African Journal of Biotechnology Vol. 10 (4), pp. 580-583, 24 January, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1444
ISSN 1684–5315 © 2011 Academic Journals

Full Length Research Paper

In-silico single nucleotide polymorphisms (SNP) mining of *Sorghum bicolor* genome

Divya Singhal*, Priyanka Gupta, Pragati Sharma, Neha Kashyap, Siddharth Anand and Himanshu Sharma

Department of Biotechnology, Ambala College of Engineering and Applied Research, Mithapur, Ambala Cantt, 133101, Haryana, India.

Accepted 7 December, 2010

Single nucleotide polymorphisms (SNPs) may be considered the ultimate genetic markers as they represent the finest resolution of a DNA sequence (a single nucleotide), and are generally abundant in populations with a low mutation rate. SNPs are important tools in studying complex genetic traits and genome evolution. SNP mining can be done by experimental and computational methods. Computational strategies for SNP discovery make use of a large number of sequences present in public databases [in most cases as expressed sequence tags (ESTs)] and are considered to be faster and more cost-effective than experimental procedures. A major challenge in computational SNP discovery is distinguishing allelic variation from sequence variation between paralogous sequences, in addition to recognizing sequencing errors. For the majority of the public EST sequences, trace or quality files are lacking which makes detection of reliable SNPs even more difficult because it has to rely on sequence comparisons only. In the present study, online SNP and allele detection tool HaploSNPer (based on QualitySNP pipeline) and *Sorghum bicolor* genome was used. As a result, 77094 potential SNPs and 40589 reliable SNPs were detected in *S. bicolor*. In the 77094 potential SNPs detected transitions, transversions and indels were 34398, 35871 and 6825, respectively. In the 40589 reliable SNPs detected transitions, transversions and indels were 17042, 20500 and 3047, respectively.

Key words: Single nucleotide polymorphisms (SNP), expressed sequence tags (EST), HaploSNPer.

INTRODUCTION

Sorghum is the seed of a monocot plant *Sorghum bicolor* (L.) Moench, of the grass family (Poaceae). It is the fifth most widely produced crop in the world. It is largely produced in U.S.A., India, Mexico, China and Africa. U.S.A. is the largest producer of sorghum followed by India. It is a rich source of proteins, vitamins and carbohydrates for millions of poor people. In addition to this, it is one of the most economically important plants as is a vital source of not only food products but also alcohol and bio-fuel. But these applications can be better put to practice if the genes responsible for coding various agronomical characters of sorghum are determined. Determination of protein coding genes or genes responsible for causing diseases can be very helpful in producing better enhanced disease free varieties of sorghum. But determination of protein coding genes is a tedious job.

During the past few decades, much work has been done to explore genomes of various vital organisms, especially plants. In order to achieve this, many studies have been carried out to detect efficient genetic markers as these markers are usually present near the protein coding genes, which can thus be used for identification of genes of interest (Kollers et al., 2009). Many types of biological markers exist but single nucleotide polymorphisms (SNPs) have been found to be the most efficient genetic marker for gene identification. SNPs are the most common types of DNA polymorphism. SNP are co-dominant, bi-allelic, highly polymorphic and have good reproducibility that makes them an efficient biological marker. As biological markers, SNPs have been found to

*Corresponding author. E-mail: divysinghal@gmail.com. Tel: 09466290947.

Abbreviations: SNPs, Single nucleotide polymorphisms; **EST,** expressed sequence tags.

be very useful in gene mapping, gene identification, drug development, etc, (Botstein and Risch 2003). In order to use SNPs as markers, many scientists and researchers are carrying out studies to discover SNPs in genome sequences. Initially, only the conventional wet lab (rDNA) techniques were used to detect SNPs. But gradually with the development of bioinformatics, various *in-silico* tools like Polybayes, SNPhunter (Xiang et al., 2009), Haplo-SNPer (Tang et al., 2006), etc, have been designed to detect SNPs. These *in-silico* tools detect SNPs using various publically available expressed sequence tags (EST) databases. Such a detection of SNPs through *in-silico* tools has been very useful in functional genomics, pharmacogenetics studies and agronomic studies. In this view, this present study was done to determine SNPs in the economically important *S. bicolor* crop. Here, among the various tools, HaploSNPer, a web based allele and SNP detection tool has been used along with EMBLs, publically available EST database (Picoult-Newberg et al., 1999)

MATERIALS AND METHODS

In the current study, SNP mining was done for *S. bicolor* with the online allele and SNP detection tool HaploSNPer (Tang et al., 2006)

Retrieval of sequence data

The first step was the retrieval of *S. bicolor* genome sequence data through the FTP site of the National Center for Biotechnology Information (NCBI) (Paterson, 2009). NCBI is one of the largest open source primary databases which contains a complete range of biological data in form of nucleotide sequence, protein sequence and structures, genome map scientific literature etc.

Perl script to cut the large sequences

A Perl script was written to cut the large sequence of chromosome into small parts of 10,00,000 bp. The whole sequence of chromosome was provided to the computer program as input for cutting the large sequence into small parts. The size of input file was different for different chromosomes. This script cut the large sequences into small sequences of 10,00,000 bp each.

Input of sequence in HaploSNPer

Input options are flexible. Input of sequences was done by pasting the sequence in FASTA format into the input text area.

Settings of HaploSNPer's parameters

There are eight parameters required to control the performance of HaploSNPer:

1) Selecting a tagging database: In the current study, the sorghum EST sequence database was selected. The EST sequence was extracted from EMBL database.
2) Selecting a sequence alignment program: In the current study, PHRAP was used for sequence alignment. CAP3 or PHRAP could be chosen for sequence alignment. For SNP mining, CAP3 uses individual sequence overlap for constructing clusters, while PHRAP tends to extend the consensus sequence by overlap. PHRAP is much faster than CAP3.
3) Pre-processing of sequences: HaploSNPer supplies Cross_ match and RepeatMasker to clean sequences. RepeatMasker is for masking repeat fragments and Cross_match is for removing vector sequences. Sequences containing long repeat fragments can create incorrect sequence assemblies. In the current study, only RepeatMasker was used, Cross_match was not used because the input was genomic sequence instead of EST sequence.
4) Settings of parameters for BLAST and CAP3: BLAST was used to search for sequences homologous/similar to the input seed sequence. The E-value for BLAST can be set to select sequences that are similar to the seed sequence. When E-value was set high, many similar sequences were found and this resulted in many clusters; while when the E-value was set low, enough similar sequences for haplotype and SNP detection were not found. In the current study, an E-value of 1e-60 was used and this value is the default in HaploSNPer. Similarity for CAP3 was taken to be 95%, which is stringent enough to prevent most paralogous sequences and keep all available allelic sequences in a cluster.
5) Settings for haplotype reconstruction: Potential haplotype is defined as a group of sequence within a cluster that has the same nucleotide at every polymorphic site. For haplotype reconstruction, the similarity between a candidate sequence and a haplotype group at each single SNP is calculated and compared with a threshold to determine whether the candidate sequence matches the haplotype group at that SNP; then the similarity over all SNPs is compared with a second threshold to determine whether the candidate sequence can be reliably assigned to the haplotype group. By using the similarity per polymorphic site as well as the similarity over all polymorphic sites, haplotypes can even be reconstructed reliably from sequences that contain sequencing errors. The threshold value of similarity per polymorphic site was taken as 75% and that of similarity over all polymorphic sites was taken as 80%.
6) Settings for low quality regions: In the current study, low quality region of sequence were set to a weight value of 0 because the data taken were genome sequence data and not EST sequence data.
7) Other settings for SNP detection: In HaploSNPer, sequence redundancy is also used to prevent sequencing errors. In the current study, values for the minimum cluster size, minimum allele size and minimum confidence score were 4, 2 and 2, respectively. The higher confidence score is the reliability of the SNP on sequence redundancy
8) Output of HaploSNPer: This can be displayed in two ways: one can either choose to view the cluster with the seed sequence or all related clusters. In the current study, all related clusters related to seed sequence were chosen, of which (5), (6) and (7) are used to control the performance of QualitySNP (Tang et al., 2006).

RESULTS AND DISCUSSION

By using HaploSNPer tool for SNP mining in the *S. bicolor* genome, the total number of potential SNPs found was 77094 SNPs out of which reliable SNPs were 40589. In the case of both potential and reliable SNPs, different types of SNPs including transitions, transversions and indels were detected and are listed in Tables 1 and 2.

This research work is highly relevant and useful in the current scenario, where the need of efficient and informative biological markers are required for exploring the

Table 1. SNPs detected in each chromosome of *S. bicolor*.

Chromosome	SNP	C/T	A/G	Transition	A/T	A/C	C/G	T/G	Transversion	Indel
1	Potential SNPs	2769	2440	5209	1344	1384	2130	1419	6277	1200
	Reliable SNPs	1426	1406	2832	807	968	1573	909	4257	579
2	Potential SNPs	2718	2367	5085	1194	942	1233	2271	4407	956
	Reliable SNPs	1244	1160	2404	649	566	795	637	2647	430
3	Potential SNPs	1796	1524	3320	803	823	912	736	3274	733
	Reliable SNPs	926	818	1744	482	523	654	425	2084	337
4	Potential SNPs	2018	1759	3777	834	782	939	786	3341	870
	Reliable SNPs	884	787	1671	375	427	573	392	1767	413
5	Potential SNPs	1340	1178	2518	681	573	700	629	2583	418
	Reliable SNPs	631	560	1191	381	356	472	380	1589	210
6	Potential SNPs	1876	1662	3538	859	725	839	727	3150	760
	Reliable SNPs	891	775	1666	474	457	577	413	1921	353
7	Potential SNPs	1713	1553	3266	952	729	807	807	3295	428
	Reliable SNPs	787	747	1534	479	439	517	459	1894	161
8	Potential SNPs	2188	1953	4141	1035	784	978	926	3723	608
	Reliable SNPs	893	919	1812	463	386	562	528	1939	266
9	Potential SNPs	1472	1322	2794	768	552	577	640	2537	497
	Reliable SNPs	625	596	1221	401	318	366	360	1445	162
10	Potential SNPs	1194	1029	2223	558	385	416	452	1811	355
	Reliable SNPs	521	446	967	249	198	255	255	957	136

Table 2. SNPs in whole genome of *S. bicolor*.

SNP	C/T	A/G	Transition	A/T	A/C	C/G	T/G	Transversion	Indel
Potential SNPs	19084	16787	35871	9028	7679	9531	9393	34398	6825
Reliable SNPs	8828	8214	17042	4760	4638	6344	4758	20500	3047

enormous astonishing applications offered by various organisms, especially plants. For many years, much work has been done to explore the genome of many vital plants. Many researchers have been working on the detection of efficient genetic markers, in this respect.

This research work has been carried out to detect SNPs which can be applied not only for making genetic maps but also for exploring the astonishing features (genes with special features) of a genome sequence. SNPs including insertion/deletion (indels) serve as effective genetic markers. In addition to this, the unique ability of SNPs to facilitate gene identification has increased the interest of scientists in the development of SNP markers. In fact such features of SNP have recently brought a flurry of SNP discovery and detection.

This research work which involves successful mining of genome of *S. bicolor* can be useful in applying SNPs as biological markers for exploring various vital features of *S. bicolor* for economic use. Mining of SNPs in sorghum will be a great success especially since it is a model for functional genomics of saccharine (sugarcane) and other C4 grasses. In fact, the relevance of the project is relatively more, where sorghum, an economically important crop, serves as a very good source for production of alcohol and bio-fuel. SNP mining will also be useful for producing high yield of agricultural improved varieties of sorghum. Discovered SNPs can also be used for determining disease causing genes in sorghum.

Therefore, though a trodden work (significant amount of work has been done before in this area) has been followed, it is apparent that this research work can definitely contribute to the functional genomics, agricultural sciences and crop improvement studies, especially in C4 grasses.

REFERENCES

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

Kollers S, Kerstens HHD, Kommadath A, Rosario MD, Dibbits B, Kinders SM, Crooijmans RP, Groenen MAM (2009). Mining for single nucleotide polymorphisms in pig genome sequence data BMC Genomics, 10: p. 4.

Paterson AH (2009). The Sorghum bicolor genome and the diversification of grasses. Nature, 457: 551-556.

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: 167-217

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JAM (2006). QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species BMC Bioinformatics, 7: p. 438 doi: 10.1186/1471-2105-7-438

Xiang W, Can Y, Qiang Y, Hong X, Nelson L, Weichuan Y (2009). MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study BMC Bioinformatics, 10: p. 13. doi: 10.1186/1471-2105-10-13.