Molecular Ecology (2007) **16**, 907–924 doi: 10.1111/j.1365-294X.2006.03195.x

The molecular ecologist's guide to expressed sequence tags

AMY BOUCK*† and TODD VISION†
*Department of Biology, Box 90338, Duke University, Durham, NC 27708, USA, †Department of Biology, Campus Box 3280, UNC Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

Genomics and bioinformatics have great potential to help address numerous topics in ecology and evolution. Expressed sequence tags (ESTs) can bridge genomics and molecular ecology because they can provide a means of accessing the gene space of almost any organism. We review how ESTs have been used in molecular ecology research in the last several years by providing sequence data for the design of molecular markers, genome-wide studies of gene expression and selection, the identification of candidate genes underlying adaptation, and the basis for studies of gene family and genome evolution. Given the tremendous recent advances in inexpensive sequencing technologies, we predict that molecular ecologists will increasingly be developing and using EST collections in the years to come. With this in mind, we close our review by discussing aspects of EST resource development of particular relevance for molecular ecologists.

Keywords: bioinformatics, gene expression, gene families, marker design, simple sequence repeats (SSRs), SNPs

Received 3 August 2006; revision accepted 7 October 2006

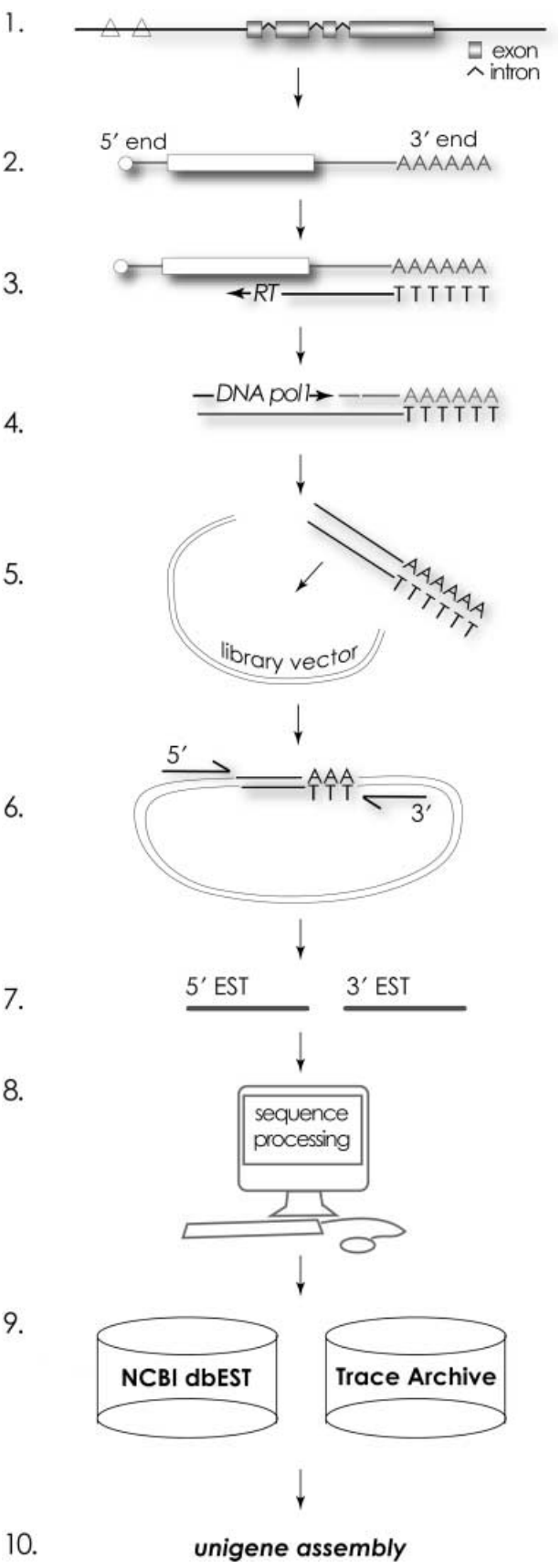
Introduction

The field of molecular ecology is built upon a well-developed body of work on the application of genetic markers to the study of population structure, gene flow, parentage analysis, biogeography and systematics. Molecular ecology also aims to elucidate the genetic basis of ecologically important phenotypic variation, and to understand the distribution of phenotypic and genotypic variation in natural populations in terms of fundamental evolutionary forces such as drift, selection, mutation, and migration (Feder & Mitchell-Olds 2003; Vasemagi & Primmer 2005; Lee & Mitchell-Olds 2006). Developments from the fields of genomics and bioinformatics have great potential to help address numerous topics in molecular ecology, not only by providing sequence information and comparative data useful in designing markers, but also by opening up entirely new methods to study the genetic basis of adaptation. The challenge in realizing genomics' impact on the field of molecular ecology lies in finding a way to extend genomics to the vast diversity of both organisms and questions studied by molecular ecologists (Feder & Mitchell-Olds 2003).

Correspondence: Amy Bouck, Fax: 919 6607293;
E-mail: bouck@duke.edu

While it is not feasible to invest heavily in genome sequence resources for every species or natural population, expressed sequence tags (ESTs) are a relatively inexpensive genomic resources that can be developed for almost any organism. ESTs are already among the most diverse (in terms of phylogenetic coverage) and abundant type of sequence data available. ESTs can serve as a source of molecular markers, and can also provide an entrée into gene and genome-level questions, even for studies of nonmodel organisms that lack other sequence resources and have no history of functional genetics. As we will illustrate, EST collections can serve as a bridge between the genomic resources of model organisms and diverse species of interest to ecologists and evolutionary biologists. Even if ESTs are not available from the organism under study, EST data from related organisms can be used in a variety of ways to study the ecology and evolution of diverse wild species.

EST collections sample the gene space of an organism by providing a snapshot of the transcribed mRNA population within a given set of tissues, developmental stages, environmental conditions and genotypes (see Rudd 2003; Dong *et al*. 2005 for reviews). In brief, ESTs are single-read sequences produced from partial sequencing of a bulk mRNA pool. Reverse transcriptase is used to produce bulk cDNA, which is then cloned into a vector library, and each



clone is individually end-sequenced (Fig. 1). The process of EST sequencing can be highly automated and is usually conducted on a scale of thousands to tens of thousands of sequence reads per library. By measuring the relative abundance of different transcripts, one can obtain information about gene expression within and between different tissues, life stages, or environmental conditions (an application termed 'expression profiling'). One may also obtain information about alternatively spliced forms of the same gene transcript. For most purposes [e.g. identification of orthologues, simple sequence repeat (SSR) primer design], raw EST sequences from the same transcript are assembled into a single consensus sequence, often called a unigene (Pontius *et al*. 2003).

EST data have a number of limitations. First, transcripts that are in low abundance in the particular tissues sampled may not be sequenced at all. Thus, the absence of a particular transcript is not strong evidence for its absence from the genome: the gene may actually be expressed at a very low level. ESTs give no information about genomic position, gene order, introns, or regulatory motifs. As we will show, some of these, such as intron position, can sometimes be inferred by comparison to genomic sequences from related organisms. ESTs typically represent only partial sequences of the original transcripts, and even unigenes seldom cover the full-length transcript. 'Raw', or unedited, ESTs are subject to a substantial rate of base-call errors and contamination. Even in the absence of error, the clustering of ESTs into unigenes is compromised by the difficulty of distinguishing alleles and alternative splice forms from paralogues (Wang *et al*. 2004). Unigene sets are difficult to evaluate when no reference genome exists for validation — the case for most organisms of interest to molecular ecologists (Dong *et al*. 2005).

Despite these limitations, we believe that ESTs are a valuable resource for molecular ecology. In this paper, we will first review how ESTs have been used to address diverse problems in evolutionary biology and ecology, including

Fig. 1 EST sequence production. Genomic DNA in the vicinity of a gene (1) contains exons, introns and upstream regulatory motifs (triangles). Introns are spliced out of mature mRNAs (2), which are capped at the 5′ end with a modified guanine nucleotide (circle) and at the 3′ end with a poly A tail. Reverse transcriptase (RT) is used to synthesize a cDNA strand from the mRNA template (3). A double stranded cDNA molecule is produced using RNAse H and DNA polymerase 1(DNA pol1) (4). cDNAs are inserted into cloning vectors to produce a cDNA library (5). This provides a means of isolating, storing, replicating and sequencing individual molecules. The inserts are sequenced from one or both ends using universal primers (6). The resulting EST sequences (7) are analysed to remove vector or other contaminating sequences and low quality base calls (8), and then deposited into public databases (9). These sequences are then clustered into unigenes (see Fig. 4) or used for other applications (10).

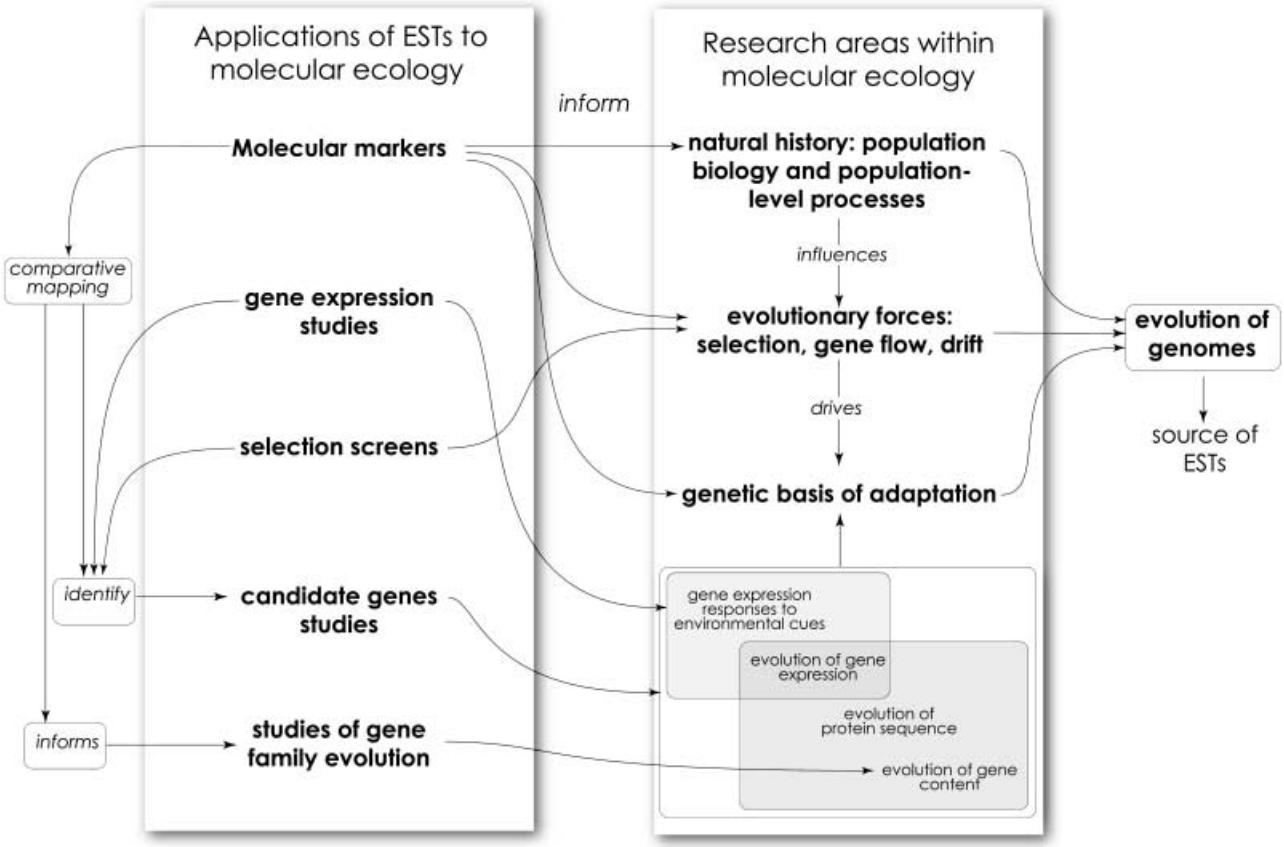


Fig. 2 The variety of ways in which ESTs can be applied to study molecular ecology. Arrows connecting the two boxes show how the types of studies described in the text address questions in molecular ecology. Mechanistically, the genetic basis of adaptation can vary, calling for different experimental approaches. Arrows to the left of the boxes show examples of how different approaches can be used together. As shown on the right, ecological and evolutionary forces ultimately shape the evolution of genomes, which are the source of EST data.

studies of natural history, evolutionary forces, and the genetic basis of adaptation (Fig. 2). Given the tremendous recent advances in inexpensive sequencing technologies, we predict that molecular ecologists will increasingly be developing and using EST collections in the coming years. With this in mind, we will close our review by discussing aspects of EST resource development of particular relevance for molecular ecologists.

ESTs in molecular ecology

Molecular marker design and discovery

The use of molecular markers has revolutionized the fields of conservation biology, population biology, population Genetics and ecology. Markers provide a means of observing otherwise hidden aspects of natural history, whether this involves population–level interactions on ecological timescales, or the evolutionary relationships of genes, populations, and taxa (Avise 2004). As a result, more effort has probably been directed at using ESTs for marker development than for any other purpose in molecular ecology. These sequences have been used to develop molecular markers tagging genomic regions ranging from highly mutable SSRs and single nucleotide polymorphisms (SNPs) to highly conserved genes, providing insights into questions ranging from population-level process such as parentage analysis, to the demarcation of orthologous genomic regions across distantly related species.

Simple sequence repeats

Simple sequence repeats, or microsatellites, have been widely used as molecular markers in ecology because of their abundance, high level of polymorphism and ease of scoring. Numerous examples of *in silico* mining of SSR markers out of EST data from diverse organisms have been published over the last several years (Scott *et al*. 2000; Scotti *et al*. 2000; Cordeiro *et al*. 2001; Rohrer *et al*. 2002; Jany *et al*. 2003; Bhat *et al*. 2005; Varshney *et al*. 2005). This approach

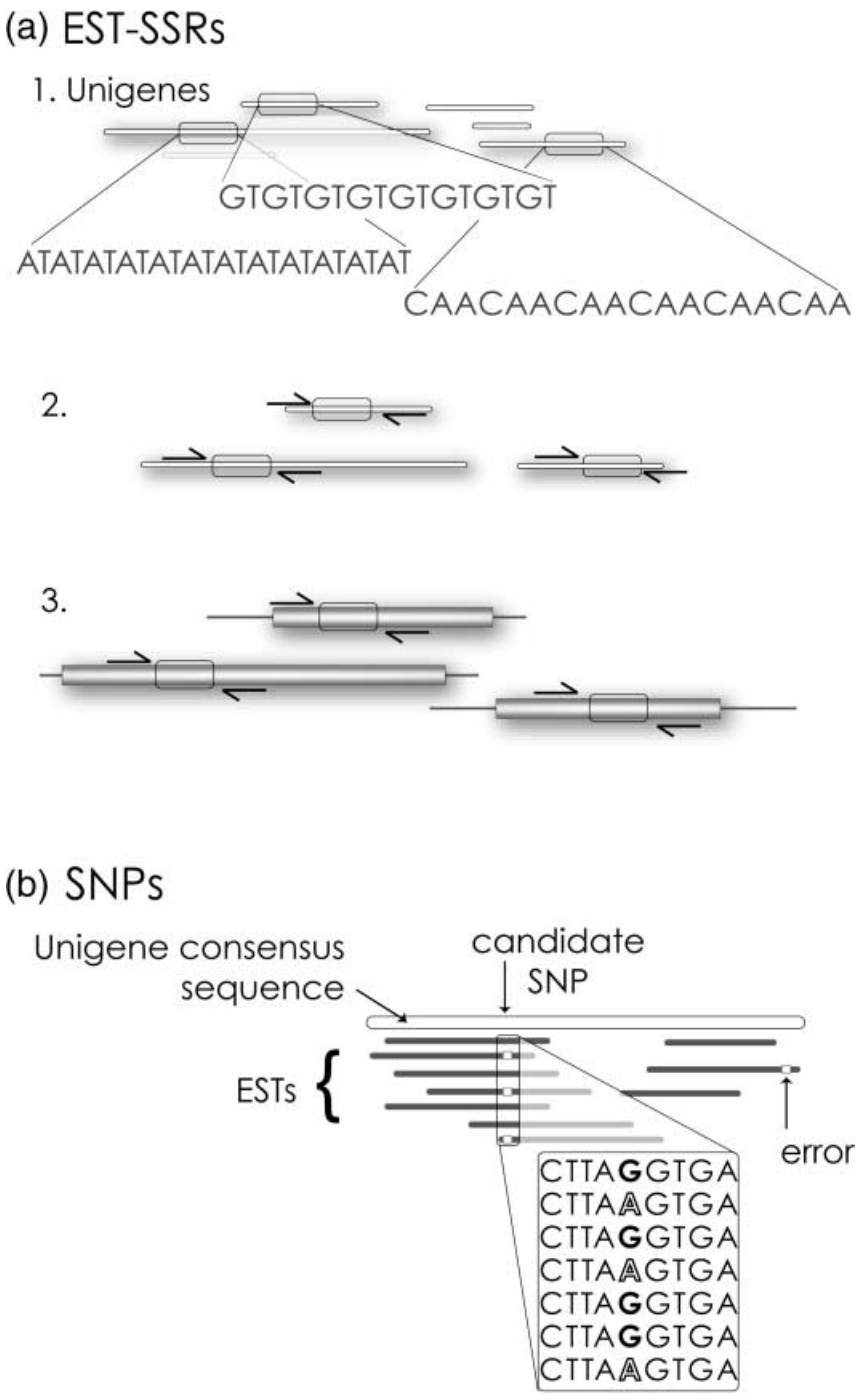


Fig. 3 Examples of molecular markers developed from ESTs. (a) EST-SSRs. Software is used to scan batches of unigenes for SSRs (1). Primers (horizontal arrows) are designed from unigene sequences flanking SSRs (2), which are then used for genotyping (3). (b) SNPs are identified directly from alignments of ESTs sequenced from different alleles, based on the occurrence of the same base call discrepancy in multiple sequences. Discrepancies occurring only once are likely to be sequencing errors. (c) EPIC markers. Unigene sequences from a species of interest (1) and genomic sequences of homologous genes from a related reference species (1) are aligned (2). Introns are inferred from gaps in the aligned unigene sequence. Note that unigenes may often be truncated relative to the complete coding sequence of the gene. Primers (horizontal arrows) are designed to flank the predicted intron positions (3) and then used for genotyping (4).

can obviate the need for the costly and time-consuming benchwork required of traditional approaches, such as library construction, enrichment and screening (Fig. 3a). It is feasible for a researcher to download a collection of sequences, identify SSRs within them, and order primers all within the space of single day.

In addition to requiring less time and money to develop, EST-derived simple sequence repeat markers (EST-SSRs) have a number of advantages over SSR markers developed by cloning and sequencing. Studies in plants, animals, and fungi have shown that EST-SSRs tend to be more widely transferable between species, and even genera, than those designed from laboratory methods such as sequencing from SSR-enriched genomic libraries (Besnard *et al*. 2003; Chagne *et al*. 2004; Boches *et al*. 2005; Coulibaly *et al*. 2005; Fraser *et al*. 2005; Ng *et al*. 2005). This may be because

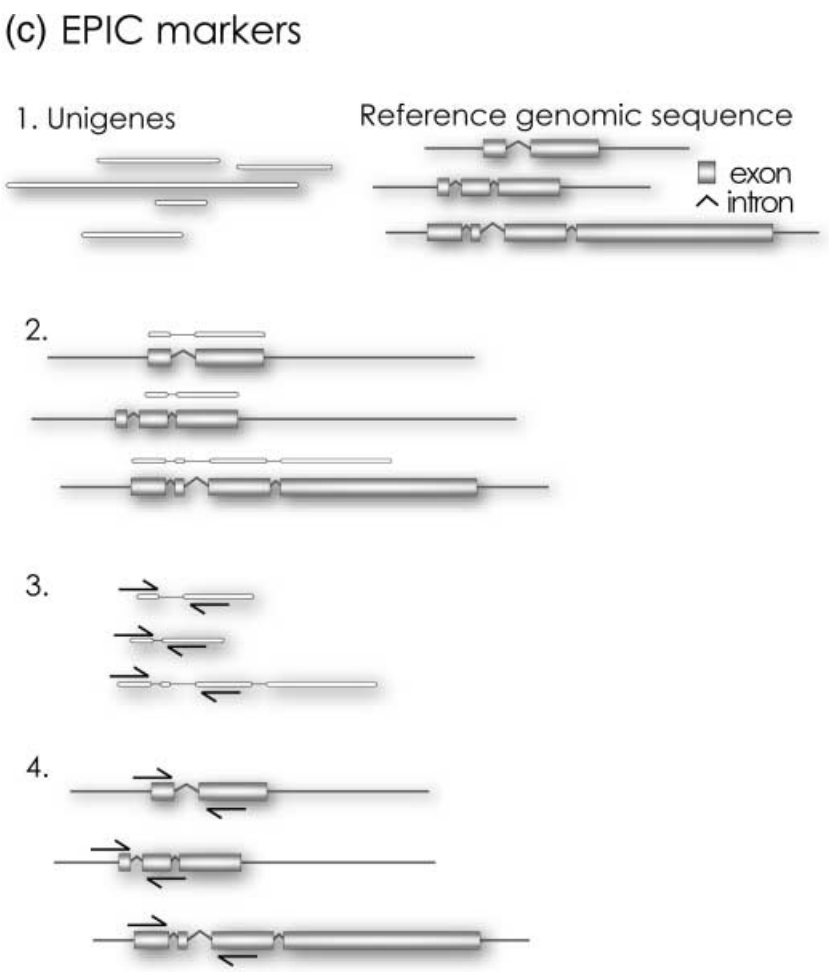


Fig. 3 *Continued*

EST-SSRs are more likely to be in gene-rich euchromatic regions of chromosomes than those developed by screening of genomic libraries (Areshchenkova & Ganal 2002), although this may not be the case in all genomes (e.g. La Rota *et al*. 2005). The high intertaxon transferability of EST-SSRs means that even if a particular organism has no EST sequence resources available, sequences from a related species can often be used for SSR development (Cordeiro *et al*. 2001; Woodhead *et al*. 2003; Barkley *et al*. 2005; Varshney *et al*. 2005). EST-SSRs are typically composed of trinucleotide repeats, which are easier to score than dinucleotide repeats (Morgante *et al*. 2002; Li *et al*. 2004). Another advantage of EST-SSRs lies in the fact that the corresponding EST sequence can be compared to protein sequence databases, possibly shedding light on the functional identity of a particular marker locus. Working with EST-SSRs may arguably provide a shortcut to a candidate gene, if markers can be designed around an SSR in a gene of interest (Vasemagi & Primmer 2005). Candidate genes may also be identified by conducting genome scans of EST-SSRs (e.g. Vasemagi *et al*. 2005): in such studies, the gene affected by selection is likely to be very close to the EST-SSR marker, perhaps the source of the EST-SSR itself. The chief disadvantage of EST-SSRs is that they do tend to show a lower rate of polymorphism (in terms of allelic richness) than those derived from genomic libraries (Eujayl *et al*. 2002; Woodhead *et al*. 2003; Chagne *et al*. 2004; Chabane *et al*. 2005).

EST sequences have been used to examine the distribution of SSRs within coding and noncoding regions of transcribed sequences, providing insights useful for those interested in using SSRs as markers. Analysis of ESTs and genomic sequence has shown that SSRs are more common in transcribed sequences (ESTs) than in nontranscribed genomic regions, are most frequent in the 5′ UTRs of genes, and that the number of repeats per SSR and the total lengths of SSRs (i.e. the number of repeats) are generally shorter in transcribed vs. nontranscribed genomic regions (Kantety *et al*. 2002; Morgante *et al*. 2002; Thiel *et al*. 2003; Li *et al*. 2004). Researchers could use this information in the design of SSR markers, by, for example, targeting SSR markers subject to varying degrees of selective constraint

based on whether the SSR is located within a coding, transcribed but untranslated, or untranscribed region of a gene, or in an intergenic region of the genome. However, it seems plausible that EST-SSRs may be generally more subject to selective constraints than SSRs in nongenic genomic regions. For example, changes in the length of SSRs in gene regulatory regions may affect the binding of transcription factors (Martin *et al*. 2005). Changes in the lengths of SSRs located in introns may also affect levels of gene expression (Chung *et al*. 2006).

SNPs

Single nucleotide polymorphisms (SNPs), though individually less polymorphic than SSRs, are even more abundant, allow for clear identification of alternate alleles, and lend themselves to highly automated genotyping. As a result, there is increasing interest in developing this class of markers for use in molecular ecology. SNPs can be identified directly from alignments of ESTs sequenced from different alleles (Picoult-Newberg *et al*. 1999; Batley *et al*. 2003), and several software programs have been designed for this purpose (e.g. Picoult-Newberg *et al*. 1999; Barker *et al*. 2003; Batley *et al*. 2003; Kota *et al*. 2003; Savage *et al*. 2005) (Fig. 3b). These programs identify SNPs where the same base call discrepancy occurs in multiple EST sequences, based on the premise that redundant discrepancies represent actual SNPs rather than simply sequencing errors. Some programs use the raw sequencing trace files as input, which has the advantage of allowing SNP identification to take the quality of each base call into account (Chevreux *et al*. 2004; Le Dantec *et al*. 2004). EST collections made from outcrossed (heterozygous) individuals, multiple genotypes, or hybrid genotypes are typically used for this kind of *in silico* SNP detection (Batley *et al*. 2003; Lai *et al*. 2005).

Even if no sequence polymorphisms are present in the EST collection for a particular transcript, EST sequences can be used to develop primers for SNP detection using denaturing gradient electrophoresis (DDGE, Sheffield *et al*. 1989) or other heteroduplex discrimination techniques [e.g. single strand conformation polymorphism (SSCP), Orita *et al*. 1989; or denaturing high-performance liquid chromatography (DHPLC), Oefner & Underhill 1995]. For instance, DDGE scoring of SNPs within EST-derived markers has recently been used to develop a comparative linkage map for oak and chestnut tree species (Casasoli *et al*. 2006). SNPs can also be assayed in noncoding sequences flanking genes, using a marker system that combines a restriction site-ligated linker primer [as in amplified fragment length polymorphisms (AFLPs), Vos *et al*. 1995] and another primer anchored within the EST (Cato *et al*. 2001). Even when EST sequences are not available for a given organism, EST collections from related species can serve as the basis for designing SNP-detection assays of species and populations of interest, an approach recently applied to issues of stock composition and migration detection in salmonid fish (Smith *et al*. 2005a, b).

SNPs discovered through analysis of ESTs can be used for the application of high-throughput SNP genotyping methods (reviewed in Tsuchihashi & Dracopoli 2002), such as pyrosequencing (Ronaghi *et al*. 1996; Ronaghi *et al*. 1998). These methods can be performed at a cost within the reach of many molecular ecology research programs. EST data alone have also been used to create gene chip assays that combine high-throughput discovery and genotyping of SNPs, as demonstrated by recent work in barley (Rostoks *et al*. 2005). The development of such gene chips is more costly and requires an extensive amount of EST sequence data. Although technological and analytical challenges remain in discriminating signal from noise in chip genotyping assays, which rely on allele-specific differences in hybridization to the probes on the gene chip, the sheer density of markers makes some applications [such as linkage and quantitative trait loci (QTL) mapping] fairly robust to the noise (e.g. Wolyn *et al*. 2004).

Identifying orthologous genes across taxa: extending the reach of model systems

One promise of genomics is that the wealth of functional genetic information developed in model systems can be extended to distantly related organisms. For example, the gene *SLC24A5* has a functionally conserved role in pigmentation in zebrafish and humans, lineages separated by 400 million years of evolution (Lamason *et al*. 2005). The ability to apply functional genetics from model organisms to distantly related, ecologically interesting lineages rests on the ability to identify homologous genes and/or chromosomal regions between them. This allows the identification of, for example, candidate genes in a model organism based on positional homology to the location of QTL in a related wild species (e.g. as proposed by Doust & Kellogg 2006), which might help researchers prioritize the hundreds to thousands of genes typically contained in QTL regions. Identification of homologous genomic regions requires the availability of markers for multiple conserved loci.

EST data can be used in a number of different ways to produce markers useful for comparative mapping. For example, EST data from different species have been combined to design cross-species EST-SSRs (Kantety *et al*. 2002; Rexroad *et al*. 2005) and software has been developed for designing primers in conserved regions of cross-species DNA alignments (Jarman 2004; Gadberry *et al*. 2005). EST data have also been used specifically to create orthologous anchor loci for comparative mapping in both animals and plants. Lyons *et al*. (1997) analysed EST sequence data from

several mammal species in order to create markers for single copy genes that are highly conserved across mammals. Lyons *et al*. called their collection of orthologous markers Comparative Anchor Tagged Sequences (CATS). Fulton *et al*. (2002) used a collection of EST sequences from tomato to develop conserved orthologue set (COS) markers for use in comparative mapping across flowering plants. Tomato ESTs were aligned to the *Arabidopsis* genome sequence in order to identify over 1000 genes that appeared to be conserved in sequence and low in copy number, and markers were designed from these. Because tomato and *Arabidopsis* represent two lineages that diverged at the base of the core eudicot clade, it is intended that by screening genes in this way, COS markers will be conserved in sequence and low in copy number across the thousands of plant species within the core eudicots. A similar approach has proven extremely successful in developing anchor loci for comparative mapping within the grass family (Van Deynze *et al*. 1998).

Exon-primed, intron-spanning markers

Comparative analysis of EST data has also been used to design genetic markers based on intron polymorphisms. Because they are noncoding, not only SNPs but also insertion/deletion (indel) polymorphisms are more common within introns than in coding exons. This property makes introns an attractive target for molecular marker design. In exon-priming, intron-crossing markers (EPICs), primers are designed to anneal in two different conserved exons will amplify the more variable intron(s) in between, and polymorphisms in intron length can be resolved by fragment analysis (Lessa 1992; Cortereal *et al*. 1994; Palumbi & Baker 1994; Wydner *et al*. 1994) (Fig. 3c). The paradox in using EST data to develop such markers lies in the fact that introns are spliced out of mature mRNA molecules — the source of EST sequences. However, the positions of introns within genes tend to be highly conserved (e.g. Ku *et al*. 2000). This means that intron positions can be inferred by aligning EST sequences to homologous genes or proteins in which the intron positions are known (such as the genomic sequence of a related species) (Fig. 3c). The success of this approach may be limited by the evolutionary divergence between the species of interest and the genomic sequence used as a reference.

This approach has been successfully applied to marker design in a variety of systems. For example, Bierne *et al*. (2000) compared penaeid shrimp ESTs to *Drosophila* genomic sequences in order to identify intron/exon boundaries and design EPIC markers that assay intron length polymorphism in shrimp populations. We (Vision, Fishman and Willis, unpublished, in preparation) have designed EPIC markers for use in *Mimulus* species, using *Arabidopsis* proteins to predict intron/exon boundaries in *Mimulus guttatus* unigenes. This has so far resulted in the design of over 800 EPIC markers, over 95% of which contain the predicted intron.

As is the case for EST-SSR markers, even if no EST resources exist for a species of interest, ESTs from related taxa can be used to develop EPIC markers, although this process can be long and tedious. EST sequences from taxa that phylogenetically circumscribe the species of interest can be aligned in order to identify conserved coding regions within closely related homologues and a more distantly related genomic sequence can be used to identify intron/exon boundaries. Degenerate primers designed from the close homologues can then be used to amplify the target sequence from the organism under study. The sequences obtained may then be cloned and sequenced so that species-specific, EPIC primers can be developed. This approach has been used successfully in the development of molecular markers for phylogeographical studies of the plant *Platystemon californicus* (Papaveraceae). The Phytome database (Hartmann *et al*. 2006) was used to identify gene families represented both in the EST collection of a related species (*Eschscholzia californicus*) and in the annotated gene set of *Arabidopsis*, so that degenerate EPIC primers could be developed to the target locus in *Platystemon* (Poslusny & Vision, in preparation). Of the 17 small gene families examined, one or more pairs of degenerate primers could be designed for 13 of them, and the target locus was successfully amplified for five *Platystemon* genes.

Studies of gene expression in molecular ecology

Genome-wide studies of gene expression hold great potential for shedding light on complex ecological phenomena such as phenotypic plasticity, host shifts, and the evolution of specialized life histories (reviewed in Gibson 2002). Such studies can be used to identify candidate genes underlying phenotypic differentiation, and they also provide a genome-wide means of studying the genetic basis of the mechanisms by which organisms respond to environmental cues. EST collections can be used to construct microarrays (Schena *et al*. 1995) for gene expression studies of organisms that otherwise lack any genomic resources (Chen *et al*. 2004). Microarrays can be fabricated using a variety of technologies, all of which essentially consist of an array of DNA probes. Since only complementary mRNA will specifically bind to each probe, microarrays can be used to detect the up- or down-regulation of specific mRNAs in contrasting biological samples. Spotted microarrays are made by affixing long oligonucleotide or cDNA (see Glossary) probes to glass plates. Arrays in which oligonucleotide probes are synthesized directly on the chip (e.g. Affymetrix or Agilent) have higher probe densities, so typically contain multiple DNA probes for each target mRNA and can provide absolute measures of mRNA abundance in an experimental sample. Both types of

microarrays provide a means of assaying the expression of thousands of genes in a single, highly parallel experiment.

Research is currently underway in developing microarrays for use across multiple, related species (e.g. Rise *et al*. 2004; Gilad *et al*. 2006). Comparative microarray analysis of closely related species can be used to identify differences in gene expression that correlate with ecological differentiation of species, populations, or genotypes. For example, *Arabidopsis halleri* is a close relative of the model plant species *Arabidopsis thaliana* that is adapted to live in heavy metal-contaminated environments such as serpentine soils. Weber *et al*. (2004) investigated the genes underlying the metal hyper-accumulation abilities of *A. halleri* using an oligonucleotide Affimetrix gene chip designed for *A. thaliana*. Although the microarray technology used in this study was designed using the considerable genomic resources available for *A. thaliana*, similar studies of the genetics of adaptation can be carried out using microarrays designed from EST collections alone. Kobayashi *et al*. (2006) used a microarray designed from an EST collection to identify genes differentially expressed during jaw development in closely related but morphologically and ecologically distinct species of cichlid fish. Le Quere *et al*. (2004) used a microarray to tie variation in gene expression patterns among strains of ectomycorrhizal fungi to variation in host specificity, and to subsequently clone and sequence the genes identified. Oleksiak *et al*. (2002) used a microarray to examine differences in gene expression between populations of *Fundulus* (killifish). Other researchers have used microarrays to study how gene expression patterns respond to environmental cues. Carsten *et al*. (2005) used a cDNA microarray to investigate how gene expression changes in response to diet in *Drosophila melanogaster*. Evans & Wheeler (2000) used microarrays to identify genes underlying the polyphenism of different honeybee castes, which is driven by differences in the larval environment.

Expression studies can also be accomplished by directly comparing EST collections created from contrasting biological samples, such as organisms exposed to contrasting environmental conditions. In such studies, ecological and evolutionary questions direct the development of EST collections. For example, Torres *et al*. (2005) created an EST collection enriched in gene transcripts associated with plant parasitism, specifically the development of haustoria — the structures that invade the roots of host plants. Microarrays may be used in the process of designing EST collections of ecological or evolutionary interest. For instance, homoploid and allopolyploid hybrid species in the plant genus *Senecio* bear striking differences in floral morphology. Hegarty *et al*. (2005) used anonymous cDNAs (i.e. cDNAs that had not been sequenced or analysed) to create a microarray that was then used to compare gene expression in hybrid and nonhybrid lineages differing in ploidy and floral morphology. Only those cDNAs that showed changes in expression patterns were sequenced, producing an EST collection enriched for transcripts of genes affected by speciation and polyploidy.

There are several challenges inherent to microarray studies, in addition to the expense (Gibson 2002). Results of these experiments can be highly variable from one biological replicate to another. Only transcripts with sufficiently high expression can even be detected. Accurately measuring expression of genes in large gene families can be difficult if there is cross-hybridization among paralogues, leading to a confounding effect between the number of hybridized transcripts and their actual proportion in the transcriptome. This problem can be addressed by carefully designing probes to discriminate between paralogues (Chen *et al*. 2004), and by consulting an EST collection enriched in 3′-ESTs, which allow for better discrimination between paralogues (Rise *et al*. 2004). Cross hybridization is of special concern when attempting to extend the use of a microarray designed for one species to other, related species (Gilad *et al*. 2005), although in some cases a single microarray has been successfully used to measure transcription in several species (Moore *et al*. 2005). In addition, allelic variation may affect the binding of mRNAs to cDNA probes, confounding the differences in signal intensity caused by differences in expression. This can be overcome by experimental design, or by designing the array to include cDNA probes from multiple genotypes. Finally, microarray experiments require careful design and extensive statistical analysis of the resulting data (see Allison *et al*. 2006 for an overview). Despite these issues, microarray experiments seem unique in their potential to shed light on the functional genetics of ecologically and evolutionarily relevant traits which, in many cases, can only be studied in nonmodel organisms.

Selection screens

Ecologically important traits are likely to have been shaped by natural selection, which should be reflected in the pattern of molecular evolution of genes underlying these traits. This means that genes involved in adaptation can be detected by looking for the molecular signature of selection. Examining patterns of molecular evolution in EST collections provide a way of screening numerous genetic loci for signatures of selection in parallel.

Genes that have evolved under strong positive selection would be expected to have an unusually high ratio of fixed amino acid replacements per replacement site to fixed synonymous differences per synonymous site (K_A:K_S) at orthologous genes between recently diverged species (Yang & Bielawski 2000; Nielsen 2001; Yang 2002). Several studies have applied this approach to EST data in order to identify genes that have apparently evolved under strong positive selection, beginning with Endo *et al*. (1996). Swanson

et al. (2001, 2004) identified male reproductive proteins and female reproductive tract genes that appeared to have rapidly evolved under positive selection in *Drosophila* species. Tiffin & Hahn (2002) compared EST sequences from *Brassica rapa* to genomic sequences of *A. thaliana*, and although they did not identify any genes that appeared to have diverged under positive selection, they found evidence for a shift in codon bias since the divergence of these two lineages. In a similar study, Barrier *et al*. (2003) identified 14 genes potentially involved in adaptive divergence of *Arabidopsis lyrata* and *A. thaliana*.

These studies do suffer from a number of caveats, including the assumptions underlying models of sequence evolution at synonymous and nonsynonymous sites (McDonald & Kreitman 1991; reviewed in Li 1997). K_A:K_S ratios can also be shaped by selection on synonymous mutations if certain codons are favoured over others (codon bias), purifying selection which may decrease the rate of nonsynonymous substitution, and balancing selection which may increase the nucleotide diversity at silent sites. Verifying that particular loci truly have been the target of a particular form of selection typically requires further investigation. Even if the action of selection has been detected, the agent of selection is unknown (Vasemagi & Primmer 2005; MacCallum & Hill 2006), it may be difficult to discern which trait or traits are affected by a locus of interest, and the locus may no longer be under selection in contemporary populations (Garrigan & Hedrick 2003). In addition, the determination of appropriate statistical thresholds for broad scale selection screens is problematic (Tiffin & Hahn 2002). Despite these limitations, such screens for selected genes show great promise as a means of identifying loci involved in ecological adaptations that would be missed using other approaches. In particular, the loci detected by a selection screen would often be missed by QTL mapping, which is ineffective at identifying loci underlying fixed differences between species that cannot be crossed.

Analysis of candidate genes: pinpointing the basis of ecologically important traits

In recent years, genes identified through functional genetic studies in model organisms have been shown to underlie ecologically important variation in natural populations of distantly related wild species. Examples include adaptive variation in coat colour in rock pocket mice associated with allelic variation at the *Mc1r* gene (characterized through mutant studies of laboratory mice) (Nachman *et al*. 2003), and adaptive morphological variation among natural populations of sticklebacks caused by differences in expression of the *Pitx1* gene (identified through mutant studies in mice and chickens) (Colosimo *et al*. 2005). The flowering time genes *Frigida* and *FLC* in *Arabidopsis* have been shown to be associated with clinal variation in flowering time in natural populations of *Arabidopsis* (Caicedo *et al*. 2004; Stinchcombe *et al*. 2004). The above studies had significant genome resources and/or a large body of functional genetic studies to draw upon. There is a need for the development of research programs tailored to the issues inherent in studying candidate genes in wild organisms without extensive resources, and EST data are likely to contribute to this. For example, EST sequences can provide a means of readily identifying candidate gene homologues in wild species, allowing researchers to direct genetic studies of adaptation in natural populations toward these loci.

Such an approach was used in a study of the genetic basis of salt tolerance in the sunflower *Helianthus paradoxus* (Lexer *et al*. 2003). These authors screened EST libraries from *Helianthus* for sequences with apparent homology to candidate genes for salt tolerance that had been identified in other plants. Three of these candidate genes were found to be associated with fitness in the high-salt habitat, and one mapped to a previously identified QTL for salt tolerance. However, a conclusive demonstration that one or more substitutions at a candidate gene underlie a specific adaptation would require additional evidence. Ideally, introgression, transgenic or knockout lines (Lee & Mitchell-Olds 2006) or deficiency mapping approaches (Pasyukova *et al*. 2000) would be used to demonstrate the phenotypic effects of allelic variants. In wild organisms that are not amenable to these kinds of genetic techniques without expending decades of time and effort, perhaps correlative evidence (such as an observed correspondence between a particular genotype and a phenotype) supporting the causal role of particular loci in ecological adaptation will have to suffice.

Adaptation and the evolution of gene families and genomes

Proliferation of gene families, followed by functional diversification of paralogues, has been postulated to underlie the acquisition of new biological functions (Ohno 1970). Large-scale gene duplication events, in which large chromosomal segments, whole chromosomes or whole genomes are duplicated, have been hypothesized to correspond to major evolutionary transitions in both animals and plants (reviewed in Van de Peer 2004; De Bodt *et al*. 2005). In other words, in some cases the genetic basis of adaptation may prove to be more complicated than changes in protein sequences or expression patterns. Instead, it seems feasible that the evolution of diverse, complex ecological functions may involve the duplication of gene networks followed by selective co-option of new genes for novel or refined functions (Conant & Wolfe 2006). EST collections have been used to identify genes and gene family expansions that are unique to certain lineages

(e.g. Albert *et al*. 2005; Laitinen *et al*. 2005). EST data can also be used to detect whole or large-scale genome duplications in a species' evolutionary history (Blanc & Wolfe 2004; Van de Peer 2004; Sterck *et al*. 2005).

Microbial ecologists have produced an extensive body of work documenting how the acquisition of new genes or other changes in gene content underlie adaptations to extreme environments or new functional capabilities in prokaryotes and Archaea (e.g. Snel *et al*. 2002; Omelchenko *et al*. 2005; see Xu 2006 for a review). These studies typically examine whole genome sequences or environmental DNA samples (e.g. 'environmental genome shotgun sequencing', Venter *et al*. 2004; Strous *et al*. 2006), for a variety of reasons: microbial genomes are relatively small and inexpensive to sequence, the application of EST approaches are precluded by the fact that many microbes cannot be isolated in culture, and prokaryotic mRNAs are not easily distinguishable from other RNAs.

Ecologists and evolutionary biologists working with eukaryotes have also begun to study differences in gene repertoire between different lineages of animals, plants and fungi, and examine how gene content is shaped by natural selection (Lespinet *et al*. 2002; Hahn *et al*. 2005; Barbosa-Morais *et al*. 2006). Lespinet *et al*. (2002) compared the gene content of yeast, *Caenorhabditis elegans*, *Drosophila* and *Arabidopsis* and found that genes involved in pathogen and environmental stress responses were particularly likely to have undergone lineage-specific gene family expansions. These authors posited that gene family expansions may be a principal means by which organisms evolve new or refined patterns of gene regulation and undergo adaptation. Working on a smaller evolutionary timescale, Hahn *et al*. (2005) examined lineage specific gene family expansions among yeast species, and found significant lineage-specific expansion of the flocculin gene family in the brewer's yeast, *Saccharomyces cerevisiae*, suggesting that this may be due to selection on that species for industrial purposes. While these two studies took advantage of complete genome sequence data, similar patterns have been observed through analysis of EST collections. Analysis of *Gerbera* ESTs identified over 1000 unique gene transcripts that do not show homology to any other plant genes represented in available databases, perhaps representing genes and gene families unique to this lineage of plants (Laitinen *et al*. 2005). The ongoing floral genome project (Albert *et al*. 2005) is designed to examine patterns of gene content and gene family evolution and its relationship to floral development and morphological evolution by comparing ESTs sampled from diverse families of flowering plants. EST data have also been used to detect whole or large-scale genome duplications in a species' evolutionary history (e.g. Blanc & Wolfe 2004; Sterck *et al*. 2005; Cui *et al*. 2006). Determining whether such events have occurred and when they occurred can be important for accurate characterization of gene family expansions, and for predicting differences in gene content between model organisms and nonmodel species (Sampedro *et al*. 2005; Durand & Hoberman 2006).

There are a number of caveats to consider in conducting studies of gene family and genome evolution with EST data alone. Even the largest EST collections fall short of being complete gene sets: the absence of a sequence in an EST collection does not mean that that gene is not in the genome. Also, it is not yet clear how causal relationships between gene content and patterns of ecological or phenotypic variation will be established. And, accurately identifying and characterizing genome duplications can be problematic in the absence of positional information, without which homologous relationships between chromosomal regions (frequently referred to as synteny) cannot be conclusively established. Despite these limitations, comparisons of gene content may be particularly interesting if applied to organisms with diverse ecologies. Such studies may provide the best approach for understanding the extent to which the evolution of gene content contributes to ecological diversification.

The design of EST collections for molecular ecology

We predict that EST data will be increasingly used by molecular ecologists as a means of incorporating genomics approaches into studies of diverse species and populations. Molecular ecologists will be both using available data and directing the development of EST collections specifically for ecological research. There are a number of issues and concerns regarding the design of EST collections that are unique to nonmodel organisms, which we discuss below. In addition, we cover some basics in EST data handling, including a brief overview of sequence processing and unigene clustering, which are helpful for understanding the limitations of these resources.

The design of an EST collection involves numerous considerations, including the total number of sequences to collect, the genotypes, tissues, and life stages to be sampled, and whether any special cDNA library construction methods or sequencing strategies are to be employed (see Glossary). One of the first issues to consider is the number of sequences needed for a specific application. A few hundred to a few thousand sequences may provide sufficient data for marker design, while a few hundred thousand sequences are typically required to annotate a genome. For example, 3977 *Mytilus* ESTs, clustered into 544 unigenes, yielded 75 putative EST-SSR primer pairs. Analysis of 17 000 unigenes from the aphid *Acyrthosiphon pisum* resulted in 1641 putative EST-SSR primer pairs, excluding dinucleotide repeats (Bouck, unpublished). At the other extreme, approximately 60 000 Atlantic salmon ESTs

clustered into 29 000 unigenes (Rise *et al*. 2004). In populus, 102 019 ESTs, assembled into approximately 25 000 unigenes, are being used to annotate the populus genome sequence (Sterky *et al*. 2004). For budgeting purposes, methods exist for determining a point of diminishing returns when the sequencing of additional transcripts from a library is unlikely to capture as-yet-unsampled gene transcripts (Susko & Roger 2004; Wang *et al*. 2005).

An issue of particular importance to molecular ecologists is which genotypes or populations to sample for EST sequencing. Model organisms used for genome and EST sequencing are usually inbred lines, and EST collections created from them will have little or no allelic variation. In such systems, any discrepancy in sequence between different reads is likely to be due to a sequencing error, rather than a true polymorphism. In wild species, on the other hand, producing inbred lines may be difficult or impossible, and may be undesirable depending on the purpose of the EST project. For example, if ESTs are to be used to identify SNPs directly, sampling from heterozygotes, F_1 hybrids, or different genotypes, populations or species of interest would maximize the likelihood of capturing allelic variation (Fig. 3b) (Lai *et al*. 2005; Smith *et al*. 2005a). However, genetically heterozygous EST collections might be difficult to cluster into unigene sets. Particularly when allelic variation is of the same magnitude as differentiation among recent or highly conserved paralogous genes, it may not be possible to distinguish the two possibilities with confidence. The presence of genetic variation might also complicate applications of ESTs for expression profiling, if genetic variation in expression is confounded with variation between treatments or tissues. At the same time, including samples from multiple genotypes could arguably provide a means of capturing some of the variation in expression among individuals. For example, if genotypes or ecotypes vary in the diversity of transcripts expressed under different conditions, then generating EST data from several genotypes or ecotypes might capture a greater range of transcripts overall.

Issues of sampling also come into play when determining what biological samples to target for RNA isolation, and whether any special techniques of library design are to be used. ESTs may be generated from different tissues, life stages or ecological conditions specifically to allow a comparison of transcripts expressed under one condition vs. another (e.g. expression profiling analyses). Even if a researcher simply wants a general sample of 'gene space', this is still an important consideration. Only a subset of all genes will be expressed at any given life stage, tissue, or ecological condition, so sampling RNA from multiple tissues and life stages can enrich the diversity of transcripts captured in an EST collection. The relative abundance of different transcripts in a cDNA library can be standardized via the use of normalization procedures (Bonaldo *et al*. 1996), resulting in an EST collection with a more diverse sample of transcripts. However, normalization would be inappropriate for expression profiling applications because the relative abundance of different transcripts has been altered relative to the original biological sample, while in non-normalized libraries the relative abundance of different transcripts is more or less preserved. Normalization may also lead to under-representation of genes within closely related gene families. Another technique used in library construction is capping (Carninci *et al*. 1996; Seki *et al*. 1998; Carninci *et al*. 2000), which is a procedure designed to obtain full-length transcript sequences, and which might be useful in a library intended for candidate gene isolation or functional analysis of protein translations (Seki *et al*. 1998). However, this procedure is expensive, technically challenging, reduces the total number and diversity of transcripts sampled and biases the EST collection toward shorter transcripts (Carninci *et al*. 2000).

If the cDNA library has been directionally cloned, EST sequencing can be targeted to capture either the 5′ or 3′ ends of the cDNA clones. Sequencing both ends (producing so-called *mate pairs*; see Glossary) can link two unigenes that correspond to opposite ends of a long gene transcript but do not have sufficient sequence overlap to be joined without mate pair information (Fig. 4). Focusing on one end or the other might be useful for different situations. The 5′ end sequences tend to contain more of the protein-coding region of the transcripts. Because coding sequences tend to be more conserved, these sequences will be more useful for applications that require the establishment of homology to sequences from distantly related organisms, such as the development of orthologous markers, or for sequence analyses such as selection screens. Since the 3′ end of the cloned cDNA usually terminates at the position of the poly A tail while an uncapped 5′ end may terminate anywhere within the cDNA, multiple 5′ end sequences from different clones of the same gene will typically result in a longer unigene with greater transcript coverage than multiple 3′ end sequences. The 3′ ends of cDNAs tend to contain a long untranslated region that is relatively tolerant of mutation. Because of this, 3′ end sequences may provide a means of discriminating between highly similar paralogous genes, and are also generally more likely to contain more SNPs and other polymorphisms.

Computational analysis of ESTs

Whether you will be using publicly available EST data or generate ESTs yourself, it is important to understand the computational steps involved in EST processing in order to be a wise consumer (reviewed more fully in Dong *et al*. 2005). The raw trace data from the sequencer is typically processed by base-calling software which assigns a quality

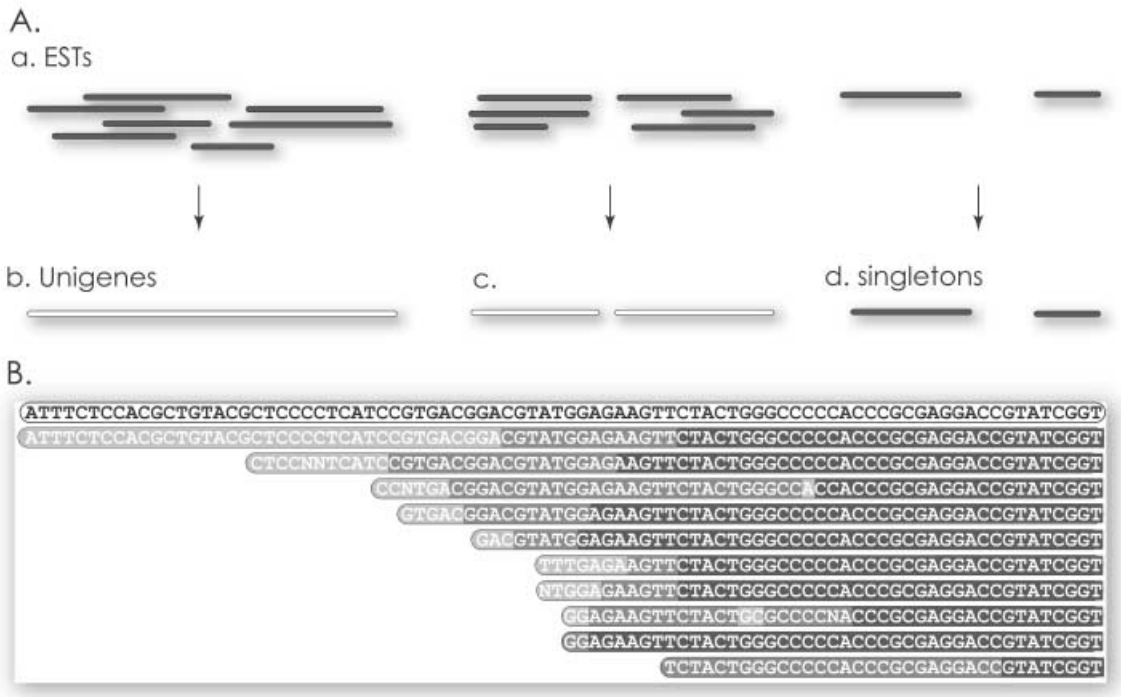


Fig. 4 (A) **Unigene assembly.** EST sequences from the same transcript (a) are clustered and assembled to produce a set of unigene consensus sequences (b and c), representing a nonredundant sample of the transcripts present in a particular EST collection. A transcript that is not fully spanned by ESTs may be represented by two or more unigenes (c). In such cases, mate pair information from paired end sequencing (see Glossary) can provide a means of linking the ends. Solo transcripts which cannot be clustered into a unigene are referred to as singletons (d). (B) **Sequence quality and unigene assembly.** A portion of a unigene consensus sequence is shown (top line: black type) along with the individual EST reads (white type) contiged to produce it. In the figure, lower quality sequence data is shown as lighter shades of grey. Quality information can be useful in identifying sequencing errors if, for example, a base call discrepancy is of lower quality.

score to each base (such as PHRED: Ewing & Green 1998; Ewing *et al*. 1998). When trace files are not available, and so base quality scores are lacking (as is the case for much of the legacy sequence data in the public domain), subsequent data processing steps are far more difficult to do well, a point we return to below. Runs of low-quality base calls, common at the ends of sequences, are trimmed. ESTs are then checked for contamination by laboratory sequences (such as vectors, adaptors or primers) by comparison against a database of potential contaminants. ESTs may also contain xenocontaminant sequences from pathogenic organisms or symbionts, or even human laboratory workers, and these are often much more difficult to detect. The stage of processing at which sequences are deposited into National Center for Biotechnology Information (NCBI) EST database, DBEST (Boguski *et al*. 1993), or other public databases varies considerably, and so one should not be surprised to find some ESTs in which low-quality sequences, repeats, etc. are still present and others from the same species in which they have been removed by the contributors. This fact becomes particularly relevant when polymorphisms are being mined from a heterogeneous collection of publicly available EST data (Fig. 3b).

For some applications, researchers will use unigenes (rather than raw ESTs; Fig. 4). The creation of a unigene set entails clustering sequences that are similar (but not necessarily identical) in some region of terminal overlap. A single consensus sequence for each unigene is then derived from the multiple sequence alignment of each cluster (e.g. Aaronson *et al*. 1996; Schuler *et al*. 1996; Burke *et al*. 1999; Liang *et al*. 2000; Parkinson *et al*. 2002; Pertea *et al*. 2003; Pontius *et al*. 2003; Chevreux *et al*. 2004; Parkinson *et al*. 2004; Malde *et al*. 2005; Murray *et al*. 2005; reviewed in Rudd 2003; Dong *et al*. 2005). The most accurate methods incorporate quality scores in the clustering and consensus sequence calculations. Information on mate pair end-reads (e.g. if cDNAs were sequenced from both ends — Fig. 1) can be used to link unigenes that are not bridged by overlapping sequence reads, but this is not universally done. Repeat motifs such as SSRs and transposable element-associated sequences are typically retained, but ignored, during some or all of the unigene assembly process

(Huang & Madan 1999; Pertea *et al*. 2003), so that they do not cause spurious clusters.

Building a unigene set is computationally intensive, and researchers will likely elect to use available unigene resources such as NCBI unigene assemblies (Pontius *et al*. 2003; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = unigene) and The Institute for Genomic Research (TIGR) gene indices (Quackenbush *et al*. 2001; www.tigr.org/tdb/tgi/index.shtml, biocomp.dfci.harvard.edu/tgi/) and plant transcript assemblies (http://plantta.tigr.org/) rather than assemble their own. It is important to recognize that these different unigene builds can differ dramatically in the composition of EST clusters and consensus sequences. Although the unigene assembly procedures tend to be similar in broad outline, the final unigene set can be highly dependent on the specific assembly algorithm and parameters. As ecologists tap into the ocean of sequence data available in public databases, they are likely to be confronted with the difficult issue of keeping track of the correspondence between unigenes from different sources and different builds.

The parameter settings used in a particular unigene assembly are typically determined by trial and error and will depend on the properties of the EST data, such as the redundancy of different transcripts, the sequencing error rate, and the heterozygosity of the sample (Rudd 2003; Dong *et al*. 2005). A good assembly represents a balance between over- and under-clustering (Wang *et al*. 2005), but any given unigene assembly is likely to contain instances of both errors due to the inherent variability among genes. If unigenes are clustered under high stringency, highly similar sequences such as alleles may be assembled into separate unigenes, resulting in an inflated number of unigenes relative to the actual number of transcripts present in the biological sample. On the other hand, clustering under low stringency may erroneously join sequences from closely related paralogues. For species with sequenced genomes, unigene assembly can be informed by the alignment of ESTs to a genomic sequence (e.g. Zhu *et al*. 2003; Dong *et al*. 2005). Alignments of ESTs to a genomic sequence (e.g. Wang & Brendel 2006) can also be used to detect alternatively spliced transcripts, which are known to cause complications in unigene clustering (Dong *et al*. 2005). In some cases, such an analysis may be possible using genomic sequence from a closely related species (Kan *et al*. 2004). In the absence of genomic sequence information, alternative splice forms may also be detected by ESTs alignments alone, although not all types of alternative splicing can reliably be detected using this approach (Dong *et al*. 2005).

Once a unigene set is available, it is common practice to computationally annotate and predict the function of each unigene. The advent of large databases of functionally characterized conserved sequence motifs (e.g. PFAM: Bateman *et al*. 2004), sophisticated software for identifying such motifs (e.g. INTERPROSCAN: Zdobnov & Apweiler 2001) and controlled vocabularies for describing gene function (e.g. GO or gene ontology; The Gene Ontology Consortium 2000) have greatly advanced the reliability of gene functional predictions based on primary sequence. But caution is still warranted when interpreting the functional assignments of ESTs or unigenes that are provided by many public databases, which are in some cases still produced by fairly crude approaches (such as reporting the identity of the top BLAST hit in GenBank), thus potentially propagating inaccurate functional assignments from one sequence to another.

ESTs as a community resource

The vast number of EST sequences already available in the public domain are widely used by the scientific community. The EST database at NCBI (DBEST) currently contains over 29 million nonhuman EST sequences. Nonetheless, these data are likely to be dwarfed by ESTs yet to come. Future sequences will increasingly be derived from a phylogenetically diverse constellation of multiple closely related species, from different populations and genotypes of the same species, and generally from samples that require more documentation than is necessary for standard laboratory strains of model organisms. Documenting the provenance of EST sequences will be critically important to ensure the future utility of ecologically motivated EST collections. For starters, such metadata will need to include the genotype, subspecies or ecotype and georeferenced source population of the sample (including a voucher specimen, if available), whether the sample was inbred or crossed in the laboratory, the tissue and life stage, and the environmental conditions (in nature or the laboratory) under which the organisms were reared and collected. As suggested by Graham *et al*. (2004), merging the type of information provided by natural history collections with EST and other sequence databases will vastly increase the value of both types of resources.

Base-calling technology continues to improve, and so quality scores are not fixed entities. Because EST sequences continue to have productive lives after they are deposited, and because the base-call quality scores are so important for their reusability, every EST should ideally be associated with its original trace file. This is now considerably easier than it once was, due to the advent of the NCBI Trace Archive (www.ncbi.nlm.nih.gov/Traces/trace.cgi) and Ensembl Trace Server (race.ensembl.org/).

Conclusion

Genomics has had far-reaching impacts on numerous fields in biology, including ecology, evolution and

population biology. Technological advances will make the development of additional EST collections or other sequence-based data sets increasingly inexpensive. New sequencing methods such as 454 sequencing (Margulies *et al*. 2005) entirely circumvent the library construction and cloning process, and this method may be extendable to EST sequencing in the future. As a result, opportunities to extend genomic approaches to the vast diversity of organisms and populations that are studied by ecologists will only increase, and we predict that ESTs are likely to play a significant role in these endeavors. Genomics approaches that make use of ESTs, such as expression studies, selection screens, and analyses of gene family and genome evolution may prove integral to the study of complex ecological phenomena, the genetic basis of which are likely to be equally complex, and impossible to study using model organisms and/or classical single gene, mutant screen functional studies.

Acknowledgements

The authors wish to thank John Willis and members of the Willis laboratory, Stefanie Hartmann, Cynthia Riginos, Robin Hopkins, Mario Vallejo-Marin, and Gina Baucom for useful discussion and comments, as well as the comments of two anonymous reviewers. Amy Bouck was supported by National Science Foundation Postdoctoral Fellowship in Biological Informatics DBI-0434666 and Todd Vision was supported by National Science Foundation grants DBI-0227314 and EF-0328636.

References

Aaronson JS, Eckman B, Blevins RA *et al*. (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Research*, **6**, 829–845.

Albert V, Soltis D, Carlson J *et al*. (2005) Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biology*, **5**, 5–5.

Allison DB, Cui XQ, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**, 55–65.

Areshchenkova T, Ganal MW (2002) Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. *Theoretical and Applied Genetics*, **104**, 229–235.

Avise C (2004) *Molecular Markers, Natural History, and Evolution*. Sinauer Associates, Sunderland, Massachusetts.

Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Research*, **16**, 66–77.

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.

Barkley NA, Newman ML, Wang ML, Hotchkiss MW, Pederson GA (2005) Assessment of the genetic diversity and phylogenetic relationships of a temperate bamboo collection by using transferred EST-SSR markers. *Genome*, **48**, 731–737.

Barrier M, Bustamante CD, Yu JY, Purugganan MD (2003) Selection on rapidly evolving proteins in the *Arabidopsis* genome. *Genetics*, **163**, 723–733.

Bateman A, Coin L, Durbin R *et al*. (2004) The PFAM protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology*, **132**, 84–91.

Besnard G, Achere V, Rampant PF, Favre JM, Jeandroz S (2003) A set of cross-species amplifying microsatellite markers developed from DNA sequence databanks in Picea (Pinaceae). *Molecular Ecology Notes*, 380–383.

Bhat PR, Krishnakumar V, Hendre PS, Rajendrakumar P, Varshney RK, Aggarwal RK (2005) Identification and characterization of expressed sequence tags-derived simple sequence repeats, markers from robusta coffee variety 'C × R' (an interspecific hybrid of *Coffea canephora* × *Coffea congensis*). *Molecular Ecology Notes*, 80–83.

Bierne N, Lehnert A, Bedier E, Bonhomme F, Moore SS (2000) Screening for intron-length polymorphisms in penaeid shrimps using exon-primed intron-crossing (EPIC)-PCR. *Molecular Ecology*, 233–235.

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667–1678.

Boches PS, Bassil NV, Rowland LJ (2005) Microsatellite markers for *Vaccinium* from EST and genomic libraries. *Molecular Ecology Notes*, 657–660.

Boguski MS, Lowe TMJ, Tolstoshev CM (1993) DBEST: database for expressed sequence tags. *Nature Genetics*, 332–333.

Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, **6**, 791–806.

Burke J, Davison D, Hide W (1999) D2_CLUSTER: a validated method for clustering EST and full-length cDNA sequences. *Genome Research*, **11**, 1135–1142.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between *Arabidopsis* FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences, USA*, **101**, 15670–15675.

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

Carninci P, Shibata Y, Hayatsu N *et al*. (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Research*, **10**, 1617–1630.

Carsten LD, Watts T, Markow TA (2005) Gene expression patterns accompanying a dietary shift in *Drosophila melanogaster*. *Molecular Ecology*, **14**, 3203–3208.

Casasoli M, Derory J, Morera-Dutrey C *et al*. (2006) Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. *Genetics*, **172**, 533–546.

Cato SA, Gardner RC, Kent J, Richardson TE (2001) A rapid PCR-based method for genetically mapping ESTs. *Theoretical and Applied Genetics*, **102**, 296–306.

Chabane K, Ablett GA, Cordeiro GM, Valkoun J, Henry RJ (2005) EST versus genomic derived microsatellite markers for

genotyping wild and cultivated barley. *Genetic Resources and Crop Evolution*, **52**, 903–909.

Chagne D, Chaumeil P, Ramboer A *et al.* (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. *Theoretical and Applied Genetics*, **109**, 1204–1214.

Chen YA, Mckillen DJ, Wu S *et al.* (2004) Optimal cDNA microarray design using expressed sequence tags for organisms with limited genomic information. *BMC Bioinformatics*, **5**, 53–59.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the MIRAEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Chung BYW, Simons C, Firth AE, Brown CM, Hellens RP (2006) Effect of 5′ UTR introns on gene expression in *Arabidopsis thaliana*. *BMC Genomics*, **7**.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Conant GC, Wolfe KH (2006) Functional partitioning of yeast co-expression networks after genome duplication. *PloS Biology*, 545–554.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* s) ESTs cross transferable to *erianthus* and *sorghum*. *Plant Science*, **160**, 1115–1123.

Cortereal H, Dixon DR, Holland PWH (1994) Intron targeted PCR — a new approach to survey neutral DNA polymorphism in bivalve populations. *Marine Biology*, **120**, 407–413.

Coulibaly I, Gharbi K, Danzmann RG, Yao J, Rexroad CE (2005) Characterization and comparison of microsatellites derived from repeat-enriched libraries and expressed sequence tags. *Animal Genetics*, **36**, 309–315.

Cui L, Wall PK, Leebens-Mack JH *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Research*, **16**, 738–749.

De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, **20**, 591–597.

Dong QF, Kroiss L, Oakley FD, Wang BB, Brendel V (2005) Comparative EST analysis in plant systems. *Methods in Enzymology: Producing the Biochemical Data*, **395**, 400–418.

Doust AN, Kellogg EA (2006) Effect of genotype and environment on branching in weedy green millet (*Setaria viridis*) and domesticated foxtail millet (*Setaria italica*) (Poaceae). *Molecular Ecology*, **15**, 1335–1349.

Durand D, Hoberman R (2006) Diagnosing duplications — can it be done? *Trends in Genetics*, **22**, 156–164.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution*, **13**, 685–690.

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics*, **104**, 399–407.

Evans JD, Wheeler DE (2000) Expression profiles during honeybee caste determination. *Genome Biology*, **2**, 1–6.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Research*, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Research*, 175–185.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics*, **4**, 649–655.

Fraser LG, McNeilage MA, Tsang GK, Harvey CF, De Silva HN (2005) Cross-species amplification of microsatellite loci within the dioecious, polyploid genus *Actinidia* (Actinidiaceae). *Theoretical and Applied Genetics*, **112**, 149–157.

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, **14**, 1457–1467.

Gadberry MD, Malcomber ST, Doust AN, Kellogg EA (2005) PRIMACLADE: a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21**, 1263–1264.

Garrigan D, Hedrick PW (2003) Detecting adaptive molecular polymorphism, lessons from the MHC. *American Journal of Human Genetics*, **73**, 375–375.

Gibson G (2002) Microarrays in ecology and evolution: a preview. *Molecular Ecology*, **11**, 17–24.

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Research*, **15**, 674–680.

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, **15**, 1153–1160.

Hartmann S, Lu D, Phillips J, Vision TJ (2006) Phytome: a platform for plant comparative genomics. *Nucleic Acids Research*, **34**, D724–D730.

Hegarty MJ, Jones JM, Wilson ID *et al.* (2005) Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. *Molecular Ecology*, **14**, 2493–2510.

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research*, 868–877.

Jany JL, Bousquet J, Khasa DP (2003) Microsatellite markers for Hebeloma species developed from expressed sequence tags in the ectomycorrhizal fungus *Hebeloma cylindrosporum*. *Molecular Ecology Notes*, 659–661.

Jarman SN (2004) AMPLICON: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, **20**, 1644–1645.

Kan Z, Castle J, Johnson JM, Tsinoremas NF (2004) Detection of novel splice forms in human and mouse using cross-species approach. *Pacific Symposium on Biocomputing*, 42–53.

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology*, **48**, 501–510.

Kobayashi N, Watanabe M, Kijimoto T *et al.* (2006) magp4 gene may contribute to the diversification of cichlid morphs and their speciation. *Gene*, **373**, 126–133.

Kota R, Rudd S, Facius A *et al.* (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Molecular Genetics and Genomics*, **270**, 24–33.

Ku H, Vision T, Liu J, Tanksley SD (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of

synteny. *Proceedings of the National Academy of Sciences, USA*, **97**, 9121–9126.
La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics*, 23–32.
Lai Z, Livingstone K, Zou Y *et al.* (2005) Identification and mapping of SNPs from ESTs in sunflower. *Theoretical and Applied Gentics*, **111**, 1532–1544.
Laitinen RAE, Immanen J, Auvinen P *et al.* (2005) Analysis of the floral transcriptome uncovers new regulators of organ determination and gene families related to flower organ differentiation in *Gerbera hybrida* (Asteraceae). *Genome Research*, **15**, 475–486.
Lamason RL, Mohideen M, Mest JR *et al.* (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **310**, 1782–1786.
Le Dantec L, Chagne D, Pot D *et al.* (2004) Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant Molecular Biology*, **54**, 461–470.
Le Quere A, Schutzendubel A, Rajashekar B *et al.* (2004) Divergence in gene expression related to variation in host specificity of an ectomycorrhizal fungus. *Molecular Ecology*, **13**, 3809–3819.
Lee CE, Mitchell-Olds T (2006) Preface to the special issue: ecological and evolutionary genomics of populations in nature. *Molecular Ecology*, **15**, 1193–1196.
Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research*, **12**, 1048–1059.
Lessa EP (1992) Rapid surveying of DNA-sequence variation in natural populations. *Molecular Biology and Evolution*, 323–330.
Lexer C, Welch ME, Durphy JL, Rieseberg LH (2003) Natural selection for salt tolerance quantitative trait loci (QTLs) in wild sunflower hybrids: implications for the origin of Helianthus paradoxus, a diploid hybrid species. *Molecular Ecology*, **12**, 1225–1235.
Li (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts.
Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, **21**, 991–1007.
Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Research*, **28**, 3657–3665.
Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, O'Brien SJ (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, **15**, 47–56.
MacCallum C, Hill E (2006) Being positive about selection. *PLoS Biology*, 293–295.
Malde K, Coward E, Jonassen I (2005) A graph-based algorithm for generating EST consensus sequences. *Bioinformatics*, **21**, 1371.
Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER (2005) Microsatellite instability regulates transcription factor binding and gene expression. *Proceedings of the National Academy of Sciences, USA*, **102**, 3800–3804.
McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.
Moore S, Payton P, Wright M, Tanksley S, Giovannoni J (2005) Utilization of tomato microarrays for comparative gene expression analysis in the Solanaceae. *Journal of Experimental Botany*, **56**, 2885–2895.
Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics*, **30**, 194–200.
Murray CG, Larsson TP, Hill T, Bjorklind R, Fredriksson R, Schioth HB (2005) Evaluation of EST-data using the genome assembly. *Biochemical and Biophysical Research Communications*, **331**, 1566–1576.
Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences, USA*, **100**, 5268–5273.
Ng SHS, Chang A, Brown GD, Koop BF, Davidson WS (2005) Type I microsatellite markers from Atlantic salmon (*Salmo salar*) expressed sequence tags. *Molecular Ecology Notes*, 762–766.
Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**, 641–647.
Oefner PJ, Underhill PA (1995) Comparative DNA sequencing by denaturing high performance liquid chromatography (DHPLC). *American Journal of Human Genetics*, **57**, 1547–1547.
Ohno S (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.
Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–266.
Omelchenko MV, Wolf YI, Gaidamakova EK *et al.* (2005) Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and radiation resistance. *BMC Evolutionary Biology*, **5**, 73–78.
Orita M, Suzuki Y, Sekiya T, Hayashi K (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics*, 874–879.
Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.
Parkinson J, Guiliano DB, Blaxter M (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31–39.
Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M (2004) PARTIGENE: constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.
Pasyukova EG, Vieira C, Mackay TFC (2000) Deficiency mapping of quantitative trait loci affecting longevity in *Drosophila melanogaster*. *Genetics*, **156**, 1129–1146.
Pertea G, Huang XQ, Liang F *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
Picoult-Newberg L, Ideker TE, Pohl MG *et al.* (1999) Milling SNPs from EST databases. *Genome Research*, 167–174.
Pontius JU, Wagner L, Schuler GD (2003) UNIGENE: a unified view of the transcriptome. In: *The NCBI Handbook (Online Book)*.
Quackenbush J, Cho J, Lee D *et al.* (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.
Rexroad CE, Rodriguez MF, Coulibaly I *et al.* (2005) Comparative mapping of expressed sequence tags containing microsatellites in rainbow trout (*Oncorhynchus mykiss*). *BMC Genomics*, **6**, 54–62.
Rise ML, von Schalburg KR, Brown GD *et al.* (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*, **14**, 478–490.

Rohrer GA, Fahrenkrug SC, Nonneman D, Tao N, Warren WC (2002) Mapping microsatellite markers identified in porcine EST sequences. *Animal Genetics*, **33**, 372–376.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**, 84–89.

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–363.

Rostoks N, Borevitz JO, Hedley PE *et al.* (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biology*, **6**.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequence? *Trends in Plant Science*, 321–329.

Sampedro J, Lee Y, Carey RE, dePamphilis C, Cosgrove DJ (2005) Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant Journal*, **44**, 409–419.

Savage D, Batley J, Erwin T *et al.* (2005) SNPSERVER: a real-time SNP discovery tool. *Nucleic Acids Research*, **33**, W493–W495.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schuler GD, Boguski MS, Stewart EA *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.

Scott KD, Eggler P, Seaton G *et al.* (2000) Analysis of SSRs derived from grape ESTs. *Theoretical and Applied Genetics*, **100**, 723–726.

Scotti I, Magni F, Fink R, Powell W, Binelli G, Hedley P (2000) Microsatellite repeats are not randomly distributed within Norway spruce (*Picea abies* K.) expressed sequences. *Genome*, **43**, 41–46.

Seki M, Carninci P, Nishiyama Y, Hayashizaki Y, Shinozaki K (1998) High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper. *Plant Journal*, **15**, 707–720.

Sheffield VC, Cox DR, Lerman LS, Myers RM (1989) Attachment of a 40 base pair G+C rich sequence (GC clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single base changes. *Proceedings of the National Academy of Sciences, USA of the United States of America*, **86**, 232–236.

Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005a) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology*, **14**, 4193–4203.

Smith CT, Templin WD, Seeb JE, Seeb UW (2005b) Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of US and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management*, **25**, 944–953.

Snel B, Bork P, Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, **12**, 17–25.

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. *New Phytologist*, **167**, 165–170.

Sterky F, Bhalerao RR, Unneberg P *et al.* (2004) A POPULUS EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences, USA*, **101**, 13951–13956.

Stinchcombe JR, Weinig C, Ungerer M *et al.* (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences, USA*, **101**, 4712–4717.

Strous M, Pelletier E, Mangenot S *et al.* (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, **440**, 790–794.

Susko E, Roger AJ (2004) Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics*, **20**, 2279–2287.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **98**, 7375–7379.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics*, **168**, 1457–1465.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, **106**, 411–422.

Tiffin P, Hahn MW (2002) Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. pekinensis. *Journal of Molecular Evolution*, **54**, 746–753.

Torres M, Tomilov A, Tomilova N, Reagan R, Yoder J (2005) PSCROPH: a parasitic plant EST database enriched for parasite associated transcripts. *BMC Plant Biology*, **5**.

Tsuchihashi Z, Dracopoli C (2002) Progress in high throughput SNP genotyping methods. *Pharmacogenomics Journal*, **2**, 103–110.

Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics*, **10**, 752–763.

Van Deynze AE, Sorrells ME, Park WD *et al.* (1998) Anchor probes for comparative mapping of grass genera. *Theoretical and Applied Genetics*, **97**, 356–369.

Varshney RK, Sigmund R, Borner A *et al.* (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Science*, **168**, 195–202.

Vasemagi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vasemagi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution*, **22**, 1067–1076.

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP — a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wang BB, Brendel V (2006) Genome-wide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences, USA*, **103**, 7175–7180.

Wang JPZ, Lindsay BG, Leebens-Mack J *et al.* (2004) EST clustering error evaluation and correction. *Bioinformatics*, **20**, 2973–2984.

Wang JPZ, Lindsay BG, Cui LY *et al.* (2005) Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. *BMC Bioinformatics*, **6**.

Weber M, Harada E, Vess C, von Roepenack-Lahaye E, Clemens S (2004) Comparative microarray analysis of *Arabidopsis thaliana* and *Arabidopsis halleri* roots identifies nicotianamine synthase, a

ZIP transporter and other genes as potential metal hyperaccumulation factors. *Plant Journal*, **37**, 269–281.
Wolyn DJ, Borevitz WO, Loudet O *et al.* (2004) Light-response quantitative trait loci identified with composite interval and extreme array mapping in *Arabidopsis thaliana*. *Genetics*, **167**, 907–917.
Woodhead M, Russell J, Squirrell J *et al.* (2003) Development of EST-SSRs from the alpine lady-fern, *Athyrium distentifolium*. *Molecular Ecology Notes*, 287–290.
Wydner KS, Sechler JL, Boyd CD, Passmore HC (1994) Use of an intron length polymorphism to localize the tropoelastin gene to mouse chromosome 5 in a region of linkage conservation with human chromosome 7. *Genomics*, **23**, 125–131.
Xu JP (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Molecular Ecology*, **15**, 1713–1731.
Yang ZH (2002) Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development*, **12**, 688–694.
Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.
Zdobnov EM, Apweiler R (2001) INTERPROSCAN — an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
Zhu W, Schlueter SD, Brendel V (2003) Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiology*, **132**, 469–484.

Amy Bouck is a postdoc at UNC Chapel Hill and Duke University. Her research on the plant *Mimulus* examines how population-level processes such as mating system, population size and hybridization affect aspects of plant genome evolution, including proliferation of transposons, distribution of genetic variation, and barriers to gene flow. Todd Vision is an assistant professor at UNC Chapel Hill. His lab's work focuses on genome evolution and the genetic basis of complex traits, using both computational and molecular methods. They are currently collaborating on studies of whole genome duplication and linkage disequilibrium in outcrossing versus selfing species of *Mimulus*.

Glossary

cDNA — (complementary DNA) a DNA molecule synthesized from a mature mRNA template.

cDNA library — a collection of cDNAs created from a pool of mRNAs sampled from a particular tissue, life stage, or environmental condition. Libraries are typically constructed by inserting the cDNA molecules into plasmid vectors, which are then transformed into *Escherichia coli* cells, allowing for storage, isolation and replication of the cDNA-containing plasmids. The *E. coli* cells are then plated and screened for transformation. Positive colonies of *E. coli* (meaning a cDNA-containing plasmid was taken up) are then individually isolated (picked) and cultured, typically in 96-well plates. This is a very labourious step: one colony must be picked for each eventual EST. Projects can be scaled up significantly by using a robotic colony picker. Sequencing of the cDNAs is accomplished by extracting the plasmids (a procedure called a miniprep). This entire process of cDNA library construction requires at minimum fairly basic genetics wet laboratory equipment, and countless commercial kits are available. The entire process from tissue sampling to the production of an EST sequence collection and a unigene set can also be contracted out to one of many commercial firms.

Directional cloning — a procedure that directs the orientation of cDNAs as they are inserted into plasmid vectors. Primers specific to the different sides of the insert site in the plasmid allow EST sequencing to be targeted toward capturing the 5′ or 3′ ends gene transcripts. The inserts in a nondirectional (random) library are in no particular orientation with respect to the plasmid insert site. Commercial directional cloning kits are available.

Normalization — a procedure by which the relative abundance of different cDNA transcripts is somewhat equalized (Bonaldo *et al*. 1996). Complementary DNA libraries that are not normalized typically contain many copies of highly expressed gene transcripts, and few to no copies of genes with low expression. Normalization increases the overall diversity of transcripts included in a cDNA library and resulting EST collection. Several techniques for normalization exist, and these can be accomplished using commercially available kits or contract services.

Cap trapping — a procedure used in cDNA library construction to produce cDNAs that capture full-length mRNA transcripts. Standard cDNA library construction produces cDNAs that are often truncated at the 5′ end, due to the failure of reverse transcriptase to process all the way down the mRNA template (see Fig. 1, part 3). The cap trapping procedure (Carninci *et al*. 1996; Seki *et al*. 1998) is used after the first strand synthesis step (Fig. 1, step 3) to isolate only those molecules that have been reverse transcribed all the way to the guanine-capped 5′ end of the mRNA (Fig. 1, step 2). These full-length cDNAs are then cloned into libraries and sequenced to achieve full coverage of the original gene transcript. This process is technically challenging, expensive, and may bias the resulting cDNA collection towards shorter and more abundant transcripts.

Paired-end sequencing — cDNA library clones are sequenced from both ends. The two corresponding sequences are referred to as **mate pairs**. If the cDNA is particularly long, these sequences may not overlap, but the fact that they correspond to two ends of a single cDNA molecule can be used in the process of unigene clustering (see Fig. 4).