W656–W659 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl083

SSRPrimer and *SSR Taxonomy Tree*: Biome SSR discovery

Erica Jewell[1,2], Andrew Robinson[1,2], David Savage[1], Tim Erwin[1,2], Christopher G. Love[1,2], Geraldine A. C. Lim[1,2], Xi Li[1], Jacqueline Batley[1], German C. Spangenberg[1,2,3] and David Edwards[1,2,3,*]

[1]Department of Primary Industries, Plant Biotechnology Centre, Primary Industries Research Victoria, Victorian AgriBiosciences Centre, 1 Park Drive, Bundoora, Victoria 3083, Australia, [2]Department of Primary Industries, Victorian Bioinformatics Consortium, Plant Biotechnology Centre, Primary Industries Research Victoria, Victorian AgriBiosciences Centre, 1 Park Drive, Bundoora, Victoria 3083, Australia and [3]Department of Primary Industries, Australian Centre for Plant Functional Genomics, Plant Biotechnology Centre, Primary Industries Research Victoria, Victorian AgriBiosciences Centre, 1 Park Drive, Bundoora, Victoria 3083, Australia

Received February 13, 2006; Accepted March 6, 2006

ABSTRACT

Simple sequence repeat (SSR) molecular genetic markers have become important tools for a broad range of applications such as genome mapping and genetic diversity studies. SSRs are readily identified within DNA sequence data and PCR primers can be designed for their amplification. These PCR primers frequently cross amplify within related species. We report a web-based tool, SSR Primer, that integrates SPUTNIK, an SSR repeat finder, with Primer3, a primer design program, within one pipeline. On submission of multiple FASTA formatted sequences, the script screens each sequence for SSRs using SPUTNIK. Results are then parsed to Primer3 for locus specific primer design. We have applied this tool for the discovery of SSRs within the complete GenBank database, and have designed PCR amplification primers for over 13 million SSRs. The SSR Taxonomy Tree server provides web-based searching and browsing of species and taxa for the visualisation and download of these SSR amplification primers. These tools are available at http://bioinformatics.pbcbasc.latrobe.edu.au/ssrdiscovery.html.

INTRODUCTION

Simple sequence repeats (SSRs), also known as microsatellites, have been shown to be one of the most powerful genetic markers in biology. They are common, readily identified DNA features consisting of short (1–6 bp), tandemly repeated sequences, widely and ubiquitously distributed throughout eukaryotic genomes (1) and have been found in all prokaryotic and eukaryotic genomes that have so far been analysed (2). SSRs are highly polymorphic, owing to the mutation affecting the number of repeat units. This hypervariability among related organisms makes them informative and excellent markers for a wide range of applications including high-density genetic mapping, molecular tagging of genes, genotype identification, analysis of genetic diversity, paternity exclusion, phenotype mapping and marker assisted selection of crop plants (3,4).

SSRs were initially considered to be evolutionally neutral, (5), though recent evidence suggests an important role in genome evolution (6). SSRs are a source of abundant, non-deleterious mutations that provide variation in the face of stabilizing selection, and their recognized role in the process of evolutionary adaptation is predicted to increase as our knowledge of them expands (7). SSR stability may be correlated with overall levels of genomic stability (8) as mutations which affect SSR stability, such as those involved in DNA mismatch repair, can also influence genomic stability. The nature of SSRs gives them a number of advantages over other molecular markers; (i) multiple SSR alleles may be detected at a single locus using a simple PCR based screen, (ii) SSRs are evenly distributed all over the genome, (iii) they are co-dominant, (iv) very small quantities of DNA are required for screening, and (v) analysis may be semi-automated. Furthermore, SSRs demonstrate a high degree of transferability between species, as PCR primers designed to an SSR within one species frequently amplifies a corresponding locus in related species, making them excellent markers for comparative genetic and genomic analysis.

*To whom correspondence should be addressed. Tel: +61 0 3 94795633; Fax: +61 0 3 94793618; Email: Dave.Edwards@dpi.vic.gov.au

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

The potential biological function and evolutionary relevance of SSRs is currently under scrutiny and leading to a greater understanding of genomes and genomics (9). Initial suggestions that the majority of DNA was either 'junk' or had no biological function are being challenged by the discovery of new functions for these sequences. Various functional roles have now been attributed to SSRs. For example, SSRs are believed to be involved in gene expression, regulation and function (7,10) and there are numerous lines of evidence suggesting that SSRs in noncoding regions may also be of functional significance (7). Furthermore, SSRs provide hotspots of recombination, a variety of SSRs have been found to bind nuclear proteins and there is direct evidence that SSRs can function as transcriptional activating elements (11).

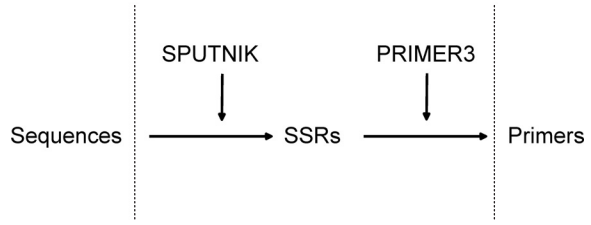

Figure 1. An overview of the SSRPrimer pipeline. Following entry of DNA sequences, each sequence is processed using SPUTNIK. If an SSR is identified, the sequence and SSR location is parsed to Primer3 for the design of suitable PCR amplification primers.

A common method for the discovery of SSR loci is to construct genomic DNA libraries enriched for SSR sequences, followed by DNA sequencing (12). This production of enriched libraries is time consuming and the specific sequencing required is expensive. Where abundant sequence data is already available, it is more economical and efficient to use computational tools to identify SSR loci. Flanking DNA

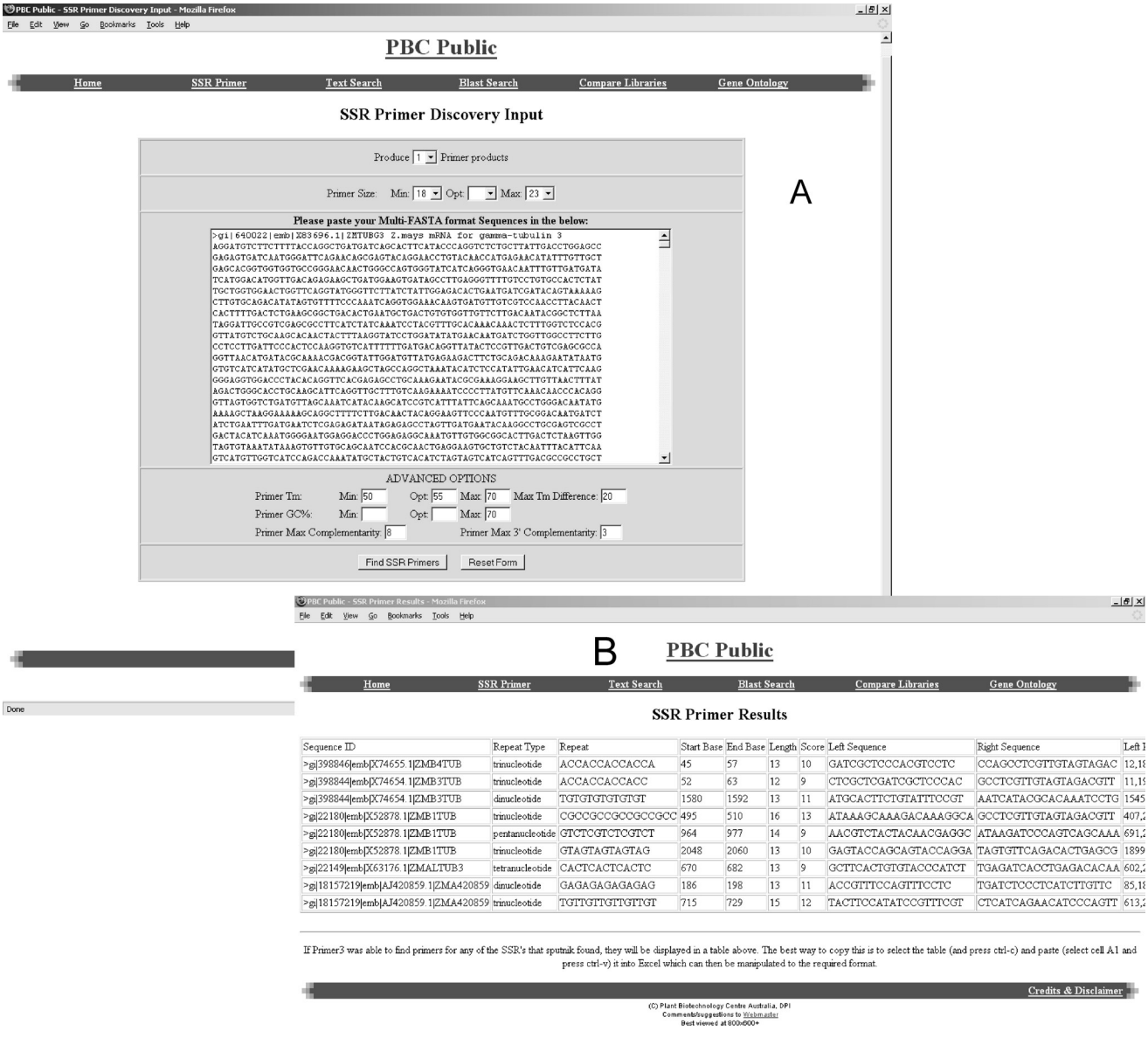

Figure 2. The SSRPrimer web server. Sequences are pasted into the entry box and PCR Primer parameters specified (**A**). The resulting identified SSRs are listed along with designed PCR primers and amplification parameters (**B**).

sequences may then be analysed for the presence of suitable forward and reverse PCR primers to assay the SSR loci. Several computational tools are currently available for the identification of SSRs within sequence data, as well as for the design of PCR primers suitable for the amplification of specific loci. We have integrated two such tools within one package SSRPrimer, enabling the simultaneous discovery of SSRs within bulk sequence data and the design of specific PCR primers for the amplification of these marker loci (13). An integrated web interface further permits the remote use of this tool.

Sequences are initially parsed to SPUTNIK (14) (http://abajian.net/sputnik/), which uses a recursive algorithm to search for repeated patterns of nucleotides of length between 2 and 5. The output of SPUTNIK is then parsed to Primer 3 (15) for PCR Primer design. Primers are designed to a defined set of constraints such as oligonucleotide melting temperature (T_m), size, GC content, primer-dimer possibilities, PCR product size and positional constraints around the SSR to identify the optimal forward and reverse primers for the SSR flanking region. The results of the application of the package to the complete GenBank database, SSR Taxonomy Tree, can be browsed and searched for SSRs and amplification primers for any species of interest.

METHODS

SSRPrimer sequence input and pipeline processing

SSRPrimer is a web-based tool that may also be run on the command line. Access to the web server version requires an internet connection and a standard web browser. The web server version of SSRPrimer acts as a web interface and wrapper for the two programs, SPUTNIK and Primer3 that make up the SSR discovery pipeline (Figure 1). The complete pipeline accepts one or more DNA sequences as input along with PCR Primer design options. Each entry sequence is processed in turn using SPUTNIK for the identification of SSRs. If an SSR is identified within a sequence, the sequence along with the SSR location is parsed to Primer3 for PCR amplification primer design. Default parameters for PCR Primer design are designed to increase primer specificity. While these and additional options may be modified on the SSRPrimer submission page (Figure 2), the authors suggest maintaining these strict criteria to ensure robust PCR amplification.

SSR Taxonomy Tree

The SSR Taxonomy Tree server provides access to over 13 million SSR Primer pairs identified through the application of SSRPrimer to the complete GenBank nucleotide sequence database (Figure 3). Default PCR Primer design parameters were one set of primer pairs designed at least 10 bp distant from either side of the identified SSR. Optimum size for the primers are 21 bases with a maximum of 23 bases. Optimum T_m is 55°C with a minimum of 50°C, maximum of 70°C and maximal difference in T_m of 20°C. The maximum GC content is 70%. Results include over 9.7 million, 1.8 million and 82 thousand SSR Primer pairs designed from mammalia, plant and fungal species, respectively. The server permits the searching of taxa by both latin and common names using standard MySQL Boolean operators and wild cards.

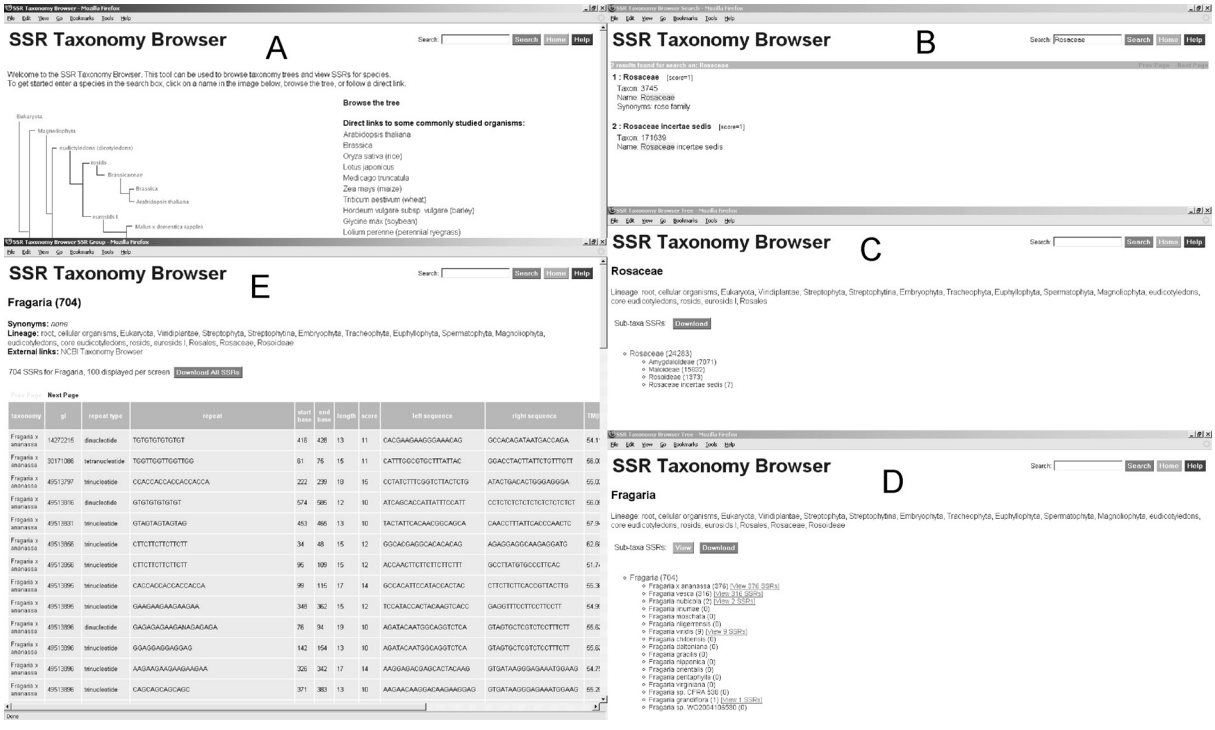

Figure 3. The SSR Taxonomy Tree server. A query (Rosaceae) is entered into the search box (**A**) identifying two matches (**B**), clicking Rosaceae displays the taxonomic branches leading to the Rosaceae sub taxa and presence of SSRs within sub taxa (**C**). Sub taxa may be browsed through Rosoideae to *Fragaria* (**D**) and identified *Fragaria* sub taxa SSR primers viewed and downloaded (**E**).

Taxa may also be browsed through a hierarchical tree. Resulting lists of SSRs and PCR primers may be viewed or downloaded as a tab-delimited text file for input into a spreadsheet. Large files (<5 Mb) may be downloaded in compressed format. Comprehensive help pages include details of how to search, view and download SSRs. Additionally, a browser log details recent and planned data updates and server down time.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Primary Research Industries Victoria (PIRVic).

Conflict of interest statement. None declared.

REFERENCES

1. Tóth,G., Gáspári,Z. and Jurka,J. (2000) Microastellites in different eukaryotic genomes:survey and analysis. *Genome Res*., **10**, 967–981.
2. Katti,M.V., Ranjekar,P.K. and Gupta,V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol*., **18**, 1161–1167.
3. Tautz,D. (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res*., **17**, 6463–6471.
4. Powell,W., Machray,G.C. and Provan,J. (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci*., **1**, 215–222.
5. Awadalla,P. and Ritland,K. (1997) Microsatellite variation and evolution in the *Mimulus guttatus* species complex with contracting mating systems. *Mol. Biol. Evol*., **14**, 1023–1034.
6. Moxon,E.R. and Wills,C. (1999) DNA microsatellites: agents of evolution. *Sci. Am*., **280**, 94–99.
7. Kashi,Y., King,D. and Soller,M. (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet*., **13**, 74–78.
8. Ross,C.L., Dyer,K.A., Erez,T., Miller,S.J., Jaenike,J. and Markow,T.A. (2003) Rapid divergence of microsatellite abundance among species of *Drosophila*. *Mol. Biol. Evol*., **20**, 1143–1157.
9. Subramanian,S., Mishra,R.K. and Singh,L. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol*., **4**, R13.
10. Gupta,M., Chyi,Y-S., Romero-Severson,J. and Owen,J.L. (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. *Theor. Appl. Genet*., **89**, 998–1006.
11. Li,Y.-C., Korol,A.B., Fahima,T., Beiles,A. and Nevo,E. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol*., **11**, 2453–2465.
12. Edwards,K.J., Barker,J.H.A., Daly,A., Jones,C. and Karp,A. (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques*, **2**, 758–760.
13. Robinson,A.J., Love,C.G., Batley,J., Barker,G. and Edwards,D. (2004) Simple sequence repeat marker loci discovery using SSR primer. *Bioinfomatics*, **20**, 1475–1476.
14. Abajian,C. (1994) SPUTNIK.
15. Rozen,S. and Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.