Expressed sequence tags: alternative or complement to whole genome sequences?

Stephen Rudd

Institut für Bioinformatik, GSF Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstraße 1, D-85764 Neuherberg, Deutschland

Over three million sequences from approximately 200 plant species have been deposited in the publicly available plant expressed sequence tag (EST) sequence databases. Many of the ESTs have been sequenced as an alternative to complete genome sequencing or as a substrate for cDNA array-based expression analyses. This creates a formidable resource from both biodiversity and gene-discovery standpoints. Bioinformatics-based sequence analysis tools have extended the scope of EST analysis into the fields of proteomics, marker development and genome annotation. Although EST collections are certainly no substitute for a whole genome scaffold, this 'poor man's genome' resource forms the core foundations for various genome-scale experiments within the as yet unsequenceable plant genomes.

Expressed sequence tags (ESTs) are currently the most widely sequenced nucleotide commodity from the plant genomes in terms of the number of sequences and the total nucleotide count. ESTs provide a robust sequence resource that can be exploited for gene discovery, genome annotation and comparative genomics. The slightly abstract nature of the EST collections, with their high levels of sequence redundancy, low-quality sequence attributes and short sequence lengths have left this enormous sequence collection as a rather under-exploited resource.

ESTs are typically unedited, automatically processed, single-read sequences produced from cDNAs (small DNA molecules reverse-transcribed from the cellular mRNA population) (Fig. 1). Libraries of cDNAs are routinely prepared that contain tens of thousands of clones, represent a variety of specific tissues types and represent a snapshot of gene expression during defined developmental stages and following specific biotic and abiotic challenges. The relative cheapness of EST sequencing and its associated automation often make EST sequencing the most attractive route for broad sampling of the transcriptome.

The concept of using cDNAs as a route to expedited gene discovery was first demonstrated in the early 1980s [1]. In 1990, Sydney Brenner proposed that an obvious method for characterizing the 'important' part of the human genome would involve looking at messengers from the expressed genes – thus advocating the application of high-throughput methods for transcriptome sampling [2]. Mark Adams first used the term EST in relation to gene discovery and the human genome project in 1991 [3]. Since these first publications, >16 million ESTs have been sequenced from more than 500 distinctly annotated species, representing a wide taxonomic variety of fungi, plants and animals (dbEST 4 April 2003 [4]).

EST sequencing initially favoured the 5′ end of directionally cloned cDNAs because the 5′ sequences are likely to contain more protein coding sequence than the 3′ ends, which often contain significant untranslated regions (UTRs). Improvements in the techniques for cDNA preparation and the advent of capillary-based sequencing have driven the evolution of high-throughput sequencing for plant ESTs. Currently, the 3′ end of the cDNA clone is often preferred because it is likely to offer more unique sequence (in many cases, the UTR) and can be used to distinguish between gene paralogues. EST sequencing strategies in which both ends of the cDNA are sequenced are also becoming widespread.

With the advent of cDNA array-based methodologies, ESTs have become a key reagent within an experiment rather than the final product [5]. In these arrays, a large collection of cDNAs is fixed to a substrate and an associated EST sequence provides the link between an experimental coordinate and a gene that might be up- or downregulated. Array experiments allow massive, parallel investigation of gene expression between tissues or following specific challenges. In the absence of complete genome sequences, the cDNA (and its EST) remains the only link back to the genome. Many of the ESTs available within the public domain are the products of cDNA sequencing associated with such array-based projects.

Contemporary uses of ESTs

Plant genome sizes extend over at least four orders of magnitude. *Arabidopsis* and *Oryza sativa* (rice), our model plants with fully sequenced genomes, have among the smallest known genomes: 125 Mbp and 430 Mbp, respectively. Tomato has a genome size of ~950 Mbp [6] and maize has a genome size of ~2670 Mbp. Cycad and wheat have genome sizes of ~14 000 Mbp and ~17 000 Mbp, respectively. The largest known genomes are currently those of *Fritillaria assyriaca*

Corresponding author: Stephen Rudd (s.rudd@gsf.de).

 1360-1385/03/$ - see front matter doi:10.1016/S1360-1385(03)00131-6

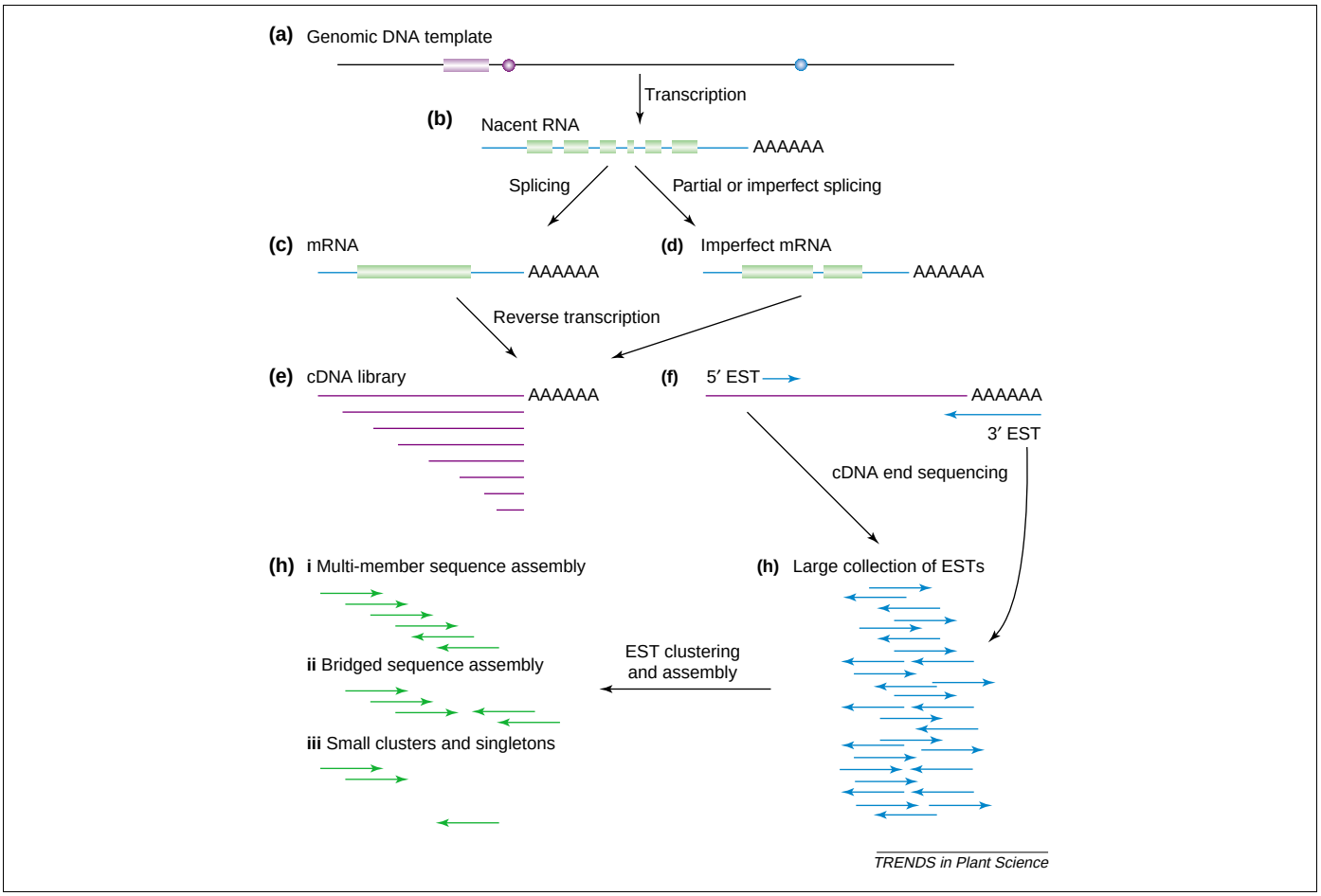

Fig. 1. Summary of cDNA cloning and expressed sequence tag (EST) sequencing. The genomic DNA template (a) contains regulatory elements and signals that define the location of a gene and recruit the DNA transcription machinery. The gene feature is transcribed and yields a nascent RNA (b). This RNA reflects the gene structure and contains untranslated regions and intron (blue lines between green boxes) and exon (green boxes) sequences. The nascent RNA is spliced perfectly into mRNAs (c) or imperfectly into aberrant or imperfect mRNAs (d). Such imperfect mRNAs might contain unspliced intron features; here, the fourth intron has not been spliced out of the RNA. RNAs prepared from cellular material are a complex mixture and both mRNAs and imperfect mRNAs are represented within the mix. A cDNA population (e) is reverse transcribed from the RNA population. The 3′ poly-A + tail is used as a selective tag for mRNA selection and so the 3′ end of the genes are more likely to be represented within the cDNA libraries than the 5′ ends are. DNA sequence is read from the ends of the cDNA, yielding 5′ and 3′ ESTs (f). These sequences join the pools of available ESTs for this organism (g). Bioinformatics-based EST clustering and sequence assembly tries to order the large EST collection into logical overlapping sequence contigs (h). Such clustering and assembly yields large assemblies that represent many individual ESTs (hi). These clusters can faithfully represent the spliced gene structure of the parental cDNA. On the basis of cDNA information, distinct sequence assemblies can be bridged into pseudo-clusters when ESTs stem from the same parental cDNA clone (hii). Several ESTs aggregate into either small clusters or persist as sequence singletons (hiii).

(125 000 Mbp) and *Psilotum nudum* (~250 000 Mbp) [7] (http://www.rbgkew.org.uk/cval/homepage.html).

The evolution of genome size has been driven in part by polyploidy and chromosomal duplication events. The expansion of genomes has mainly been the result of multiplication of retrotransposon repeat sequences. In maize, such retrotransposons have accounted for the doubling of the genome size during the past six million years [8–10]. Retrotransposons have been shown to aggregate within the gene space and their presence has been used to explain the narrow range of GC percentages within the gene space isochores [11,12]. These retrotransposon sequences are problematic within the context of complete genome sequencing. Although the main emphasis of plant genome sequencing is currently on discovering and characterizing the range of protein-coding genes present within the genome, thousands of copies of large repeats yield no information on the proteome. Such repeats additionally generate statistical issues that prevent the assembly of individual sequence reads into meaningful contigs [13,14]. The larger plant genomes are therefore currently precluded from complete sequencing.

This has directed the evolution of robust alternative methods to access the desirable protein-coding component of the genome. Techniques such as *Rescuemu* allow a genome sampling strategy similar to genome survey sequencing but that is selective to just the gene space, although retrotransposon sequences will remain highly abundant [15]. Methyl filtration of the gene space [16] works on the observation that the repetitive sequences within the genome are usually methylated. By selecting for nonmethylated islands of genomic DNA, the protein-coding portion of the genome is greatly enriched. However, cDNA sequencing remains one of the more accessible and widely used methods for sampling the actively transcribed portion of the gene space. The preparation of cDNA libraries depends on the underlying mRNA population of a cell, tissue or organism. The genomic structure of the host plant is therefore largely inconsequential. However,

the dependence on the mRNA population generates a different set of issues, problems and limitations.

The use of EST sequencing as a strategy and perspective to circumvent complete plant genome sequencing has been reviewed recently [17]. Here, I address what is already present within the sequence databases from both genomics and biodiversity standpoints. I then illustrate why ESTs are a suitable alternative to complete-genome sequencing but also demonstrate that ESTs should best form a highly suitable accompaniment to the complete genome projects. Some of the major plant EST bioinformatics resources are outlined and I present a brief view of how this 'traditional' sequence resource has recently been exploited within annotation and marker development projects.

EST sequence availability and biodiversity

With the latest release of the EMBL sequence database [18] and the weekly updates to the EST database (ftp://ftp.ebi.ac.uk/pub/databases/embl/new/), there were ~16.1 million ESTs available within the public domain by 14 April 2003. Of these, over 3.1 million are from plant species and account for 1550 Mbp sequence, with almost 200 species represented. Table 1 lists the plant species with most available ESTs ranked by the number of ESTs.

When we consider the overall biodiversity represented within the EST libraries, most sequences are attributed to either model plant species (*Arabidopsis, Chlamydomonas, Physcomitrella*) or species of agricultural or agronomic interest (rice, maize, soybean). Represented species are restricted to just a few groups within the plant evolutionary tree (euasterids I and II, eurosids I and II, the monocots and the chlorophytes), leaving both unsampled and superficially sampled groups (e.g. the gnetales and ginkos). There is no evidence upon which we can consider our currently completed plant genomes or the genomes with deeply sampled EST collections (Table 1) as being taxonomically representative beyond their most immediate clades. This naturally poses limitations on the scope and types of comparative analyses that can be performed using the currently available plant EST sequences. The taxonomically rich sequence diversity already existing within and between the individual groups certainly has the potential to be used to address specific questions about the conservation of protein families between well-sampled groups.

The need for a more even sampling of plant genomes has recently been discussed, and there are many 'ideal' genomes that could be the focus of complete genome sequencing [19,20]. With the complications of complete plant genome sequencing, deep EST sampling from a broader collection of currently unsampled taxa might offer us a better glimpse of the functional and evolutionary processes that are fundamental to plant life.

Table 1. Top 28 plant species ranked by the available number of ESTs

Species[a]	Family[a]	Number of ESTs
Triticum aestivum	Poaceae	423 632
Hordeum vulgare	Poaceae	339 034
Glycine max	Fabaceae	308 582
Zea mays	Poaceae	206 002
Oryza sativa	Poaceae	199 186
Medicago truncatula	Fabaceae	181 444
Arabidopsis thaliana	Brassicaceae	178 544
Lycopersicon esculentum	Solanaceae	148 566
Chlamydomonas reinhardtii	Chlamydomonadaceae	140 457
Populus spp.	Salicaceae	105 353
Solanum tuberosum	Solanaceae	94 423
Sorghum bicolor	Poaceae	84 712
Physcomitrella patens	Funariaceae	70 065
Lactuca sativa	Asteraceae	68 188
Pinus taeda	Pinaceae	60 226
Helianthus annuus	Asteraceae	46 951
Vitis vinifera	Vitaceae	42 093
Gossypium arboreum	Malvaceae	38 894
Lotus japonicus	Fabaceae	33 124
Ipomoea nil	Convolvulaceae	25 899
Mesembryanthemum crystallinum	Aizoaceae	25 803
Capsicum annuum	Solanaceae	22 433
Sorghum propinquum	Poaceae	21 387
Phaseolus coccineus	Fabaceae	20 120
Beta vulgaris	Chenopodiaceae	18 999
Gossypium hirsutum	Malvaceae	10 725
Porphyra yezoensis	Lamiaceae	10 354
Prunus persica	Rosaceae	10 185

[a]Taxonomic assignment was determined through the NCBI taxonomy database [54]. Only EST collections with more than 10 000 sequences are represented here. A summary for all ESTs available within the dbEST database is available from http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.

ESTs and their limitations

There are two main problems associated with EST sequences: (1) the overall representation of host genes within a library and (2) the overall quality of any individual sequence within a collection.

Underlying mRNA populations

The mRNA population within a single cell, tissue or organism represents the collection of genes that are being actively transcribed to maintain homeostasis and genes that are expressed to achieve growth, pathogen defence or any other response from a wide assortment of available effects. A collection of core housekeeping genes is ubiquitously expressed within cells, and other genes are expressed within a narrow range of cells following a specific challenge or during a subtle developmental transition. A standard cDNA library is a faithful representation of the ratio of mRNAs present within a specific tissue under exact conditions at the time of sampling. Poorly expressed genes will be poorly represented within libraries, and genes that are not expressed will be absent. For example, the 178 000 ESTs from *Arabidopsis* only match to 16 115 distinct genes, even though the *Arabidopsis* EST collection represents 61 distinct tissues, biotic and abiotic challenges, and developmental stages.

The incomplete coverage of the underlying gene collection is a complex issue to resolve. To achieve sampling of all transcriptional units would require mRNA from all cell types at all developmental stages and following all combinations of biological and environmental challenges. This is something that can only be addressed through experimental design and the deepest

sampling of the widest range of cell types and challenges – something that is not yet practical or effective. The uneven representation of cDNA clones within the underlying libraries, however, can be addressed. Both oligofingerprinting [21] and normalization/subtraction [22] of cDNA libraries have been used to equalize the relative occurrence of the common and rarer transcripts, and have recently accounted for a leap in the sequence diversity reflected within some cDNA libraries (e.g. sugar beet and zebrafish).

The limitation of gene representation within cDNA libraries is difficult to resolve. We can safely use EST data for large-scale comparative genomics experiments, for gene discovery and for genome annotation as long as we are aware of the underlying library issues. The presence of an EST is a reliable attribute, whereas the absence of an EST does not mean that the gene is not present in the genome, and does not mean that the gene is not expressed. We can only conclude that there was no measurable transcript under the conditions from which the particular tissue was sampled.

Sequence quality

Sequence quality describes the faithfulness with which an EST sequence represents the gene sequence from which it was reverse-transcribed and cloned. A low quality EST sequence is a poor representation of its cognate host gene. When we consider individual nucleotides within an EST against their cognate genomic reference nucleotide, as many as 3% of nucleotides can be incorrect [23], representing insertions, deletions and substitutions. The quality of individual nucleotides reflects the fidelity of the reverse transcriptase used within cDNA preparation [24], the fidelity of the sequencing reaction performed and the accuracy with which the sequence has been determined from the electropherogram trace file [25].

In addition to the background of incorrect nucleotides within EST sequences, there is a background of partially or completely incorrect sequence. Partially incorrect sequences contain stretches of vector or polylinker sequence. Completely incorrect sequences represent xenocontaminants (sequences from a foreign genome such as *Escherichia coli* or human) and, in many cases, structural or regulatory RNAs that should not have been cloned or sequenced on the basis of the poly-A+ tail. True xenocontaminants are rare but non-protein-coding RNAs can account for as much as 1% of all ESTs. Poly-A+ selection of the mRNAs introduces further sequence bias because the cDNA contains the complete 3′ UTR. ESTs therefore contain significant amounts of non-protein-coding UTR (typically over 100 nucleotides each).

ESTs are not the complete representation of the parental cDNA. An EST's length is limited to a few hundred nucleotides of reliable sequence. This is significantly shorter than the length of the average plant gene. Even if both ends of the cDNA have been sequenced (producing two ESTs), the ESTs will not overlap in most cases, although the clones can be physically joined into a single logical pseudomolecule.

The limitations of EST sequence that pertain to quality are merely technical issues and, within the context of most analyses, do not really detract from their value. There has been much recent development of bioinformatics-based methods that address these limitations and, in many cases, remove the problems. The only viable alternatives to EST sequencing that address the attributes of incomplete sequence coverage and nucleotide quality are the full-length cDNA sequences. Full-length cDNA sequences are obtained by shotgun sequencing cDNA clones that have been selected for both 5′ and 3′ ends [26]. Such a strategy yields many individual ESTs that can be assembled into a single contig. The underlying redundancy within the shotgun sequence means that each individual residue will be sequenced many times and the consensus will be of the highest quality. Because many sequence reads are performed, the full-length cDNA time and consumable costs are much higher than for ESTs.

Bioinformatics of plant EST collections

Bioinformatics-based sequence resources have been developed that address the quality, redundancy and partial nature of EST sequences. Sequence resources such as the dbEST database [4] and the EMBL database [18] archive all the available ESTs and provide methods to search for individual sequences on the basis of species, clone or homology attributes. However, these searches are limited to the sequence features that are supplied when the sequence is submitted.

A range of plant specific EST databases has been described in the last few years in which sequence analysis and annotation has moved beyond the EST sequence and significant value has been added. The first crucial step in adding value to EST sequences involves clustering and assembling the ESTs into a more manageably sized dataset of better quality clustered sequences. Although there are a range of methods that achieve this goal, they generally perform the same processing steps to achieve a common result. Sequences are aggressively trimmed of vector and polylinker remnants before a fast clustering method places the ESTs into buckets of similar sequences (e.g. Ref. [27]). A final assembly step places the clustered sequences into logical contigs and singletons [28,29]. The clustered sequences are typically longer than any individual EST and are of a higher quality. Cluster consensus sequences additionally merge valuable information on sequence polymorphisms that would otherwise not be observable. Figure 2 shows a graphical representation of how 20 barley EST sequences have been reduced into a single contig of higher overall quality. A clear single nucleotide sequence polymorphism can be observed.

These cluster consensus and singleton sequences form the core sequence data within several plant specific EST derived databases. A collection of these sequence resources is shown in Table 2. Most of these sequence databases have added further value to the sequences by attaching additional annotation to the sequences and by providing methods to select specific sequences or groups of sequences that satisfy specific criteria. The most valuable annotations and methods are those that assign tentative function and allow retrieval and identification of sequences on the basis of tissue or challenge specificity.

Bioinformatics-based clustering of EST sequences

yields a better quality and less redundant sequence collection – a basic approximation of the partial host genome. Bioinformatics also introduces some error into the large cluster collections. Successive rounds of polyploidization and gene duplication events associated with plant genomes have created a wide range of gene redundancy. Clustering ESTs from close paralogues will aggregate sequences from different genes into the same clusters on the basis of conserved nucleotide sequence. In addition, EST sequences that stem from the same gene might not assemble into a single cluster – the ESTs might not physically overlap or low quality sequence stretches might separate the sequences into different clusters. These problems remain subtle and do not detract from the applicability of such clustered ESTs within comparative genomics, gene mining or marker development.

ESTs as a current alternative to complete genomes

Within the field of 'reconstructomics' [30], ESTs have widely been applied as the foundation sequence of some genome-scale analyses. Such reconstructomic analyses use the EST cluster assemblies and singletons as an equivalent to a whole genome's gene collection. EST derived cluster sequences have been widely annotated with tentative functions. Sources of functional annotation have included non-redundant protein databases [31], the *Arabidopsis* genome annotation [6] and catalogues of functionally assigned proteins [30]. The annotations are homology based and EST sequences or clusters inherit the annotative attributes of their match. This approach naturally suffers from problems with the propagation of annotation errors, but manual validation of EST assignments has been shown to be consistent with such automated annotations [32]. The surrogate annotation methods have been used to crudely dissect the overall representation and distribution of functional classes of protein both within and between genomes, and functional pie charts have become common within both genome and EST papers [6,33,34].

ESTs have additional broad applications within comparative genomics. Although individual EST collections might only represent as much as 60% of the host genes [30], the overlap between collections provides a suitable hunting ground for the selection of the common genes. Before the advent of either the *Arabidopsis* or rice genome, smaller collections of ESTs were used to estimate the number of genes conserved between the genomes [35]. Thirty-six thousand *Arabidopsis* ESTs and 27 000 rice ESTs were clustered within and then between species to generate 'mixed superclusters'. The observation that ~35% of total clusters were common to *Arabidopsis* and rice is not too different to the ~50–80% that was presented in the publication of the draft rice genome, particularly when we consider than the draft genome data was based on predicted genes rather than 'validated' transcripts. However, these findings are a more potent indicator of the profound gene content differences between these two genomes, and that will undoubtedly exist between other similarly related pairs of genomes. Although the *Arabidopsis*–rice comparison used tens of thousands of sequences, such deep sequencing is not a prerequisite for such overlap detection. Shallow EST sampling (~1000 samples per genome) from seven beetle species revealed interspecies library overlaps of 5–21% [36]. Such small numbers of sequences cannot be used within a genomics context, but this rapid analysis method has applications in taxon sampling and phylogenomics.

With a selection of annotated proteins from a mixture of tissues from the same species, commonality can be observed between libraries. In potato, for example, 48 cluster consensus sequences were found to be common to nine different libraries and to contain known and probable housekeeping genes [31]. The reciprocal question asks which sequences can be observed that are restricted to a single library. Again, within the same potato sequence collection, one library represented tissue following an incompatible infection of the fungus *Phytophthora infestans* and over 400 sequences were observed to be specific to this cDNA library. Although some of the sequences corresponded to known disease resistance genes and others could be placed within defence response or other signalling pathways, ~70% of the sequences had no match to known proteins with an attributed function [31]. A similar strategy was used for to select transcripts that are expressed preferentially at different developmental stages of the potato tuber. In *Chlamydomonas*, the EST resources have been used in a similar manner to select the genes that are most likely to be involved within stress responses by performing such *in silico* subtraction on genes found within abiotically challenged cells [37]. Sequences identified as having a restricted pattern of gene expression

Table 2. Plant specific EST databases in which significant value has been added to large collections of EST sequences[a]

Plant EST database	URL	Genomes[b]	Refs
TIGR Plant Gene Indices	http://www.tigr.org/tdb/tgi/plant.shtml	All large collections of plant ESTs	[55]
NCBI Unigenes	http://www.ncbi.nlm.nih.gov/UniGene/	11 plants with largest EST collections	[54]
MIPS Sputniks	http://mips.gsf.de/proj/sputnik/	All large collections of plant ESTs	[30]
PlantGDB	http://www.zmdb.iastate.edu/PlantGDB/	All large collections of plant ESTs	
University Minnesota	http://www.ccgb.umn.edu/	Pt, Mt, Gm	[56]
B-EST barley database[c]	http://pgrc.ipk-gatersleben.de/est/login.php	Hv	
Kazusa EST databases[c]	http://www.kazusa.or.jp/en/plant/database.html	Lj, At, Py, Cr	
Solanaceae genomics network	http://sgn.cornell.edu/	Different *Lycopersicon* and *Solanum* species	[6]
Chlamydomonas resource centre	http://www.biology.duke.edu/chlamy_genome/	Cr	[37]

Abbreviations: At, *Arabidopsis thaliana*; Cr, *Chlamydomonas reinhardtii*; Hv, *Hordeum vulgare*; Mt, *Medicago truncatula*; Lj, *Lotus japonicus*; Pt, *Pinus taeda*; Py, *Porphyra yezoensis*.

[a]These resources provide annotation on a range of different large EST collections and provide methods and resources to select for alternative splice sites, tissue specific patterns of representation or functional or structural attributes, or on the basis of other annotations.

[b]The EST collections that are represented within the resource.

[c]These sequence collections perform annotation and analysis using just proprietary datasets.

Fig. 2. Schematic representation of a sequence cluster. Twenty expressed sequence tags (ESTs) from *Hordeum vulgare* have been assembled into a single consensus sequence of 1535 nucleotides. Dark-blue bars represent individual nucleotides that disagree with the consensus sequence. The overall cluster consensus sequence is ~2.6 times longer than the average EST length of 597 nucleotides and the complexity of the dataset has been greatly reduced. The conserved pattern of mismatches that can be observed in the EST sequences is an indicator of a putative single nucleotide polymorphism, which is shown in detail in the sequence close-up view and corresponds to a polymorphism between the Barke–Morex–Optic varieties and an unannotated or Haruna Nijo variety.

between tissues and challenges, and their respective controls, are good candidates for forward genetic screens to characterize their role and function. When we consider that the public plant EST sequence databases contain ~200 species, ~1000 distinct tissues, biotic and abiotic challenges and developmental stages, and three million individual ESTs, we already have an adequate resource for the selection of at least some candidate genes.

Another useful attribute of such reconstructomes is the detection of overlap between sequence collections – a more in depth comparative genomics analysis. Potato and tomato sequence collections overlap by between 70% and 80% of all sequences, depending on the similarity thresholds used as a cut off [31]. Using slightly tenuous parameters, tomato has been estimated to overlap with 70% of *Arabidopsis* sequences [6]. Of the 114 tomato sequences with no detectable *Arabidopsis* homologue but with a characterized homologue elsewhere within the protein sequence databases, 11 could be labelled as belonging to putative Solanaceae-specific gene families. Seventeen additional sequences were identified that could be found within both the euasterid and eurosid clades, suggesting gene loss within the *Arabidopsis* lineage [6]. Such analyses certainly illuminate the potential of EST sequences within the understanding of gene evolution. Although these analyses are clear examples of how ESTs can be exploited, the comparative analyses used the *Arabidopsis* genome scaffold. The safe assignment of a gene as being restricted to a single lineage on the basis of EST data alone is dangerous – sampling of the transcriptome is biased and further genome-based experimental evidence would be required to validate such absence or presence differences.

ESTs as a complement to complete genomes

Complete genome sequences have been produced for *Arabidopsis* [33] and rice [34,38]. The complete genome scaffolds for *Zea mays, Medicago truncatula, Brassica napus* and *Populus* are either within the sequencing or preparation stages and other plant genomes will follow. ESTs really spring into the limelight when we are presented with a new complete genome sequence and wish to start annotating genes to the chromosomes. Although the underlying methods and science required for the detection and modelling of eukaryotic genes have been well described elsewhere [39,40], one universal theme is the strong value and dependence placed on ESTs, first within the identification of the gene regions for training the gene prediction algorithms and, second, within the validation and correction of genes that have been predicted using the trained gene modelling algorithms [41].

Re-annotation of the *Arabidopsis* genome using a new collection of full-length cDNAs characterized 240 genes that had escaped annotation using the standard gene modelling algorithms [42]. ESTs have also demonstrated their worth in the selection of apparently unannotated proteins and putative small peptides from *Arabidopsis* [43,44]. This EST and cDNA approach has also been used to annotate the UTRs of genes, to correct the boundaries of introns and exons, and to identify new introns (especially within the UTRs) and probable micro-exons. ESTs have also been used to discover non-canonical splice sites [42,44]. On the basis of EST data, alternative splicing has been shown to be a rare occurrence within plants, although examples can be found [42,45]. This contrasts greatly with the mammalian system, in which alternative splicing is widespread.

ESTs are invaluable within genome annotation and, with the arrival of new genomes, more ESTs and full-length cDNAs are sure to follow. Issues with annotation of the rice genome have interestingly been partly attributed to the lack of high quality ESTs and full length cDNAs [46].

With the public sequence resources containing over 130 000 sequences – we can perhaps ask what else they need.

New tricks with old sequences

It is only recently that plant biologists have taken these vast EST datasets in hand and started a concerted effort to mine the data for novel attributes, started *de novo* annotation of the sequences, used the sequences within proteomics-based analysis pipelines and exploited the sequences for molecular marker development. There has recently been much interest in the field of expression profiling. By clustering and relating genes on the basis of their expression patterns, genes can be identified that are either involved within the same metabolic pathway or functional or structural complex, or are co-regulated. ESTs have potential here beyond the basic subtraction methods mentioned earlier. A crude measure of expression profile can be estimated on the basis of EST count alone [47]. A rice EST collection of 27 000 sequences (in 1999) was exceptional in the number of underlying libraries represented within the sequence collection. Expression profiles were computed for each EST cluster that was represented at least five times within any of ten different cDNA libraries. A rigorous statistical test was applied and, despite the inherent noise, coherent patterns of gene expression were observed. The expression profiles reflected the expected similarities between related libraries, and genes with related functions were observed to have correlated expression patterns [47]. With the enormous subsequent growth in the number of both plant species and distinct tissues represented, such methods will be more widely applicable and will reveal further uses for ESTs within gene discovery.

Although EST libraries are renowned for their low sequence quality and high levels of contamination, steps have been taken to exploit this dubious side of the sequence collections. Noncoding RNAs that represent structural RNAs, putative gene regulators and signal molecules can exist as polyadenylated RNAs and can appear within cDNA libraries (although the absence of a poly-A + tail does not guarantee exclusion). A screen of specific *Arabidopsis* EST collections was used to identify new classes of apparently plant specific noncoding RNAs, RNAs that represent small peptides and other uncharacterizable RNAs [48]. EST collections have applications within sequence discovery beyond the mRNAs.

Within the field of proteomics, one of the goals is to unequivocally identity proteins that have been resolved on two-dimensional polyacrylamide gels. A protein spot is excised from the gel and digested, and the molecular weights of the observed peptide fragments are matched to theoretically digested proteins from sequence databases using peptide mass fingerprinting (PMF). For optimal assignments, the complete sequence of the parental genome is required. In the absence of the parental genome, phylogenetically close relatives can be used, but increasing evolutionary distances make the process highly ineffective. Open reading frames from plant ESTs have been matched using PMF to empirically generated mass information [49]. ESTs make a useful substitute where no complete genome is available.

In the absence of complete genome sequences, the desire to generate high-density genetic maps of the different plant genomes remains a priority for the directed identification of specific genes. There is a range of contemporary genetic marker types and all have been exploited using attributes of EST data. Simple sequence repeats have been identified from the wheat genome and have applications in genotyping [50]. Single nucleotide polymorphism (SNP) markers have been preselected from various EST collections on the basis of available quality scores [51] and, more recently, SNPs have been predicted and validated from barley by screening for conserved patterns of polymorphism within EST sequence clusters (Fig. 2) (R. Kota *et al.*, unpublished). The efficiency with which SNPs have been preselected from the already available EST data is higher than with a *de novo* SNP selection strategy. The strategy is also faster because, once candidate loci have been identified, only validation and characterization steps are required. Such EST-based SNP screening projects are likely to become more common in the future. SNP markers rely upon the underlying redundancy within EST collections and assume that distinct varieties of a plant genome will be represented within a collection.

Conserved orthologue set (COS) markers (or anchored reference loci) are gene sequence markers that anchor genes to common syntenic islands between related genomes [32,52]. COS markers between both *Arabidopsis* and tomato, and between a range of *Populus* species have been identified using EST data. These marker-based methods rely on the depth of the EST collections to provide enough search space to identify the required sequence attributes. COS markers require deep sequencing from distinct but somehow related genomes. This is something that is available already for a variety of key genomes (e.g. maize, rice, wheat, and barley). Directed sequencing will be required for more-specific COS marker development.

ESTs often have robust applications outside the research areas within which they were originally sequenced. The need for more genome sequence from more genomes continues to grow. As long as complete genomes remain unsequenced and the EST collections continue to grow, new roles for ESTs will be found. ESTs certainly have a niche within marker development and have some primitive applications within proteomics and expression profiling. Will the future reveal more novel applications for ESTs?

Final comment

As long as ESTs continue to be actively sequenced to fill in knowledge gaps from the gene complement of the large plant genomes, our potential knowledge bases will continue to grow. EST sequencing certainly avoids the biggest problems associated with genome size and the accompanying retrotransposon repetitiveness. The EST sequence resources have been shown to have a wide range of applications and novel uses have been found for the resources. There are, however, some fundamental

limitations to this sequence resource. The lack of sequence information beyond the primary transcripts excludes all information about the regulation and control of the corresponding genes. This is something that can only be addressed with other gene space enrichment and sequencing methods. These methods, however, have not yet become widespread outside the maize genome sequencing project [53]. Also lacking is topological information that can be used in the detection of synteny and within colinearity studies, although these data can be obtained within subsequent experiments.

There is no real substitute for a complete genome sequence: only when presented with the completed chromosomes can we dissect the gene complement and unravel the mechanistic pathways that make the plant. Until new technologies become generally available that can produce longer sequence reads more cheaply, we will be limited to incomplete solutions. I believe that ESTs are worth their costs, and the subtleties of the rapidly growing libraries will certainly keep my research group out of mischief for the foreseeable future.

Acknowledgements

Thanks to Heiko Schoof and Wojciech Karlowski for critical appraisal of the manuscript. I am funded within the GABI project by the BMBF (0312270/4).

References

1 Putney, S.D. *et al.* (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 302, 718–721
2 Brenner, S. (1990) The human genome: the nature of the enterprise. *CIBA Found. Symp.* 149, 6–17
3 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
4 Boguski, M.S. *et al.* (1993) dbEST – database for 'expressed sequence tags'. *Nat. Genet.* 4, 332–333
5 Gress, T.M. *et al.* (1992) Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* 3, 609–619
6 Van der Hoeven, R. *et al.* (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14, 1441–1456
7 Obermayer, R. *et al.* (2002) Nuclear DNA *C*-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* 90, 209–217
8 Gaut, B.S. *et al.* (2000) Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7008–7015
9 Heslop-Harrison, J.S. (2000) Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *Plant Cell* 12, 617–636
10 Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115, 29–36
11 Carels, N. *et al.* (1995) The gene distribution of the maize genome. *Proc. Natl. Acad. Sci. U. S. A.* 92, 11057–11060
12 Barakat, A. *et al.* (1997) The distribution of genes in the genomes of Gramineae. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6857–6861
13 Hoskins, R.A. *et al.* (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol.* 3, RESEARCH0085.1–0085.16
14 Kent, W.J. and Haussler, D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res.* 11, 1541–1548
15 Raizada, M.N. *et al.* (2001) Somatic and germinal mobility of the RescueMu transposon in transgenic maize. *Plant Cell* 13, 1587–1608
16 Rabinowicz, P.D. *et al.* (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* 23, 305–308
17 Mayer, K. and Mewes, H.W. (2002) How can we deliver the large plant genomes? Strategies and perspectives. *Curr. Opin. Plant Biol.* 5, 173–177
18 Stoesser, G. *et al.* (2003) The EMBL nucleotide sequence database: major new developments. *Nucleic Acids Res.* 31, 17–22
19 Daly, D.C. *et al.* (2001) Plant systematics in the age of genomics. *Plant Physiol.* 127, 1328–1333
20 Pryer, K.M. *et al.* (2002) Deciding among green plants for whole genome studies. *Trends Plant Sci.* 7, 550–554
21 Herwig, R. *et al.* (2002) Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes. *Plant J.* 32, 845–857
22 Bonaldo, M.F. *et al.* (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 6, 791–806
23 Hillier, L.D. *et al.* (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6, 807–828
24 Curry, J. and Glickman, B.W. (1997) Moloney murine leukemia reverse transcriptase suspect in the production of multiple misincorporations during hprt cDNA synthesis. *Mutat. Res.* 374, 145–148
25 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194
26 Seki, M. *et al.* (1998) High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper. *Plant J.* 15, 707–720
27 Heumann, K. and Mewes, H.W. (1996) The hashed position tree (HPT): a suffix tree variant for large data sets stored on slow mass storage devices. In *Proceedings of the Third South American Workshop on String Processing* (Ziviani, N. *et al.*, eds), pp. 101–115, Carlton University Press
28 Gordon, D. *et al.* (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* 8, 195–202
29 Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877
30 Rudd, S. *et al.* (2003) Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Res.* 31, 128–132
31 Ronning, C.M. *et al.* (2003) Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol.* 131, 419–429
32 Fulton, T.M. *et al.* (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14, 1457–1467
33 The *Arabidopsis* Genome Initiative, (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
34 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100
35 Ewing, R. *et al.* (1999) Comparative analysis of the *Arabidopsis* and rice expressed sequence tag (EST) sets. *In Silico Biol.* 1, 197–213
36 Theodorides, K. *et al.* (2002) Comparison of EST libraries from seven beetle species: towards a framework for phylogenomics of the Coleoptera. *Insect Mol. Biol.* 11, 467–475
37 Shrager, J. *et al.* (2003) *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 131, 401–408
38 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92
39 Stormo, G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.* 10, 394–397
40 Mathe, C. *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103–4117
41 Reese, M.G. *et al.* (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10, 483–501
42 Haas, B.J. *et al.* (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* 3, RESEARCH0029.1–0029.12
43 Brendel, V. and Zhu, W. (2002) Computational modeling of gene structure in *Arabidopsis thaliana*. *Plant Mol. Biol.* 48, 49–58
44 Zhu, W. *et al.* Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.* (in press)
45 Burke, J. *et al.* (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8, 276–290
46 Schoof, H. and Karlowski, W. (2003) Comparison of rice and *Arabidopsis* annotation. *Curr. Opin. Plant Biol.* 6, 1–7
47 Ewing, R.M. *et al.* (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9, 950–959
48 MacIntosh, G.C. *et al.* (2001) Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.* 127, 765–776

49 Lisacek, F.C. *et al.* (2001) Strategy for protein isoform identification from expressed sequence tags and its application to peptide mass fingerprinting. *Proteomics* 1, 186–193
50 Eujayl, I. *et al.* (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor. Appl. Genet.* 104, 399–407
51 Picoult-Newberg, L. *et al.* (1999) Mining SNPs from EST databases. *Genome Res.* 9, 167–174
52 Brown, G.R. *et al.* (2001) Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. *Genetics* 159, 799–809
53 Chandler, V.L. and Brendel, V. (2002) The maize genome sequencing project. *Plant Physiol.* 130, 1594–1597
54 Wheeler, D.L. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31, 28–33
55 Quackenbush, J. *et al.* (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* 29, 159–164
56 Lamblin, A.F. *et al.* (2003) MtDB: a database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Res.* 31, 196–201