Plant Molecular Biology **48:** 501–510, 2002.

Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat

Ramesh V. Kantety[1], Mauricio La Rota[1], David E. Matthews[2] and Mark E. Sorrells[1,*]
[1]*Department of Plant Breeding, 252 Emerson Hall, Cornell University, Ithaca, NY 14853, USA (*author for correspondence; e-mail mes12@cornell.edu);* [2]*USDA-ARS Center for Agricultural Bioinformatics, 409 Bradfield Hall, Cornell University, Ithaca, NY 14853, USA*

Received 30 January 2001; accepted in revised form 5 September 2001

Key words: EST, comparative genomics, comparative mapping, EST, EST clustering, SSR

Abstract

Plant genomics projects involving model species and many agriculturally important crops are resulting in a rapidly increasing database of genomic and expressed DNA sequences. The publicly available collection of expressed sequence tags (ESTs) from several grass species can be used in the analysis of both structural and functional relationships in these genomes. We analyzed over 260 000 EST sequences from five different cereals for their potential use in developing simple sequence repeat (SSR) markers. The frequency of SSR-containing ESTs (SSR-ESTs) in this collection varied from 1.5% for maize to 4.7% for rice. In addition, we identified several ESTs that are related to the SSR-ESTs by BLAST analysis. The SSR-ESTs and the related sequences were clustered within each species in order to reduce the redundancy and to produce a longer consensus sequence. The consensus and singleton sequences from each species were pooled and clustered to identify cross-species matches. Overall a reduction in the redundancy by 85% was observed when the resulting consensus and singleton sequences (3569) were compared to the total number of SSR-EST and related sequences analyzed (24 606). This information can be useful for the development of SSR markers that can amplify across the grass genera for comparative mapping and genetics. Functional analysis may reveal their role in plant metabolism and gene evolution.

Introduction

Microsatellites are short (1–5 bp) repeat motifs that are usually associated with a high level of frequency of length polymorphism. Often referred to as simple sequence repeats (SSRs), they can be found in both coding and non-coding DNA sequences of all higher organisms examined to date (Tautz and Renz, 1984; Gupta *et al.*, 1996). The elevated frequency of length polymorphism associated with microsatellites provides a basis for the development of a marker system that has broad application in genetic research including studies of genetic variation, linkage mapping, gene tagging, and evolution.

Linkage maps employing microsatellites have been developed for several cereal grains such as barley (*Hordeum vulgare*) (Liu *et al.*, 1996; Ramsay *et al.*, 2000), maize (*Zea mays* L.) (Senior *et al.*, 1993), rice (*Oryza sativa* L.) (McCouch *et al.*, 1997; Temnykh *et al.*, 2000), hexaploid wheat (*Triticum aestivum* L.) (Röder *et al.*, 1998) and durum wheat (*Triticum turgidum* L. var. *durum*) (Korzun *et al.*, 1999; Nachit *et al.*, 2000). These maps have been widely used for tagging genes of agricultural importance. Similarly, the high level of variation detected using microsatellites increases the resolution of relationships in cultivated germplasm and reduces the number of markers required to distinguish all genotypes. A few examples of genetic diversity studies for Gramineae species include rice (Cho *et al.*, 2000), fingermillet (*Eleusine coracana*) (Salimath *et al.*, 1995), maize (Kantety *et al.*, 1995; Chin *et al.*, 1996), barley (Becker and Heun, 1995; Russell *et al.*, 1997) and wheat (Plaschke *et al.*, 1995; Bryan *et al.*, 1997; Eujayl *et al.*, 2000).

Earlier genetic studies on microsatellite marker development primarily utilized anonymous DNA fragments containing SSRs isolated from a genomic DNA library. mRNA transcripts also contain repeat motifs and the abundance of expressed sequence tags (ESTs) makes this an attractive potential source of microsatellite markers. The presence of SSRs in transcripts of known genes suggests that they may have a role in gene expression or function. In man, expansion of trinucleotide repeats in coding regions is sometimes associated with neuropathological diseases such as Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993). In plants, the *waxy* gene in rice has been found to contain a poly(CT) microsatellite in the 5′-untranslated region (UTR) whose length polymorphism is associated with amylose content (Ayers *et al.*, 1997).

Transferability among species demonstrated in earlier studies (Herron *et al.*, 1998; Eujayl *et al.*, 2000; Scott *et al.*, 2000; Sorrells, 2000a, b) implies that SSR-ESTs have considerable potential for comparative mapping. To be useful for comparative mapping, a molecular marker must identify orthologous loci in two or more species and exhibit a sufficient level of polymorphism within a species to facilitate determination of map location. For PCR-based markers, these criteria are conflicting because DNA sequence variation is essential for polymorphism whereas conservation of DNA sequence is essential for designing primers that function across species. Sequences containing conserved regions of a gene flanking a hypervariable region such as SSR-ESTs are most useful for designing primers that work across two or more species.

Comparative DNA sequence analysis methods can facilitate high-throughput comparative mapping (Band *et al.*, 2000; Laurent *et al.*, 2000; Sorrells, 2000a, b). By using ESTs to cross-reference genes between species maps it is possible to enhance resolution of comparative maps and facilitate interspecies gene cloning. Rebeiz and Lewin (2000) used the comparative mapping by annotation and sequence similarity (COMPASS) method and predicted the bovine chromosome assignment of about 60% of cattle unigenes based on the human unigene database and a bovine/human comparative map. Sorrells (2000a) used sequence data from oat and barley cDNA clones that had been previously mapped in rice, wheat, maize, or barley to identify similar ESTs in GenBank. For those ESTs that had been mapped in the same species, the chromosome map location of the EST was compared to the location of the locus mapped with the oat or barley clone. More than 60% (50/80) of the sequences mapped to the predicted location (within 20–30 cM) based on existing comparative maps suggesting that for the majority of the cases, sequence similarity alone can be used for enhancing comparative maps.

The primary goal of this study was to estimate the frequency of SSRs in the publicly available EST databases for barley, maize, rice, sorghum and wheat. This information is used to assess the feasibility of using SSR-ESTs for comparative mapping and to develop strategies that take advantage of DNA sequence analyses for cross-referencing genes between species and genera.

Materials and methods

EST sequence sources and processing

The grass EST sequences used in this project were FASTA-formatted files collected from several publicly available databases in October of 2000. A total of 61 928 rice ESTs and 45 264 sorghum ESTs were downloaded from dbEST/GenBank (http://ncbi.nlm.-nih/entrez), while 63 626 maize ESTs were obtained from ZmDB (http://www.zmdb.iastate.edu). We obtained 38 238 wheat ESTs and 22 869 ESTs from *Triticum* species from International Triticeae EST Consortium (ITEC) at the Wheat NSF project homepage (http://wheat.pw.usda.gov/NSF/) and ITEC homepage (http://wheat.pw.usda.gov/genome/group/pool/), respectively. Finally, the 30 706 barley ESTs where obtained from the barley EST projects at Institute for Plant Genetics and Crop Plant Research, Gatersleben, Germany (http://pgrc.ipk-gatersleben.de/) and from Clemson University Genomics Institute (http://www.genome.clemson.edu/). A relational database was used to store all information related to the EST sequences from the five species, including similarity search results, SSR presence and cluster membership.

SSR detection

All six data sets of sequences were screened for the presence of SSRs with the Perl script SSRIT. SSRIT is a microsatellite search tool available at the USDA-ARS Center for Bioinformatics and Comparative Genomics at Cornell University (http://arsgenome.cornell.edu/cgi-bin/rice/ssrtool.pl). ESTs that contained SSRs were labeled as SSR-ESTs and the exact location of the SSRs within the ESTs was

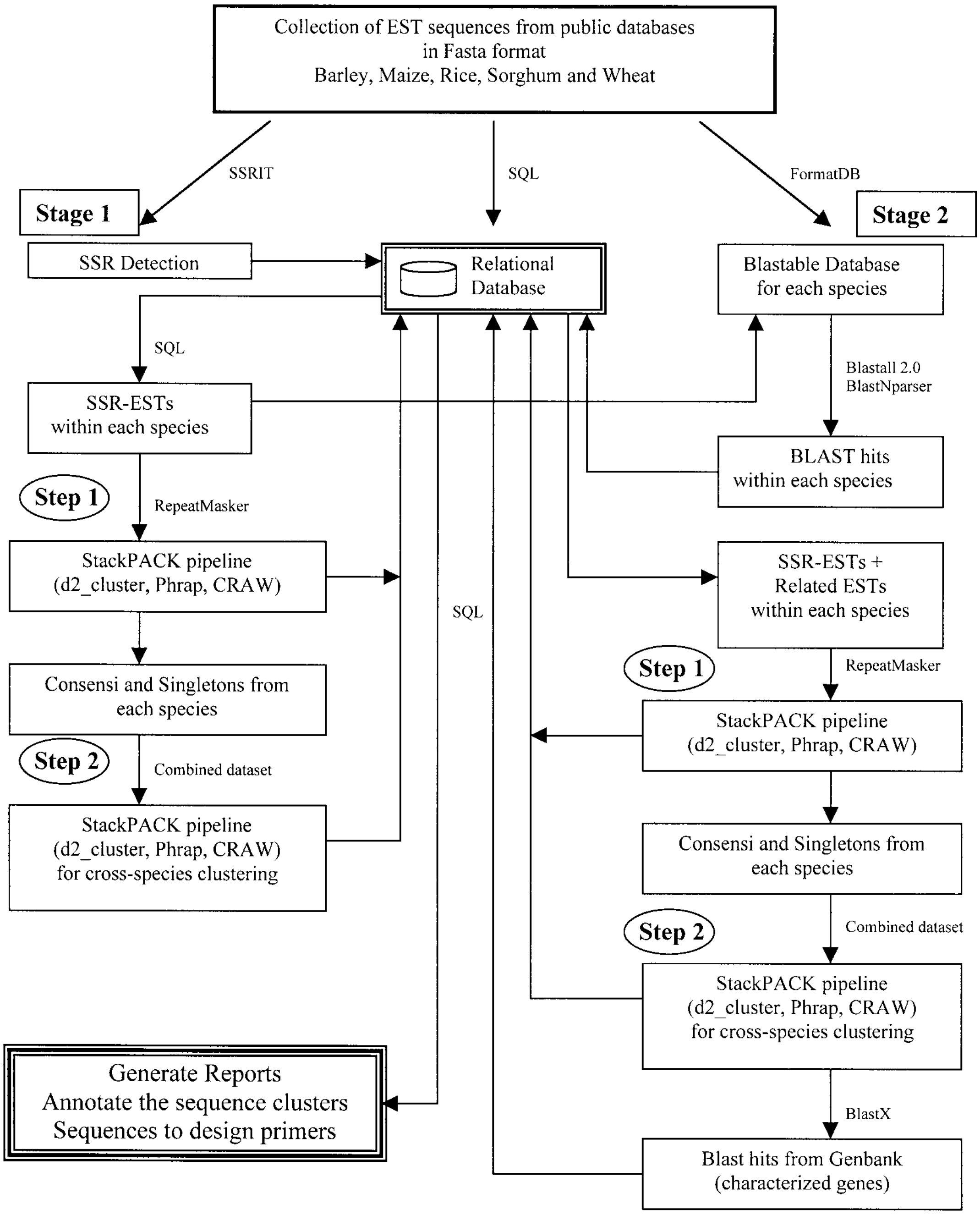
Collection of EST sequences from public databases
in Fasta format
Barley, Maize, Rice, Sorghum and Wheat
SSRIT
SQL
FormatDB
Stage 1
Stage 2
SSR Detection
Relational Database
Blastable Database for each species
SQL
Blastall 2.0
BlastNparser
SSR-ESTs within each species
BLAST hits within each species
Step 1
RepeatMasker
StackPACK pipeline (d2_cluster, Phrap, CRAW)
SSR-ESTs + Related ESTs within each species
SQL
Consensi and Singletons from each species
Step 1
RepeatMasker
StackPACK pipeline (d2_cluster, Phrap, CRAW)
Step 2
Combined dataset
StackPACK pipeline (d2_cluster, Phrap, CRAW) for cross-species clustering
Consensi and Singletons from each species
Step 2
Combined dataset
StackPACK pipeline (d2_cluster, Phrap, CRAW) for cross-species clustering
Generate Reports
Annotate the sequence clusters
Sequences to design primers
BlastX
Blast hits from Genbank (characterized genes)

Figure 1. Flow chart of the clustering and annotation of ESTs from multiple species.

annotated in the database together with repeat motif and microsatellite size information.

Clustering of the SSR-ESTs

The identified SSR-ESTs were subjected to masking of other repeat sequences from the Poaceae family (MITES, etc.) as well as low-complexity sequences, including the SSRs and interspersed repeats, with the RepeatMasker Program (Smit, 1999; http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) prior to clustering. The masked SSR-ESTs were then clustered using the StackPACK 2.0 system (Miller *et al.*, 1999; http://www.egenetics.com).

Table 1. EST analysis based on the SSRIT script, which identified di-, tri- and tetra-nucleotide repeat motifs that are at least 18 to 20 bases in length.

Organism	Number of ESTs	Number of SSR-ESTs	% of ESTs with SSRs	Number of repeats			% of repeats over all SSR-ESTs		
				di-	tri-	tetra-	di-	tri-	tetra-
Barley	30 706	1049	3.4	304	622	123	29	59	12
Wheat	38 238	1212	3.2	125	1000	87	10	83	7
ITEC	22 869	742	3.2	154	534	54	21	72	14
Maize	63 626	983	1.5	291	556	136	30	57	14
Sorghum	45 264	1634	3.6	269	1145	220	17	70	12
Rice	61 928	2894	4.7	463	2279	152	16	79	5
Total	262 631	8514	3.2	1606	6136	772	19	72	9

Clustering of the SSR-ESTs and their related ESTs

After the first exploration of the SSR-ESTs, it was apparent that additional information could be obtained by returning to the EST databases to search for the ESTs overlapping with those that carry SSRs. The purpose of this analysis was to enlarge the consensus and to reduce redundancy by joining the singletons into new or existing clusters. The new overlapping ESTs do not contain SSRs but are expected to bridge or extend existing clusters or singletons containing SSR-ESTs. A local copy of the program BLAST2.0 (Altschul *et al.*, 1997) was used to make pairwise comparisons between the SSR-ESTs and the full EST data set for each species separately. All significantly similar ESTs (with an E-value <1 e^{-5}, or above a 90% similarity over a stretch of more than 30 bases) were added to the SSR-EST data set for further clustering. Another cycle of repeat masking and filtering of low-complexity sequences was performed for each of the species sets of SSR-ESTs and their overlapping sequences, and they were clustered again with the StackPACK system. A combined cross-species set of non-redundant (NR) sequences, which include consensus sequences and singletons from each species, was analyzed in addition to the individual species-specific sets.

Because the first stage of clustering defines loose clusters of both SSR containing ESTs and normal ESTs, some of the sequences that were included into clusters at the first stage may be separated in the more stringent second (PHRAP) or third (CRAW) stages, forming independent consensus sequences that do not contain any SSRs. These clusters were filtered out in a subsequent step. The consensus sequences containing ESTs from two or more species were compared for type of SSR motif and putative function. Lower-stringency settings (85% similarity threshold) were used to cluster EST sequences from more than one species.

Results and discussion

Search for ESTs containing SSR motifs

We used the SSRIT Perl script to identify ESTs from the data sets for barley, wheat, ITEC, maize, sorghum and rice, which contain di, tri- and tetra-nucleotide SSRs with a minimum length of 18 (di- and tri-) to 20 (tetra-) bases. These parameters were chosen because the polymorphism level in rice decreases with SSRs shorter than 18 bases (Cho *et al.*, 2000). The SSRIT program identified 8514 SSRs containing ESTs (SSR-ESTs) from a total of 262 631 ESTs. The frequency of finding a SSR motif in the EST collection ranged from 1.5% in maize to 4.7% in rice, with an average of 3.2% over all the data sets tested (Table 1). The relative abundance of di-, tri-, and tetra-nucleotide repeat motifs in the EST collection was 19%, 72% and 9% respectively (Table 1).

The di-nucleotide repeat motifs were present in similar frequencies in ESTs from different species. Among the di-nucleotide motifs, GA/CT was the most abundant and GC/CG was the least abundant motif in all the species (Table 2). Other researchers also found the GA/CT motif to be one of the most abundant di-nucleotide motifs in rice genes (Temnykh *et al.*, 2000). A di-nucleotide motif can represent multiple codons depending on the reading frame and translate into different amino acids. For example, the GA/CT motif

Table 2. Distribution of the SSR-ESTs based on the motif sequence.

SSR Motif	Barley	Wheat	ITEC	Maize	Sorghum	Rice	Total
GA/CT	244	81	112	198	147	373	1155
CA/GT	37	31	24	40	70	24	226
AT/TA	14	13	8	51	52	65	203
GC/CG	9	0	10	2	0	1	22
AGC/TCG	106	125	84	109	199	195	818
AGT/TCA	7	7	3	21	16	25	79
CTA/GAT	14	13	11	13	24	22	97
AAG/TTC	44	65	27	24	40	226	426
AAC/TTG	33	402	171	21	14	21	662
AAT/TTA	5	3	1	42	7	20	78
GGC/CCG	259	271	159	141	452	1095	2377
CCT/GGA	73	55	41	59	175	435	838
Total	845	1066	651	721	1196	2502	6981

can represent GAG, AGA, UCU and CUC codons in a mRNA population and translate into the amino acids Arg, Glu, Ala and Leu respectively. Ala and Leu are present in proteins at high frequencies of 8% and 10%, respectively (Lewin, 1994). This could be one of the reasons why GA/CT motifs are present at such high frequencies in EST collections. Temnykh *et al.* (2000) found GA/CT repeats to be the most polymorphic class of SSRs in rice. In our datasets, we identified over 1150 ESTs with GA/CT motif. The most abundant tri-nucleotide repeat motif was GGC/CCG in all the species except in wheat where AAC/TTG motif was the most common tri-nucleotide repeat motif. GC-rich motifs in rice were reported to be mostly present in the coding regions (Cho *et al.*, 2000).

Variation in SSR length and stability varies greatly between species and among loci within species (Chakraborty *et al.*, 1997; Schug *et al.*, 1998; Cho *et al.*, 2000; Temnykh *et al*, 2000) and microsatellites derived from genomic libraries have been found to be considerably more polymorphic than those from ESTs (Becker and Heun, 1996; Cho *et al.*, 2000; Eujayl *et al.*, 2000). Microsatellite markers derived from rice genomic libraries were more polymorphic than those from ESTs for all motif classes examined, especially when the SSRs were located in exons or open reading frames rather than untranslated regions (UTRs) (Cho *et al.*, 2000). In grape ESTs, SSRs from the 3′-UTR were more polymorphic among cultivars than those from the 5′-UTR while SSRs from coding regions were polymorphic only between species and genera (Scott *et al.*, 2000). In a survey of genetic variation among 64 durum lines, landraces, and varieties, Eujayl *et al.* (2000) reported that polymorphism was 25% for microsatellites derived from ESTs compared to 53% for those derived from genomic libraries.

However, SSR-ESTs represent a unique opportunity, in which sequence information is readily available and one can bypass the need of creating and sequencing SSR-enriched genomic libraries. Perhaps, the most important feature is that the primer pairs designed from SSR-ESTs are more likely to function in distantly related species than SSR primer pairs derived from genomic libraries. In our laboratory, 11 of the 21 SSR-EST primer pairs originated from wheat EST collection (Eujayl *et al.*, 2000) amplified in cultivated tef, *Eragrostis tef*, and 7 of these were mapped. In contrast, we were unable to amplify tef DNA using 180 SSR primer pairs originated from wheat genomic libraries (data not shown).

EST clustering

A cluster is defined here as a group of overlapping EST sequences. The purpose of clustering was to eliminate redundancy in the data set and to improve the quality and length of the reads by obtaining a consensus sequence after EST sequence alignment (similar to genomic sequence assembly). Previously, two main methodologies for clustering were used predominantly by research groups: (1) a generalized simple clustering based on sequence similarity that did not intend to derive a consensus from the clusters (NCBI Uni-

Table 3. Clustering analyses of SSR-ESTs and the ESTs that are related to SSR-ESTs. The analysis of each data set was performed at the default stringency of 96% similarity.

Source of ESTs	Number of SSR-ESTs[1]	Consensus sequences	Singletons	Total number of NR sequences	% Reduction in redundancy
Barley	3 462	269	600	869	75
Wheat	3 624	335	302	637	83
ITEC	2 277	230	457	687	70
Maize	4 107	408	385	793	81
Sorghum	3 976	430	498	928	77
Rice	7 160	752	1001	1753	76
Total	24 606	2424	3243	5667	77

[1] Includes the SSR-ESTs and their related ESTs selected based on the BLAST search against EST database for each of the data sets.

gene set, Schuler *et al.*, 1996), and (2) a competing methodology from TIGR that derived a high-quality consensus after applying very stringent clustering settings (Quackenbush *et al.*, 2000). Both procedures have advantages and disadvantages so we used instead a methodology that incorporates features of both, called the StackPACK system (Miller *et al.*, 1999).

The StackPACK system divides the clustering process into three main steps with increasing levels of stringency, to combine the information gained from clustering with the simplicity of a consensus sequence (La Rota, 2000). A preliminary step removes any remaining cloning vector sequences. Then, in the first step, loose clusters are defined on the basis of similarity (using the 'd2_cluster' program in StackPACK). Some ESTs may be left as single sequences if they do not match any other sequences. A few spurious similarities may form, but the first stage allows for EST members of the same gene families to be clustered. A second, more stringent step uses the PHRAP program (Green, 1999) and assembles the ESTs present in each cluster, to verify the cluster and generate a contig, or to divide the cluster into two or more contigs on singleton sequences, if the differences in the sequences exceed the preset criteria. The alignment of the overlapping sequences inside each contig generates a new consensus sequence and if two or more contigs were created, they maintain a linked relationship due to their participation in the same original cluster (this relationship is not maintained in the methodology used by TIGR). Further refinements to the consensus sequence are possible with the aid of the third step that uses the CRAW program of the StackPACK system, which allows the identification of alternative splicing sequences or similar variants of the same gene.

Clustering of ESTs reduces the redundancy in the dataset and creates a longer consensus sequence by deriving a longer segment of a particular gene than what is covered by a single EST. Our data set of grass ESTs comprised over a quarter-million sequences from five species and over 60 different cDNA libraries. Of these about 8500 sequences contained SSR motifs. To reduce redundancy we employed EST clustering in two stages with two major steps within each stage (Figure 1). In the first step of each stage, sequences were clustered by species to generate EST consensus and singleton sequences for each of the 5 species. The result of this step is a collection of non-redundant (NR) sequences containing SSRs for each of the species. In the second step, the NR sequences from all the species were pooled and analyzed with the StackPACK clustering procedure, with the goal of obtaining cross-species matches to identify evolutionarily conserved genes.

Stage 1

In the first stage of analyses, the number of SSR-EST sequences for the data sets ranged from 742 for ITEC to 2894 for rice (Table 1). We performed clustering analysis of the individual data sets in the first step. At a default stringency of 96% minimum similarity, the reduction of the redundancy was about 50% (data not shown). Over all the species, there were 1489 consensus sequences and 2742 singletons with a total of 4231 non-redundant (NR) sequences. In the second step, we clustered these 4231 NR sequences at both the default and a lower stringency of 85% similarity in order to match as many sequences across species as possible. As the default stringency, 169 consensus sequences were formed across two or more species. The majority

Table 4. Clustering analysis of non-redundant (NR) sequences pooled from all the data sets and analyzed at 96% and 85% similarity levels.

Source	Number of sequences	Consensus sequences	Singletons	Total number of NR sequences across species	% reduction in redundancy
Pooled NR ESTs (96%)	5667	366	4017	4383	23
Pooled NR ESTs (85%)	5667	647	2922	3569	37

of these were formed between the most closely related species, for example, barley and wheat. A total of 286 consensus sequences were formed at the reduced stringency of 85% similarity. Although few more matches between distantly related species, such as barley and rice, were found by reducing the stringency, there was no significant improvement in the number of cross-species consensus sequences at the reduced stringency with this data set (data not shown).

Stage 2

As discussed earlier, a good reason to extend the length of the sequence of an EST is to facilitate the design of primers when the SSR motif is near the end and no room is left for an anchoring oligonucleotide. In addition, the quality of sequence at the EST borders improves as ESTs are added to existing clusters and longer consensus sequences can be calculated. Clustering was limited in the first stage of analyses because there were only a handful of SSR-ESTs that represent a particular transcript. Since the SSRIT program identifies ESTs solely based on the fact that they contain a certain SSR motif, we were missing other ESTs that represent a different part of the same transcript. In the first stage of analyses there were only a few cross-species matches between SSR-EST consensus sequences because the sequences flanking the SSR were probably not long enough to find a matching sequence from another species.

The strategy in Stage 2 of the analysis was to identify ESTs that represent other parts of the same transcripts represented by SSR-ESTs. We created BLAST databases for all the ESTs from each data set. Then we used the SSR-EST sequences (masked for low-complexity regions including the SSR itself) from each species to query the corresponding BLAST database of EST sequences. The resulted in the identification of many ESTs (usually three to four times more) that have matches to the original SSR-ESTs. The newly identified ESTs are only relatives to the SSR-ESTs as they do not contain SSR motifs in them but will contribute to the cross-species sequence matching. These were pooled, along with the SSR-ESTs for each species, and were introduced into the clustering analysis.

The second stage of clustering within each data set resulted in a near 2-fold increase in the number of consensi compared to the clustering of only the SSR-EST sequences (Table 3). This was due to one of the following reasons: (1) some of the sequences left out in the first stage as singletons formed new clusters in the second stage, or (2) some of the newly introduced sequences formed clusters of their own (they may not necessarily include a SSR-EST because the default stringency of StackPACK, unlike the BLAST threshold we used, does not allow lower similarity or shorter sequence matches). We eliminated those consensus sequences and singletons that did not contain SSR motifs from each of the data sets, but for the cross-species experiment we retained some sequences without SSR motifs because they formed a cluster with SSR-ESTs from two or more species. A total of 670 singletons over all the data sets from the first stage joined clusters in the second stage of clustering analysis.

EST clustering across grass species

The default stringency of 96% similarity resulted in few clusters that formed mostly between the sequences from closely related species. Therefore the stringency was reduced to 85% similarity. About a third of the NR sequences from each species formed clusters with NR sequences from other species at the lower stringency. At a minimum of 85% similarity in Stage 2 analyses, 647 consensus sequences and 2922 singletons were formed from a total of 5667 NR sequences (Table 4). The total reduction in redundancy of the NR sequences at the lower stringency was 37%. The SSR motif remained the same in multi-species clusters. In some cases more than one SSR region was found in a

Table 5. Number of cross-species clusters representing a pair-wise match between sequences from any two datasets. The numbers presented above the diagonal are for the 96% similarity level and the numbers below the diagonal are for the 85% similarity level. The numbers along the diagonal are for each species against all the other species.

96% *85%*	Barley	Wheat	Maize	Sorghum	Rice
Barley	*100* *131*	45	5	5	12
Wheat	60	*113* *130*	5	7	10
Maize	17	16	*29* *59*	18	9
Sorghum	20	20	34	*30* *63*	9
Rice	35	34	22	24	*27* *73*

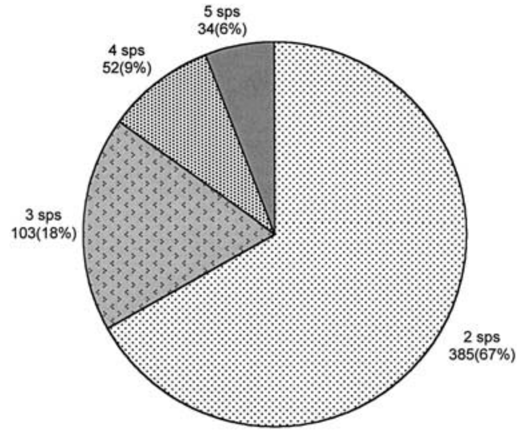

Figure 2. Distribution of the cross-species consensus sequences based on the number of species present in each consensus. Each section represents the proportion of all consensus sequences that match a sequence in at least 2, 3, 4, or 5 species (sps).

single consensus. The ITEC sequences in our analysis were treated as a single data set but they contain EST sequences from closely related species such as barley and wheat. Therefore the number of consensus sequences formed between ITEC ESTs and barley or wheat sequences could not be treated as cross-species matches.

The number of consensus sequences that formed across multiple species was increased by 2- to 3-fold when the stringency was reduced (Table 5). However, the increase in distantly related species was more than the increase in closely related species suggesting that we were clustering orthologues from distantly related species and not paralogues from closely related species. A search of GenBank non-redundant (nr) database using the new consensus sequences derived from the distantly related species revealed that these genes encode evolutionarily conserved genes such as ubiquitin, myb-class genes and metabolic enzymes.

Over all the analyses, it was evident that our clustering approach was highly successful in identifying orthologues from most closely related species. There were very few consensus sequences that formed between all the species (Figure 2). The software programs we used were originally designed to assemble genomic sequences at very high stringency or to assemble ESTs from a single species. These stringent parameters ensure that a high-quality consensus is derived for that species. However, matching of a pool of sequences from different species involves several challenges. Determining the stringency level to match the orthologues from multiple species is difficult because genes evolve at different rates in different species. And at lower stringencies, paralogous sequences from the same species will joint the clusters and reduce the quality of a cluster. In our analyses, individual species datasets were processed a high stringency and the NR data set was first analyzed at the high stringency and then at a lower stringency where there were still an insignificant number of within-species clusters. Identification of orthologues will benefit from refinement of sequence matching parameters.

When BLAST analysis alone was used on the NR data set of sequences that had been clustered at both 85% and 96% similarity using StackPACK, at a E-value of e^{-10}, a total of 785 clusters were formed (data not shown). Any clusters obtained using BLAST are generally considered as loose clusters or preclusters. However, this can provide useful information because we will be able to look at more matches, albeit loose, which is expected in a cross-species comparison. Although we may not have matched all genes across all species, it is possible to collect the related genes from different species based on BLAST results to perform a multiple alignment and identify the conserved regions flanking SSR sequences. This information may lead to development of new markers that amplify in multiple species. Individual species-specific primers can be designed in cases where a single set of primers could not be designed for two or more species.

Function of the genes that contain SSR motifs

The functions of genes that contain SSRs and the role of the SSR motif in the function of the plant genes was poorly documented in the literature. Expansions of trinucleotide repeats in some human genes were reported to be associated with neurological disorders in man (Sasaki *et al.*, 1996; Sanpei *et al.*, 1996; Pulst *et al.*, 1996; Neri *et al.*, 1996; Pujana *et al.*, 1997). Variation in number of GA/CT repeats in the 5′-UTR of the *waxy* gene is correlated with amylose content in rice (Ayers *et al.*, 1997). In addition, Cho *et al.* (2000) reported 27 genes from rice that contain a SSR in the exons (8), introns (5), 5′-UTR (8) or 3′-UTR regions (5). Our large-scale analysis identified many genes or transcripts that contain SSR motifs and a review of their similarity to known genes indicated that they have a range of functions, such as metabolic enzymes, structural proteins, storage proteins, disease signaling and transcription factors. Our goal is to characterize all the genes that correspond to the SSR-ESTs and analyze their putative function. The collective information can be used for functional analyses of genes, *in silico* mapping and/or development of polymorphic, highly transferable anchor markers for comparative mapping in grasses. The NR sequences from our analyses can be used to predict their map position and putative function by matching to the full genome sequence of rice.

Conclusions

Our analyses indicated that there are many transcripts that contain SSR motifs and that they can be effectively identified and reconstructed into a longer consensus sequence by the strategy we implemented. The SSR-containing transcripts from a species can be used to identify their orthologues in other species and have direct use for comparative mapping. The clustering approach within and across species is a powerful tool to reduce the redundancy in the data set, forming multi-species sequence alignments and extending the sequence information in order to design primers for most of the SSR-ESTs. The sequence information can be used to assign putative map location by matching with the sequences of mapped genes or markers. In addition, when the complete rice genomic sequence is revealed, the SSR-EST sequences can be assigned to a more precise map/physical location in the genome from which comparative map location can be derived for other grass species.

Acknowledgements

We are grateful for the financial support from USDA (NRI Competitive Grant 2001-35301-10612), Hatch project (149419), and support from McKnight Foundation to R.V.K., and support from Syngenta to M.L.R.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389–3402.

Ayers, N.M., McClung, A.M., Larkin, P.D., Bligh, H.F.J., Jones, C.A. and Park, W.D. 1997. Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. Theor. Appl. Genet. 94: 773–781.

Band M.R., J.H. Larson, M. Rebeiz, C.A. Green, D.W., Heyen, J., Donovan, R., Windish, C., Steining, P., Mahyuddin, J.E., Womack and H.A. Lewin. 2000. An ordered comparative map of the cattle and human genomes. Genome Res. 10: 1359–1368.

Becker, J. and Heun, M. 1995. Barley microsatellites: allele variation and mapping. Plant Mol. Biol. 27: 835–845.

Bryan, G.J., Collins, A.J., Stephenson, P., Orry, A., Smith, J.B. and Gale, M.D. 1997. Isolation and characterisation of microsatellites from hexaploid bread wheat. Theor. Appl. Genet. 94: 557–563.

Chakraborty, R., Kimmel, M., Strivers, D.N., Davison, L.J. and Deka, R. 1997. Relative mutation rates at di-, tri- and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA 94: 1041–1046.

Chin, E.C.L., Senior, M.L., Shu, H. and Smith, J.S.C. 1996. Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39: 866–873.

Cho, Y.G., Ishii, T., Temnykh, S., Chen, X., Lipovich, L., McCouch, S.R., Park, W.D., Ayer, N. and Cartinhour, S. 2000. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa*). Theor. Appl. Genet. 100: 713–722.

Eujayl, I., Sorrells, M.E., Baum, M., Wolters, P. and Powell, W. 2000. Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. Theor. Appl. Genet.

Green, P. 1999. SWAT/Crossmatch/PHRAP package, University of Washington. URL: http://www.phrap.org.

Gupta, P.K., Balyan, H.S., Sharma, P.C. and Ramesh, B. 1996. Microsatellites in plants: a new class of molecular markers. Curr. Sci. 70: 45–54.

Herron, B.J., Silva, G.H. and Flaherty, L. 1998. Putative assignment of ESTs to the genetic map by use of the SSLP database. Mammal. Genome 9: 1072–1074.

Kantety, R.V., Zeng, X., Bennetzen, J.L. and Zehr, BE. 1995. Assessment of genetic diversity in dent and popcorn (*Zea mays* L.) inbred lines using inter-simple sequence repeat (ISSR) amplification. Mol. Breed. 1: 365–373.

Korzun, V., Röder, M.S., Wendekake, K., Pasqualone, A., Lotti, C., Ganal, M.W. and Blanco, A. 1999. Integration of dinucleotide microsatellites from hexaploid bread wheat into a genetic linkage map of durum wheat. Theor. Appl. Genet 98: 1202–1207.

La Rota, C.M. 2000. EST clustering for database simplification and candidate gene discovery in rice. M.S. Thesis, Cornell University, New York.

Laurent, P., Elduque, C., Hayes, H., Saunier, K., Eggen, A. and Levéziel, H. 2000. Assignment of 60 human ESTs in cattle. Mammal. Genome 11: 748–754.

Lewin, B. 1994. Genes V. Oxford University Press, New York.

Liu, Z.W., Biyashev, R.M. and Maroof, M.A.S. 1996. Development of simple sequence repeat DNA markers and their integration into a barley linkage map. Theor. Appl. Genet. 93: 869–876.

McCouch, S.R., Chen, X., Panaud, O., Temnykh, S., Xu, Y., Cho, Y.G., Huang, N., Ishii, T. and Blair, M. 1997. Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol. Biol. 35: 89–99.

Miller, R.T., Christoffels, A.G. *et al.* 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res. 9: 1143–55.

Nachit, M., Elouafi, I., Pagnotta, M.A., El Saleh, A., Iacono, E., Labhilili, M., Asbati, A., Azrak, M., Hazzam, H., Benscher, D., Khairallah, M., Ribaut, J., Tanzarella, O.A., Porceddu, E. and Sorrells, M.E. 2001. Molecular linkage map for an intraspecific recombinant inbred population of durum wheat (*Triticum turgidum* L. var. *durum*). Theor. Appl. Genet. 102: 177–186.

Neri, C., Albanese, V., Lebre, A-S, Holbert, S., Saada, C., Bougueleret, L., Meier-Ewert, S., Le Gall, I., Millasseau, P., Bui, H., Giudicelli, C., Massart, C., Guillou, S., Gervy, P., Poullier, E., Rigault, P., Weissenbach, J., Lennon, G., Chumakov, I., Dausset, J., Lehrach, H., Cohen, D. and Cann, H.M. 1996. Survey of CAG/CTG repeats in human cDNAs representing new genes: candidates for inherited neurological disorders. Human Mol. Genet. 5: 1001–1009.

Plaschke, J., Ganal, M.W. and Röder, M.S. 1995. Detection of genetic diversity in closely related bread wheat using microsatellite markers. Theor. Appl. Genet. 92: 1078–1084.

Pujana, M.A., Gratacos, M., Corral, J., Banchs, I., Sanchez, A., Genis, D., Cervera, C., Volpini, V. and Estivill, X. 1997. Polymorphisms at 13 expressed human sequences containing CAG/CTG repeats and analysis in autosomal dominant cerebellar ataxia (ADCA) patients. Human Genet. 101: 18–21.

Pulst, S.-M., Nechiporuk, A., Nechiporuk, T., Gispert, S. Chen, X.-N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., DeJong, P., Rouleau, G.A., Aurburger, G., Korenberg, J.R., Figueroa, C. and Sahba, S. 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nature Genet. 13: 269–276.

Quackenbush, J., Liang F. *et al.* 2000. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucl. Acids Res 28: 141–145.

Ramsay, L., Macaulay, M., degli Ivannissevich, S., MacLean, K., Cardle, L., Fuller, J., Edwards, K.J., Tuvesson, S., Morgante, M., Massari, A., Maestri, E., Marmiroli, N., Sjakste, T., Ganal, M., Powell, W. and Waugh, R. Genetics 156: 1997–2005.

Rebeiz, M. and Lewin, H.A. 2000. COMPASS of 47 787 cattle ESTs. Animal Biotechnol. 11: 175–241.

Röder, M.S., Korzun, V., Wandehake, K., Planschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. 1998. A microsatellite map of wheat. Genetics 149: 2007–2023.

Russell, J., Fuller, J., Young, G., Tomas, B., Taramino, G., Macaulay, M., Waugh, R. and Powell, W. 1997. Discriminating between barley genotypes using microsatellite markers. Genome 40: 442–450.

Salimath, S.S., de Oliveira, A.C., Godwin, I.D. and Bennetzen, JL. 1995. Assessment of genomic origins and genetic diversity in the genus *Eleusine* with DNA markers. Genome 38: 757–763.

Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, T., Ishida, Y., Ikeuchi, T., Koide, R., Saito, M., Sato, A., Tanaka, T., Hanyu, S., Takiyama, Y., Nishizawa, M., Shimizu, N., Nomura, Y., Sagawa, N., Iwabuchi, K., Eguchi, T., Tanaka, H., Takanashi, H. and Tsuji, S. 1996. Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. Nature Genet. 13: 277–284.

Sasaki, T., Billet, E., Petronis, A., Ying, D., Parsons, T., Macci ardi, F.M., Meltzer, H.Y., Lieberman, J., Joffe, R.T., Ross, C.A., McInnis, M.G., Li, S.H. and Kennedy, J.L. 1996. Psychosis and genes with trinucleotide repeat polymorphism. Human Genet. 97: 244–246.

Schug, M.D., Hutter, C.M., Wetterstrand, K.A., Gaudette, M.S., Mackay, T.P.C. and Aquadro, C.F. 1988. The mutation rates of di-, tri-, and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol. 5: 1751–1760.

Schuler, G.D., Boguski, M.S. *et al.* 1996. A gene map of the human genome. Science 274 (5287): 540–546.

Scott, K.D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E.M., Lee, L.S. and Henry, R.J. 2000. Analysis of SSRs derived from grape ESTs. Theor. Appl. Genet. 100: 723–726.

Senior, M.L., Chin, E.C.L., Lee, M. and Smith, J.S.C. 1996. Simple sequence repeat markers developed from maize found in the GenBank database: map construction. Crop Sci. 36: 1676–1683.

Smit, A. 1999. RepeatMasker. University of Washington, Seattle, WA. URL: http://www.phrap.org.

Sorrells, M.E. 2000a. The evolution of comparative plant genetics. In: J.P. Gustafson (Ed.) Genomes. Proceedings 22nd Stadler Symposium (6–8 June 1998, Columbia, MO), Kluwer Academic Publishers, Boston, MA.

Sorrells, M.E. 2000b. Comparative genomics for tef improvement. In: H. Tefera (Ed.) Proceedings of the International Workshop for tef Improvement (13–16 October 2000, Addis Ababa, Ethiopia).

Tautz, D. and Renz, M. 1984. Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. Nucl. Acids Res 12: 4127–4138.

Temnykh, S., Park, W.D., Ayers, N., Cartinhour, S., Hauck, N., Lipovich, L., Cho, Y.G., Ishii, T. and McCouch, S.R. 1999. Mapping and genome organization of microsatellites in rice (*Oryza sativa* Theor. Appl. Genet. 100: 698–712.

The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971–983.