Tree Genetics & Genomes (2014) 10:1755–1757
DOI 10.1007/s11295-014-0785-2

SHORT COMMUNICATION

Standalone EST microsatellite mining and analysis tool (SEMAT): for automated EST-SSR analysis in plants

Naganeeswaran Sudalaimuthu Asari · Manimekalai Ramaswamy · Elain Apshara Subbian · Manju Kalathil Palliyarakkal · S. K. Malhotra · Anitha Karun

Received: 4 July 2013 / Revised: 4 June 2014 / Accepted: 31 July 2014 / Published online: 12 August 2014

Abstract Public databases contain large datasets of plant expressed sequence tags (ESTs) that can be used for mining microsatellite/simple sequence repeat markers. The identification and annotation of these markers take considerable time. Here, we describe an efficient, high-throughput microsatellite mining, and analysis pipeline, standalone EST microsatellite mining and analysis tool (SEMAT). The pipeline bundles sequence trimming, assembly, microsatellite identification, primer selection, and blast annotation, for which it consecutively uses SeqClean, CAP3, MISA, Primer3, and Blast. SEMAT is written using Perl scripts, and it runs under Ubuntu and Fedora Linux. SEMAT is an efficient and time-saving bioinformatics tool to accomplish the high throughput EST-SSR analysis. It is freely available from http://semat.cpcribioinformatics.in/.

Keywords EST · Microsatellite · Pipeline · Tool

Communicated by J. L. Wegrzyn

N. S. Asari · M. Ramaswamy · M. K. Palliyarakkal
DIT- Agribioinformatics Promotion centre, Central Plantation Crops Research Institute, P. O. Kudlu, Kasaragod 671124, Kerala, India

N. S. Asari
e-mail: naganeeswaran@gmail.com

M. K. Palliyarakkal
e-mail: manjusathyan@gmail.com

E. A. Subbian
Crop Improvement Division, Central Plantation Crops Research Institute, Regional station, Vittal 574 243, Karnataka, India
e-mail: elain_apshara@yahoo.co.in

A. Karun
Crop Improvement Division, Central Plantation Crops Research Institute, P. O. Kudlu, Kasaragod 671124, Kerala, India
e-mail: anithakarun2008@gmail.com

S. K. Malhotra
Indian Council of Agricultural Research, Krishi Bhavan, New Delhi, India
e-mail: malhotraskraj@yahoo.com

M. Ramaswamy (✉)
Biotechnology, Sugarcane Breeding Institute, Veerakeralam, Coimbatore 641007, Tamil Nadu, India
e-mail: rmanimekalaiicar@gmail.com

Introduction

Expressed sequence tags (ESTs) are single-pass, redundant, partial nucleotide sequences that represent the transcribed portion of the genome. Advanced high-throughput sequencing technologies have generated vast number of ESTs and most of the sequences are available in public EST database (73,360,923, dbEST release 120701). Using EST sequences, one can perform gene discovery, transcriptome annotation, gene structure identification and proteomic exploration (Jongeneel 2000; Dong et al. 2005; Rudd 2003). Microsatellite or simple sequences repeats (SSRs) are short (1–6 bp) tandemly repeated DNA motifs and occurs abundantly in eukaryotic genomes. They have been successfully used as molecular markers by targeting the repeat region with unique flanking primer pairs. Presently, EST sequences are one of the main sources of data for mining SSRs. As most of the public EST datasets consists of raw sequences, different computer-based methods are required for processing, assembly, SSR mining, and corresponding annotation. We have developed an analysis pipeline "standalone EST microsatellite mining and analysis tool" (SEMAT) for high throughput discovery (mining and annotation) of SSR loci in EST sequence and automated design of primers for their PCR amplifications.

Implementation and methods

SEMAT combines a set of third party, freely available bioinformatics softwares and locally developed Perl (http://www.perl.org) scripts for automated processing of ESTs and analysis of the results. SEMAT follows the general analysis steps like preprocessing of the sequence data, clustering, and assembly, SSR detection and primer designing. The work flow of SEMAT is given in Fig. 1. Without any manual interaction, data is transferred automatically among consecutive subroutines. Parameters for the softwares are stored in flat files (flat files are stored in folder-named file) and users can change the parameter according to their need. SEMAT takes multiFASTA file as an input and automatically performs EST processing, assembly, SSR prediction, annotation, and designing primers encompassing the SSR regions.

Preprocessing and assembly

The removal of vectors and low-quality regions from the EST data set (pre-processing) are carried out using SeqClean program (http://compbio.dfci.harvard.edu/tgi/software/) (Parameter for SeqClean : -c 1, -n 2000, -l 100, -x 96, and –y 11) embedded with NCBI's UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). The cleaned EST sequences are subjected to de novo assembly using CAP3 program (Huang and Madan 1999) with default parameters (parameter: -a 20, -b 20, -c 12, -d 200, -e 30, -f 20, -g 6, -h 20, -i 40, -j 80, -k 1, -m 2, -n -5, -o 40, -p 90, -v 1, -s 900, -t 300, -u 3, -v 2, -y 100, and –z 3). CAP3 does the assembly based on mutual percent identity over a minimum number of overlapping bases. Separate Perl scripts were incorporated to combine the resultant contigs and singleton files to a single file for further analysis.

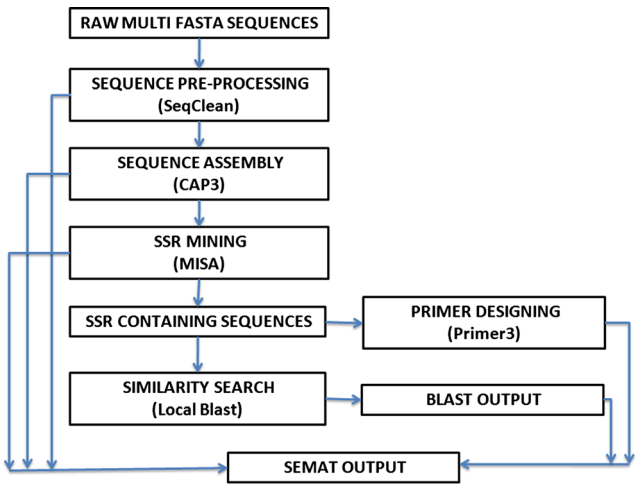

Fig. 1 Workflow of the tool SEMAT. The tasks performed and the tools used are given inside the *rectangles*

Microsatellite mining, primer designing, and annotation

Assembled ESTs were then transferred to microsatellite mining by MISA program (MIcroSAtellite analysis tool) (http://pgrc.ipk-gatersleben.de/misa/). Here, monomers with at least ten repetitions, dimers with minimum of six repetitions and trimers, tetramers, pentamers, and hexamers with minimum of five repetitions were considered as a valid SSRs. MISA can identify the simple and compound microsatellites. In the present pipeline, the code of MISA program was modified to retrieve all the SSR containing sequences. In order to design primers flanking the SSR region, the modified MISA program generates an informative file to be used as an input for Primer3 (Rozen and Skaletsky 2000). Primer3 program incorporated in the pipeline generates oligonucleotides flanking the SSR region based on predefined parameter values. For annotation, SSR containing sequences were subjected to similarity search against the preformatted plant specific protein database using standalone BlastX program (Altschul et al. 1997). The plant protein databases contain the sequence of *Arabidopsis thaliana*, *Oryza sativa*, *Vitis vinifera*, *Populus trichocarpa*, *Glycine max*, *and Zea mays*. Results obtained from the similarity search (e-value $<1e^{-15}$) are organized into tabular format. SEMAT generates the summary file containing the information about the EST analysis.

Results and discussion

We have benchmarked SEMAT with a test set of cocoa EST sequences. A total of 2,685 EST sequences (Argout et al. 2008) were organized as multiFASTA file and given as input for SEMAT. EST processing resulted in 2,682 good quality EST sequences (3 sequences were completely trashed and 1,831 were trimmed). These sequences were then subjected into sequence assembly using CAP3 program. One thousand four hundred seventy-one nonredundant EST (246 contigs and 1,225 singletons) sequences were obtained. The contig summary file (.ace) was analyzed and showed that assembled sequences were in the range of 2–434 sequences. The complete set of 1,471 nonredundant sequences was further subjected to microsatellite discovery. MISA detected 313 SSRs within 210 sequences and the sequence length ranged between 135 to 1,222 nucleotides and GC of 42 %. Fifty-three sequences had more than one SSR. A total of 72 compound SSRs were detected. The similarity search carried out for the 210 SSR containing sequences using standalone Blast which showed 92 orthologous genes. The result of the similarity search is organized in tabular format. A total of 129 primer sets were designed for the possible SSR regions. All the benchmarked results are available in the URL: http://semat.cpcribioinformatics.in/.

A few other automated applications are available for EST functional analysis like EST-PAGE (Matukumalli et al. 2004), ParPEST (D'Agostino et al. 2005), and EST2uni (Javier et al. 2008). But the SEMAT is specially designed for EST-SSR analysis. We compared SEMAT with similar pipeline, SSRPrimer (Robinson et al. 2004). The SSRPrimer tool performs SSR search and primer design, but the process was not automated. In this way, SEMAT is a unique tool for automated EST processing and annotation.

Advantage and availability

SEMAT pipeline was developed using Perl language and tested in Linux operating systems (Ubuntu and fedora). It is freely available at http://semat.cpcribioinformatics.in/. SEMAT completed the analysis of cocoa ESTs (2,685 sequences) within 15 min. Further, we tested SEMAT pipeline with all the available cocoa ESTs (159,996 sequences) in both workstation with Linux operating system (Sun Ultra 24, 4GB RAM, Intel (R) Core (TM) 2 Quad, 2. 33 GHz) and server system (SunFire x220 computers, 8 GB RAM, Quad-Core AMD Opteron (TM)). The time taken to complete the analysis by workstation was 14 h and by server machine was 11 h. The time taken for the analysis depends on the input sequence size and computer capacity (RAM and processor).

Conclusion

SEMAT is an efficient and time-effective bioinformatics tool to accomplish the high-throughput application of EST-SSR analysis. Its source code is open to all; anyone can edit the code and incorporate specific tools according to their need.

Acknowledgments This work was supported by grants from Department of Information Technology (DIT), India. Our sincere thanks to Dr. George. V Thomas, Director, Central Plantation Crops Research Institute, Kasaragod, India, for his guidance and support.

Data archiving statement The manuscript contains no data that has to be submitted to public database. We uploaded all the data of the results of cocoa EST analysis and the SEMAT tool in our web server http://semat.cpcribioinformatics.in/ for open access.

References

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Argout X, Fouet O, Wincker P, Gramacho K et al (2008) Towards the understanding of the cacao transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. BMC Genomics 9:1–19

D'Agostino N, Aversano M, Chiusano ML (2005) ParPEST: a pipeline for EST data analysis based on parallel computing. BMC Bioinforma 6:S9

Dong Q, Kroiss L, Oakley FD et al (2005) Comparative EST analyses in plant systems. Methods Enzymol 395:400–408

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Javier F, Francisco G, Antonio R (2008) EST2uni: an open, parallel tool for automated EST analysis and database creation, with data mining web interface and microarray expression data integration. BMC Bioinforma 9:5

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. Brief Bioinform 1:76–92

Matukumalli LK, Grefenstette JJ, Sonstegard TS et al (2004) EST-PAGE – managing and analyzing EST data. Bioinformatics 20:286–288

Robinson AJ, Love CG, Batley J, Barker G, Edwards D (2004) Simple sequence repeat marker loci discovery using SSR primer. Bioinformatics 20:1475–1476

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 8:321–329