BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3379–3386
doi:10.1093/bioinformatics/bth410



A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences

Mehmet Bilgen†*, Mehmet Karaca**,†*, A. Naci Onus and Ayşe Gül Ince*

Faculty of Agriculture, Akdeniz University, 07059 Antalya, Turkey

Received on April 22, 2004; revised on June 30, 2004; accepted on July 6, 2004
Advance Access publication July 15, 2004

ABSTRACT

Motivation: One of the most interesting features of genomes (both coding and non-coding regions) is the presence of relatively short tandemly repeated DNA sequences known as tandem repeats (TRs). We developed a new PC-based standalone software analysis program, combining sequence motif searches with keywords such as organs, tissues, cell lines or development stages for finding exact, inexact and compound, TRs. Tandem Repeats Analyzer 1.5 (TRA) has several advanced repeat search parameters/options over other repeat finder programs as it does not only accept GenBank, FASTA and expressed sequence tag (EST) sequence files but also does analysis of multifiles with multisequences. Advanced user-defined parameters/options let the researchers use different motif lengths search criteria for varying motif lengths simultaneously. The outputs show statistical results to be evaluated by the user. The discovery of TRs in ESTs could be useful for both gene mapping and association studies and discovering TRs located in coding regions of important genes that are expressed under various conditions of environment, stress, organ, tissue and development stage.

Results: In this paper, we demonstrated applications of TRA using 175 899 ESTs sequences for three *Arabidopsis* spp. downloaded from GenBank. The EST-SSRs/ESTs ratios were found 43.1%, 15.3% and 2.34% in *A.lyrata*, *A.thaliana* and *A.halleri*, respectively. Analysis revealed that organs, tissues and development stages possessed different amounts of repeats and repeat compositions. This indicated that the distribution of TRs among the tissues or organs may not be random differing from the untranscribed repeats found in genomes.

Availability: The program can be obtained free by anonymous FTP from ftp.akdeniz.edu.tr/Araclar/TRA

Contact: mkaraca@akdeniz.edu.tr

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

INTRODUCTION

One of the most interesting features of genomes (both coding and non-coding regions) is the presence of relatively short tandem repeats (TRs). These repeated DNA sequences are found in both prokaryotes and eukaryotes, distributed almost at random throughout the genomes (Jeffreys *et al*., 1985; Heslop-Harrison, 2003). Some of the TRs play important roles in the regulation of gene expression and some others may not have any biological function; however, they are proven to be very beneficial in DNA profiling and genetic linkage analysis studies (Scott *et al*., 2000; Toth *et al*., 2000).

Repeats containing DNA sequences have attracted many researchers since (i) their significant presence in genomic sequences have been shown to be important in the formation of hairpin structures that may provide some structural or replication mechanism (McMurray, 1999; Keniry, 2000; Shafer and Smirnov, 2000), (ii) a growing number of neurological disorders associated with the repeated DNA (Reddy and Housman, 1997; Timchenko and Caskey, 1999) and (iii) their use in DNA-marker technologies, such as microsatellites or simple sequence repeats (SSRs), inter simple sequence repeats (ISSRs) and directed amplification of minisatellite DNA (DAMD–PCR) in marker assisted selection (MAS), positional cloning, identification of quantitative and qualitative loci and mapping for breeding and evolutionary studies (van Belkum *et al*., 1998; Scott *et al*., 2000; Karaca *et al*., 2002). Evidences have lately emerged that some variable number of TRs (VNTRs) and SSR sequences play significant roles in the regulation of transcription, and that some may also influence the translational efficiency or stability of mRNA, or modify the activity of proteins by altering their structure (Klintschar and Wiegand, 2003).

Expressed sequence tags (ESTs) are single-pass DNA sequences usually ∼300–500 nt in length, obtained from mRNA (cDNA) representing genes expressed in a given tissue and/or at a given development stage. A typical EST usually contains only a portion of the coding region (either or both translated or untranslated) of the original gene transcript. One

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

of the useful and interesting applications of ESTs is the study of the gene expression pattern in response to a given organ, tissue or development stage. The composition of a tissue-specific EST population, therefore, offers an overall overview of the expressed genes and, consequently, is a novel tool in gene discovery and in understanding the biochemical pathways involved in physiological responses. ESTs have been mined for single nucleotide polymorphism (SNP) (Schmid *et al*., 2003) and SSR (Thiel *et al*., 2003).

Single sequence repeats (SSRs) are stretches of DNA consisting of exact simple tandemly repeated short motifs of 1–6 bp in length. SSRs are ideal DNA markers because they are highly polymorphic between individuals and highly abundant dispersing evenly throughout the genomes (Klintschar and Wiegand, 2003). SSRs are also inherited in a codominant fashion, fast and easy to assay by PCR using two unique primer pairs flanking the TRs. Moreover, they can serve as sequence-tagged sites for anchoring in genetic and physical maps (Karaca *et al*., 2002). The standard procedure for developing SSRs involves the construction of a small-insert genomic library, the subsequent hybridization with tandemly repeated oligonucleotides and the sequencing of candidate clones; thus, making the process time consuming and labor-intensive. An alternative strategy for the development of SSRs arises from increasing information available in genomic DNA and EST databases. Owing to the rapid increase in sequence information, the generation of EST–SSR becomes an attractive alternative to complement-existing genomic SSR collections (Thiel *et al*., 2003). The development of SSR primer pairs can be performed at significantly reduced costs, as EST–SSRs are free by-product of the currently expanding EST databases. Since ESTs represent the transcribed part of the genome, EST–SSR markers lead to the direct mapping of the genes. SSRs located in coding regions of important genes that are expressed under various conditions of environment, stress, organ, tissue and development stage would also lead to the development of tissue/organ/development stage-specific SSRs and that would be very valuable to understand the repeat function in gene and mapping for breeding and evolutionary studies.

There are several programs to locate repeat strings in sequences such as Tandem Repeats Finder (TRF) (Benson, 1999), REPuter (Kurtz *et al*., 2001), Simple Sequence Repeat Identification Tool (SSRIT) (Kantety *et al*., 2002), Simple Sequence Repeat Finder (SSRF) (Sreenu *et al*., 2003), Search for TRs IN Genomes (STRING) (Parisi *et al*., 2003), Microsatellite Search (MISA) (Thiel *et al*., 2003). Although these repeat finding programs are very useful they have several disadvantages that limit their use. Important limiting aspects of these programs are the number of sequences that programs accept, the length of the repeats they find and acceptable DNA sequence formats. None of these repeat-finding programs informs researchers about the organisms, organs, tissues or cell types or development stages when multisequences or organs are used. Furthermore, they treat a compound repeat as two different exact or inexact repeats. A compound repeat is another kind of TR that contains two or more different TRs united.

A new program will be very useful for those dealing with a huge sequence data and wishing to compare the repeat composition and contents among the organisms, organs, tissue types and development stages. TRA searches for exact TRs and exact compound TR in one of the two modules. The other module of the program searches for exact–inexact TRs and exact–inexact compound repeats. An exact TR can be defined as a single exact tandem repetition of a suitable motif. If an exact TR undergoes a small number of point mutations, it becomes an inexact TR. Variations in repeats possibly take three main forms. Repeat numbers can vary due to repeat unit insertion–deletion (indels). These kinds of repeats are collectively called exact TRs and they result in changes in the length of repeated unit, and therefore can be easily detected by PCR analysis (Klintschar and Wiegand, 2003). Second, base substitutions or small insertions or deletions (indels) may occur within the motif repeats. As base substitutions or equal amounts of insertions and deletions do not change the length of repeating units, thus this type of variation is usually hidden and requires direct sequencing to be detected. These repeats are known as inexact or mismatch containing repeats. A third type of variation can occur at compound repeats that contain two or more different TRs. Repeat number variation at such loci can vary, which can be either detected by PCR analysis or direct sequencing (Saha *et al*., 2003). Studies indicated that some of the compound repeats (whole compound repeats) show higher level of polymorphisms compared to that of exact and inexact repeats (Wuthisuthimethavee *et al*., 2003). In our preliminary studies using some crops, human and chicken ESTs, we found that there exists two types of compound repeats, exact and inexact compound repeats in EST database.

In this paper, we described TRA program written in C++ using Microsoft Visual C++ software running on Windows 98, Windows NT, Windows ME and Windows XP. The program and the sample datasets are available free of charge and can be obtained by anonymous FTP from ftp.akdeniz.edu.tr/Araclar/TRA. Main aim of this study was to develop a PC-based program for finding and characterizing EST–SSRs and TR–ESTs specific for organisms, organs/tissues/development stage in terms of frequency and distributions for further analysis.

MATERIALS AND METHODS

A total of 20 685 791 EST sequences from GenBank (ftp://ftp.ncbi.nih.gov/repository/dbEST/) were scanned and a total of 175 899 ESTs from *Arabidopsis thaliana*, *A.lyrata* and *A.halleri* subspp. *halleri* were processed. In this study using the exact module of the program, we searched for EST–SSRs. SSRs are considered to contain motifs that are

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

between 1 and 6 nt in size. The minimum motif length criteria were defined as being 10 repeats for mononucleotides, 6 repeats for dinucleotides and 5 repeats for all higher-order motif length according to Thiel *et al*. (2003). A total of 10 organs/tissues/development stages were used as keywords. To identify unique keywords representing only one particular tissue, we used a homemade program (unpublished data) that searched for annotated ESTs and picked the keywords used in this analysis. Whole analysis of 175 210 ESTs from *A.thaliana*, 561 ESTs from *A.lyrata* and 128 ESTs from *A.halleri* subspp. *halleri* has been completed in 3 min and 32 s using a standard PC (Pentium 4™ 1.4 GHz CPU, 396 MB DD RAM, 80 GB, 7200 RPM HDD). This indicated that TRA can be used for analyzing huge datasets, however, running time will be dependant on the search parameters/options and computer hardware used.

For InExact module of the TRA, we used the score of 50 for alignment and limited the motif lengths up to 200. Our preliminary studies indicated that motif length greater than 200 is very rare or inexistent in EST database. We strongly recommend that users interested in inexact repeats should aware of the fact that motif lengths longer than 200 bp will take hours to be analyzed by the InExact module for data consisting of multifiles with multisequences. However, single files can be analyzed in minutes. Simple regression coefficient analysis was utilized to investigate whether the motif contents as well as motif lengths of organisms, organs/tissues/development stages were dependent on the EST numbers and we also investigated the occurrence of exact, inexact and compound TRs in ESTs.

ALGORITHM

TRA uses two different algorithms independently for detecting the repeats in DNA sequences. The main criteria required for a computer program for the repetitive structure of a large number of sequences should first: (i) identify both exact and inexact (mismatch-containing) TRs, (ii) be fast and (iii) allow to perform all analyses in one single run. The Exact module of the TRA is fast but it misses inexact repeats. The InExact module captures all the exact and inexact TR, but it does not allow to search for a particular motif length and repeat number. Detailed discussion was provided in the following sections of the manuscript.

Exact Repeats (SSRs): TRA uses a simple algorithm for detecting exact repeats. Briefly, TRA searches for S_n, a string of repeated units in a DNA sequence $w_n \cdot S_1 = w_1[i_1, j_1]$ symbolizes the S_1 starting with the i_1-th and ending with the j_1-th bases of the DNA sequence w_1. The distance between i_1 and j_1 will be equal to $m_1 \times r_1$ where m_1 and r_1 refer to a type of DNA motif length and the number of repeats in S_1 string of each w of a fixed length, respectively. When applicable, strings in a sequence of w are referred to as S_1, S_2, S_3 and S_n for each consecutive string in a w. The distance between S_1 and S_2 is referred to as d_1, and the distance between S_2 and S_3 is d_2 (S_n, d_n). Currently, TRA allows w in a maximum of 1 Mb in length with an infinite number of w. TRA calls the repeats as compound repeats when d equals to 0.

Inexact Repeats (Mismatch Repeats or Minisatellites): TRA basically uses an algorithm defined by Parisi *et al*. (2003) performed by means of a dynamic programming procedure for locating the repeats in inexact module. Repeat finding algorithm of STRING: finding TRs in DNA sequences was kept the same but a compound repeat finding option was included as stated previously. For inexact and exact TR analyses, an alignment score between 30 and 150, can be selected, higher scores will denote 'better' alignments.

OVERVIEW OF TRA

TRA is involved in locating and characterizing string(s) containing repeat(s) in a given DNA sequence formatted in FASTA, GenBank or EST sequence format. TRA performs repeats finding and classification tasks in basically four major steps as follows: (i) it searches the user-defined organism(s) (or user can select all organisms in the dataset) and/or keywords (organs, cell lines, tissue types or development stages) analyzing the whole dataset provided in a data folder; (ii) isolates TRs by determining their types, lengths and sequence locations in string within DNA sequences; (iii) characterizes the repeats containing sequences based on the user-defined parameters/options; and (iv) displays the results according to the user's parameters/options. The results of keyword (source) searches herein we called digital differential display (DDD) show the repeat frequency and contents among the keywords. With DDD, users will find TRs in only organs/tissues or developmental lines they are interested in.

Tandem repeats analyzer (TRA) has two modules, Exact and InExact, to detect and locate the TRs. The two modules have advantages and disadvantages. Mismatch-containing TRs will not be detected in the Exact module of TRA and this is the major disadvantage of the Exact module over the InExact module of TRA. However, the Exact module has more parameters/options. Users would not only specifically search a specific motif length with a defined repeat number but also search several motif lengths with different repeat numbers utilizing the Detail options. Furthermore, the Exact module is very fast and could analyze a huge data in seconds. The user who would like to quickly analyze exact TRs and design primer pairs for SSR, ISSR and DAMD–PCR for breeding and fingerprinting studies would like to utilize the Exact module of the TRA. Since most of the longer TRs are almost never perfectly conserved, Exact TR module of TRA will fail in detecting these structures. InExact module of the TRA locates and detects those mismatch-containing repeats and exact repeats. However, users specifically interested in a specific motif length such as 18 nt-long would have to analyze all the motif length up to 18 nt. The Exact module of TRA has

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

Table 1. Distribution of EST–SSRs among 10 *Arabidopsis* spp. organs/tissues/development stages

Sources/keywords	EST (processed)	EST–SSRs	Repeat percentage	Repeat strings	Repeat index[a]
Root	26 504	4215	15.90	5148	0.194
Rosette	19 650	2523	12.84	2835	0.144
Leaf	12 836	889	6.93	956	0.074
Silique	18 896	2762	14.62	3162	0.167
Adult plants, mixed stresses	12 530	3740	29.85	6060	0.484
Developing seeds	11 238	637	5.67	704	0.063
Seedling	7726	1649	21.34	2079	0.269
Inflorescence	2957	960	32.47	1336	0.452
Adult vegetative tissue	2417	433	17.91	516	0.213
Hormone-treated callus	2317	457	19.72	572	0.247

[a]Repeat Index is a numerical value showing how many repeat strings were found among the all sequences processed and it can be calculated by dividing the value of repeat string numbers to the number of ESTs processed. Repeat percentage (%) shows how many sequences were processed and how many of them contained repeats.

a 'Detail' button for the users who would like to set other minimum repeat numbers criteria for different motif lengths ranging from 1 to 10, and also those users who like to conduct simultaneous analyses for different motif lengths. Users interested in finding a specific motif length distribution in a DNA sequence population do not have to use this 'Detail' option.

Comprehensive help file was provided within the program. TRA differs from the other repeat finder programs because it (i) simultaneously analyses multifiles with multisequences; (ii) accepts multisequence formats, such as EST, GenBank and FASTA formats; (iii) analyzes the distributions of the SSRs and other TRs among the organism(s), organ(s), cell line(s) and development stage(s); and (iv) provides the data on the type of SSRs and TRs. TRA shows analyzed results in tabular formats: distribution of repeats among organisms, organisms by keywords and keywords (source). The results showing the organism by keyword show how many keywords were found in each organism. This will be useful when many organisms and many keywords are simultaneously used. Repeat containing sequences are grouped according to: (i) motif length, (ii) motif length and motif content, (iii) motif length, repeat number and motif content, and (iv) compound repeats. By clicking the related button, users would see the individual repeats as per the user-defined options/parameters filled. We also included another table showing the used options/parameters. This will be useful for comparison of the individual runs performed at different dates.

There is another unique option in TRA showing the sources not provided within the keyword search file. This is useful to see the distribution of the defined repeats (exact and inexact) among other sources. TRA recalls all the parameters/options used in the previous run, making it much easier to start a new analysis. The minimum motif length that program currently analyze can vary from 1 to 1000 nt. TRA also has [poly(A/T)] filter for either both upstream and/or downstream of the repeat ends. Sequences containing poly(A/T) tails (in either one or both ends) are not considered A or T repeats by the program. At least 1000 scientific names of organisms can be given in organism list file. Keywords (organs, tissue types, cell lines and development stages) can be given in the keyword list file.

RESULTS

Analyses used two different modules, Exact and InExact TR modules. The repeat finding criteria of the Exact module used were as per Thiel *et al*. (2003). This was done in order to compare the findings of our study with that of the previous studies. The repeat finding criteria of the InExact module has been defined in detail by Parisi *et al*. (2003). However, these criteria have not been applied to ESTs or any plant genomes.

Exact repeats (SSRs)

A total of 175 899 ESTs were screened for EST–SSR and 26 996 EST–SSR (15.4%) were detected. There were considerable variations in the amounts of EST–SSRs among the three *Arabidopsis* spp. The highest numbers of EST–SSRs were found in *A.lyrata* (561 ESTs and 43.1% of which were EST–SSRs). *A.thaliana* consisted of 175 210 ESTs and the EST–SSRs ratio was 15.3%. *A.halleri* on the other hand had the lowest ESTs (128) and the lowest EST–SSRs (2.34%). Overall, the occurrence of the individual SSR motifs was not evenly distributed within and between the *Arabidopsis* spp: 21 801 (76.6%) were mononucleotide, 4702 (14.2%) dinucleotide, 6386 (19.4%) trinucleotide, 81 (0.2%) tetranucleotide, 6 penta- and 23 hexa-nucleotide microsatellites.

The results revealed that organs, tissues and cell lines possessed different amounts of repeats indicating the distribution of EST–SSRs among the tissues or organs were not random differing from the untranscribed repeats found in the genomes. SSRs were the most dominant in ESTs derived from inflorescence (32.5%) followed by (29.9%) in *Arabidopsis* spp. adult plant. These plants have been treated 24 h with various stresses (Schmid *et al*., 2003). On the other hand, developing seeds contained the lowest amounts of EST–SSRs (Table 1). Simple

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

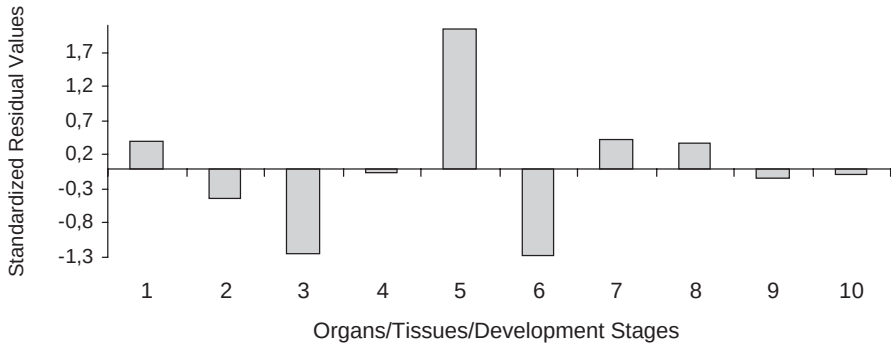


Fig. 1. Standardized residual values (negative or positive) indicated that the repeat contents (EST–SSRs) of the organs/tissues/development stages of *Arabidopsis* spp. showed considerable variations. Keywords (organs/tissues/development stages): 1, root; 2, rosette; 3, leaf; 4, siliques; 5, mixed stressed adult plants; 6, developing seeds; 7, seedlings; 8, inflorescence; 9, vegetative tissue; and 10, callus (hormone-treated).

regression coefficient analysis indicated that repeat contents of several organs/tissues/developmental stages were either more or less than expected (Fig. 1). The results indicated that EST–SSR contents of adult plants that have been given various stresses were much higher, and the repeat contents of the leaf and developing seed ESTs were much lower than expected, independent of the ESTs studied.

The distribution of A/T motifs [either poly(A)s or poly(T)s] among the keywords (organs/tissues/development stages) did not only differ in amounts but also differed in compositions varying from 99% in leaf tissues to 79% in callus. With the exception of seedling EST–SSRs, which were rich in AT/TA (33.1%), other keywords were rich in AG/TC motifs ranging from 55.5% in developing seeds to 31.6% in seedlings. For trinucleotide SSRs, callus, vegetative tissues and rosette EST–SSRs were rich in CTT/GAA, whereas all other keywords were rich in AGA/TCT EST–SSRs (Table 2). These findings indicated that distributions of EST–SSRs motifs were not random and EST derived SSR primer pairs could be used in fingerprinting of organisms, organ and sources. The composition and distribution of SSRs and/or TRs specific to a keyword (source) could be used to differentiate experimentally a tissue from another by making specific primer pairs. For instance, a cDNA library constructed from root tissues should contain root-specific SSRs and/or TRs.

Exact and inexact TRs

Again a total of 175 899 ESTs were screened for TRs containing ESTs (TR–ESTs) and 11 929 TR–EST (6.8%) were identified. TR–ESTs consisted of mononucleotide to 144 nt. The distribution and the occurrence of the TR–ESTs were not random at the keyword levels (Table 3). Simple regression coefficient analysis indicated that TR–EST contents of keywords were again either more or less than expected (Fig. 2). TR–EST contents of stressed adult plants were quite high as in EST–SSRs. Most of the TR–ESTs with unit length a multiple of 3 (9, 12, 15, etc.) bp were the most frequent among units between 7 and 144 (89%), probably reflecting that they are part of coding sequences.

Repeat polymorphisms in TR–ESTs (different ESTs containing the same TR) generally decreased as the motif lengths increased. For instance, motif length of 24 (GGGATGCAGCACCAGGGCGGGCAC) varied from 2 to 15 repetitions, whereas motif length of 84 varied from 2 to 5 repetitions. Variations in the EST–SSRs in most cases were due to the repeat number variations, whereas variations in the TR–ESTs were mostly due to mismatches or base substitution in motifs. There were just 22 compound repeats of which 63.6% were exact and 36.4% were inexact compound repeats.

DISCUSSION

We developed and implemented a PC program for identifying SSRs (exact TRs algorithm) and modified an exact–inexact repeats algorithm for studying tandemly repeated DNA sequences.

Earlier approaches, although very powerful, miss either compound repeats or/and detect only exact or inexact TRs in a particular data format. Moreover, they do not implement keywords and organisms options. TRA detects compound microsatellite loci, i.e. those containing stretches of two or more different repeats, which appear to comprise ∼10% of SSRs (Bull *et al*., 1999). Since compound SSRs show significant variation, they are very valuable tools in plant breeding studies (Klintschar and Wiegand, 2003). TRA accepts unlimited numbers of DNA sequences, therefore, offers researchers to compare many sequences at once and informs about the distribution of TR and SSRs in many organisms. TRA can also analyze the whole DNA sequences of chloroplast and mitochondria genomes and those genomes <1 Mb in length. This is disadvantage of TRA since it cannot proceed large genome sequences, and therefore, it is not appropriate for whole genome analysis. The DDD option of TRA could be utilized to locate the SSRs in those genes expressed in various stress conditions using the Exact and InExact TR modules.

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

Table 2. Distribution of EST–SSRs motifs among *Arabidopsis* spp. organs/tissues/development stages

Keywords	Mono	%	Di	%	Tri	%	Tetra	%
Root	A/T	94.2	AG/TC	42.9	AGA/TCT	16.16	GAGC	33.3
			CT/GA	26.5	AAG/TTC	15.93	GTTT	66.7
			AT/TA	15.1	CTT/GAA	13.13		
Rosette	A/T	91.5	AG/TC	39.1	CTT/GAA	17.8	ATCT/TAGA	46.2
			CT/GA	27.6	AGA/TCT	16.9	ATAG	23.1
			AT/TA	22.6	AAG/TTC	10.8	AAAC	7.7
Leaf	A/T	99	AG/TC	33.7	AGA/TCT	16.8	ATCT	75
			AT/TA	32.6	CTT/GAA	13	ATAG	12.5
			CT/GA	20.9	CAT	9.16	CTTT	12.5
Siliques	A/T	98.1	AG/TC	38.9	AGA/TCT	20.1	AAAG	40
			AT/TA	33.1	CTT/GAA	14.2	AAAC	20
			CT/GA	17.9	AAG/TTC	10.1	CTTT	20
Mixed stressed adult plants	A/T	90.2	AG/TC	50.6	AGA/TCT	19.1	ATCT	50
			CT/GA	27.3	CTT/GAA	13.1	CAAT	25
			AT/TA	12.9	AAG/TTC	12.8	CGAG	25
Developing seeds	A/T	95.8	AG/TC	55.5	AGA/TCT	20.7	AGAA	100
			CT/GA	34.5	CTT/GAA	20.2		
			AC/TG	4.1	AAG/TTC	14.1		
Seedlings	A/T	91.6	AT/TA	33.1	AGA/TCT	24.1	ATCT	50
			AG/TC	31.6	TCA	15.6	TTTG	33.3
			CT/GA	28.6	AAG/TTC	10.5	GAGC	16.7
Inflorescence	A/T	94.6	AG/TC	51.3	AGA/TCT	29.2		
			CT/GA	28.2	CTT/GAA	17.7		
			AT/TA	8.9	AAG/TTC	8.5		
Vegetative tissue	A/T	85.6	AG/TC	50.4	CTT/GAA	19.1	AAAC	50
			CT/GA	31.7	AGA/TCT	17.6	ATAC	50
			CA/GT	8.9	AAG/TTC	15.2		
Callus (hormone-treated)	A/T	78.9	AG/TC	49.3	CTT/GAA	19.1	CATA	33.3
			CT/GA	24.7	AAG/TTC	13.4	TTCT	33.3
			AT/TA	12.3	AGA/TCT	11.5	TTTC	33.3
Other keywords	A/T	94.6	AG/TC	39.1	AGA/TCT	17.3	ATAG	29.7
			CT/GA	26.7	CTT/GAA	15.9	CTTT/GAAA	13.5
			AT/TA	24.1	AAG/TTC	13.2	AAAG	13.5

Table 3. Distribution of *Arabidopsis* spp. tandemly repeated ESTs (TR–ESTs) among 10 *Arabidopsis* spp. organs/tissues/development stages

Sources/keywords	EST (processed)	TR–ESTs	Repeat percentage	Repeat strings	Repeat index
Root	26 504	2023	7.63	2238	0.084
Rosette	19 650	884	4.50	930	0.047
Leaf	12 836	395	3.08	407	0.032
Silique	18 896	880	4.66	910	0.048
Adult plants, mixed stresses	12 530	2588	20.65	2816	0.225
Developing seeds	11 238	354	3.15	406	0.036
Seedling	7726	1113	14.41	1194	0.155
Inflorescence	2957	664	22.46	738	0.250
Adult vegetative tissue	2417	150	6.21	165	0.068
Hormone-treated callus	2317	157	6.78	173	0.075

Repeat Index is a numerical value showing how many repeat strings were found among the all sequences processed and it can be calculated by dividing the value of repeat string numbers to the number of ESTs processed. Repeat percentage (%) shows how many sequences were processed and how many of them contained repeats.

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

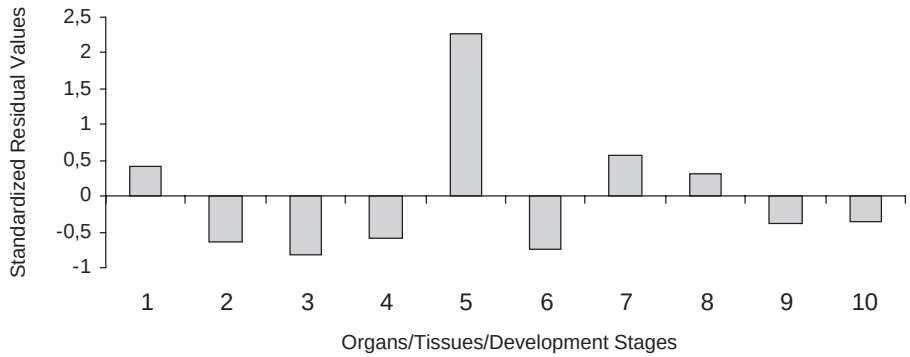


Fig. 2. Standardized residual values (negative or positive) indicated that the TRs in ESTs (TR–ESTs) of the organs/tissues/development stages of *Arabidopsis* spp. showed considerable variations. Keywords (organs/tissues/development stages): 1, root; 2, rosette; 3, leaf; 4, siliques; 5, mixed stressed adult plants; 6, developing seeds; 7, seedlings; 8, inflorescence; 9, vegetative tissue; and 10, callus (hormone-treated).

The Exact module of TRA misses mismatch-containing TRs but InExact module detects both mismatch-containing and exact TRs. However, when TRs are compared with the exact and inexact TRs run under the two different modules, some discrepancies are evident. This is probably due to fact that (1) the repeat finding criteria which are due to the different algorithms applied and (2) some of the exact repeats could be converted into inexact repeats, indels or mismatches may extend the exact repeat either in both ends. In our analyses, because of the above-mentioned reasons we found only 11 929 TR–ESTs detected although there were 26 996 SSR–ESTs. However, the main source of the discrepancy was due to the repeat finding criteria. The repeat finding criteria of the Exact module were as per Thiel *et al*. (2003) in order to compare the findings of our study with that of the Thiel *et al*. (2003).

The occurrences of SSRs or TRs in transcribed genes (ESTs) accepted by scientific community and the use of EST-derived SSR in Differential Display has also been reported (Saha *et al*., 2003). In this study, we confirmed that a significant portion of ESTs contained SSR and other TRs in model flowering plant, *Arabidopsis* spp. Moreover, we also reported that the distribution of TRs among the tissues or organs may not be random. We used a SSR-based Differential Display study as per Saha *et al*. (2003) to confirm the preferential expression of some SSRs in particular tissues. Our preliminary findings indicated that some SSR expressed in cotton leaves and some expressed in fiber tissues. Further analyses will be helpful to confirm the biased distribution of repeat containing ESTs among organs and tissues. In order to obtain an idea about the putative functions of SSR- or TR-containing genes (ESTs), TRA results could be compared with the SWISSPIRPLUS protein database using the BlastX2 program after being analyzed by BLAST program (Altschul *et al*., 1997). However, BLASTing of TRs does not work well: repeated units often produce multiple matches due to their low complexity, making the results analysis very difficult. Further studies can also be performed using BlastN searches to address the question which proportion of the EST–SSR loci have homologies to genomic SSRs or TRs within and between species.

The results of the present study indicated that SSR motifs were not randomly distributed in *Arabidopsis* spp. coding genes. Similar results about unevenly distributed individual SSR motifs occurrence in ESTs have also been observed in *Hordeum vulgare* L. by Thiel *et al*. (2003). The proportion of specific motif contents in *Arabidopsis* spp. SSRs were different from the results of Thiel *et al*. (2003). This discrepancy is probably due to the number of ESTs used between our experiments (175 210) and that of Thiel *et al*. (2003) who implemented only 24 595 ESTs. Using the simple regression coefficient analysis, our results clearly indicated that there was a strong relationship between the SSR-containing ESTs and the number of ESTs used. Mononucleotides, poly(A) and poly(T) repeats were the most common (93.6%), which was not surprising due to the fact that eukaryotic mRNAs contain poly(A) tails. Dimeric SSRs, the motifs AG/TC (41.6%) and CT/GA (26.6%) were the most common ones, whereas CA/GT and CG/GC microsatellites were present only at low abundance (4.8 and 0.12%, respectively). These findings are in accordance with earlier reports (Chin, 1996; Temnykh *et al*., 2000; Cardle *et al*., 2000; Thiel *et al*., 2003). The most common tetrameric microsatellite motifs were ATCT/TAGA (22.2%), ATAG (17.3%) and AAAC/TTTG (12.3%). The dominance of trimeric EST–SSRs and TR–ESTs in *Arabidopsis* EST database can be explained by the suppression of non-trimeric SSRs in coding regions (ESTs) probably due to the risk of frameshift mutations (Metzgar *et al*., 2000). Trimeric motifs, TAA (0.13%), TAG (0.06%) and TGA (3.1) rarely appeared, probably because they code for stop codons that have a direct effect on protein synthesis (Chin, 1996).

In conclusion, TRA will be very useful to those researchers interested in (i) identifying the repeat (exact, inexact and compound) containing EST those from small genomes (some bacteria with <1 Mb genome size and most of the chloroplast and mitochondria genomes) for further studies

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021

[for instance; instead of genomic DNA mapping, transcribed gene (EST) mapping on the chromosomes will be very helpful in breeding studies], (ii) data mining (*in silico*) for repeat containing sequences and characterizing the repeats, compositions and distributions among the organisms [for instance; Vergnaud and Denoeud (2000) describe a TRs database that allow the comparison of TR distributions between genomes], among the organs, development stages or tissues. Information gained from such studies will be very useful for understanding the expression, regulation and evolution of repeats in DNA. In our preliminary studies, we utilized TRA to develop EST–SSR and TR–ESTs primer pairs (using one of the primer design programs) and amplified genomic DNAs. To date, there is a limited number of research on TR for their theoretical comparison and importance in ESTs. Therefore, further analysis using TRA may be very useful to obtain advanced knowledge about the functions of the repeated DNA sequences in transcribed genes of organs/tissue types/cell lines in various organisms.

ACKNOWLEDGEMENTS

This research was funded by the Scientific Research Projects Administration Unit of Akdeniz University (Project No: 2003.01.0104.001).

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,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. *Nucleic Acids Res.*, **25**, 3389–3402.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Bull,L.N., Pabon-Pena,C.R. and Freimer,N.B. (1999) Compound microsatellite repeats: practical and theoretical features. *Genome Res.*, **9**, 830–838.

Cardle,L., Ramsay,L., Milbourne,D., Macaulay,M., Marshall,D. and Waugh,R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*, **156**, 847–854.

Chin,E.C.L. (1996) Maize simple repetitive DNA sequences: abundance and allele variation. *Genome*, **39**, 866–873.

Heslop-Harrison,J.S. (2003) Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. *Chromosome Res.*, **11**, 241–253.

Jeffreys,A.J., Wilson,V. and Thein, S.L. (1985) Hypervariable minisatellite regions in human DNA. *Nature*, **314**, 67–73.

Kantety,R.V., La Rota,M., Matthews,D.E. and Sorrells,M.E. (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.*, **48**, 501–510.

Karaca,M., Saha,S., Jenkins,J.N., Zipf,A., Kohel,R. and Stelly,D.M. (2002) Simple sequence repeat (SSR) markers linked to the *Ligon Lintless* (Li_1) mutant in cotton. *J. Hered.*, **93**, 221–224.

Keniry,M.A. (2000) Quadruplex structures in nucleic acids. *Biopolymers*, **56**, 123–146.

Klintschar,M. and Wiegand,P. (2003) Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. *Forensic Sci. Int.*, **135**, 163–166.

Kurtz,S., Jomuna,V.C., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.

McMurray,C.T. (1999) DNA secondary structure: a common and causative factor for expansion in human disease. *Proc. Natl Acad. Sci. USA*, **96**, 1823–1825.

Metzgar,D., Bytof,J. and Wills,C. (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.*, **10**, 72–80.

Parisi,V., Fonzo,V. D. and Aluf-Pentini,F. (2003) STRING: finding tandem repeats in DNA sequences. *Bioinformatics*, **19**, 1733–1738.

Reddy,P.S. and Housman,D.E. (1997) The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.*, **9**, 364–372.

Saha,S., Karaca,M., Jenkins,J.N., Zipf,A.E., Reddy,O.U.K., Pepper,A.E. and Kantety,R. (2003) Simple sequence repeats as useful resources to study transcribed genes of cotton. *Euphytica*, **130**, 355–364.

Schmid,K.J., Sorensen,T.R., Stracke,R., Torjek,O., Altmann,T., Mitchell-Olds,T. and Weisshaar,B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana. Genome Res.*, **13**, 1250–1257.

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.

Shafer,R.H. and Smirnov,I. (2000) Biological aspects of DNA/RNA quadruplexes. *Biopolymers*, **56**, 209–227.

Sreenu,V.B., Vishwanath,A., Nagaraju,J. and Nagarajaram,H.A. (2003) MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res.*, **31**, 106–108.

Temnykh,S., Park,W.D., Ayres,N., Cartinhour,S., Hauck,N., Lipovich,L., Cho,Y.G., Ishii,T. and McCouch,S.R. (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.*, **100**, 697–712.

Thiel,T., Michalek,V. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.*, **106**, 411–422.

Timchenko,L.T. and Caskey,C.T. (1999) Triplet repeat disorders: discussion of molecular mechanisms *Cell. Mol. Life Sci.*, **55**, 1432–1447.

Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.

van Belkum,A., Scherer,S., van Alphen,L. and Verbrugh,H. (1998) Short sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.*, **62**, 275–293.

Vergnaud,G. and Denoeud,F. (2000) Minisatellites: mutability and genome architecture. *Genome Res.* **10**, 899–907.

Wuthisuthimethavee,S., Lumubol,P., Vanavichit,A. and Tragoonrung,S. (2003) Development of microsatellite markers in black tiger shrimp (*Penaeus monodon Fabricius*). *Aquaculture*, **224**, 39–50.

Downloaded from https://academic.oup.com/bioinformatics/article/20/18/3379/202188 by Akdeniz Universitesi user on 20 September 2021