JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 7, 2016
© Mary Ann Liebert, Inc.
Pp. 597–602
DOI: 10.1089/cmb.2015.0215

CsSNP: A Web-Based Tool for the Detecting of Comparative Segments SNPs

YI WANG,[1] SHUANGSHUANG WANG,[2] DONGJIE ZHOU,[2] SHUAI YANG,[2]
YONGCHAO XU,[2] CHAO YANG,[1] and LONG YANG[2,3]

ABSTRACT

SNP (single nucleotide polymorphism) is a popular tool for the study of genetic diversity, evolution, and other areas. Therefore, it is necessary to develop a convenient, utility, robust, rapid, and open source detecting-SNP tool for all researchers. Since the detection of SNPs needs special software and series steps including alignment, detection, analysis and present, the study of SNPs is limited for nonprofessional users. CsSNP (Comparative segments SNP, http://biodb.sdau.edu.cn/cssnp/) is a freely available web tool based on the Blat, Blast, and Perl programs to detect comparative segments SNPs and to show the detail information of SNPs. The results are filtered and presented in the statistics figure and a Gbrowse map. This platform contains the reference genomic sequences and coding sequences of 60 plant species, and also provides new opportunities for the users to detect SNPs easily. CsSNP is provided a convenient tool for nonprofessional users to find comparative segments SNPs in their own sequences, and give the users the information and the analysis of SNPs, and display these data in a dynamic map. It provides a new method to detect SNPs and may accelerate related studies.

Key words: comparative genome, single nucleotide polymorphism, web-based software.

1. INTRODUCTION

The development of next generation sequencing technologies significantly accelerated the collection of genome sequences by providing a series of new methods to explore the mysteries of life, such as SNPs (single nucleotide polymorphisms). SNPs are an excellent tool for the study of genetic diversity, evolution relationship, and gene functions because they are high density and convenient (Hirakawa et al., 2013; Qi et al., 2013; Rasic et al., 2014). In the past few years, popular studies based on SNPs were construction of high density maps and the defining of evolution relationships (Chung et al., 2014; Cronin et al., 2014; Kaur et al., 2014). Comparative segments SNPs is a new popular tool for the study of evolution and gene function.

Comparative segments SNPs are those that exist in the synteny segments of different varieties or even different species, such as the function gene and the conserved segments. A recent study shows that these

[1]Key Laboratory of Crop Biology of China, [2]College of Plant Protection, and [3]Agricultural Big-Data Research Center, Shandong Agricultural University, Taian, China.

SNPs are closely related with human disease and gene functions (Marzolini et al., 2004). Some studies have even analyzed the evolution relationship of the comparative SNPs between different species (Chimpanzee Sequencing and Analysis Consortium, 2005), and more related studies are in progress.

To make SNPs more useful, a series of software and databases have been developed in recent years. GATK (McKenna et al., 2010) is a classic one. BWA (Li and Durbin, 2009) combined with SAM tools (Li et al., 2009) is one of the more widely used SNP detectors, while Soapsnp (Li et al., 2009) is the fastest detector. All these tools are based on the Linux (UNIX) operating system and mainly analyze resequencing data. At the same time, some web-based platforms also provide service about SNPs, such as FunctSNP (Goodswen et al., 2010), snp-search (Al-Shahib and Underwood, 2013), and SNPsea (Slowikowski et al., 2014).

All these software are of great help to researchers in the study of SNPs. But they are not enough. SoapSNP and most of other SNP detectors only can detect SNPs between the reference sequence and the resequencing data, which must be in the format of FASTAQ, and most of them also need professional knowledge of Linux and professional computer language. These are not easily used for nonprofessional researchers, and also limit the number of users. In addition, a tool to detect the comparative segments SNPs is lacking.

CsSNP (http://biodb.sdau.edu.cn/cssnp/) aims to provide a convenient tool for nonprofessional users to find comparative segments SNPs in their own sequences, and give the users the information and the analysis of SNPs and display these data in a dynamic map. These SNPs may related to the evolution, disease, or other phenotype, so it can make the study easier and provide new insight to the genomic function studies. CsSNP also provides a new method to detect SNPs and may accelerate the related studies.

2. METHODS

CsSNP contains genome sequences and coding sequences of 60 plant species (Supplementary Table S1; supplementary material is available online at www.liebertpub.com/cmb), including almost all the sequenced species. These sequences are treated with four steps, including alignment, detection, filter and statistics, and the results present (Fig. 1). All these steps are based on the Blat, Blast and a series of Perl scripts.

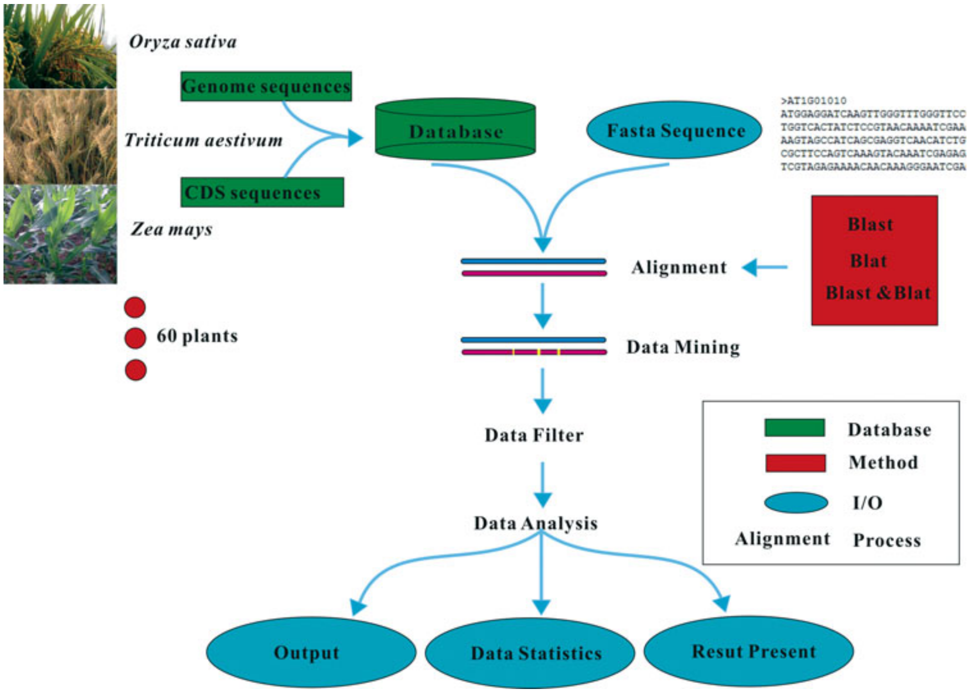
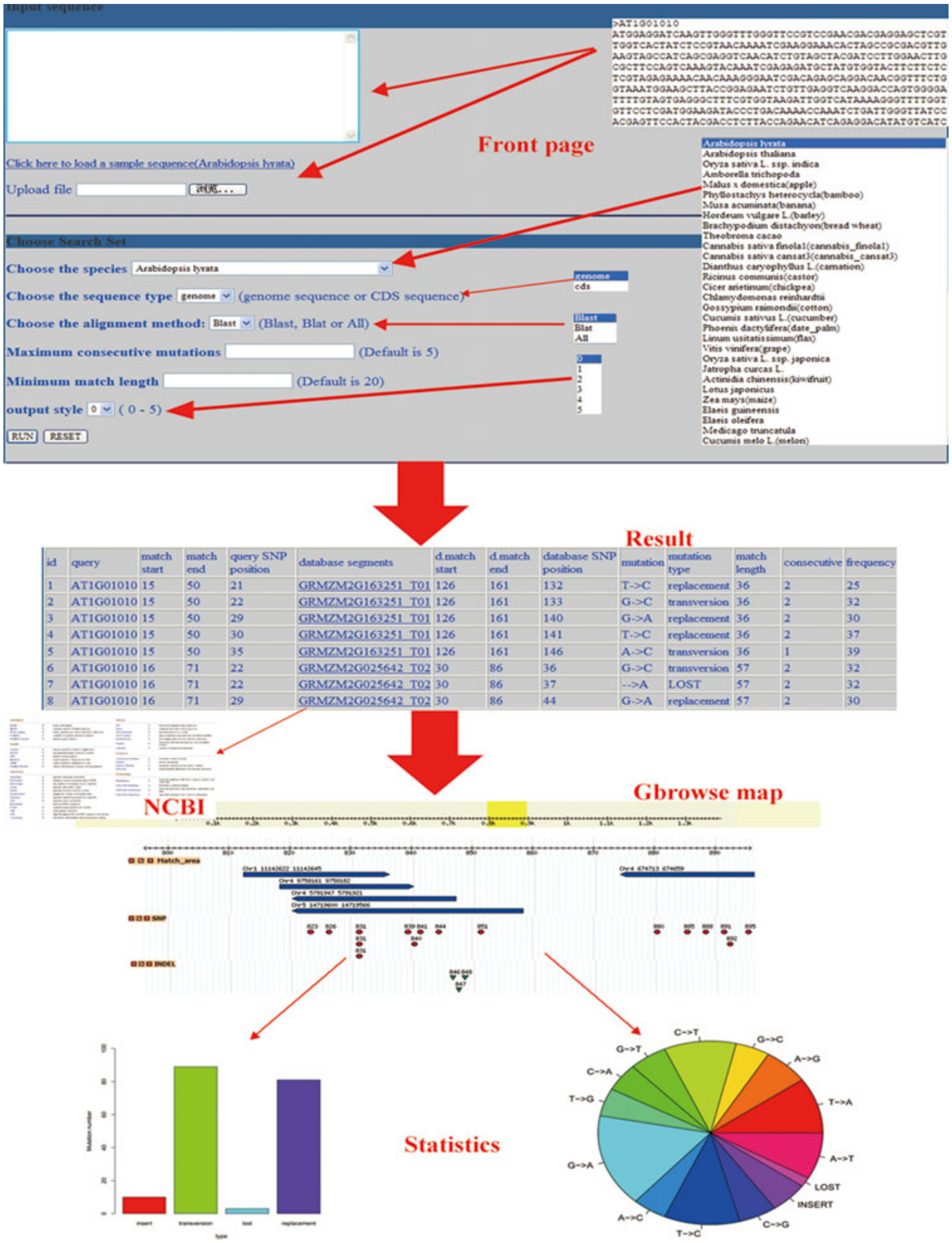


FIG. 1. The flowchart of CsSNP. The sequences from the users will be treated with four steps including alignment, detection, filter and statistics, and the results present.



Input sequence

```
>AT1G01010
ATGGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAACGACGAGGAGCTCGT
TGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTG
AAGTAGCCATCAGCGAGGTCAACATCTGTAGCTACGATCCTTGGAACTTG
CGCTTCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTC
TCGTAGAGAAAACAACAAAGGGAATCGACAGAGCAGGACAACGGTTTCTG
GTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTCAAGGACCAGTGGGGA
TTTTGTAGTGAGGGCTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGT
GTTCCTCGATGGAAGATACCCTGACAAAACCAAATCTGATTGGGTTATCC
ACGAGTTCCACTACGACCTCTTACCAGAACATCAGAGGACATATGTCATC
```

Click here to load a sample sequence(Arabidopsis lyrata)

Upload file

Choose Search Set

Choose the species: Arabidopsis lyrata

Choose the sequence type: genome (genome sequence or CDS sequence)

Choose the alignment method: Blast (Blast, Blat or All)

Maximum consecutive mutations (Default is 5)

Minimum match length (Default is 20)

output style 0 (0 - 5)

id	query	match start	match end	query SNP position	database segments	d.match start	d.match end	database SNP position	mutation	mutation type	match length	consecutive	frequency
1	AT1G01010	15	50	21	GRMZM2G163251_T01	126	161	132	T->C	replacement	36	2	25
2	AT1G01010	15	50	22	GRMZM2G163251_T01	126	161	133	G->C	transversion	36	2	32
3	AT1G01010	15	50	29	GRMZM2G163251_T01	126	161	140	G->A	replacement	36	2	30
4	AT1G01010	15	50	30	GRMZM2G163251_T01	126	161	141	T->C	replacement	36	2	37
5	AT1G01010	15	50	35	GRMZM2G163251_T01	126	161	146	A->C	transversion	36	1	39
6	AT1G01010	16	71	22	GRMZM2G025642_T02	30	86	36	G->C	transversion	57	2	32
7	AT1G01010	16	71	22	GRMZM2G025642_T02	30	86	37	-->A	LOST	57	2	32
8	AT1G01010	16	71	29	GRMZM2G025642_T02	30	86	44	G->A	replacement	57	2	30

FIG. 2. The dynamic present of CsSNP.

2.1. Sequence alignment

CsSNP mainly deals with the ordinary sequence in the FASTA format and the sequence should just be a scaffold, such as gene, CDS sequence, or even an unknown segment. The sequences from the users will be first aligned with the reference sequence by Blat (Doerks et al., 2002), Blast, or both. The alignment will find the high similar sequences that are more likely to be the homologous sequences in the query and database. The Blat alignment will find fewer SNPs with a higher similarity and longer match sequence. The Blast alignment can find more SNPs, but sometimes it is not as effective as Blat.

2.2. Variation detecting

Using a Perl pipeline, the alignment sequence that is shorter than the user parameter or the default will be ignored. The difference between the high quality alignments will be detected and stored according to the

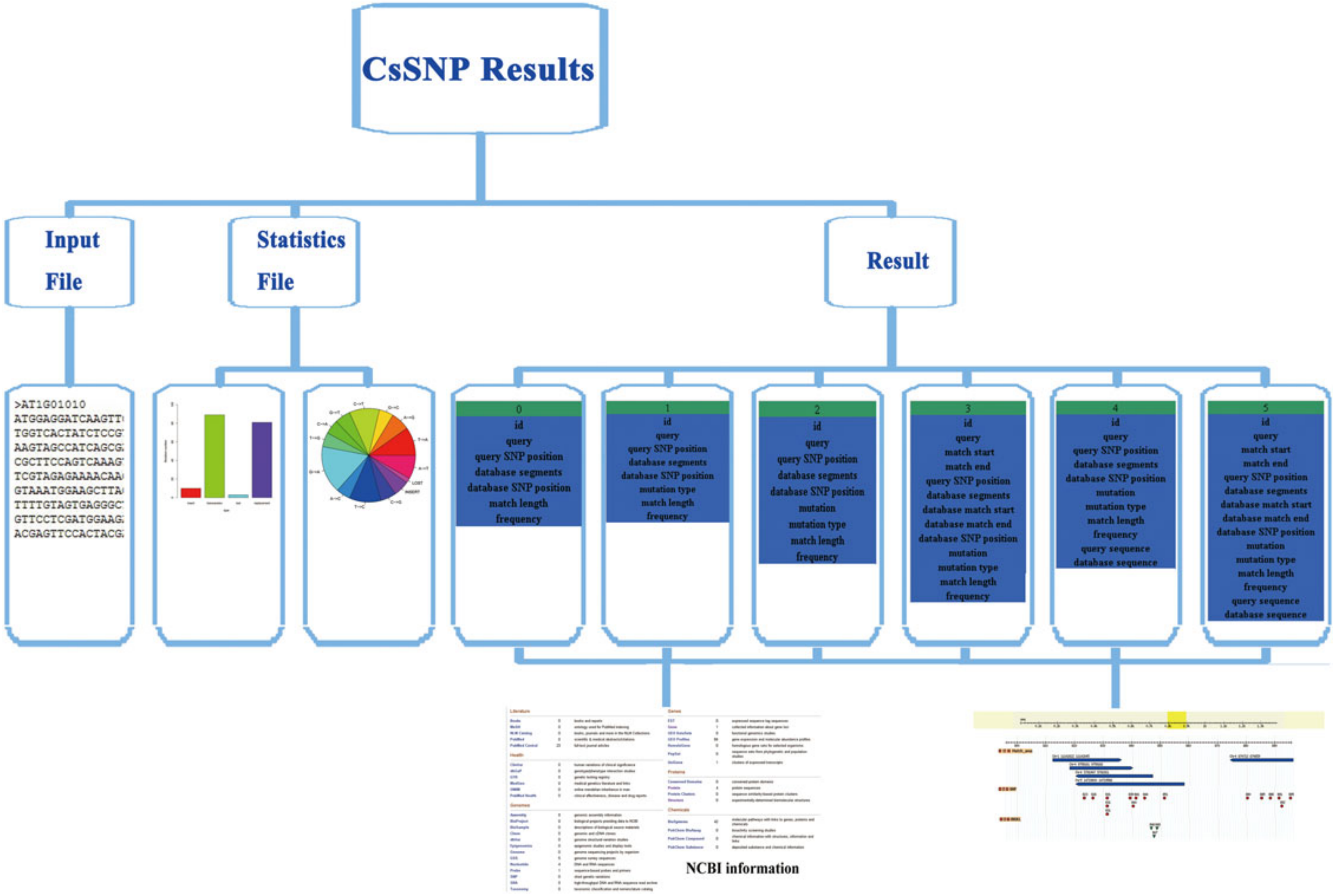


FIG. 3. The output formats in different types of detailed information of CsSNP.

Table 1. Blast and Blat Method Detection Between CDS Sequence of Melon and Watermelon

	Alignment number	*Alignment length*	*Variation times*	*Variation number*	*Frequency*
Blat	139,459	34,546,267	3,096,045	2,177,808	1.42
Blast	1,129,156	1,610,755,772	15,941,130	6,067,579	2.63

parameter of this program. In this tool, the position of the insert position is the left base of this insert on the normal sequence. All differences obey the parameter that is selected to do the further filter and analysis.

2.3. Data filter and analysis

After getting the variation data, the results are filtered by several rules. The first rule is that the continuous SNPs must not be longer than the parameter. The second rule is that the ambiguous base will be ignored, such as the ‘‘N’’ and other words. The third rule is that the insert and lost will be treated as short indels, and SNPs only contain replacements and transversions. The variation will be classified into four types: replacement, transversion, insert, and lost. The number of each variation is counted and presented. At the same time, the frequency of each variation and the continuous variations are also annotated to the detail information.

2.4. Dynamic present

All the various information is presented as a Gbrowse map (Fig. 2), which is composed of three parts. The first part contains blue blocks that indicate the alignment area on the reference sequence. The second part contains red circles that indicate the positions of SNPs. The third part shows some green triangles, which are the symbol of the insert and lost. If there are several reference sequences involved, the sequence will be assembled into one sequence with a 100 bp gap between every two.

There are also two figures in the format of PDF that can be downloaded from the result page which have been drawn by the R based on the statistics. The first figure is a pie chart of 14 types of variation, including the insert and lost. The second one is a bar chart showing the number of replacement, transversion, insert, and lost.

2.5. Output style

The output package contains three files of the sequence analyzed, SNPs information, and the output report. For the file of SNPs information, six output formats will be provided to show different types detail information of the SNPs (Fig. 3); they provide different types of data to the users. Of all these six output styles, 5 has the most detailed description, and 3 may be more fit for programmed handle. This report file contains the analysis of the input sequences and the statistics of SNPs. The report will remain for 20 minutes for the user to download. There are two usage introductions in Chinese and English, both of which can be downloaded from the web. This platform also provides four Perl scripts for the users to pretreat and analyze the sequence.

2.6. Verification in silico

The two methods have been used to detect the variations between the CDS sequences of melon and watermelon (Table 1). The results show that Blast gets more alignments and more variations, the alignment number of Blast is nearly 8 times more than the Blat, and the alignment length is about 46 times, but the variation times are less than 3 times. This because that Blast can put up with more mismatches and gaps in the alignment, so the aligned sequence is more and longer. But the higher tolerance means that most variation is out of the selected conditions, and the detected variations are smaller. Therefore, Blast can detect more variations but the alignment quality is not as high as the Blat.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No.31201231), Foundation of Big Data of Shandong Agricultural University. The authors thank Dr. Jiajie Wu from the College of Agronomic Science, Shandong Agricultural University, for helpful suggestions on the manuscript.

AUTHOR DISCLOSURE STATEMENT

The authors declare that there are no conflicting financial interests.

REFERENCES

Al-Shahib, A., and Underwood, A. 2013. Snp-search: Simple processing, manipulation and searching of SNPs from high-throughput sequencing. *BMC Bioinformat* 14, 326.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.

Chung, W.H., Jeong, N., Kim, J., et al. 2014. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res* 21,153–167.

Cronin, M.A., Rincon, G., Meredith, R.W., et al. 2014. Molecular phylogeny and SNP variation of polar bears (Ursus maritimus), brown bears (U. arctos), and black bears (U. americanus) derived from genome sequences. *J Hered* 105, 312–323.

Doerks, T., Copley, R.R., Schultz, J., et al. 2002. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res* 12, 47–56.

Goodswen, S.J., Gondro, C., Watson-Haigh, N.S., et al. 2010. FunctSNP: An R package to link SNPs to functional knowledge and dbAutoMaker: A suite of Perl scripts to build SNP databases. *BMC Bioinformat* 11, 311.

Hirakawa, H., Shirasawa, K., Ohyama, A., et al. 2013. Genome-wide SNP genotyping to infer the effects on gene functions in tomato. *DNA Res* 20, 221–233.

Kaur, S., Kimber, R.B., Cogan, N.O., et al. 2014. SNP discovery and high-density genetic mapping in faba bean (Vicia faba L.) permits identification of QTLs for ascochyta blight resistance. *Plant Sci* 217–218, 47–55.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., et al. 2009. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, R., Li, Y., Fang, X. et al. 2009. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19, 1124–1132.

Marzolini, C., Paus, E., Buclin,T., et al. 2004. Polymorphisms in human MDR1 (P-glycoprotein): Recent advances and clinical relevance. *Clin Pharmacol Ther* 75, 13–33.

McKenna, A., Hanna, M., Banks, E., et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20,1297–1303.

Qi, J., Liu, X., Shen, D., et al. 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat Genet* 45, 1510–1515.

Rasic, G. Filipovic, I. Weeks, A.R., et al. 2014. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, Aedes aegypti. *BMC Genomics* 15, 275.

Slowikowski, K., Hu, X., and Raychaudhuri, S. 2014. SNPsea: An algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* 30, 2496–2497.

Address correspondence to:
Dr. Long Yang
College of Plant Protection
Shandong Agricultural University
Taian 271018
China

E-mail: lyang@sdau.edu.cn