academicJournals

Vol. 12(30), pp. 4713-4721, 24 July, 2013
DOI: 10.5897/AJB12.1410
ISSN 1684-5315 ©2013 Academic Journals
http://www.academicjournals.org/AJB

African Journal of Biotechnology

Review

Bioinformatics tools for development of fast and cost effective simple sequence repeat (SSR), and single nucleotide polymorphisms (SNP) markers from expressed sequence tags (ESTs)

Sushmita Gupta[1*], Raju Bharalee[2], Ranjita Das[1] and Debajit Thakur[1]

[1]Institute of Advanced Study in Science and Technology, Guwahati – 781035, India.
[2]Department of Biotechnology, TERI, Guwahati - 781 036, India.

Accepted 17 July, 2013

The development of current molecular biology techniques has led to the generation of huge amount of gene sequence information under the expressed sequence tag (EST) sequencing projects on a large number of plant species. This has opened a new era in crop molecular breeding with identification and/or development of a new class of useful DNA markers called genic molecular markers (GMMs). These markers represent the functional component of the genome in contrast to all other random DNA markers (RMMs). Many recent studies have demonstrated that GMMs may be superior to RMMs for use in the marker assisted selection, comparative mapping and exploration of functional genetic diversity in the germplasms adapted to different environment. Therefore, identification of DNA sequences which can be used as markers remains fundamental to the development of GMMs. Amongst others; bioinformatics approaches are very useful for development of molecular markers, making their development much faster and cheaper. Already, a number of computer programs have been implemented that aim at identifying molecular markers from sequence data. A revision of current bioinformatics tools for development of genic molecular markers is, therefore, crucial in this phase. This mini-review mainly provides an overview of different bioinformatics tools available and its use in marker development with particular reference to SNP and SSR markers.

Key words: Genic molecular marker, simple sequence repeat (SSR), and single nucleotide polymorphisms (SNP) markers from expressed sequence tags (ESTs).

INTRODUCTION

Most of the agriculturally important traits such as yield, quality and tolerance and/or resistance to biotic and abiotic stress are polygenic in nature and are often termed as ‘quantitative traits’. The regions within genomes that contain genes associated with a particular quantitative trait are known as ‘quantitative trait loci’ (QTLs) (Collard et al., 2005). Genetic markers are specific loci in chromosomes of particular organisms associated with a trait and can be used as tool for marker assisted selection (MAS) in plant breeding. Genetic marker assisted breeding is more efficient, effective, reliable and cost effective as compared to conventional plant breeding (Collard et al., 2005). Genetic marker system can be broadly classified into three types: (i) morphological markers, (ii) biochemical markers and (iii) molecular (DNA) markers (Winter and Kahl, 1995). The phenotypic

*Corresponding author: E-mail: sgpakhi27@gmail.com

traits which can be visually characterized such as leaf colour, seed shape and size, flower colour, etc., are termed as morphological markers (Winter and Kahl, 1995). Isozymes are the most common biochemical markers used in plant breeding. The major disadvantages of morphological and biochemical markers are that they are limited in number and also in some cases influenced by environmental factors (Varshney et al., 2005). Moreover, their expression may be restricted to specific developmental stages or tissues. Biochemical markers are superior to morphological markers in that they are generally independent of environmental growth conditions (Varshney et al., 2005). The third and the most advance form of genetic markers are molecular markers which reveal DNA sequence variations called 'polymorphisms' (Collard et al., 2005). Polymorphic markers can be dominant or co-dominant markers based on whether markers can discriminate between homozygotes and heterozygotes loci (Collard et al., 2005). Molecular markers are broadly classified into three classes based on the method of their detection: (i) hybridization based markers such as restriction fragment length polymorphisms (RFLP), (ii) PCR based markers such as random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphisms (AFLP) and microsatellite or simple sequence repeat (SSR), and (iii) sequence based markers such as single nucleotide polymorphisms (SNP) (Gupta and Rustgi, 2004).

With the recent advancement of functional genomics, several gene discovery projects such as genome sequencing, EST generation and analysis has resulted in the accumulation of enormous amount of sequence data from complete or partial genes (Varshney et al., 2005). ESTs are short DNA sequences corresponding to a fragment of a complimentary DNA (cDNA) molecule and which may be expressed in a cell at a particular given time. ESTs are currently used as a fast and efficient method of profiling genes expressed in various tissues, cell types or developmental stages (Adams et al., 1991). These sequences are mainly stored into three databases which are again interconnected. These three databases are (i) GenBank in the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), (ii) the European Molecular Biology Laboratory nucleotide sequence database (EMBL, http://www.ebi.ac.uk/embl/) and (iii) the DNA database bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/). Also, recently many specific databases are set up for specific researches or specific species, for example databases in The *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org/). These nucleotide sequences have become a valuable and cheap source for developing molecular markers which has opened up a new chapter in molecular markers as genic molecular markers (GMMs) which are developed directly from coding sequences like ESTs or fully characterized genes (Anderson and Lubberstedt, 2003).

The identification of sequences among all others which can be used as markers thus is fundamental to development of GMMs. Amongst others, bioinformatics approaches are very useful for the development of GMMs, making their development much faster and cheaper (Anderson and Lubberstedt, 2003). A number of software programs have been implemented for identification of molecular markers from sequence data. SSRs and SNPs markers are abundant in genomic sequences as well as in ESTs which can be detected automatically with the help of different programs and pipelines developed for mining these markers from public sequences (Ching et al., 2002). An understanding of the different tools and bioinformatics techniques for marker identification and/or development will enable plant breeders and researchers working in other relevant disciplines to work together towards a common goal of increasing the efficiency of global food production.

SNP MARKER IDENTIFICATION AND DEVELOPMENT

Single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide -A-T-C or G- in the genome differs between members of a species (or between paired chromosomes in an individual) (Ching et al., 2002). There are three different categories of SNPs: transitions (C/T or G/A), transversions (C/G, A/T, C/A, or T/G) and small insertions/deletions (indels). SNPs at any particular site could be principle in bi-, tri- or tetra-allelic, however tri- and tetra-allelic SNPs are rare, and in practice SNPs are generally biallelic (Doveri et al., 2008). SNPs may occur in the coding, non-coding and intergenic regions of the genome, thus enabling the discovery of genes as a result of the differences in the nucleotide sequences. In recent years, many research papers have reported SNPs as excellent markers for association mapping of polygenic traits with highest map resolution (Botstein and Risch, 2003; Brookes, 1999; Bhattramakki et al., 2002). Also, SNPs are reported to be the most frequent type of variation found in DNA (Brookes, 1999; Cho et al., 1999), with their discovery together with insertions/deletions has formed the basis of most differences between alleles. In *Arabidopsis*, over 37, 000 SNPs have been identified through the comparison of two accessions (Jander et al., 2002). Ching et al. (2002) reported that they occur in a frequency of one non-coding SNP per 31 bp and 1 coding SNP per 124 bp in 18 maize genes assayed in 36 inbred lines. A number of EST collections have been used to describe and detect SNPs in maize (*Zea mays* L.) (Ching et al., 2002) and Soybean (*Glycine max* L.) (Zhu et al., 2003).

Different strategies used for development of new SNP markers can be broadly classified under two categories. The first is a wet lab method (experimental) and the other is the computational (bioinformatics) methods. The experimental method of SNPs discovery is expensive and

time consuming (Schlotterer, 2004; Useche et al., 2001). Also, the infra structure needed for, may be unavailable to laboratories in the under and developing world. In contrast, a computational approach to discover potential SNPs from publicly available sequences makes the development of SNP markers rapid and less expensive. For computational SNP discovery, two important points should be considered. First, the program should be able to distinguish allelic variation from sequence variation between paralogous sequences (Marth et al., 1999; Le Dantec et al., 2004; Batley et al., 2003). Secondly, the program should be able to recognize sequencing errors which are usually caused by poor quality sequences, especially for EST data (Picoult-Newberg et al., 1999; Garg et al., 1999; Batley et al., 2003; Matukumalli et al., 2006).

Mining of SNPs from EST sequences is an attractive method for marker development in plants where genome sequences are not yet available. The steps involved in SNP discovery from EST sequences include clustering, sequence assembly and SNP detection (Batley et al., 2003). There is several bioinformatics software to handle each of these steps. A number of methods used to identify SNPs in aligned sequence data rely on sequence trace file analysis to filter out sequence errors by their dubious trace quality (Marth et al., 1999). The major drawback to this approach is that the sequence trace files required are rarely available for large sequence datasets collected from a variety of sources. In cases where trace files are unavailable, two complementary approaches have been adopted to differentiate between sequence errors and true polymorphisms: (i) assessing redundancy of the polymorphism in an alignment, and (ii) assessing co-segregation of SNPs to define a haplotype. The most important limitation for use of EST for SNP marker development is that EST data provides very limited polymorphisms (Matukumalli et al., 2006). Also, other factors such as alternative splicing, reverse transcription errors and RNA editing interfere with the predictions even after including sequence quality scores. But SNP discovery from EST sequences was successfully implemented for maize (Rafalski, 2002) and pine (Le Dantec et al., 2004) species by constructing a software data analysis pipeline. Thus, the selection of optimal tool for SNP identification and/or discovery basically depends on the nature of input sequences. A number of pipelines have been developed to automatically detect SNPs in sequences which have been listed in Table 1.

TOOLS REQUIRING TRACE FILES

In late 1990s, efforts were being made to develop computer programs to automate base calling (Phred), sequence assembly (Phrap) and sequence assembly editing (Consed) to analyze the results of fluorescence based sequencing. Nickerson et al. (1997) came forward with a program called 'PolyPhred' that automatically detects the presence of heterozygous single nucleotide substitutions by fluorescence - based sequencing of PCR products. When sequences containing known variants were analysed using this program, approximately 99% accuracy was found. Polyphred is widely used because it can detect heterozygous bases from two alleles within an individual (Matukumalli et al., 2006). This was one of the major developments with regard to automated detection of SNPs. Another tool which requires sequence trace files is PolyBayes which uses a Bayesian-statistical model to find differences within assembled sequences based on the depth of coverage, the base quality values and the expected rate of polymorphic sites in the region (Marth et al., 1999).

Another software tool, which came forward in due time for automated identification of SNPs and mutations in fluorescence-based re-sequencing reads is SNPdetector (Zhang et al., 2005). This software tool was designed to model the process of human visual inspection with a very low false positive and false negative rate. The author states superior performance of SNPdetector in SNP and mutation analysis by comparing its results with those derived by human inspection, PolyPhred and independent genotype assays in three different large-scale investigations (Zhang et al., 2005). SNPdetector runs on Unix/Linux platform and is available publicly (http://lpg.nci.nih.gov). Another user friendly, freely available software tool for inspecting SNP based genetic variations is novoSNP (Weckx et al., 2005) and InSNP (Manaster et al., 2005). The author of both software tool states it to perform better than that of PolyPhred and PolyBayes.

An improved version of novoSNP (Weckx et al., 2005) came as novoSNP3 (Rijk et al., 2007) that along with discovering SNPs and indels polymorphisms in sequence trace files, can also be used to create databases containing annotated reference sequences, add and align trace data, keep track of validation status of variants, annotate variants, and produce reports on validated variants and genotypes. novoSNP is available from http://www.molgen.ua.ac.be/bioinfo/novosnp. There are versions for MS Windows as well as Linux. Software tool SNP-PHAGE (SNP discovery Pipeline with additional features for identification of common haplotypes within a sequence tagged site (Haplotype Analysis and GenBank (-dbSNP) submissions) was applied for analyzing sequence traces from diverse soybean genotypes to discover over 10,000 SNPs (Matukumalli et al., 2006). This package is being made available at open source at http://bfgl.anri.barc.usda.gov/ML/snp-phage/. SNP-PHAGE uses PolyBayes (Marth et al., 1999) and PolyPhred (Nickerson et al., 1997) for analysis, storing and editing of polymorphisms information in a relational database through a user friendly web interface. SNP-PHAGE was used to analyze sequences from diverse soybean genotypes with discovery of 10,000 SNPs. SNP-PHAGE is freely available at http://bfgl.anri.barc.usda.gov/ML/snp-phage/.

Table 1. Tools for single nucleotide polymorphisms Identification.

Program	Website	Reference
PolyPhred	http://droog.mbt.washington.edu/	Nickerson et al., 1997
PolyBayes	http://bioinformatics.bc.edu/marthlab/PolyBayes	Marth et al., 1999
autoSNP	http://acpfg.imb.uq.edu.au	Batley et al., 2003
SNPdetector	http://lpg.nci.nih.gov	Zhang et al., 2005
InSNP	at www.mucosa.de/insnp/	Manaster et al., 2005
novoSNP	http://www.molgen.ua.ac.be/bioinfo/novosnp	Weckx et al., 2005
SNPServer	http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html	Savage et al., 2005
SNP-PHAGE	http://bfgl.anri.barc.usda.gov/ML/snp-phage	Matukumalli et al., 2006
QualitySNP	http://www.bioinformatics.nl/tools/snpweb/	Tang et al., 2006
HaploSNPer	http://www.bioinformatics.nl/tools/haplosnper/	Tang et al., 2008
Seq-SNPing	(http://bio.kuas.edu.tw/Seq-SNPing	Chang et al., 2009
SNiPlay	http://sniplay.cirad.fr/	Dereeper et al., 2011

Tools detecting SNPs without trace files

AutoSNP software program was developed to detect SNPs and indels from EST sequences (Batley et al., 2003). This program uses d2cluster (Burke et al., 1999) and cap3 (Huang and Madan, 1999) to cluster and align EST sequences, and uses redundancy to differentiate between candidate SNPs and sequence errors. Candidate polymorphisms are identified as occurring in multiple reads within an alignment. AutoSNP calculates two associated measurements of confidence in the validity of SNPs for each polymorphism. The frequency of occurrence of a polymorphism at a particular locus provides a primary measure of confidence in the SNP representing a true polymorphism and is referred to as the SNP redundancy score. The co-segregation of multiple SNPs within an alignment to define a haplotype provides a second measure of confidence in SNP validity and is referred to as the co-segregation score.

QualitySNP (http://www.bioinformatics.nl/tools/snpweb/) was reported to be an efficient tool for SNP detection, storage and retrieval in diploid as well as polyploidy species. It can be run on Linux or UNIX system (Tang et al., 2006). It uses a haplotype-based strategy to detect reliable synonymous and non-synonymous SNPs from public EST data without the requirement of trace/quality files or genomic sequence data. Haplotypes in this context represent the different alleles of a gene in a dataset. The haplotype reconstruction is based on a mathematical algorithm. It uses three filters for the identification of reliable SNPs. Filter 1 screens for all potential SNPs and identifies variation between or within genotypes. Filter 2 is the core filter that uses a haplotype-based strategy to detect reliable SNPs. Clusters with potential paralogs as well as false SNPs caused by sequencing errors is identified. Filter 3 screens SNPs by calculating a confidence score, based upon sequence redundancy and quality. Non-synonymous SNPs are subsequently identified by detecting open reading frames of consensus sequences (contigs) with SNPs. The pipeline includes a data storage and retrieval system for haplotypes, SNPs and alignments. QualitySNP's versatility was demonstrated by the identification of SNPs in EST datasets from potato, chicken and humans (Tang et al., 2006).

HaploSNPer (http://www.bioinformatics.nl/tools/haplosnper/) is web-based SNP discovery and allele detection tool based on QualitySNP (Tang et al., 2008). It is a flexible web-based tool for detecting SNPs and alleles in user-specified input sequences from both diploid and polyploidy species. It includes BLAST for finding homologous sequences in public EST databases, CAP3 or PHRAP for aligning them, and QualitySNP for discovering reliable allelic sequences and SNPs. Also, HaploSNPer provides a user friendly interface for visualization of SNP and alleles. Singhal et al. (2011) used HaploSNPer and found 40589 reliable SNPs in *Sorghom bicolor* genome. Although, HaploSNPer is able to detect SNP, allele, haplotype reconstruction but it does not extend the analysis to diversity, linkage disequilibrium or haplotype network study. Another web based tool which fulfil this need came forward in 2011 called SNiPlay (Dereeper et al., 2011) which is expected to assist biologists in extracting and analyzing polymorphism data in a simple and robust way. SniPlay (http://sniplay.cirad.fr/) is reported to be a user-friendly and integrative web-based tool dedicated to polymorphism discovery and analysis. It integrates pipeline which is freely accessible through the internet, combining existing software's with new tools to detect SNPs and to compute different types of statistical indices and graphical layouts for SNP data. It is able to detect SNPs and indels from standard sequence alignments, genotyping data or Sanger sequencing. Furthermore, the pipeline allows the use of external data (such as phenotype, geographic origin, taxa, stratification) to define groups and compare statistical indices. It also integrates database for storing polymorphisms, genotyping data and grapevine sequences released by public and private projects which allows the user to retrieve SNPs using various filters (such as genomic position, missing data, polymorphism type, allele frequency). Also, it can

be used to compare SNP patterns between populations (Dereeper et al., 2011).

SNPServer (Savage et al., 2005) is a real time implementation of the autoSNP method, accessed via a web server. It uses autoSNP software by providing a web interface for sequence input, comparison and assembly and permits rapid discovery of SNPs. SNPServer (http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html) uses BLAST to identify related sequences, and CAP3, to cluster and align these sequences. The alignments are parsed to the SNP discovery software autoSNP.

All the above mentioned tools were developed to discover single nucleotide polymorphisms (SNPs) derived from re-sequencing. Whether an identified SNP is indeed a novel SNP or is already contained in dbSNP was a big question and sometimes confusing. Chang et al. (2009) came forward with freely available software called 'Seq-SNPing' (http://bio.kuas.edu.tw/Seq-SNPing), which is Java-based software for SNP discovery, and ID identification and editing and visualization of sequence alignments. According to its author, it is easy to use, fast, and provides an accurate method for searching and organizing SNP IDs from multiple sequence inputs, thereby greatly facilitating genetic studies.

Different software tools described above were designed based on the needs of different developers. InSNP is windows based package and can be helpful for users not familiar with Linux. SNPdetector scripts work only on Unix/Linux platforms and use the Smith-Waterman algorithm for aligning reads, as well as a modified version of the NQS (Altshuler et al., 2000) method for detecting homozygous SNPs among different individuals. Also, SNP detector requires a minimum of a 30% threshold for secondary peak intensity for detecting heterozygous SNPs. NovoSNP works on windows as well as Unix/Linux based platforms. NovoSNP uses BLAST (Altschul et al., 1990) for aligning sequence reads and uses a series of filters to reduce false positives. This package is configured to work with a database, and, hence, it makes polymorphism discovery and data storage convenient. Other polymorphism discovery software, such as auto SNPrely on redundancy and co-segregation of markers within a sequence are useful when trace data are not available.

SSR MARKER IDENTIFICATION AND DEVELOPMENT

Microsatellites or SSRs are shot tandem repeats of 1-6 nucleotides that occur with high frequency throughout the genomes of many organisms (Weber, 1990). There polymorphisms consists of variations in the number of repeats, which was suggested to be due to slippage of polymerase (Kruglyak et al., 1998). SSRs have been reported to be superior to other molecular markers because (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 (Varshney et al., 2005). Due to these features, SSRs have become valuable genetic markers for linkage mapping, QTL mapping, association mapping and diversity analysis (Jones et al., 1997; Powell et al., 1996; Varshney et al., 2005). Conventional methods for developing SSRs is laborious, time consuming and expensive (Powell et al., 1996) which involves construction of genomic libraries and subsequent screening for the presence of SSR repeat motifs in the clones (Powell et al., 1996). With the recent advancement and establishment of EST sequencing projects in several plant species, a wealth of DNA sequence information has been generated and deposited in public databases (Rudd, 2003). Also, sequence data for many fully characterized genes and full length cDNA clones have been generated for some plant species (Varshney et al., 2005). Genic SSRs have certain noticeable advantages over genomic SSRs. They are (i) quickly obtained by electronic sorting, (ii) represents functional region of the genome and (iii) more transferable between related species (Gao et al., 2003; Cordeiro et al., 2001; Decroocq et al., 2003; Yu et al., 2004; Varshney et al., 2005; Tang et al., 2006). The presence of SSR in expressed region of genomes suggests that they may have a role in gene expression or function. For example the *waxy* gene in rice has been found to contain a poly(CT) microsatellite in the 5'-untranslated region (UTR) whose length polymorphisms is associated with amylase content (Ayres et al., 1997). In general, approximately 5% of plants ESTs contain SSRs with a minimum length of 20 nucleotides (Varshney et al., 2005; Kantety et al., 2002; Ghislain et al., 2004; Poncet et al., 2006). Thus, *in silico* approaches for screening SSRs from sequences have become efficient and inexpensive alternative for plant species. Different software tools that have been developed to detect SSRs are listed in Table 2.

Sputnik is a C language program that searches DNA sequence file in FASTA format for microsatellite repeats. It uses a recursive algorithm to search for repeated patterns of nucleotides of length between 2 and 5 (Abajian, 1994) and finds perfect, compound and imperfect repeats. The output is a file of SSRs in tabular format. Unix, Linux and windows versions of sputnik are available from http://espressosoftware.com/pages/sputnik.jsp and http://cbi.labri.fr/outils/Pise/sputnik.html. Sputnik has been applied for SSR identification in many species including *Arabidopsis* and barley (Cardle et al., 2000). Tandem Repeats Finder (TRF) (Benson, 1999) (http://tandem.bu.edu/trf/trf.html) can find very large SSR repeats, up to a length of 2000 bp. It uses a set of statistical tests for reporting SSRs, which is based on four distributions of pattern length, the matching probability, the indel probability and the tuple size. TRF finds perfect, imperfect and compound SSRs, and is available for Linux. TRF has been used for SSR identification in cowpea (Chen et al., 2007).

Table 2. Bioinformatics tools for microsatellites identification.

Program	WebSite	Reference
Sputnik	http://espressosoftware.com/pages/sputnik.jsp	Abajian, 1994
Tandem repeat Finder (TRF)	http://tandem.bu.edu/trf/trf.html	Benson, 1999
SSR identification Tool (SSRIT)	http://www.gramene.org/db/searches/ssrtool	Kantety et al., 2002
Tandem repeat Occurrence Locator (TROLL)	http://wsmartins.net/webtroll/troll.html	Castelo et al., 2002
MicroSAtellite (MISA)	http://pgrc.ipk-gatersleben.de/misa/	Thiel et al., 2003
RepeatFinder	http://www.cbcb.umd.edu/software/RepeatFinder/	Volfovsky et al 2001
SSR Locator	http://www.ufpel.edu.br/	Maia et al., 2008
SSRPoly	http://acpfg.imb.uq.edu.au/ssrpoly.php	Tang et al., 2008

The tool Simple Sequence Repeat Identification Tool (SSRIT) (http://www.gramene.org/db/searches/ssrtool , Temnykh et al., 2001) uses Perl script to find perfect SSR repeats (2 to 10 bp in length) within a sequence. Kantety et al. (2001) used SSRIT to mine SSR in ESTs from Barley, maize, rice, sorghum and wheat. Singh et al. (2011) used SSRIT to mine SSRs in wheat rust *Puccinia* sp. Another SSR identification tool is TROLL (Tandem Repeat Occurrence Locator, Castelo et al., 2002) which draws a keyword tree and matches it with a technique adapted from bibliographic searches, based on the *Aho-Corasick* algorithm. One of the major disadvantages of TROLL is that it cannot handle very large sequences and cannot process large batches of sequences as the tree takes up large amounts of memory.

The microsatellite (MISA) tool (http://pgrc.ipk-gatersleben.de/misa/) identifies perfect, compound and interrupted SSRs. It requires a set of sequences in FASTA format and a parameter file that defines unit size and minimum repeat number of each SSR. The output includes a file containing the tables of repeat found, and a summary file. MISA can also design PCR amplification primers either side of SSR. The tool is written in Perl and is therefore platform independent, but it requires as installation of Primer3 for primer search (Thiel et al., 2003). MISA has been applied for SSR identification in coffee (Aggarwal et al., 2007), barley (Thiel et al., 2003; Kota et al., 2001), wheat (Yu et al., 2004), rye (Khlestkina et al., 2004) and peanut (Liang et al., 2009). Another SSR search tool called as 'Repeat Finder' (Volfovsky et al., 2001) (http://www.cbcb.umd.edu/software/RepeatFinder/) finds SSRs in four steps. The first step is to find all exact repeats using Repeat Match or REPuter. The second step merges repeats together into repeat classes and the third step includes merging all of the other repeats that match those already merged, into the same classes. Finally, step four matches all repeats and classes against each other in a non-exact manner using BLAST. The input is a genome or set of sequences, and the output is a file containing the repeat classes and number of merged repeats found in each class. Repeat Finder can finds repeats of any length. Also it finds perfect, imperfect and compound repeats and runs on Unix or Linux. It has been used to identify SSRs in peanut (Jayashree et al., 2005).

SSRPrimer combines Sputnik and the PCR primer design software Primer3 to find SSRs and associated amplification primers (Robinson et al., 2004, Jewell et al., 2006). It takes multiple sequences in FASTA format as input and produce lists of SSRs and associated PCR primers in tabular format. SSRPrimer has been applied to a wide range of species including shrimp (Perez et al., 2005), citrus (Chen et al., 2006), mint (Lindqvist et al., 2006), strawberry (Keniry et al., 2006), *Brassica* (Batley et al., 2007; Burgess et al., 2006; Hopkins et al., 2007; Ling et al., 2007), *Sclerotinia* (Winton et al., 2007) and *Eragrostiscurvula* (Cervigni et al., 2008).

Maia et al. (2008) came with an interesting tool for SSR discovery integrated with primer design and PCR simulation called SSR Locator (http://www.ufpel.edu.br/). SSR Locator detects SSR and minisatellite motifs between 1 and 10 bp, design primer for each locus found, amplify fragments with different primer pairs from a given set of FASTA files, produce global alignment between amplicons generated by the same primer pair and estimates alignment scores and identities between amplicons thus generating information on primer specificity and redundancy. Victoria et al. (2011) used SSR Locator to study the pattern of EST derived microsatellite markers for model plants.

All the SSR identification tool described above are not able to identify polymorphic SSRs. The only tool which is capable of identifying polymorphic SSRs from DNA sequence data is SSRPoly (http://acpfg.imb.uq.edu.au/ssrpoly.php). The input is a file of FASTA format sequences. SSRPoly includes a set of Perl scripts and MySQL tables that can be implemented on UNIX, Linux and Windows platforms (Tang et al., 2008).

CONCLUSION

The recent advances in bioinformatics role in genic molecular marker development will assist molecular biologists to address many evolutionary, ecological and taxonomic research questions. The development of bioinformatics tools will improve marker identification with reducing cost

and therefore will help plant breeders to include more diverse species and a greater variety of traits. Bioinformatics tools have been developed to mine sequence data for markers and present these in a biologist friendly manner. The SNP and SSR marker have many uses in plant genetics such as the detection of alleles associated with disease, genome mapping, association studies, genetic diversity and inferences of population history. With the help of these tools molecular plant breeders will be able to develop new/novel markers and use these markers in diverse applications. The availability of large sequence data makes it an economical choice to develop SSR and SNP marker from it. EST SSR and SNP are gene specific and thus functional molecular markers. Several computational tools described here for the identification of SNPs and SSRs in sequence data as well as for the design of PCR amplification primers will help plant breeders new to molecular breeding and marker assisted selection to opt SSR and SNP marker to solve crop breeding related problems.

ACKNOWLEDGEMENTS

The authors acknowledge the generous funding of Department of Biotechnology, Government of India and also the encouragement and support of the Director of Institute of Advanced Study in Science and Technology (IASST), Guwahati and The Energy and Resource Institute (TERI).

REFERENCES

Abajian C (1994) Sputnik,: (http://espressosoftware.com/sputnik /index.html)

Adams MD, Kelly JM, Gocayne JD, Dubnick M, Polymeropoulus MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991). Links complementary DNA sequencing: expressed sequence tags and human genome projects. Science. 252:1651-1656.

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor. Appl. Genet. 114: 359-372.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nat. 407:513-516

Anderson JR, Lubberstedt T (2003) Functional markers in plants. Trend Plant Sci. 8:554-560.

Ayres MM, Mcclung AM, Larkin PD, Bligh FJ, Jonesc A, Park WD (1997) Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. Theor. Appl. Genet. 94:773-781

Batley J, Barker G, O' Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Pl. Physiol. 132: 84-91.

Batley J, Hopkins CJ, Cogan NOI, Hand M, Jewell E, Kaur J, Kaur S, Li X, Ling AE, Love C, Mountford H, Todorovic M, Vardy M, Walkiewicz M, Spangenberg, Edwards D. (2007) Identification and characterization of simple sequence repeat markers from Brassica napus expressed sequences. Mol. Ecol. Notes. 7:886-889.

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

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (2002). Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol. Biol. 48: 539-547.

Botstein D, Risch N (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat. Genet. 33: 228-237.

Brookes AJ (1999). The essence of SNPs. Gene 234:177-186.

Burgess B, Mountford H, Hopkins CJ, Love C, Ling AE, Spangenberg GC, Edwards D, Batley J (2006) Identification and characterization of simple sequence repeat (SSR) markers derived in silico from Brassica oleracea genome shotgun sequences. Mol. Ecol. Notes. 1191-1194.

Burke J, Davison D, Hide W (1999). d2_cluster: A validated method for clustering EST and full-length cDNA sequences. Genome Res. 9:1135-1142.

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

Castelo AT, Martins W, Gao GR (2002) TROLL-Tandem Repeat Occurrence Locator. Bioinf.18:634-636.

Cervigni GD, Paniego N, Díaz M, Selva JP, Zappacosta D, Zanazzi D, Landerreche I, Martelotto L, Felitti S, Pessino S, Spangenberg G, Echenique V (2008) Expressed sequence tag analysis and development of gene associated markers in a near-isogenic plant system of Eragrostiscurvula. Plant. Mol Biol. 67:1-10.

Chang H, Chuang L, Cheng Y, Ho C, Wen C, Yang C (2009) Seq-SNPing: Multiple-Alignment Tool for SNP Discovery, SNP ID Identification, and RFLP Genotyping. OMICS: A J. Int. Biol. 13(3):253-260.

Chen CX, Zhou P, Choi YA, Huang S, Gmitter FG(2006) Mining and characterizing microsatellites from citrus ESTs. Theor. Appl. Genet. 112:1248-1257.

Chen XF, Laudeman TW, Rushton PJ, Spraggins TA, Timko MP (2007) CGKB: an annotation knowledge base for cowpea (Vignaunguiculata L.) methylation filtered genomic genespace sequences. BMC Bioinf. 8: 112-116.

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002).SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines.BMC Genet. 3:19-24.

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubel E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, AusubelFM, Davis RW, Oefner PJ (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat. Genet. 23: 203-207

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005)An introduction to markers, quantitative trait loci (QTL) mappintg and marker-assisted selection for crop improvement: The basic concepts. Euphytica. 142:169-196.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (Saccharum spp.) EST cross transferable to erianthus and sorghum. Plant. Sci. 160: 1115-1123.

Decroocq V, Fave MG, Hagen L, Bordenave L, Decroocq S (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor. Appl. Genet. 106: 912-922.

Dereeper A, Nicolas S, Cunff LL, Bacilieri R, Doligez A, Peros J, Ruiz M, This P (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC Bioinfo. 12:134-140

Doveri S, Lee D, Maheswaran M, Powell W (2008). Molecular markers: History, features and applications. In Principles and Practices of Plant Genomics, Volume 1, C.K.a.A.G. Abbott, ed. (Enfield, USA: Science Publishers), pp. 23-68.

Gao LF, Tang JF, Li HW, Jia JZ (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol. Breed.12:245-261.

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Gen. Res. 9: 1087-1092.

Ghislain M, Spooner DM, Rodríguez F, Villamón F, Núñez J, Vásquez

C, Waugh R, Bonierbale M (2004) Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. Theor. Appl. Genet. 108:881-890.
Gupta PK, Rustgi S (2004). Molecular markers from the transcribed /expressed region of the genome in higher plants.Funct.Int. Gen. 4:139-162.
Hopkins CJ, Cogan NOI, Hand M, Jewell E, Kaur J, Li X, Lim GAC, Ling A, Love C, Mountford H, Todorovic M, Vardy M, Spangenberg GC, Edwards D, Batley J (2007) Sixteen new simple sequence repeat markers from Brassica juncea expressed sequences and their cross-species amplification. Mol. Ecol. Notes. 7: 697-700.
Huang XQ, Madan A.(1999) CAP3: A DNA sequence assembly program. Genome Res. 9: 868-877.
Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002). Arabidopsis map-based cloning in the plant-genome era. Plant Physiol. 129: 440-450.
Jayashree B, Ferguson M, Ilut D, Doyle J, Crouch JH (2005) Analysis of genomic sequences from peanut (Arachishypogaea). Electron. J. Biotech. 8.
Jewell E, Robinson A, Savage D, Erwin T, Love CG, Lim GAC, Li X, Batley J, Spangenberg GC, Edwards D(2006) SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery. Nucleic Acids Res. 34:W656-W659.
Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, van de Wiel C, Bredemeijer G, Vosman B, Matthes M, Daly A (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol. Breed. 3: 381-390.
Kantety RV, Rota ML, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barely, maize, rice, sorghum and wheat. Plant Mol. Biol. 48: 501-510.
Keniry A, Hopkins CJ, Jewell E, Morrison B, Spangenberg GC, Edwards D, Batley J (2006) Identification and characterization of simple sequence repeat (SSR) markers from Fragaria x ananassa expressed sequences. Mol. Ecol. Notes. 6: 319-322.
Khlestkina EK, Than MHM Pestsova EG, Röder MS, Malyshev SV, Korzun V, Börner A (2004) Mapping of 99 new microsatellite-derived loci in rye (Secalecereale L.) including 39 expressed sequence tags Theor. Appl. Genet. 109:725-732
Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (Hordeumvulgare L.). Hereditas 135:145-151.
Kruglyak S, Durrett R, Schug MD, Aquadro CF (1998) Equilibrium distribution of microsatellite repeats length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA. 95:10074-10078.
Le Dantec L, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio JM, Chaumeil P, Leger P, Garcia V, Legrait F, de Daruvar A, Plomion C (2004) Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Mol. Biol. 54: 461-470.
Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, Guo B (2009) Utility of EST-derived SSR in cultivated peanut (Arachishypogaea L.) and Arachis wild species BMC Plant Biol. 9:35.
Lindqvist C, Scheen AC, Yoo MJ, Grey P, Oppenheimer DG, Leebens-Mack JH, Soltis DE, Soltis PS, Albert VA (2006) An expressed sequence tag (EST) library from developing fruits of an Hawaiian endemic mint (Stenogynerugosa, Lamiaceae): characterization and microsatellite markers. BMC Plant Biol. 6: 16.
Ling AE, Kaur J, Burgess B, Hand M, Hopkins CJ, Li X, Love CG, Vardy M, Walkiewicz M, Spangenberg G, Edwards D, Batley J (2007). Characterization of simple sequence repeat markers derived in silico from Brassica rapa bacterial artificial chromosome sequences and their application in Brassica napus. Mol. Ecol. Notes. 7: 273-277.
Maia LC, Palmieri DA, Souza VQ, Kopp MM, Carvalho FI, Oliveira AC (2008) SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. Int. J. Plant Gen.
Manaster C, Zheng W, Teuber M, Wächter S, Döring F, Schreiber S, Hampe J (2005). InSNP: A tool for automated detection and visualization of SNPs and InDels. Human Mutation. 26(1):11-19
Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR (1999). A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23: 452-456.
Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP (2006). Application of machine learning in SNP discovery. BMC Bioinf. 7:44-51.
Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 25: 2745-2751.
Perez F, Ortiz J, Zhinaula M, Gonzabay C, Calderon J, Volckaert F (2005) Development of EST-SSR markers by data mining in three species of shrimp: Litopenaeusvannamei, Litopenaeusstylirostris, and Trachypenaeus birdy. Marine Biotechnol. 7:554-569.
Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167-174.
Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, Kochko A, Hamon P (2006) SSR mining in coffee tree EST database: potential use of EST-SSRs as markers for the coffea genus. Mol. Gen. Genomics 276:436-449
Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1:215-222.
Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics.Curr. Op. Plant Biol.5:94-100.
Rijk PD, Del-Favero J (2007). novoSNP3: variant detection and sequence annotation in resequencing projects. Methods Mol. Biol. 396:331-344.
Robinson AJ, Love CG, Batley J, Barker G, Edwards D (2004) Simple sequence repeat marker loci discovery using SSR primer. Bioinfor. 20: 1475-1476.
Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci. 8, 321–329.
Savage D, Batley J, Erwin T, Logan E, Love CG, Lim GAC, Mongin E, Barker G, Spangenberg GC, Edwards D (2005) Nucleic Acids Research. 33: W493-W495.
Schlotterer C (2004) The evolution of molecular markers - just a matter of fashion? Nat Rev Genet 5: 63-69.
Singhal D, Gupta P, Sharma P, Kashyap N, Anand S, Sharma H (2011) In-silico single nucleotide polymorphisms (SNP) mining of Sorghum bicolor genome. African Journal of Biotechnology. 10(4):580-583.
Tang J, Leunissen JAM, Voorrips RE, Linden CG, Vosman B (2008) HaploSNPer: a web-based allele and SNP detection tool. BMC Genet.9:23
Tang JF, Gao LF, Cao YS, Jia JZ (2006) Homologous analysis of SSR-ESTs and transferability of wheat SSR-EST markers across barley, rice and maize. Euphytica 151:87-93.
Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11:1441-1452.
Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeumvulgare L.). Theor. Appl. Genet. 106:411-422.
Useche FJ, Gao G, Harafey M, RafalskiA (2001). High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform 12: 194-203.
Varshney RK, Sigmund R, Boerner A, Korzun V, Stein N, Sorrells M, Langridge P, Graner A (2005). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Science. 168:195-202
Victoria FC, Maia LC, Oliveira AC (2011). Insilico comparative analysis of SSR markers in plants. BMC Plant Biol. 11:15
Volfovsky N, Haas BJ, Salzberg SL (2001). A clustering method for repeat analysis in DNA sequences. Genome Biol. 2. (8)
Weber JL (1990) Genomics 7:524-530.
Weckx S, Del Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 15:436-442
Winter P, Kahl G (1995). Molecular marker technologies for plant improvement.World J. Microbiol. Biotechnol. 11:438-448.

Winton LM, Krohn AL, Leiner RH (2007). Microsatellite markers for Sclerotiniasubarctica nom. prov., a new vegetable pathogen of the High North. Mol. Ecol. Notes. 7:1077-1079.
Yu Jk, Dake TM, Singh S, Benscher D, Li W, Gill BS, and Sorrells ME (2004). Development and mapping of EST-derived simple sequence repeat (SSR) markers for hexaploid wheat. Genome 47:805-818.
Zhang J, Wheeler DA, YakubI,Wei S, Sood R, Rowe W, Liu PP, Gibbs RA, Buetow KH (2005). SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. PLOS Comput. Biol. 1 (5):395-404.
Zhu YL, Song QJ, Hyten DL, Van Tasell, CP, Matukumali, LK, Grim DR, Hyatt, SM, Fickus EW, Young ND, Cregan PB (2003). Single nucleotide polymorphisms in soybean. Genetics. 163:1123-1134.