RESEARCH ARTICLE

Exact Tandem Repeats Analyzer (E-TRA): A new program for DNA sequence mining

MEHMET KARACA*, MEHMET BILGEN, A. NACI ONUS, AYSE GUL INCE and SAFINAZ Y. ELMASULU

Akdeniz University, Faculty of Agriculture, 07059 Antalya, Turkey

Abstract

Exact Tandem Repeats Analyzer 1.0 (E-TRA) combines sequence motif searches with keywords such as 'organs', 'tissues', 'cell lines' and 'development stages' for finding simple exact tandem repeats as well as non-simple repeats. E-TRA has several advanced repeat search parameters/options compared to other repeat finder programs as it not only accepts GenBank, FASTA and expressed sequence tags (EST) sequence files, but also does analysis of multiple files with multiple sequences. The minimum and maximum tandem repeat motif lengths that E-TRA finds vary from one to one thousand. Advanced user defined parameters/options let the researchers use different minimum motif repeats search criteria for varying motif lengths simultaneously. One of the most interesting features of genomes 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 genome. Some of the tandem repeats play important roles in the regulation of gene expression whereas others do not have any known biological function as yet. Nevertheless, they have proven to be very beneficial in DNA profiling and genetic linkage analysis studies. To demonstrate the use of E-TRA, we used 5,465,605 human EST sequences derived from 18,814,550 GenBank EST sequences. Our results indicated that 12.44% (679,800) of the human EST sequences contained simple and non-simple repeat string patterns varying from one to 126 nucleotides in length. The results also revealed that human organs, tissues, cell lines and different developmental stages differed in number of repeats as well as repeat composition, indicating that the distribution of expressed tandem repeats among tissues or organs are not random, thus differing from the un-transcribed repeats found in genomes.

[Karaca M., Bilgen M., Onus A. N., Ince A. G. and Elmasulu S. Y. 2005 Exact Tandem Repeats Analyzer (E-TRA): A new program for DNA sequence mining. *J. Genet.* **84**, 49–54]

Introduction

One of the most interesting features of prokaryotic and eukaryotic genomes (both coding and non-coding regions) is the presence of relatively short perfect tandemly repeated DNA sequences. These repeated DNA sequences are distributed almost at random throughout the genome (Huang *et al.* 1998; Richard *et al.* 1999; Heslop-Harrison 2003). Repeats containing DNA sequences have attracted much attention from researchers since (i) they play important roles in the formation of hairpin structures that may provide some structural or replication mechanism (Huang *et al.* 1998; Richard *et al.* 1999; McMurray 1999), (ii) they are often associated with neurological disorders (Lalioti *et al.* 1997; Wren *et al.* 2000), and (iii) they are used as DNA markers, such as microsatellites or Simple Sequence Repeats (SSR), Inter Simple Sequence Repeats (ISSR) 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

*For correspondence. E-mail: mkaraca@akdeniz.edu.tr.

Keywords. expressed sequence tag (EST); motifs; simple sequence repeats (SSR); bioinformatics; Exact Tandem Repeats Analyzer.

studies (van Belkum *et al*. 1998; Scott *et al*. 2000; Karaca *et al*. 2002a,b). Recent evidence also suggests that some Variable Number of Tandem Repeats (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 about 300–500 nucleotides in length, obtained from mRNA (cDNA) representing genes expressed in a given tissue and/or at a given development stage (Bilgen *et al*. 2004). A typical EST usually contains only a portion of the coding region (either translated or untranslated, or both) of the original gene transcript. One of the useful applications of ESTs is in the study of the gene expression pattern in a given organ, tissue or development stage in response to a particular treatment. 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 also been mined for Single Nucleotide Polymorphisms (SNP) (Schmid *et al*. 2003) and SSR (Thiel *et al*. 2003; Ince *et al*. 2004).

Microsatellites or SSRs are stretches of DNA consisting of exact simple tandemly repeated short motifs of 1–6 base pairs in length. SSRs are one of the best DNA markers because they are highly polymorphic, inherited in a co-dominant fashion, and highly abundant, being dispersed evenly throughout the genome (Klintschar and Wiegand 2003). They can serve as sequence-tagged sites for anchoring in genetic and physical maps (Karaca *et al*. 2002a). The standard procedure for developing SSRs involves the construction of a small-insert genomic library, its subsequent hybridization with tandemly repeated oligonucleotides, and the sequencing of candidate clones. Unfortunately, this process is time consuming and labour-intensive. An alternative strategy for development of SSRs arises from increasing information available in DNA sequence databases. Due to the rapid increase of sequence information, the generation of DNA database derived SSRs 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 database derived SSRs are a free by-product of the currently expanding EST databases (Ince *et al*. 2004).

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 Tandem Repeats IN Genomes, STRING (Parisi *et al*. 2003), and Microsatellite Search, MISA (Thiel *et al*. 2003). Although these repeat finding programs are 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. With the exception of Tandem Repeats Analyzer (Bilgen *et al*. 2004), none of these repeat finding programs informs researchers about the distribution of repeats among organisms, organs, tissues, cell types or development stages when multi-sequences or organs are used.

Repeat types found in DNA sequences (GenBank databases) have been described in Bilgen *et al*. (2004). Briefly, an exact tandem repeat is a single exact tandem repetition of a suitable motif. If an exact tandem repeat undergoes a small number of point mutations, it becomes an inexact tandem repeat. A third type of variation can occur at compound repeats that contain two or more different tandem repeats. Bilgen *et al*. (2004) reported two different kinds of compound repeats, exact compound and inexact compound repeats. Our preliminary studies using GenBank DNA databases indicated that other kinds of compound repeats in DNA sequences also existed. These repeats cannot be detected by TRA and other programs (Bilgen *et al*. 2004). These repeats are termed in this paper as compound, imperfect, and extended compound repeats. Compound repeats are those repeats with two or more repeat strings run of the same or different uninterrupted repeats shown as $(AGAAG)_{r1}$ $(AGATAA)_{r2}$. Imperfect repeats are those sequences having at least two or more exact simple repeats separated by non-repeated nucleotides varying in size, shown by $(TCTTC)_{r1}CACATAA-(AGAAG)_{r2}$ (CACATAA nucleotides are non-repeated sequences in the given example). Extended compound repeats are sequences having at least two or more compound repeats, but one of the compound repeats is interrupted by non-repeating units of adjacent sequences shown as $(CTTCT)_{r1}(AGAAG)_{r2}TCTTATGA(TATA)_{r3}$.

In this paper, we describe the Exact Tandem Repeats Analyzer (E-TRA) program written in C++ using Microsoft Visual C++ software. The program can be run on Windows 98, Windows NT, Windows ME and Windows XP. The program, and the sample data sets, are available free of charge and can be obtained by anonymous FTP from ftp.akdeniz.edu.tr/Araclar/e-TRA, or from the authors. The main aims of this study were (i) to develop a PC-based program for finding and characterizing DNA sequences specific to organisms or organs/tissues/development stages in terms of frequency and distributions for further analysis, and (ii) to demonstrate the use of E-TRA, using 5,465,605 human EST sequences derived from 18,814,550 GenBank EST sequences. The annotation files were downloaded on June 10, 2004. These ESTs are non-redundant, annotated files available in GenBank.

Materials and methods

A total of 18,814,550 Expressed Sequence Tags (ESTs) derived from publicly available GenBank data (ftp://ftp.ncbi.nih.gov/repository/dbEST/) were scanned and 5,465,605 human ESTs were processed. We used all the human ESTs available in the GenBank on 10 June, 2004. A total of 25 organ/tissue type/cell lines were used as keywords. The whole analysis was completed in 10 h using a standard PC (Pentium 4^{TM} 1.4 GHz CPU, 396 MB DD RAM, 80 GB, 7200 RPM HDD). This indicates that E-TRA can be used for analyzing huge data sets in a very reasonable time-frame. However, running time will be dependent on the search parameters/options and computer hardware used.

E-TRA uses one of the algorithms of TRA described in Bilgen *et al*. (2004). Briefly, E-TRA searches for S_n, a string of repeated units in a DNA sequence w_n. $S_1 = w_1$ $[i_1, j_1]$ 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 therefore 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, $S_3 \ldots S_n$ for each consecutive string in a sequence w. The distance between S_1 and S_2 is referred to as d_1, the distance between S_2 and S_3 is d_2 and so on till S_n, d_n. Currently, TRA allows a maximum length of 1 Mb for any w, with an infinite number of sequences w. E-TRA calls the repeats as compound repeats when d equals 0. The value of d can be provided by the user. E-TRA can detect compound, imperfect repeats, and extended compound repeats. Users are allowed to select different minimum repeat numbers for particular DNA motifs, or all the motifs can be analysed using the same minimum or maximum repeat number criteria.

Identification of simple exact tandem repeats and non-simple repeats (compound, imperfect repeats, and extended compound repeats) involves both locating and characterizing strings in given DNA sequences formatted in either FASTA, GenBank or EST in databases. E-TRA accomplishes repeats finding and classification tasks basically in four major steps as follows: (i) searches the user defined organism (s) and/or keywords (organs, cell lines, tissue types or development stages) analyzing the whole data set provided in a data folder; (ii) isolates simple and non-simple (compound, imperfect and extended compound) tandem repeats by determining their type, lengths, and sequence location in a given DNA strings 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.

In this paper we used human ESTs and searched repeats using the following criteria: repeated DNA (cDNA) sequence had to be sixteen units for a monomer, eight units long for dimers, whereas trimers, tetramers, pentamers to nanomers and repeat units of lengths equal or greater than 10 required seven, six, five, and four repeats, respectively (Fondon *et al*. 1998; Wren *et al*. 2000). Simple regression coefficient analysis was utilized to investigate whether the repeat content of organ/tissue/cell lines were dependent on the ESTs studied.

Results and discussion

Our results using the E-TRA demonstrated that about 12.44% of the human ESTs have repeats, indicating that a significant portion of the transcribed human mRNAs contained repeats. The most common motif type found in the human ESTs was the mononucleotide A/T (table 1). This is not surprising because eukaryotic mRNAs contain poly A tails. The most common dinucletide repeat type was AC/GT. Riley and Krieger (2004) reported that AC/GT was 8 times more common in membrane-function mRNAs. This indicated that AC/GT repeats preferentially express in membrane proteins. The most abundant trinucleotide repeat type was CTG which possibly codes the amino acid lysine. Repeat lengths greater than 4 were less common in human ESTs. Human transcribed genes had 3010 ESTs containing repeat lengths between 10 and 126 nucleotides (table 1). The most abundant repeat length was 18 nucleotides (691 ESTs). The second most abundant repeat lengths were 15 followed by 12, 16, 17, 20, 30 and 27 nucleotides. Non-simple repeats, that is

Table 1. Distribution of simple and non-simple repeats in human ESTs.

Repeat type	Repeat containing EST strings	% (percentage)
Simple repeats		
Mononucleotides	615398	83.62
Dinucleotides	29790	4.05
Trinucleotides	13517	1.84
Tetranucleotides	4535	0.62
Pentanucleotides	2442	0.33
Hexanucleotides	1017	0.14
Heptanucleotides	23	0.00
Octanucleotides	35	0.00
Nanonucleotides	44	0.01
Deca and up to 126 nucleotides	3010	0.41
Total	669811	91.01
Non-simple repeats		
Compound repeats	8328	1.13
Imperfect repeats	53731	7.30
Extended compound repeats	4115	0.56
Total	66174	8.99
Overall total	735985	

compound, imperfect and extended compound repeats were found in significant numbers in human ESTs (table 1). Among the non-simple repeats, imperfect repeats were the most abundant repeat type and they were followed by compound and extended compound repeats. This indicates that non-simple repeats may play some important role(s) in the transcribed portion of the genome.

We studied about 25 human organs/tissues or cell lines (table 2). The highest EST numbers analysed were for brain, followed by lung tissues. The distribution and the number of repeats varied considerably among the organs/tissue or cell lines. Simple regression analysis indicated that human brain, liver, colon, stomach and heart contained significantly much less repeat containing transcribed genes than lung, nerve and stem cells (figure 1). On the other hand, cord blood and thymus contained repeat containing transcribed genes as expected. The highest and lowest repeat contents were observed in lung and heart, respectively. There were significant variations in the cell lines for repeat numbers. Stem cells showed the highest number of repeats and they were followed by B cells and germ cell lines. T-cells had the lowest number of repeats among the cell lines.

Mononucleotide repeats were the most abundant in nerve and lung while they were the least abundant in thymus among the human organ and tissues we studied. Dinucleotide repeats were prominent in heart and liver, tri-, tetra-, penta-hexa-hepta, octa and nano-nucleotide repeats were the most abundant in testis and stomach, heart and muscle, testis and heart, stomach and heart, T-cells and breast, brain and testis, and testis and embryo, respectively. EST containing repeats ranging from deca and up to 126 repeat lengths were the most abundant in cord blood cell and breast while they were not significant in the lungs. The distribution of the non-simple repeats also varied among the human organs/tissues or cell lines. Thymus and T-cells showed the highest number of compound repeats (table 2). Imperfect repeats were prominent in T-cells and thymus and also thymus and testis had the most abundant extended compound repeat as per our repeat finding criteria.

Repeat rates (repeat containing EST/total EST) also varied among the organ and tissue. Repeat rates typically decreased with increasing repeat length from mono to deca and up to 126 nucleotides. However, trinucleotide repeat rates of T-cells, embryo and testis were higher than their dinucleotide repeat rates. Also, pentanucleotide repeat rate of T cells was higher than its tetranucleotide repeat rate. Mononucleotide repeat rate was the highest in the lungs and it was the lowest in thymus. Heart and bone marrow showed the highest and lowest dinucleotide repeats, respectively. Trinucleotide repeat rate was the most prominent in testis and nerve cells showed the least amount of trinucleotide rate. Muscle had the highest amount of tetranucleotide repeat rate while cord blood,

Table 2. Distribution of simple (1–10 repeat number) and non-simple (CR: compound repeats, IR: imperfect repeats, ECR: extended compound repeats) repeats among 25 human organs/tissue/cell lines.

Material	1	2	3	4	5	6	7	8	9	10	CR	IR	ECR
B-Cells	15162	606	302	110	45	24	1	0	0	10	158	1194	85
Bone marrow	8301	137	59	27	7	6	0	0	0	5	65	830	15
Brain	33814	2571	1636	456	291	129	5	7	1	120	746	5222	403
Breast	10342	556	198	109	59	19	2	0	3	173	201	1086	74
Colon	12434	869	322	139	69	34	0	1	1	120	218	1095	73
Cord Blood	1102	59	8	3	0	3	0	0	0	126	1	132	2
Embryo	3321	109	121	26	13	4	0	0	1	2	29	599	14
Germ cells	7491	268	98	64	14	10	0	0	0	5	34	217	7
Heart	3886	458	107	83	37	16	0	0	0	2	47	228	18
Kidney	14403	805	361	127	72	22	0	1	0	18	221	1332	78
Liver	13093	1420	408	256	114	18	0	2	2	31	250	1665	132
Lung	106563	2530	1161	286	172	78	2	1	2	43	701	3472	248
Muscle	7016	730	272	159	60	31	0	1	1	16	122	1060	58
Nerve	24445	417	88	71	19	6	0	1	0	22	95	449	16
Ovary	15259	651	345	99	76	26	2	2	1	191	206	1230	113
Pancreas	21385	1173	562	121	56	34	1	0	4	67	398	1768	198
Placenta	27981	1398	963	128	142	45	1	2	0	34	654	4831	458
Prostate	16336	649	321	86	42	30	0	0	1	115	175	1437	81
Stem cells	11556	324	74	33	13	4	0	0	0	6	100	1500	115
Stomach	7425	611	408	60	33	42	1	1	0	28	221	876	77
T-cells	2406	105	117	13	22	3	1	0	0	5	89	651	46
Testis	5861	429	481	79	77	25	1	1	11	23	210	1130	113
Thymus	1062	79	38	14	11	3	0	0	0	6	53	276	30
Tumour	33335	1547	433	239	87	50	1	1	2	31	579	1945	308
Uterus	21692	1312	453	161	84	55	0	0	1	99	487	2070	290

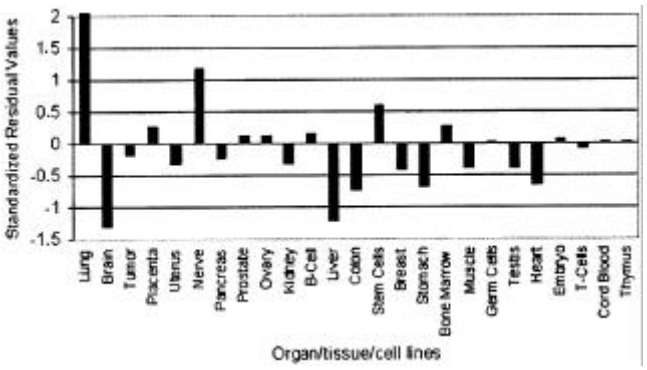

Figure 1. Standardized residual values (negative or positive) indicating that the repeat contents of different human organ/ tissue and cell lines showed considerable variation.

lung and stem cells showed the least amount of tetranucleotide repeat rate. There were no significant repeat rates for penta, hexa, hepta, octa and nano nucleotide repeats among the human organ and tissues. Rate of repeats ranging from 10 to 126 nucleotides was significant for brain, breast, colon, liver, lung, muscle, ovary, pancreas, placenta, stomach, T-cells, testis, thymus, tumour and uterus.

Research efforts in recent years have yielded considerable numbers of genomic and transcribed ESTs from many species, and researchers are now able to access the information contained within these sequences. The Institute for Genomic Research (TIGR) gene indexes (Quackenbush *et al*. 2001) and the NCBI UniGene sets (Wheeler *et al*. 2003) are the two important genomic databases. Previous studies indicated that a significant portion of ESTs consisted of short tandem repeats (Kantety *et al*. 2002; Saha *et al*. 2003; Bilgen *et al*. 2004). Our preliminary study on human organ and tissue and cell lines indicates that it might be possible to identify the genes that contribute to a cell's unique characteristics. Further, analysis using E-TRA may be very useful to obtain advanced knowledge about the functions of the repeated DNA sequences in transcribed genes of organs and tissue types/ cell lines in various organisms. The program E-TRA will be very useful to those researchers interested in (i) identifying the repeat containing ESTs as well as those small genomes (bacteria, chloroplast and mitochondria) for further studies (for instance; instead of genomic DNA mapping, transcribed gene (ESTs) mapping on the chromosomes will be very helpful in breeding studies), and (ii) data mining for repeat containing sequences and characterizing the repeats, compositions and distributions among the organisms (for instance; animal versus plant, trees versus annual crops), among the organs (for instance; flower versus seed), tissue types (for instance adipose tissues versus muscle tissues or tumour cells versus normal cells), cell lines (for instance; B-cell versus T-cells or stem cells), development stages (for instance; seed setting versus seed maturation). Information gained from such studies will be very useful for understanding the expression, regulation and evolution of repeats in DNA.

In conclusion, we developed a new program and found that ESTs derived from different tissue/organ or development stage contained different amounts of repeats. E-TRA program was also utilized to develop EST–SSR primer pairs (using one of the primer design programs) and amplified genomic DNAs from various plant species including *Gossypium* and *Capsicum spp*. E-TRA would be utilized to develop EST-SSR primer pairs for other plant and animal species too Also to date, there is a limited research on tandem repeats for their theoretical comparison and importance in ESTs. Therefore, further analysis using E-TRA may be very useful to obtain advanced knowledge about the functions of the repeated DNA sequences in transcribed genes of organs and 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

Benson G. 1999 Tandem repeats finder: a program to analyse DNA sequences. *Nucl. Acids Res*. **27**, 573–580.

Bilgen M., Karaca M., Onus A. N. and Ince A. G. 2004 A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. *Bioinformatics* **20**, 3379–3386.

Fondon J. W., Mele G. M., Brezinschek R. I., Cummings D., Pande A. and Wren J. *et al*. 1998 Computerized polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. *Proc. Natl. Acad. Sci. USA* **95**, 7514–7519.

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

Huang C., Lin Y., Yang Y., Huang S. and Chen C. 1998 The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol. Microbiol*. **28**, 905–916.

Ince A. G., Onus A. N., Elmasulu S. Y., Bilgen M. and Karaca M. 2004 *In silico* data mining for development of *Capsicum* microsatellites. *Proc. Int. 3rd Balkan Symposium on vegetables and potatoes*. Bursa, Turkey, *Acta Horticulturae* (in press).

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. 2002a Simple Sequence Repeat (SSR) markers linked to the *Ligon lintless* (Li_1) mutant in cotton. *J. Heredity* **93**, 221–224.

Karaca M., Saha S., Zipf A., Jenkins J. N. and Lang D. J. 2002b Genetic diversity among Forage Bermuda grass (*Cynodon* spp.): evidence from chloroplast and nuclear DNA fingerprinting *Crop Sci*. **42**, 2118–2127.

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. *Nucl. Acids Res.* **29**, 4633–4642.

Lalioti M. D., Scott H. S., Buresi C., Bottani A., Norris M. A., Malafosse A. and Antonarakis S. E. 1997 Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**, 847–852.

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.

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

Quackenbush J., Cho D., Lee F. L., Holt I., Karamycheva S. and Parvizi B. *et al*. 2001 The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucl. Acids Res.* **29**, 159–164.

Richard G. F., Hennequin C., Thierry A. and Dujon B. 1999 Trinucleotide repeats and other microsatellites in yeasts. *Res. Microbiol.* **150**, 589–602.

Riley D. E. and Krieger J. N. 2004 Short tandem repeats are associated with diverse mRNAs encoding membrane-targeted proteins. *Bioassays* **26**, 434–444.

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.

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

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.

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.

Wheeler D. L., Church D. M., Federhen S., Lash A. E., Madden T. L. and Pontius J. U. *et al*. 2003 Database resources of the national center for biotechnology. *Nucl. Acids Res.* **31**, 28–33.

Wren J. D., Forgacs E., Fondon J. W., Pertsemlidis A., Cheng S. Y. and Gallardo T. *et al*. 2000 Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.* **67**, 345–356.

Received 25 September 2004; in revised form 12 January 2005