BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software
Volume 9(10)

GMATo: A novel tool for the identification and analysis of microsatellites in large genomes

Xuewen Wang*, Peng Lu & Zhaopeng Luo

China Tobacco Gene Research Center, Zhengzhou Tobacco Research Institute, NO.2 Fengyang Street, Hi-tech zone, Zhengzhou 450001, China; Xuewen Wang – Email: xwwang@ymail.com; *Corresponding author

Received January 01, 2013; Accepted January 13, 2013; Published June 08, 2013

Abstract:
Simple Sequence Repeats (SSR), also called microsatellite, is very useful for genetic marker development and genome application. The increasing whole sequences of more and more large genomes provide sources for SSR mining *in silico*. However currently existing SSR mining tools can't process large genomes efficiently and generate no or poor statistics. Genome-wide Microsatellite Analyzing Tool (GMATo) is a novel tool for SSR mining and statistics at genome aspects. It is faster and more accurate than existed tools SSR Locator and MISA. If a DNA sequence was too long, it was chunked to short segments at several Mb followed by motifs generation and searching using Perl powerful pattern match function. Matched loci data from each chunk were then merged to produce final SSR loci information. Only one input file is required which contains raw fasta DNA sequences and output files in tabular format list all SSR loci information and statistical distribution at four classifications. GMATo was programmed in Java and Perl with both graphic and command line interface, either executable alone in platform independent manner with full parameters control. Software GMATo is a powerful tool for complete SSR characterization in genomes at any size.

Availability: The soft GMATo is freely available at http://sourceforge.net/projects/gmato/files/?source=navbar or on contact.

Keywords: Genome, Microsatellite, SSR, Marker development, Software.

Background:

Simple Sequence Repeats (SSR) or microsatellite is a relative short tandem repeats of DNA **[1, 2]**. Its length polymorphism is specie specific and inheritable, which makes SSR very useful for developing genetic SSR marker widely used in linking genome sequence with traits, diversity investigation, map-base cloning and molecular breeding **[2]**. There are some useful software and tools developed for SSRs discovery and marker development in silico. However, they were designed before large genome era and have two major limitations: i) too low sequence processing capability and slow speed as pointed by Sharma **[2]** *et al*. to deal with large genomes while more large genomes i.e. those from crops become important sources for SSR characterization with the benefit from the advanced next generation sequencing technology, ii) no or simple statistical function provided such as TROLL **[3]**. In addition, some tools have platform dependence i.e. SSR Locator **[4]** and SciRoko **[5]**. Most command tools have no graphic interface and very limited other functions, i.e. tool MISA **[6, 7]**. In order to overcome those limitations mentioned above, novel software named GMATo was developed for faster and accurate SSR discovery and comprehensively statistical analyzing especially for large genomes running at multiple platforms with both graphic and command interface.

Methodology:

The soft GMATo was written in Perl and Java language. Java was used for developing graphic interface. Perl was used to discover the microsatellite and perform statistical analyzing. In GMATo DNA sequences are formatted first and the long DNA sequence is chunked to small segments at several Mb for easy processing. All microsatellite motifs consisting of A, T, G and C nucleotide of DNA at user controlled length are generated using Perl meta-characters and regular expression pattern. All motifs are searched greedily through each DNA chunk using Perl powerful pattern matching function. The returned values are used to generate SSR loci information at each chunk and the final SSR loci data at a chromosome after merging data from chunks. This method allows microsatellite discover efficiently in any genome with any size theoretically. Statistical classification and summarization were performed at four levels i.e. motif length, motif composition, grouped complementary motifs and chromosome/scaffold. A flowchart was shown in **(Figure 1).**

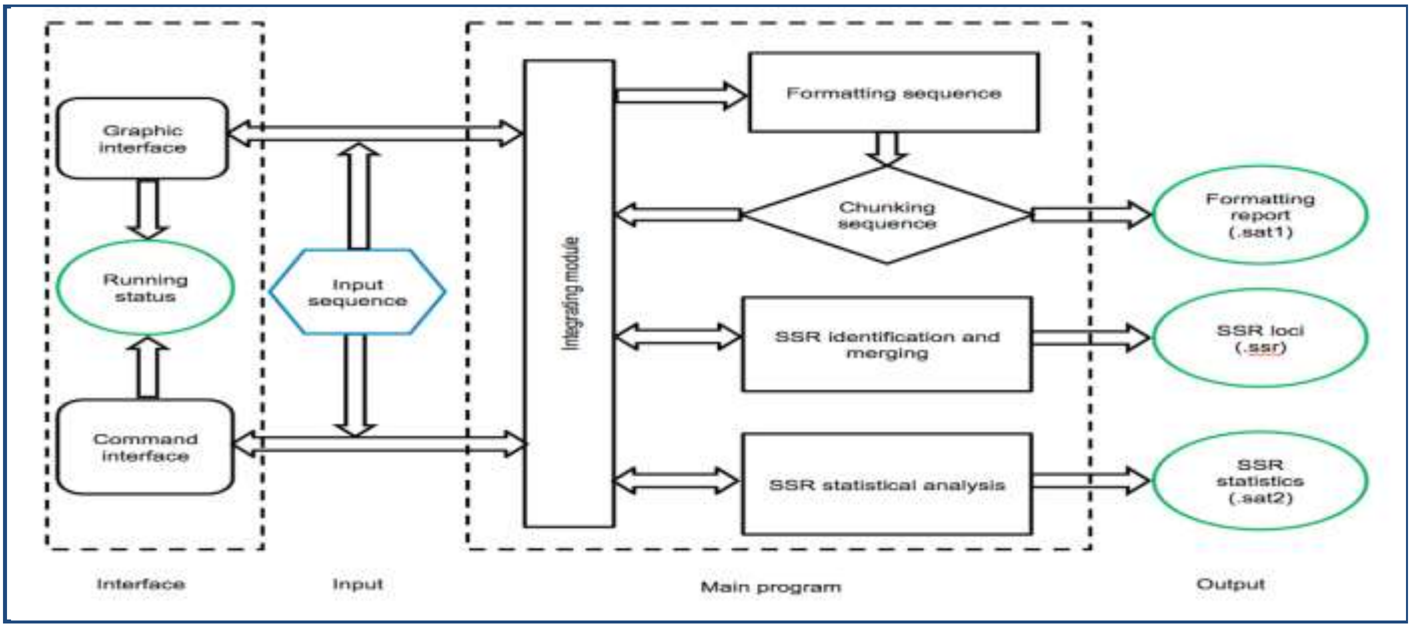
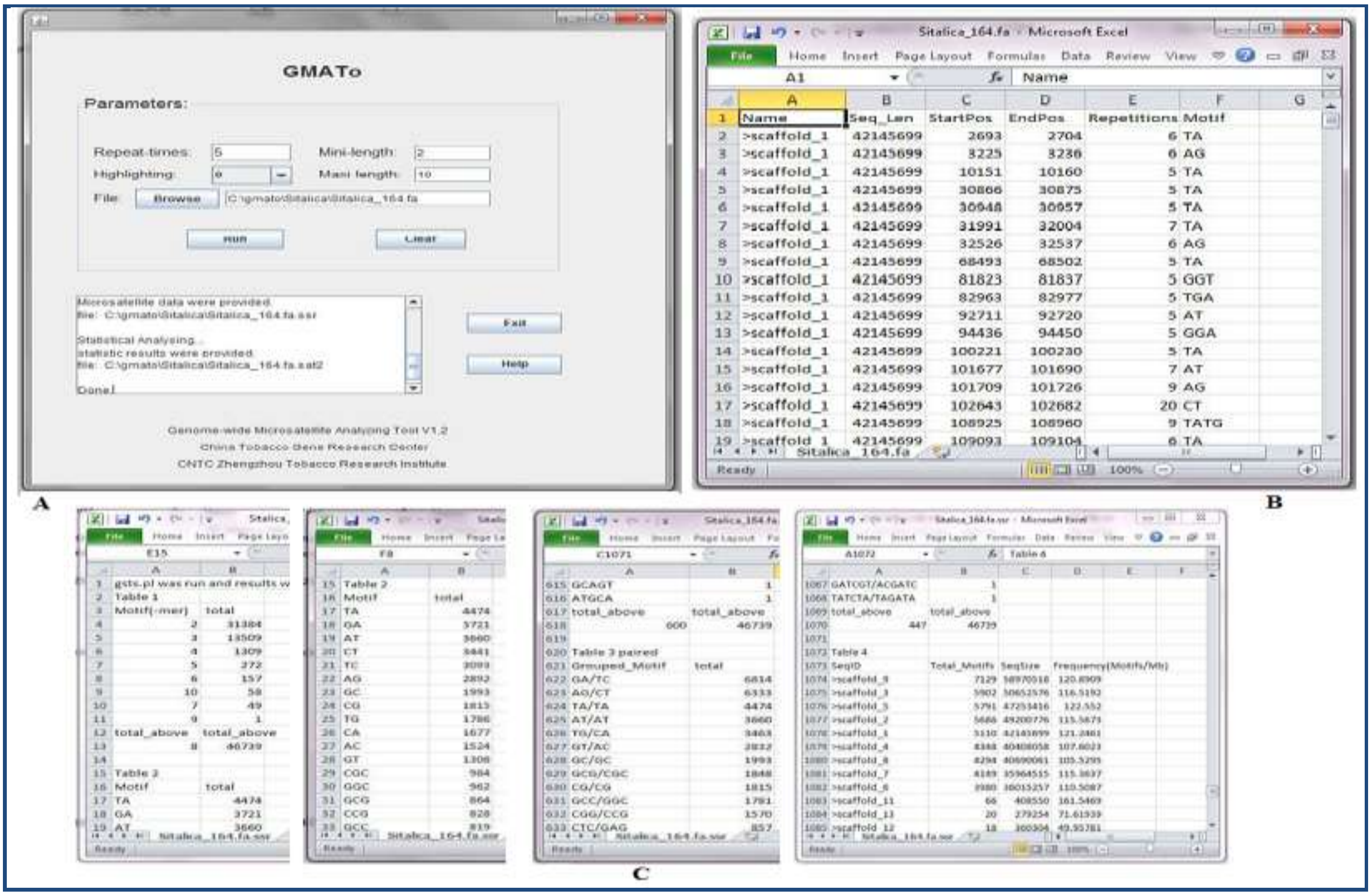


Figure 1: A flowchart of software GMATo.



GMATo

Parameters:

Repeat-times: 5
Mini-length: 2
Highlighting: 6
Maxi length: 10
File: Browse C:\gmato\Sitalica\Sitalica_164.fa

```
Microsatellite data were provided.
file:  C:\gmato\Sitalica\Sitalica_164.fa.ssr

Statistical Analysing...
statistic results were provided.
file:  C:\gmato\Sitalica\Sitalica_164.fa.sat2

Done!
```

Genome-wide Microsatellite Analyzing Tool V1.2
China Tobacco Gene Research Center
CNTC Zhengzhou Tobacco Research Institute

Sitalica_164.fa - Microsoft Excel

	A	B	C	D	E	F
1	Name	Seq_Len	StartPos	EndPos	Repetitions	Motif
2	>scaffold_1	42145699	2693	2704	6	TA
3	>scaffold_1	42145699	3225	3236	6	AG
4	>scaffold_1	42145699	10151	10160	5	TA
5	>scaffold_1	42145699	30866	30875	5	TA
6	>scaffold_1	42145699	30648	30657	5	TA
7	>scaffold_1	42145699	31991	32004	7	TA
8	>scaffold_1	42145699	32526	32537	6	AG
9	>scaffold_1	42145699	68493	68502	5	TA
10	>scaffold_1	42145699	81823	81837	5	GGT
11	>scaffold_1	42145699	82963	82977	5	TGA
12	>scaffold_1	42145699	92711	92720	5	AT
13	>scaffold_1	42145699	94436	94450	5	GGA
14	>scaffold_1	42145699	100221	100230	5	TA
15	>scaffold_1	42145699	101677	101690	7	AT
16	>scaffold_1	42145699	101709	101726	9	AG
17	>scaffold_1	42145699	102643	102682	20	CT
18	>scaffold_1	42145699	108925	108960	9	TATG
19	>scaffold_1	42145699	109093	109104	6	TA

E15

	A	B
1	gsts.pl was run and results w	
2	Table 1	
3	Motif(-mer)	total
4	2	31384
5	3	13509
6	4	1309
7	5	272
8	6	157
9	10	58
10	7	49
11	9	1
12	total_above	total_above
13	8	46739
14		
15	Table 2	
16	Motif	total
17	TA	4474
18	GA	3721
19	AT	3660

F8

	A	B
15	Table 2	
16	Motif	total
17	TA	4474
18	GA	3721
19	AT	3660
20	CT	3441
21	TC	3093
22	AG	2892
23	GC	1993
24	CG	1815
25	TG	1786
26	CA	1677
27	AC	1524
28	GT	1308
29	CGC	984
30	GGC	962
31	GCG	864
32	CCG	828
33	GCC	819

C1071

	A	B
615	GCAGT	1
616	ATGCA	1
617	total_above	total_above
618	600	46739
619		
620	Table 3 paired	
621	Grouped_Motif	total
622	GA/TC	6814
623	AG/CT	6333
624	TA/TA	4474
625	AT/AT	3660
626	TG/CA	3463
627	GT/AC	2832
628	GC/GC	1993
629	GCG/CGC	1848
630	CG/CG	1815
631	GCC/GGC	1781
632	CGG/CCG	1570
633	CTC/GAG	857

Sitalica_164.fa.ssr - Microsoft Excel

A1072 Table 4

	A	B	C	D	E	F
1067	GATCGT/ACGATC	1				
1068	TATCTA/TAGATA	1				
1069	total_above	total_above				
1070	447	46739				
1071						
1072	Table 4					
1073	SeqID	Total_Motifs	SeqSize	Frequency(Motifs/Mb)		
1074	>scaffold_9	7129	58970518	120.8909		
1075	>scaffold_3	5902	50652576	116.5192		
1076	>scaffold_5	5791	47253416	122.552		
1077	>scaffold_2	5686	49200776	115.5671		
1078	>scaffold_1	5110	42145699	121.2461		
1079	>scaffold_4	4348	40408058	107.6023		
1080	>scaffold_8	4294	40690061	105.5295		
1081	>scaffold_7	4189	35964515	115.3637		
1082	>scaffold_6	3980	36015257	110.5087		
1083	>scaffold_11	66	408550	161.5469		
1084	>scaffold_13	20	279254	71.61939		
1085	>scaffold_12	18	360304	49.95781		

Sitalica_164.fa.ssr

Figure 2: Image showing input and output of soft GMATo; **(A)** graphic input interface; **(B)** SSR loci information produced by GMATo; **(C)** SSR statistical data produced by GMATo.

ISSN 0973-2063 (online) 0973-8894 (print)

Validation:

The performance of microsatellite identification in recent published Setaria Italica entire genome**[8]** showed GMATo ran faster than either of most widely used tools, SSR Locator and MISA, in all three platforms **Table 1 (see supplementary material).** It was also easily to mine SSR in the genome in a normal computer because processing one chunked segment at a time in GMATo required less computing memory. A total number of 46,739, 46,625 and 46,782 microsatellite loci were identified by GMATo, SSR Locator and MISA respectively **(Table 1)**, suggesting more accurate SSR mining than SSR Locator. Manually comparison of these loci revealed that the extra loci from MISA are mined redundantly in the overlapped microsatellites.

Software input:

Both graphic user and command line interface were provided in GMATo, either executable independently in Windows, Linux or Mac OS system. Only one input file containing DNA sequence(s) in (raw) (multi-) FASTA format is required to be chosen in graphic mode or typed in command mode if taken the default parameters. The parameters are the motif length range, the minimum repeated times and an option for highlighting microsatellite **(Figure 2 A).** The motif length can be set to any range instead of 1-10 bp given in most SSR mining tools.

Software output:

The output files generated by GMATo are one formatting report, one file containing SSR loci information and another file containing statistical distribution of SSR. All output files are in a tab delimited plain text format for easily importing to other applications i.e. spread sheet for viewing or other manipulation **(Figure 2 B, C).**The formatting report summarizes the input sequence(s). The SSR loci file lists the input sequence ID and its length, starting and ending position of a microsatellite, the repeated times and the motif sequence.

The statistical distribution file provides statistical data at four different classifications at genome aspect. A summary of total is generated in the end of each classification. Classification I is the motif length statistics, providing overview information for the type, abundance in rank order. Classification II is the motif statistics based on sequence composition, i.e. motif composition, occurrence in ranked order. Classification III is the statistics of grouped complementary motifs, providing distribution data for complementary motifs such as TC/GA in a group and their occurrence in ranked order. Classification IV is the statistics of chromosome level distribution, providing the total occurrence of motif(s) and SSR frequency (loci/Mb) at each chromosome or super-scaffold.

Utility:

GMATo can be used for efficient and faster microsatellite sequence identification from any given DNA sequences or genomes at any size. Detailed statistic distribution of microsatellites can be used for genome analysis.

Caveat and future development:

Current version provides each perfect SSR loci information. The compound and long imperfect microsatellites can be calculated from the SSR loci output using additional script. For a future development, more functions including displaying statistical data graphically, primer designing, marker generation and electronic mapping markers into a genome will be added. The final goal is to develop an integrated powerful toolkit facilitating microsatellite characterization and marker development in large genomes.

References:

[1] Ellegren H, *Nat Rev Genet.* 2004 **5:** 435 [PMID: 15153996]

[2] Sharma P, *Trends Biotechnology.* 2007 **25**: 490 [PMID: 17945369]

[3] Castelo AT *et al. Bioinformatics*. 2002 **18**: 634 [PMID: 12016062]

[4] da Maia LC *et al. Int J Plant Genomics*. 2008 **412:** 696 [PMID: 18670612]

[5] Kofler R *et al. Bioinformatics*. 2007 **23**: 1683 [PMID: 17463017]

[6] Thiel TT *et al. Theor Appl Genet.* 2003 **106**: 411 [PMID: 12589540]

[7] Sonah H *et al. PLoS ONE.* 2011 **6**: e21298 [PMID: 21713003]

[8] Bennetzen JL *et al. Nat Biotechnol*. 2012 **30**: 555 [PMID: 22580951]

Edited by P Kangueane

Citation: **Wang *et al.*** Bioinformation 9(10): 541-544 (2013)

Supplementary material:

Table 1: Performance comparison of microsatellite mining software GMATo, SSR Locator and MISA in *Setaria Italica* whole genome

***Setaria Italica* whole genome**				
Soft		**Time**		**SSR Loci**
	Windows*	Linux#	Mac&	
GMATo	10m0s	8m40s	8m6s	46,739
SSR Locator	>12h+	Not executable	Not executable	46,625
MISA	16m11s	12m14s	15m13s	46,782

SSR motif length ranging from 2 to 10 bp, minimum repeated times at least 5. SSR locator V1.1 and MISA was downloaded from official site http://www.ufpel.tche.br/ and http://pgrc.ipk-gatersleben.de/misa/ respectively.

Foxtail millet (*Setaria Italica*) whole genome sequence (~515Mb) file Sitalica_164.fa was downloaded from phytozome http://www.phytozome.net/.

*environment: HP 8000 Elite 32 byte Windows 7， Inter core2 CPU 2.83 GHz, 4G RAM， disk space 500 G； # environment: Linux gridview 2.6.18, 64 byte, AMD Opteron Processor 612 CPU 2.0 GHz, 66G RAM， disk space 500G; & environment: Mac Pro OS 10.7.5, Intel Xeon CPU 2.66 GHZ, 12G RAM， disk space 1T; + Summary of SSR loci was produced at 29m49s but it took more than 12 hours to export mined SSR data.