols being followed by different groups for EST analysis and our own shortcut through the EST analysis maze. Wherever possible the most useful and extensively used proven resources are identified. Individual tools and pipelines for EST analysis are compared and a detailed list of available web resources pertaining to EST analysis is maintained at http://biolinfo.org/EST/ [9].

EST GENERATION

Messenger RNA (mRNA) sequences in the cell represent copies from expressed genes. As RNA cannot be cloned directly, they are reverse transcribed to double-stranded cDNA using a specialized enzyme, the reverse transcriptase. The resultant cDNA is cloned to make libraries representing a set of transcribed genes of the original cell, tissue or organism. Subsequently, these cDNA clones are sequenced randomly from both the directions in a single-pass run with no validation or full-length sequencing to obtain 5′ and 3′ ESTs. These ESTs usually range in size from 100–800 bp. The resultant set of ESTs is redundant, as the cDNA template used can be of partial or full length. Bonaldo *et al.* [7] provide a detailed description of cDNA library construction and normalization applied to remove redundancies.

Simpson and co-workers [10] have developed a novel cost-effective method for generating high-throughput ESTs called ORESTES (open reading frame expressed sequence tags). This method differs from conventional EST generation by providing sequence data from the *central* protein coding region, and thus the most informative and desired portion, of transcripts. ORESTES representing highly, moderately and rarely expressed transcripts have been derived from several species with more than a million human sequences and thousands from other species such as cow and honey bee deposited in the Expressed Sequence Tags database, dbEST [11–13].

ERRORS ASSOCIATED WITH EST GENERATION

A typical EST sequence (Figure 1A) is only a very short copy of the mRNA itself and is highly error prone, especially at the ends. The overall sequence quality is usually significantly better in the middle (Figure 1B). Vector and repeat sequences either in the end or rarely in the middle are excised during EST pre-processing.

As ESTs are sequenced only once, they are susceptible to errors. Generally, the quality of base reads in individual EST sequences is initially poor (upto 20% or ~50–100 bp), gradually improves and then diminishes once again towards the end [14].

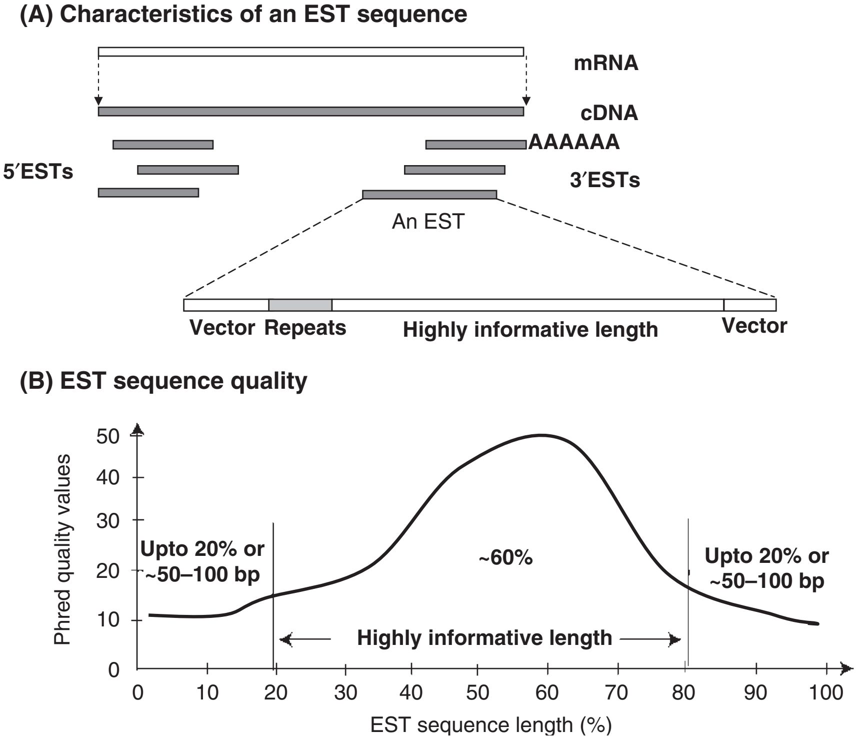

Figure 1: Characteristics of EST sequences. **A**. An EST sequence usually starts and ends with vector-contaminated bases, interspersed with possible repeats or low-complexity regions. **B**. Phred quality scores are plotted as a function of sequence length for a hypothetical EST sequence shown in A.

The overall sequence quality is usually significantly better in the middle ('highly informative length,' Figure 1B). Phred scores [15] provide a measure of sequence quality with higher values corresponding to better sequence quality. Phred examines the peaks around each base call to assign a quality score to each base call that are logarithmically linked to error probabilities. Quality scores range from 4 to 40 and are estimated as

$$q = -10 \times \log_{10}(P)$$

where q is the Phred quality score and P is the estimated error probability of that call. A Phred score of 20 represents 1/100 chance of being incorrect or 99% accurate base calling while $q = 30$ denotes 1/1000 chance of being incorrect or 99.9% base calling accuracy. Phred scores can therefore be used to extract either entire sequences or segments of specified quality, based on the biological question being addressed in subsequent analysis.

Redundancy, under-representation and over-representation of selected host transcripts are inherent problems with EST data due to the variable protocols used in their generation. Sequencing artefacts, such as base-calling errors as high as 5% [14], base stuttering (repeated bases, specifically G and T) and low quality sequences are some of the frequently observed errors in ESTs. There can be possible contaminations from vector, linker, adaptor and chimeric sequences, as also from genomic DNA fragments. In addition, low quality sequence attributes, repeats (simple or tandem), short sequence length and annotation errors can pose problems during downstream analysis. Moreover, natural sequence variations, such as RNA editing and genomic variations, due to SNPs will bring about additional challenges as it is not trivial to distinguish between sequencing artefacts and naturally-occurring substitutions, and insertion/deletion of events in a given EST data set.

ESTs and untranslated regions (UTRs)

The 5′ and 3′ UTRs of eukaryotic mRNA have been experimentally shown to contain sequence elements essential for gene regulation, expression and translation [16]. In this context, EST data has proven to be important for mining UTRs as both 5′ and 3′ ESTs contain significant sections of the UTRs along with protein coding regions. The CORG

(COmparative Regulatory Genomics) resource [17] supports promoter analysis using assembled ESTs, while more than half of the Eukaryotic Promoter Database [18] entries are based on 5′ EST sequences. Mach [19] has developed the PRESTA (PRomoter EST Association) algorithm for promoter verification and identification of the first exon, by mapping EST 5′ ends.

Polyadenylation or poly(A) tails found in 3′ UTR of the majority of mRNA transcripts are implicated in mRNA metabolism [16]. Gene boundaries have been predicted using poly(A) sites from 3′ EST clusters [20]. Differential poly(A) produces mRNAs with specific properties, attributable to post-transcriptional regulation mechanisms. Computational analyses of alternative poly(A) [21–23] have advanced our understanding of mRNA regulation. Gautheret *et al*. [22] identified previously unreported poly(A) sites in human mRNAs. Yan and Marr [23] used 3′ ESTs with poly(A) tails and demonstrated that at least 49% of human) 31% of mouse and 28% of rat polyadenylated transcription units show alternative poly(A) sites resulting in new protein isoforms. The presence of poly(A) tails in ESTs can also be used to distinguish untranslated mRNAs from productive transcripts, leading to protein isoforms.

EST DATA RESOURCES

The largest, freely-available repository of EST data (32 889 225 ESTs from 559 different organisms; as on Feb 2006) is dbEST [24, 25]. UniGene [25] from the National Center for Biotechnology Information, USA (NCBI) stores unique genes and represents a non-redundant set of gene-oriented clusters generated from ESTs. Other specialized EST resources created for specific organisms include the The Institute for Genome Research, USA (TIGR) Gene Indices [26], the Rat EST project (University of Iowa) and the Cancer Genome Anatomy Project. Table 1 alphabetically lists key EST resources. These resources and EST analysis programs, discussed in subsequent sections, have been categorized as F (free for academic users), D (data available for download), C (commercial package) and W (web interface available).

OVERVIEW OF EST SEQUENCE ANALYSIS

An individual raw EST has negligible biological information. Analysis using different combinations of computational tools augments this weak signal and when a multitude of ESTs are analysed, the results enable the reconstruction of transcriptome of that organism. While diverse research groups have used different combinations of tools for extraction of data from specific databases followed by analyses [32–37], a generic protocol of the different steps in the analysis of EST data sets is shown in Figure 2.

Chromatograms or EST sequences extracted from databases are pre-processed (Step 1, Figure 2) into high-quality ESTs wherein they are screened for sequence repeats, contaminants and low-complexity sequences, which are eliminated. Subsequently, high-quality ESTs are grouped into ‘clusters’ (Step 2, Figure 2) based on sequence similarity. The maximum informative consensus sequence is generated by ‘assembling’ these clusters, each of which could represent a putative gene. This step serves to elongate the sequence length by culling information from several short EST sequences simultaneously. Database similarity searches are subsequently performed against relevant DNA databases (Step 3, Figure 2) and possible functionality is assigned for each query sequence if significant database matches are found. Additionally, a consensus sequence can be conceptually translated to a putative peptide (step 4, Figure 2) and then compared with protein sequence databases (step 5, Figure 2). Protein centric functional annotation, including domain and motif analysis, can be carried out using protein analysis tools. It must be noted that the entire transcriptome is not translatable into protein products.

Each of these steps is briefly described subsequently, with special emphasis on the software tools available, followed by EST analyses tools for specific applications such as open reading frame (ORF) prediction, gene finding and detection of SNPs and alternative splicing.

EST pre-processing

Pre-processing reduces the overall noise in EST data to improve the efficacy of subsequent analyses. Vector contamination is prevalent in ESTs, and often a part of the vector is also sequenced along with the EST sequences (Figure 2). These vector fragments have to be clipped before ESTs are clustered. Comparing the ESTs with non-redundant vector databases, such as UniVec and EMVEC with

Table 1: Resources for EST data

Resource#	**Web site**	**Contents of EST resource**	**Organisms**	**Category***
ApiEST-DB [27]	http://www.cbil.upenn.edu/apidots/	Raw	Apicomplexan parasites	F, D
dbEST at NCBI [24]	http://www.ncbi.nlm.nih.gov/dbEST/	Raw	All	F, D
Diatom EST database [28]	http://avesthagen.sznbowler.com.	Raw and clusters	Diatoms	F, D
ESTree [29]	http://www.itb.cnr.it/estree/	Raw and clusters	Peach	F, D
Fungal genomics project	https://fungalgenomics.concordia.ca/home/index.php	Raw	Fungal	F, D
Honey bee brain EST project [30]	http://titan.biotec.uiuc.edu/bee/honeybee.project.htm	Raw and clusters	Honey Bee	F, D
Nematode ESTs at the Sanger Institute	ftp://ftp.sanger.ac.uk/pub/pathogens/nem.ests/	Raw and clusters	Parasitic nematodes	F, D
NEMBASE- parasitic nematode ESTs	http://www.nematodes.org	Raw and clusters	Parasitic nematodes	F, D
Parasitic and free-living nematode EST resource	http://www.nematode.net/	Raw and clusters	Nematodes	F, D
Phytopathogenic Fungi and Oomycete EST database	http://cbr-rbc.nrc-cnrc.gc.ca/services/cogeme/	Plant pathogenic fungi ESTs	Fungi and oomycetes	F, D
Plant Gene Research, Kazusa DNA Research Institute	http://www.kazusa.or.jp/en/plant/database.html	Raw	Heterogeneous set	F, D
Plant Genome database [31]	http://www.plantgdb.org/	Raw and clusters	Plants	F, D
Rat EST data at University of Iowa	http://ratest.eng.uiowa.edu	Raw	Rat	F, D
Sanger Institute *Xenopus tropicalis* EST project	http://www.sanger.ac.uk/Projects/X.tropicalis/	Raw and clusters	Xenopus	F, D
The TIGR Gene Indices [26]	http://www.tigr.org/tdb/tgi/	Raw and gene indices	All	F, D
UniGene database at NCBI	www.ncbi.nlm.nih.gov/UniGene	Raw and clusters	All	F, D

#Very useful resources are shown in bold.

*F, free for academic users; D, data available for download.

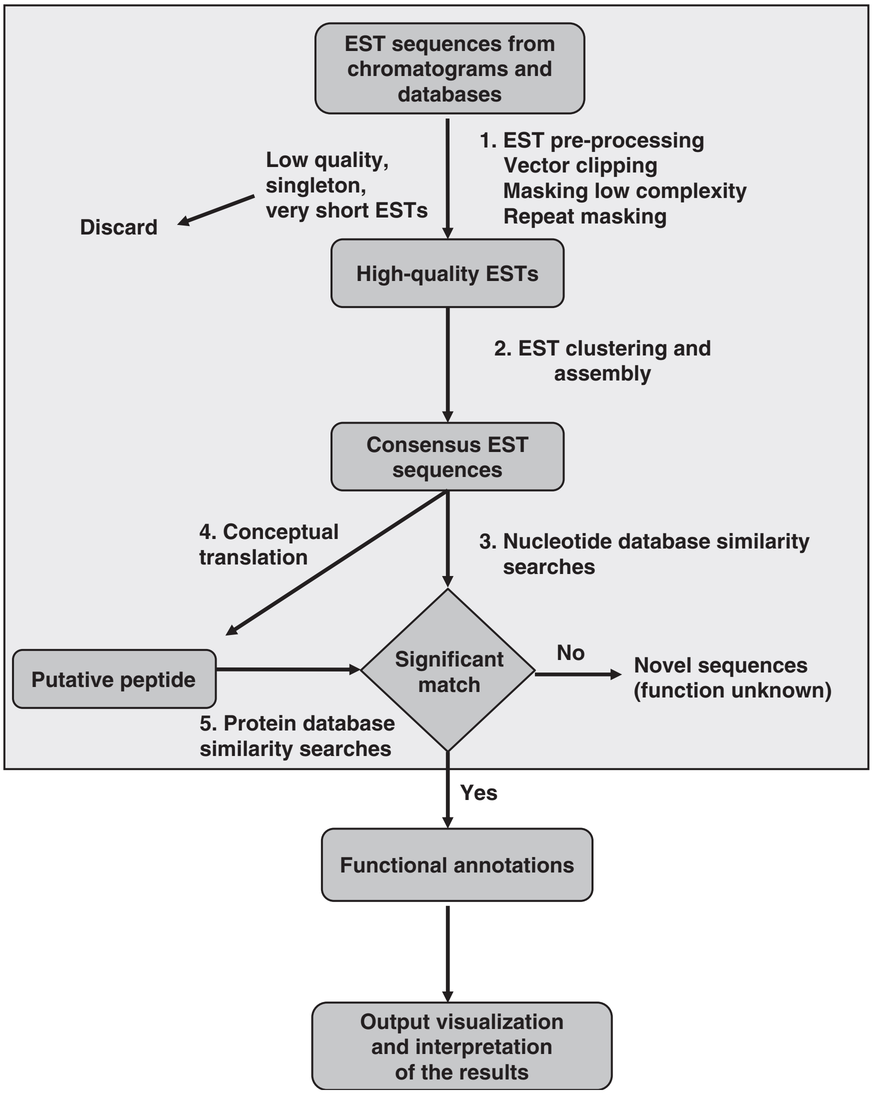
EST sequences from chromatograms and databases
1. EST pre-processing
Vector clipping
Masking low complexity
Repeat masking
Low quality, singleton, very short ESTs
Discard
High-quality ESTs
2. EST clustering and assembly
Consensus EST sequences
4. Conceptual translation
3. Nucleotide database similarity searches
Putative peptide
Significant match
No
Novel sequences (function unknown)
5. Protein database similarity searches
Yes
Functional annotations
Output visualization and interpretation of the results

Figure 2: Generic steps involved in EST analysis. 1. Raw EST sequences are checked for vector contamination, low complexity and repeat regions, which are excised or masked. Low quality, singleton and very short sequences are also removed. 2. ESTs are then clustered and assembled to generate consensus sequences ('putative transcripts'). 3. DNA database similarity searches are carried out to assign, identify homologues and sign possible function. 4. Putative peptides are obtained by conceptual translation of consensus sequences. 5. Protein database similarity searches are performed to assign putative function(s). The analysis is extended to functional annotation followed by visualization and interpretation of results. The steps enclosed by the grey box alone are implemented in the currently available pipelines.

the locally installed tools such as BLAST [38] or Cross_Match (Smith and Green, unpublished work [39]), can identify vector contamination for removal. Any low complexity regions in EST data can be detected and masked using DUST from NCBI or nseg [40]. Repetitive elements, such as LINEs (Long interspersed elements), SINEs (Short interspersed elements), LTRs (Long terminal repeat) and SSRs (Short simple repeats), can lead to erroneous assembly of sequences. Therefore, they should be 'repeat masked' during the analysis using either RepeatMasker [39] or MaskerAid [41] to screen DNA sequences for low complexity DNA sequences and interspersed repeats.

Poly(A) is not encoded in the genomic sequence and should be trimmed to retain a few adenines (usually 6–10) to get high-quality ESTs for clustering and assembly process. A list of web resources related to EST pre-processing is given in Table 2, grouped according to their functionality.

Table 2: Resources for EST pre-processing

Name#	Website	Description	Category*
UniVec	http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html	Vector database	F
EMVec	http://www.ebi.ac.uk/blastall/vectors.html	Vector database	F
Phred/Cross.Match	http://www.phrap.org/	Base caller/vector trimming and removal	F
Trimest	http://emboss.sourceforge.net/apps/#Apps	Poly(A) tail trimmed	F
Trimseq	http://emboss.sourceforge.net/apps/#Apps	Ambiguous ends trimmed	F
VecScreen	http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html	Vector, linker and adapter identification	F
Vector cleaning	http://www.aborygen.com/products/biOpen/tools-for-biOpen/vector-cleaning.php	Vector cleaning	C
Vectorstrip	http://emboss.sourceforge.net/apps/#Apps	DNA between vector sequences extracted	F
Paracel	http://www.paracel.com/	EST pre-processing package	C, W
Lucy2 [42]	http://www.complex.iastate.edu/download/Lucy2/index.html	Sequence trimming and visualization	F
Dust	ftp://ftp.ncbi.nih.gov/blast/	Low-complexity regions masked	F
The TIGR Plant Repeat databases	http://www.tigr.org/tdb/e2k1/plant.repeats/	Plant repeats database	F
MaskerAid [41]	http://blast.wustl.edu/maskeraid/	Repeats masked	F
RepeatMasker	http://www.repeatmasker.org/	Repeats masked	F

#Very useful resources are shown in bold.
*F, free for academic users; C, commercial package; W, web interface available.

EST clustering and assembly

The purpose of EST clustering is to collect overlapping ESTs from the same transcript of a single gene into a unique cluster to reduce redundancy. An EST cluster is a fragmented data, which can be consolidated and indexed using gene sequence information, such that all the expressed data arising from a single gene is grouped into a single index class, and each index class contains information for only that particular gene [43]. A simple way to cluster ESTs is by measuring the pair-wise sequence similarity between them. Then, these distances are converted into binary values, depending on whether there is a significant match or not, such that the sequence pair can be accepted or rejected from the cluster being assembled. Different steps for EST clustering are described in detail by Ptitsyn and Hide [44]. In their formalism, there are two approaches for EST clustering, 'stringent' and 'loose.' The stringent clustering method is conservative, uses single-pass grouping of ESTs resulting in relatively accurate clusters, but generates shorter-sequence consensuses with low coverage of expressed genes. In contrast, loose clustering is 'liberal' and repeats low quality EST sequence alignments many times to generate less accurate but longer-sequence consensuses. Consequently, there is a better coverage of expressed gene data and alternatively spliced transcripts, but there is a risk of including paralogues in the clusters. stackPACK [36] is designed to use loose clustering, while TIGR Gene Indices [26] adopt stringent clustering. UniGene clusters lie between the two extremes.

Phrap [45] and CAP3 [46] are among the most extensively used programs for sequence clustering and assembly. Quackenbush and co-workers [47] have compared Phrap, CAP3 and TIGR Assembler programs with a benchmarking data set of 1 18 000 rat ESTs. They assessed the effects of sequencing errors on EST assembly and critically evaluated the tools for EST clustering, assembly and relative accuracy of algorithms. Their results demonstrate that CAP3 consistently out-performed other similar programs, producing high-fidelity consensus sequences and maintaining a high level of sensitivity to gene family members while effectively handling sequencing errors. A newer study by Wang *et al.* [48] have recognized two types of errors accruing from CAP3 EST clustering, where ESTs from the same gene do not form a cluster (Type I) and ESTs from distinct genes are wrongly clustered together (Type II). They propose a novel statistical approach for more accurate estimates of the true gene cluster profile.

Table 3 provides a list of resources available for EST clustering, assembly and consensus generation.

The STACK (Sequence Tag Alignment and Consensus Knowledge Base) [36] system was created to cluster and assemble ESTs. The major difference between STACK and other resources lies in its initial tissue-specific classification (15 tissue-based categories and one disease category) of EST data, for differential expression analysis. STACK uses a loose, unsupervised clustering strategy to group pre-processed ESTs for a wider gene coverage. The d2_cluster (a non-pairwise alignment algorithm) [43], Phrap and CRAW programs are incorporated into STACK to cluster, assemble and analyse EST alignments, respectively. Consensus contigs are then merged based on clone-identification data to obtain the best putative gene representation.

Table 3: Programs for EST sequence assembly and consensus generation

Name#	Website	Category*
CAP3 [46]	http://genome.cs.mtu.edu/cap/cap3.html	F, W
CLOBB [49]	http://zeldia.cap.ed.ac.uk/CLOBB/	F
CLU [44]	http://compbio.pbrc.edu/pti	F
ESTate	http://www.ebi.ac.uk/~guy/estate/	F
ESTs aSSEmbly using Malig	http://alggen.lsi.upc.es/recerca/essem/frame-essem.html	F
megaBLAST	ftp://ftp.ncbi.nih.gov/blast/	F, W
miraEST [50]	http://www.chevreux.org/projects.mira.html	F
Paracel Transcript Assembler	http://www.paracel.com/	C, W
Phrap [45]	http://www.phrap.org/	F
stackPACK [36]	http://www.sanbi.ac.za/Dbases.html#stackpack	F
Xsact and Xtract [5l]	http://www.ii.uib.no/~ketil/bioinformatics/	F

#Very useful resources are shown in bold.

*F, free for academic users; C, commercial package; W, web interface available.

Database similarity searches

Once consensus sequences (putative genes) are obtained from assembled ESTs, possible functions can be assigned through downstream annotation, achieved via database similarity searches, employing familiar freely available tools and databases.

Different flavours of BLAST [38] programs from NCBI serve as a universal tools for database similarity searches. BLASTN can be used to search ESTs against nucleotide sequence database and BLASTX to search against protein databases. BLASTX translates a consensus EST sequence (query) into protein products in six reading frames followed by comparisons with protein databases. In addition, one can scan for protein domains by selecting the (CDD) Conserved Domain Database [52] and the COG (Cluster of Orthologous Groups) [53] database using RPS-BLAST (Reverse PSI-BLAST) [52]. High-throughput EST analysis and annotation involve the generation and interpretation of thousands of BLAST output results. In such cases, BLAST parsers such as MuSeqBox can be used [54]. Another option is to use SSAHA (Sequence Search and Alignment by Hashing Algorithm) [55], a fast and efficient DNA database searching tool. For transcriptome analysis, ESTs are additionally aligned to the genome sequence of the organism itself (if available) or the closest relative, using specialized alignment programs (Table 4) to facilitate genomic mapping and gene discovery. BLAT, GMAP and MGALIGN are considered to be reliable methods for this process [56].

Conceptual translation of ESTs

EST data can be correlated with protein-centric annotations by accurate and robust polypeptide translations, since polypeptides are better templates for identifying domains and motifs, to study protein localization and to assign gene ontologies (GOs).

The first step in translating EST sequences is in identifying the protein-coding regions or ORFs, from consensus EST sequences, to enhance the process of gene discovery and gene boundary predictions. Some tools have been explicitly created for this purpose. For example, OrfPredictor [61] has been designed specifically to identify protein-coding regions in EST-derived sequences, wherein the program provides six frames of translation and predicts most probable coding regions in all frames. ESTScan [62] and DECODER [63] can detect and extract coding regions from low-quality ESTs or partial cDNAs while correcting for frame shift errors, and provide conceptual translations. Table 5 lists key programs related to protein-coding region predictions from EST data. One can also use Prot4EST [64], a pipeline with six tier polypeptide prediction tool, to translate ESTs into polypeptides. Prot4EST effectively incorporates DECODER, ESTScan and BLASTX for more accurate predictions. The putative peptides obtained can be compared with protein databases using BLASTP from the BLAST suite of programs.

FUNCTIONAL ANNOTATIONS

Once a putative polypeptide is obtained, its function can be predicted by matching against non-redundant protein sequence, motif and family databases using an integrated tool such as Interproscan [66]. Protein sequences are better templates for functional annotation, particularly for the construction of multiple sequence alignments, profile and HMM generation, phylogenetic analysis, creation of protein-mass fingerprint libraries for proteomics applications, and domain and motif analysis using Pfam [67] and SMART [68].

Table 4: Programs for EST/cDNA sequence to genome DNA alignment

Name#	Website	Category*
BLAT [57]	http://genome.ucsc.edu/cgi-bin/hgBlat	F, W
est2genome [58]	http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html	F, W
GMAP [56]	http://www.gene.com/share/gmap/	F
MGAlign [59]	http://origin.bic.nus.edu.sg/mgalign	F, W
SSAHA [55]	http://www.sanger.ac.uk/Software/analysis/SSAHA/	F
Sim4 [60]	http://globin.cse.psu.edu/html/docs/sim4.html	F
Splign	http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi	F, W

#Very useful resources are shown in bold.
*F, free for academic users; W, web interface available.

Table 5: ORF prediction and conceptual translation for ESTs

Name#	Website	Category*
DECODER [63]	http://www.gsc.riken.go.jp/	F
DIANA-EST [65]	artemis@pcbi.upenn.edu.(upon request from the author)	F
Diogenes	http://analysis.ccgb.umn.edu/diogenes/index.html	F
ESTScan2 [62]	http://www.ch.embnet.org/software/ESTScan2.html	F, W
OrfPredictor [61]	https://fungalgenome.concordia.ca/tools/OrfPredictor.html	F, W
TargetIdentifier	https://fungalgenome.concordia.ca/tools/TargetIdentifier.html	F

#Very useful resources are shown in bold.
*F, free for academic users; W, web interface available.

EST ANALYSIS PIPELINES

In any large-scale sequencing project, in which thousands of ESTs are generated daily, it is extremely important to be able to store, organize and annotate using an automated analysis pipeline. Here, a protocol is required to transfer data efficiently between programs without human intervention, based on carefully parameterized threshold criteria. Given the ESTs from a database or from raw sequence chromatograms, such a pipeline would automatically clean, cluster, assemble, generate consensus sequences, conceptually translate and assign putative function based on various DNA and protein similarity searches. Table 6 summarizes the salient features of some of the EST analysis pipelines [32–37], which comprise Steps 1–5 of Figure 2.

The TIGR Gene Indices are a rich resource for the freely-available high-quality gene contigs derived from their in-house EST analysis pipeline, TGCIL (TGI Clustering tools) [71]. A high stringency approach to establish sequence similarities during the initial clustering process is employed, that groups closely related genes into distinct consensus sequences. This process also allows the identification and separation of splice variants and then assembles individual clusters to generate longer, virtual transcripts or tentative consensus (TC) sequences, often referenced to mRNA or genomic sequences [72]. TCs are then annotated with tools for ORF and SNP prediction, long oligonucleotide prediction for microarrays, putative annotation using a controlled vocabulary, GO and enzyme commission number assignments, and mapping to complete or draft genomes and available genetic maps. Presently, TCs have been generated for 77 species and populate resources including EGO and RESOURCERER [26].

Two other successful and widely accessed EST 'warehouses' with high-quality data are Unigene and ENSEMBL. We have briefly described the data and software used by these warehouses, since an appreciation of these systems, if it were developed, will form the basis of future EST analysis pipelines.

Unigene uses GenBank mRNA and coding sequence data as 'reference' or 'seed' sequences for supervised cluster generation. Sequence alignment programs identify contaminant (linker, vector, bacterial, mitochondrial and ribosomal) sequences. Repeat and low-complexity regions are suppressed by RepeatMasker and DUST, respectively. ESTs are compared pairwise, using megaBLAST [38] with different stringency levels, and then grouped into clusters. UniGene stores all gene isoforms in a single cluster, but does not generate consensus sequences. Singletons, not belonging to any cluster, are reprocessed at lower stringency and stored separately. With EST data accumulating daily in GenBank, UniGene clusters are updated each week for progressive data management [25].

EST-based gene model validation forms one of the core components of the ENSEMBL automatic gene annotation system [73]. The ENSEMBL EST gene build process involves two steps. In the first step, EXONERATE [74] or BLAT is employed for the pairwise alignment of ESTs and unspliced genomic alignments, respectively. Clusters resulting from these alignments are then submitted to the ClusterMerge program [74] which derives a minimal set of non-redundant transcripts compatible with the splicing structure of a set of ESTs. Mapping these clusters onto a genome leads to 'putative genes' called 'ESTgenes' [73, 74].

APPLICATIONS OF ESTS

ESTs are versatile and have multiple applications. ESTs were first used to construct maps of the human genome [75], followed by assessment of the gene coverage from EST sequencing alone [14] and mapping of gene-based site markers [25]. With the exponential rise in genomic data from global sequencing projects, EST databases are used for gene structure prediction [20, 76], to investigate alternative splicing [77–80], to discriminate between genes exhibiting tissue or disease-specific expression [81] and for the discovery and characterization of candidate SNPs [82–84].

EST-based gene expression protocols have been used in the identification and analysis of coexpressed genes on a large scale [85, 86]. Ewing *et al.* [85] have generated digital gene expression profiles of the rice genome from dbEST classified tissue types and organs. The *in silico* differential display technique allows the identification of genes which are expressed differentially in various tissues [87] and between normal and diseased individuals [81]. Recently, applying DNA 'bar coding' in EST projects, Qiu *et al.* [88] categorized maize cDNA libraries using distinct 6-bp DNA sequences to track the origin of ESTs from specific mRNA pools. ESTs have also become invaluable resources in the area of proteomics for peptide identification and proteome characterization of proteomes, particularly in the absence of complete genome sequence information [89, 90]. EST sequencing strategies together with other genomic and proteomics methodologies (transcription profiling and peptide fingerprinting) have been employed for gene and allele identification [91, 92]. Overall, the usefulness of EST data has extended well beyond its original application in gene finding and in transcriptome analysis, and a comprehensive list of specialized EST application software is provided in Table 7.

A ROADMAP FOR EST ANALYSIS

In order to make sense of the bewildering array of tools available for so many different EST analysis steps, we present a simplified solution as a roadmap, with a small set of selected options for each step. Specific resources have been selected for each step during the analysis, based on their usefulness and performance evaluation with heterogeneous EST data sets. Some of the most reliable tools, such as CAP3 or Phrap, have already been used as base algorithms to either develop more advanced methods for the analysis of ESTs or as a part of EST analysis pipelines (Table 6). The resources selected for the roadmap can be freely downloaded (by the academic community), installed and run locally for high-throughput EST analysis. Also, these programs have been demonstrated to be consistent independently, or efficient as a part of different EST analysis pipelines and we wish to recommend these as a general '*Modus operandi*' for small or large-scale EST analysis projects.

For EST pre-processing, the Univec database with Cross_Match [39] can effectively identify and eliminate vector sequences. RepeatMasker or MaskerAid [41] (for large-scale EST analysis) can mask repeat sequences. CAP3 [46] or Phrap [45] can cluster, assemble and generate consensus contigs from pre-processed EST fragments and ESTscan2 [62] will then detect coding regions in consensus contigs. In addition, GOs can be linked to consensus contigs to assign a possible function using Blast2GO [99].

For specific analysis, a brief selection is considered. miraEST [50] can be used for the detection and classification of SNPs, by reconstructing mRNA transcripts from ESTs. Prot4EST [64] outputs peptide sequences for protein-centric downstream annotations.

We have checked this roadmap on a test data set of 20 000 ESTs from a plant parasitic nematode, *Meloidogyne incognita* (S.H.N., R.B.G. and S.R., unpublished results). We pre-processed the raw ESTs using a combination of UniVec database and Cross_Match program to remove vectors, followed by masking of repeats with RepeatMasker. We then used Phrap and CAP3 to cluster, assemble and generate consensus contigs, and found high similarity in the quality and number of contigs generated by the two programs. BLASTX (to search for similar sequences in various relevant databases) and Blast2GO (to assign GOs) were then applied for functional annotation of the EST contigs. Partigene [35], ESTAP [33] and ESTAnnotator [32] are integrated sequence analysis suites which we consider to be useful for high-throughput EST analysis and annotation. ESTs also have a multitude of specialized applications, listed in Table 7. Describing all of them is beyond the scope of this review.

CONCLUSIONS

The International Nucleotide Sequence Database Collaboration recently reached a milestone with 100 000 000 000 bases (100 Gigabases) of the genetic code, representing individual genes and partial and complete genomes of more than 165 000 organisms, most of which are represented by EST data sets. A cornucopia of ESTs will continue to be generated

Table 6: Characteristics of EST pipelines

Name#	Website	Pre-processing programs	Clustering and assembly programs	Translation programs	Category*
Edinburgh EST-Pipeline [35]	http://zeldia.cap.ed.ac.uk/ PartiGene/index.html	Phred; Cross.Match	CLOBB; Phrap	DECODER; ESTSCAN	F
EST Analysis Pipeline (ESTAP) [33]	http://staff.vbi.vt.edu/estap/	Phred; Cross.Match	D2cluster; CAP3	BLASTX	F
ESTAnnotator [32]	http://genome. dkfz-heidelberg.de/menu/ biounit/dev.shtml#estannotator	Phred; RepMask with UniVec data	CAP3	BLASTX	W
ESTIMA [34]	http://titan.biotec.uiuc.edu/ ESTIMA/	Information not available	BlastClust; CAP3	BLASTX	F
ESTweb [37]	http://bioinfo.iq.usp.br/estweb/	Phred; Cross.Match	None	None	F
Nematode.net [69] NemaGene Clusters	http://nematode.net/	Phred; Consed	Phrap	BLASTX	F
PipeOnline[70]	http://bioinfo.okstate.edu/pipeonline/	Phred; Cross.Match	Phrap	BLASTX	F
The TIGR Gene Indices [26] (TGICL)	http://www.tigr.org/tdb/tgi/	SeqClean; megaBLAST	CAP3; Paracel TranscriptAssembler®	DIANA-EST; ESTscan; Framefinder	F (except paracel)

#Very useful resources are shown in bold.
*F, free for academic users; W, web interface available.

Table 7: EST applications and visualization resources

Name	Web site	Description	Application area	Category
ESTgene [74]	http://www.ebi.ac.uk/~guy/estate/	Alternative splicing detection	Alternative splicing	F
galaxieEST [93]	http://galaxie.cgb.ki.se/galaxieEST.html	Automated phylogenetic analysis	Evolutionary studies	F
GBA server [94]	http://gba.cbi.pku.edu.cn:8080/gba/	EST-based digital gene expression profiling	Gene expression profiles	F
ESTminer [91]	http://www.soybase.org/publication.data/Nelson/ ESTminer/ESTminer.html	Gene and allele identification	Gene structure prediction and alternative splicing	F
ESTminer [95]	ftp://cggc.agtec.uga.edu/estMiner/	Web application and database schema for mining of EST clusters	Candidate gene discovery	F
GeneSeqer [96]	http://bioinformatics.iastate.edu/cgi-bin/gs.cgi	Alternative splicing detection	Gene structure prediction and alternative splicing	F, W
Transcript Assembly program [20]	http://sapiens.wustl.edu/~zkan/TAP/	Predominant and alternative gene structures identified	Gene structure prediction and alternative splicing	F,W
prot4EST	http://zeldia.cap.ed.ac.uk/PartiGene/	ESTs to protein multiple sequence alignments	Protein alignment	F
miraEST [50]	http://www.chevreux.org/projects.mira.html	Human SNP discovery	SNP discovery	F, W
SNP discovery from ESTs	http://www.atgc.org/	To find SNP candidates in EST assemblies	SNP discovery	F, W
JESAM [97]	http://corba.ebi.ac.uk/EST/	EST alignments and clusters	Visualization of clusters	F, W
SpliceNest [98]	http://splicenest.molgen.mpg.de/	Visualization of gene splicing from ESTs	Visualization	F, W

*F, free for academic users; W, web interface available.

for many organisms as a low-cost alternative to genome sequencing or to complement genome sequencing projects by early characterization of the transcriptome. There are currently numerous databases to store EST data and an overwhelming number of tools for their analyses. However, a critical evaluation of different procedures and methods, including EST clustering, assembly, consensus generation and tools for DNA or protein downstream annotation on benchmark data sets, is lacking. Such an evaluation will guide researchers to choose appropriate tools, depending on the nature and extent of the EST data sets being analysed and the biological questions being addressed, such that they can adopt or develop in-house, complete or semi-automated approaches.

In the case of high-throughput EST analysis there is a need for integrated, automated approaches enabling EST data mining for the biologically useful information across disciplinary boundaries. Moreover, ESTs have diverse applications, and the question being addressed will determine the choice of methods or pipelines to be used. As the objective of individual methods and tools can vary substantially, it is difficult to evaluate all of them using a common platform and choose the most appropriate ones for individual projects. This is particularly true of assembly programs (CAP3 and Phrap), several of which have been developed for genome sequences generated by a shotgun approach rather than for EST data. New generation algorithms, such as Xsact/Xract [51] and CLU [44] have been developed specifically for EST clustering and assembly and will continue to play a central role in the analysis of large data sets, although they are still to be incorporated into existing pipelines.

With the rapid convergence of various technologies crossing the 'omics' barrier for holistic solutions to complex biological questions, the analysis of EST data sets will continue to be indispensable in many areas of biomedicine and biotechnology.

Key Points

- ESTs are short, unedited, randomly selected single-pass sequence reads derived from cDNA libraries, providing a low-cost alternative to whole genome sequencing.
- ESTs are error prone and require multi-step pre-processing followed by clustering, assembly, database matching and functional annotation to yield transcriptome information.
- ESTs can be directly used for gene discovery, gene structure identification, alternative splicing, SNP characterization and proteome analysis.
- Our road map provides a quick solution for ESTanalysis, picking out the best computational methods available for each step,

Acknowledgements

Partial funding from Genetic Technologies Limited and the Australian Research Council (LP0667795), and the award of a Macquarie University research scholarship to S.H.N. is acknowledged.

References

1. Kirschner MW. The meaning of systems biology. *Cell* 2005; **121**:503–4.
2. Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991;**252**:1651–6.
3. Jongeneel CV. Searching the expressed sequence tag (EST) databases: panning for genes. *Brief Bioinform* 2000;**1**: 76–92.
4. Dong Q, Kroiss L, Oakley FD, *et al*. Comparative EST analyses in plant systems. *Methods Enzymol* 2005;**395**:400–8.
5. Rudd S. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci* 2003;**8**: 321–9.
6. Bennetzen JL. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 2002;**115**: 29–36.
7. Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 1996;**6**:791–806.
8. Barbazuk WB, Bedell JA, Rabinowicz PD. Reduced representation sequencing: a success in maize and a promise for other plant genomes. *Bioessays* 2005;**27**:839–48.
9. http://biolinfo.org/EST/. Web resources for EST data and analysis (17 February 2006, date last accessed).
10. Dias Neto E, Correa RG, Verjovski-Almeida S, *et al*. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci USA* 2000;**97**: 3491–6.
11. da Mota AF, Sonstegard TS, Van Tassell CP, *et al*. Characterization of open reading frame-expressed sequence tags generated from *Bos indicus* and *B. taurus* mammary gland cDNA libraries. *Anim Genet* 2004;**35**:213–19.
12. Camargo AA, Samaia HP, Dias-Neto E, *et al*. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc Natl Acad Sci USA* 2001;**98**:12103–8.
13. Nunes FM, Valente V, Sousa JF, *et al*. The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BMC Genomics* 2004;**5**:84.
14. Aaronson JS, Eckman B, Blevins RA, *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* 1996;**6**:829–45.
15. http://www.phrap.com/phred/. Website for CodonCode Corporation, information on phred quality scores can be found here (27 February 2006, last date accessed).
16. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. *Genome Biol* 2002;**3**:REVIEWS0004.
17. Dieterich C, Grossmann S, Tanzer A, *et al*. Comparative promoter region analysis powered by CORG. *BMC Genomics* 2005;**6**:24.

18. Schmid CD, Perier R, Praz V, Bucher P. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* 2006;**34**:D82–5.
19. Mach V. PRESTA: associating promoter sequences with information on gene expression. *Genome Biol* 2002; **3**:research0050.
20. Kan Z, Rouchka EC, Gish WR, States DJ. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* 2001;**11**:889–900.
21. Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* 2001;**11**:1520–6.
22. Gautheret D, Poirot O, Lopez F, *et al*. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* 1998;**8**:524–30.
23. Yan J, Marr TG. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* 2005;**15**:369–75.
24. Boguski MS, Lowe TM, Tolstoshev CM. dbEST–database for ''expressed sequence tags''. *Nat Genet* 1993;**4**:332–3.
25. Boguski MS, Schuler GD. ESTablishing a human transcript map. *Nat Genet* 1995;**10**:369–71.
26. Lee Y, Tsai J, Sunkara S, *et al*. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* 2005;**33**:D71–4.
27. Li L, Brunk BP, Kissinger JC, *et al*. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res* 2003;**13**:443–54.
28. Maheswari U, Montsant A, Goll J, *et al*. The Diatom EST Database. *Nucleic Acids Res* 2005;**33**:D344–7.
29. Lazzari B, Caprera A, Vecchietti A, *et al*. ESTree db: a tool for peach functional genomics. *BMC Bioinformatics* 2005;**6 [Suppl 4]**:S16.
30. Whitfield CW, Band MR, Bonaldo MF, *et al*. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res* 2002; **12**:555–66.
31. Dong Q, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res* 2004; **32**:D354–9.
32. Hotz-Wagenblatt A, Hankeln T, Ernst P, *et al*. ESTAnnotator: A tool for high throughput EST annotation. *Nucleic Acids Res* 2003;**31**:3716–19.
33. Mao C, Cushman JC, May GD, Weller JW. ESTAP – an automated system for the analysis of EST data. *Bioinformatics* 2003;**19**:1720–2.
34. Kumar CG, LeDuc R, Gong G, *et al*. ESTIMA, a tool for EST management in a multi-project environment. *BMC Bioinformatics* 2004;**5**:176.
35. Parkinson J, Anthony A, Wasmuth J, *et al*. PartiGene–constructing partial genomes. *Bioinformatics* 2004;**20**: 1398–404.
36. Miller RT, Christoffels AG, Gopalakrishnan C, *et al*. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res* 1999;**9**:1143–55.
37. Paquola AC, Nishyiama MY Jr, Reis EM, *et al*. ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics* 2003;**19**:1587–8.
38. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**: 3389–402.
39. http://www.phrap.org/. Phred, Phrap and Consed. Laboratory of Phil Green, Department of Genome Sciences,University of Washington (17 February 2006 date last accessed).
40. Wan H, Li L, Federhen S, Wootton JC. Discovering simple regions in biological sequences associated with scoring schemes. *J Comput Biol* 2003;**10**:171–85.
41. Bedell JA, Korf I, Gish W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* 2000;**16**: 1040–1.
42. Li S, Chou HH. LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics* 2004;**20**:2865–6.
43. Burke J, Davison D, Hide W. d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome Res* 1999;**9**:1135–42.
44. Ptitsyn A, Hide W. CLU: a new algorithm for EST clustering. *BMC Bioinformatics* 2005;**6 [Suppl 2]**:S3.
45. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**:186–94.
46. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res* 1999;**9**:868–77.
47. Liang F, Holt I, Pertea G, *et al*. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res* 2000;**28**: 3657–65.
48. Wang JP, Lindsay BG, Leebens-Mack J, *et al*. EST clustering error evaluation and correction. *Bioinformatics* 2004;**20**: 2973–84.
49. Parkinson J, Guiliano DB, Blaxter M. Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* 2002;**3**:31.
50. Chevreux B, Pfisterer T, Drescher B, *et al*. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004;**14**:1147–59.
51. Malde K, Coward E, Jonassen I. A graph based algorithm for generating EST consensus sequences. *Bioinformatics* 2005;**21**:1371–5.
52. Marchler-Bauer A, Panchenko AR, Shoemaker BA, *et al*. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 2002;**30**:281–3.
53. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
54. Xing L, Brendel V. Multi-query sequence BLAST output examination with MuSeqBox. *Bioinformatics* 2001; **17**:744–5.
55. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res* 2001;**11**: 1725–9.
56. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**:1859–75.
57. Kent WJ. BLAT the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.

58. Churbanov A, Pauley M, Quest D, Ali H. A method of precise mRNA/DNA homology-based gene structure prediction. *BMC Bioinformatics* 2005;**6**:261.
59. Lee BT, Tan TW, Ranganathan S. MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res* 2003;**31**:3533–6.
60. Florea L, Hartzell G, Zhang Z, *et al*. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 1998;**8**:967–74.
61. Min XJ, Butler G, Storms R, Tsang A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* 2005;**33**:W677–80.
62. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol* 1999;138–48.
63. Fukunishi Y, Hayashizaki Y. Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol Genomics* 2001;**5**:81–7.
64. Wasmuth JD, Blaxter ML. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* 2004;**5**:187.
65. Hatzigeorgiou AG, Fiziev P, Reczko M. DIANA – EST: a statistical analysis. *Bioinformatics* 2001;**17**:913–19.
66. Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 2001;**17**:847–8.
67. Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;**32**: D138–41.
68. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 1998;**95**: 5857–64.
69. Wylie T, Martin JC, Dante M, *et al*. Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes. *Nucleic Acids Res* 2004;**32**:D423–6.
70. Ayoubi P, Jin X, Leite S, *et al*. PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Res* 2002;**30**:4761–9.
71. Pertea G, Huang X, Liang F, *et al*. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 2003;**19**: 651–2.
72. Quackenbush J, Liang F, Holt I, *et al*. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* 2000;**28**:141–5.
73. Curwen V, Eyras E, Andrews TD, *et al*. The Ensembl automatic gene annotation system. *Genome Res* 2004;**14**: 942–50.
74. Eyras E, Caccamo M, Curwen V, Clamp M. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res* 2004;**14**:976–87.
75. Wilcox AS, Khan AS, Hopkins JA, Sikela JM. Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res* 1991;**19**: 1837–43.
76. Jiang J, Jacob HJ. EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Res* 1998;**8**:268–75.
77. Brett D, Hanke J, Lehmann G, *et al*. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 2000;**474**:83–6.
78. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res* 1999;**9**:1288–93.
79. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 2001;**29**:2850–9.
80. Kan Z, Castle J, Johnson JM, Tsinoremas NF. Detection of novel splice forms in human and mouse using cross-species approach. *Pac Symp Biocomput* 2004;42–53.
81. Schmitt AO, Specht T, Beckmann G, *et al*. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res* 1999;**27**: 4251–60.
82. Buetow KH, Edmonson MN, Cassidy AB. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet* 1999;**21**:323–5.
83. Picoult-Newberg L, Ideker TE, Pohl MG, *et al*. Mining SNPs from EST databases. *Genome Res* 1999;**9**:167–74.
84. Useche FJ, Gao G, Harafey M, Rafalski A. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform SerWorkshop Genome Inform* 2001;**12**:194–203.
85. Ewing RM, Ben Kahla A, Poirot O, *et al*. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res* 1999;**9**:950–9.
86. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999;**9**:1106–15.
87. Hawkins V, Doll D, Bumgarner R, *et al*. PEDB: the Prostate Expression Database. *Nucleic Acids Res* 1999;**27**:204–8.
88. Qiu F, Guo L, Wen TJ, *et al*. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. *Plant Physiol* 2003;**133**:475–81.
89. Lisacek FC, Traini MD, Sexton D, *et al*. Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. *Proteomics* 2001;**1**: 186–93.
90. Kim SI, Kim JY, Kim EA, *et al*. Proteome analysis of hairy root from Panax ginseng C.A. Meyer using peptide fingerprinting, internal sequencing and expressed sequence tag data. *Proteomics* 2003;**3**:2379–92.
91. Nelson RT, Grant D, Shoemaker RC. ESTminer: a suite of programs for gene and allele identification. *Bioinformatics* 2005;**21**:691–3.
92. Mooney BP, Krishnan HB, Thelen JJ. High-throughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of UniGene expressed sequence tag databases for protein identification. *Phytochemistry* 2004;**65**:1733–44.
93. Nilsson RH, Rajashekar B, Larsson KH, Ursing BM. galaxieEST: addressing EST identity through automated phylogenetic analysis. *BMC Bioinformatics* 2004;**5**:87.
94. Wu X, Walker MG, Luo J, Wei L. GBA server: EST-based digital gene expression profiling. *Nucleic Acids Res* 2005;**33**: W673–6.
95. Huang Y, Pumphrey J, Gingle AR. ESTminer: a Web interface for mining EST contig and cluster databases. *Bioinformatics* 2005;**21**:669–70.

96. Usuka J, Zhu W, Brendel V. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* 2000;**16**:203–11.

97. Parsons JD, Rodriguez-Tome P. JESAM: CORBA software components to create and publish EST alignments and clusters. *Bioinformatics* 2000;**16**:313–25.

98. Krause A, Haas SA, Coward E, Vingron M. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res* 2002;**30**:299–300.

99. Conesa A, Gotz S, Garcia-Gomez JM, *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;**21**:3674–6.