

Manual for software GMATo

Genome-wide Microsatellite Analyzing Tool (GMATo)

Xuewen Wang Ph.D

Program scripts written by Xuewen Wang (PhD) and Peng Lu

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

2013-1-7th

Content

1. Introduction	1
2. Installation	2
2.1. Hardware environment.....	2
2.2. Software environment	2
3. GMATo Interface.....	3
4. Running	4
4.1. Graphic mode.....	4
4.2. Command line mode.....	8
Linux /Mac.....	8
Windows.....	9
5. Input file.....	12
6. Output files	14
6.1. Formatting report	14
6.2. SSR location file	14
6.3. SSR distribution statistical file	15
7. Support and Contact	19

1. Introduction

This file documents soft Genome-wide Microsatellite Analyzing Tool (GMATo) graphic and command line application. The soft GMATo is freely available at <http://sourceforge.net/projects/gmato/files/?source=navbar> or on contact.

Genome-wide Microsatellite Analyzing Tool (GMATo) is a novel soft for faster and accurate microsatellite mining at any length and comprehensive statistical analysis at DNA sequences in any genome at any size, with easily customized parameters control for biologists and bio-informatician, running easily at normal computers with Windows, Linux, MAC OS etc multiple platforms (platform independently) with both graphic and command interface programmed in Java and Perl computing language.

Software GMATo is a one-step tool for Simple Sequence Repeats (SSR), also called microsatellite, characterization in any genome and for facilitating SSR marker designing in genome scale. The program was easily analyzed all SSRs in 2G Zea mays genome.

Only one input file is required which contains DNA sequences in raw fasta format and output files in tabular format list all SSR loci information and statistical distribution at four biologist interested classifications.

The program consists of three major functional modules: i) format sequence module, ii) SSR search and processing module and iii) statistic module. Each module can run independently or sub-sequentially. Module GMAT in Java or Perl integrates all above modules.

2. Installation

The program is distributed in source code and/or executable code. Latest version is GMATo V1.2.

2.1. Hardware environment

The computer should be capable to run Perl 5.14 or above and Java 7. For GMATo installation, minimum disk space is 4 M. For running GMATo, minimum CPU 1G or higher, minimum 512M RAM, available disk space is at least 2.5X of the input sequence file size.

2.2. Software environment

Before running GMATo, You want to make sure your computer has Perl and Java installed. If not, Install Perl version 5.14 or above and Java 7 or above in your computer first according to their instructions. The Perl is available at <http://www.activestate.com/> and Java is available at <http://www.java.com/>.

For installing, just download latest GMATo zip package from <http://sourceforge.net/projects/gmato/files/?source=navbar> and unzip it to a directory of a disk or any other movable device. Then run GMATo in command lines or graphic interface in you systems. The GMATo was tested and worked perfectly at computing system Windows xp, Window 7, MS Windows DOS 6.1, Mac OS X 10.7, Unix/Linux 5.5. It may run very well in the other system. If you have test GMATo in other systems, you are welcome to share your opinion with us by email or discuss in the sourceforge forum at <http://sourceforge.net/projects/gmato/files/?source=navbar>.

For executable code, it may be available for Windows only at the moment. Just download and install it.

3. GMATo Interface

GMATo offers graphic user interface mode and command line interface mode, either executable alone tested in Windows, Linux or Mac OS system.

Only one input file name containing DNA sequence(s) was needed to be chosen in graphic mode or typed in command mode if taken the default parameters. The input file should be a plain text file containing DNA sequences in (raw) fasta format.

The parameters allow user to set the motif minimum length (-m) and maximum length (-x), the minimum motif repeated times (-r), an option for highlight microsatellite sequence in original sequence. The motif length can be set to any length wanted instead of the range of 1-10 given in most SSR mining tools.

Parameters :

Before explaining the parameters, let's give a description of several concepts used in this soft. Repeat sequence here refers to whole length of repeated sequence or SSR or microsatellite, ie (CT)₂₈. The minimum unit of a microsatellite is called motif, ie CT. Correspondently, the length of motif CT is 2 here.

-r : minimum repeated times (Repeat-times) of motif, integer value, ≥ 1 , default value 5

-m : minimum length (Mini-length) of motif, integer value, ≥ 1 , default value 2

-x : maximum length (Maxi-length) of motif, default value 10,

value of x greater or equal to value of m,

if value x= value m, mining motif at only this given length, meaning a certain motif with one length

-s : highlighting repeat sequence or not, integer value 0 or 1.

1 meaning output the sequence and the highlighting the SSRs in upcase letters while other part in lowercase.

0 meaning no sequence output in SSR loci file

-i : file (File) name of input source sequences, plain text file, (raw) fasta format

4. Running

There are two running modes: graphic or command. Either produces the same result.

4. 1. Graphic mode

The mode can be run in any system platform if the Java run time environment is installed. These steps are:

step 1. Double click the “gmat.jar” file located in the program directory to start.

Or use the command to start Java. If java graphic can't start via click in some computer due to Java path setting issue when Java was installed, please try the following command to start graphic running mode. Go to the directory where GMATo was installed and then type to following and then hit return key to run GMATo.

```
java -jar gmat.jar
```

for Chinese version, type:

```
java -jar gmat._Cn.jar
```

A graphic window will appear (Figure 1).

A Chinese language version is also available by click “gmat_Cn.jar”.

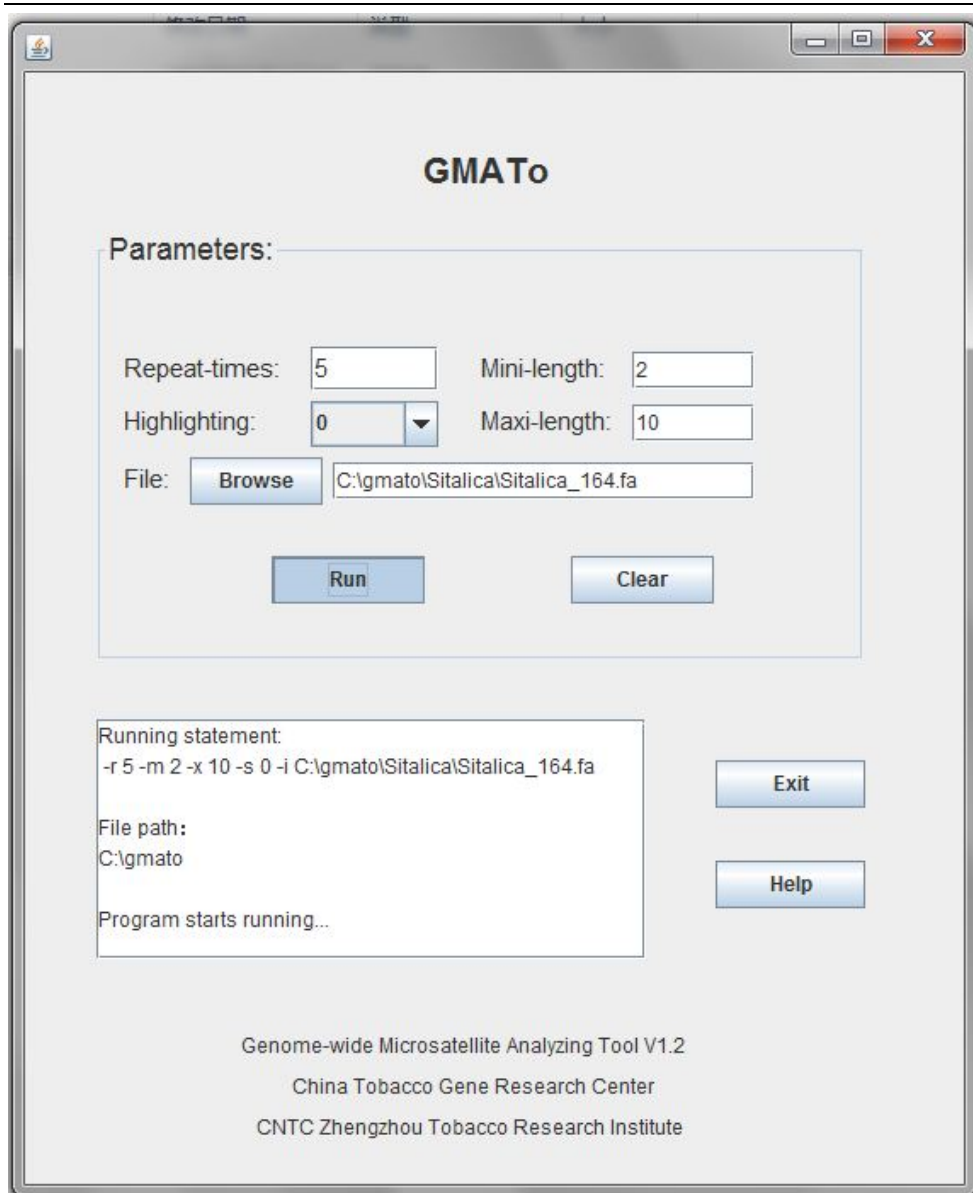


Figure 1 GMATo graphic interface---starting running

- step 2. Set up the parameter values or use the default values.
- step 3. Click the button “browse” to select the input DNA sequence file, followed by click “open” in popup window.
- step 4. Click button “Run” to run the program. The program will format the input sequences, then mine SSR and generate statistical results. Running status information and results files location will be shown in the text box below (Figure 2). The result files usually will be located in the same folder of the input sequence file.

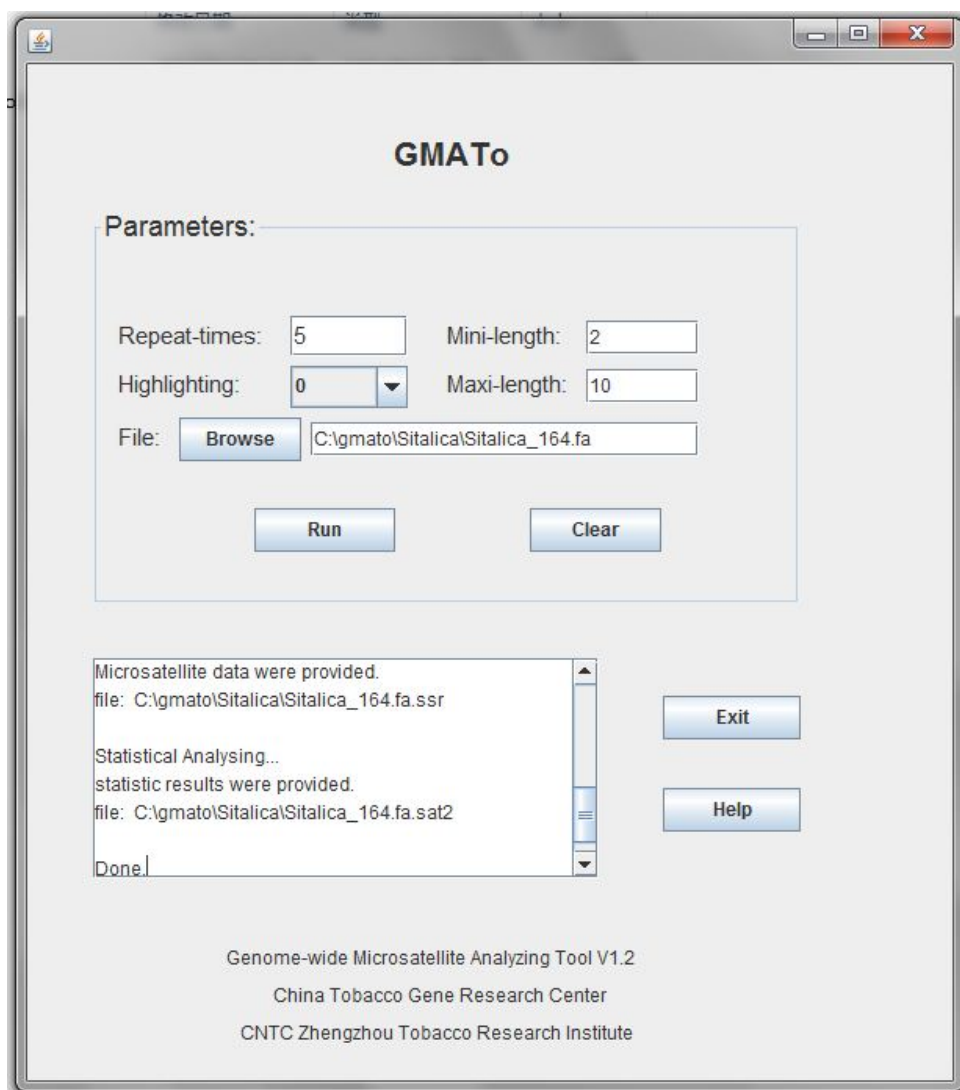


Figure 2 GMATo graphic interface--- running completed

- step 5. When analyzing is done, a message will be given. Click “OK” to finish.
- step 6. Open result files in text editor or any other program such as spread sheet. For details, refer to section 0 at page 14 .
- step 7. Click “clear” reset values and start a new round of running.
- step 8. Click “Exit” to exit GMATo.

Example:

Install the program GMATo in folder “H:\gmatoV1.0”.

Copy the sequence file “testseq.fasta” to a folder called “data” (any name you can set) in the

directory "gmatov1.0". If the test file comes with the program so just choose it in the browse step.

The test sequence consist of 3226 real DNA sequence obtained from NCBI.

Click "gmat.jar". A window looks like Figure 1 will appear. Followed the steps describe above. The text box of the interface will show the running details:

The following information was given in when running.

Running statement:

```
-r 5 -m 2 -x 10 -s 0 -i H:\gmatov1.0\data\testseq.fasta
```

File path:

H:\gmatov1.0

Program starts running...

Please wait during running...

Sequences were successfully formatted.

file: H:\gmatov1.0\data\testseq.fasta.fms.

file: H:\gmatov1.0\data\testseq.fasta.sat1

Microsatellite data were provided.

file: H:\gmatov1.0\data\testseq.fasta.ssr

statistic results were provided.

file: H:\gmatov1.0\data\testseq.fasta.sat2

Done.

4. 2. Command line mode

Here shows how to run GMATo in Windows, Unix and MAC OS. The test sequence data file (testseq.fasta) coming with the program is located in a folder called "data". The test sequence consists of 3226 real DNA sequences obtained from NCBI.

Same as graphic interface, the output files are located in the same folder as the input DNA sequence file.

The command mode provides one step running module by calling "gmat.pl", and the details were given below. For some special complicated purpose, the program offers running independently or sub-sequentially format sequence module (formatchunk.pl), SSR search module (gssr.pl, gssrtrim.pl) and statistic module (gsts.pl) if necessary. For example, the SSR search module can be run for two kinds of specific motif separately and then the SSR result can be merged followed by statistical analysis if necessary.

Linux /Mac

In the command terminal, type the following command followed by return to run GMATo. There is a single space between words. For the meaning of the parameters, please read related information in section 0 at page 3.

Command :

```
perl [PATH]/gmat.pl parameters -i [PATH]/SequenceFileName
```

Alternatively, go to the soft directory first using command "cd", then type the following command to run. The [directory] is the directory name of where the sequence file located.

```
perl gmat.pl parameters -i [directory]\SequenceFileName
```

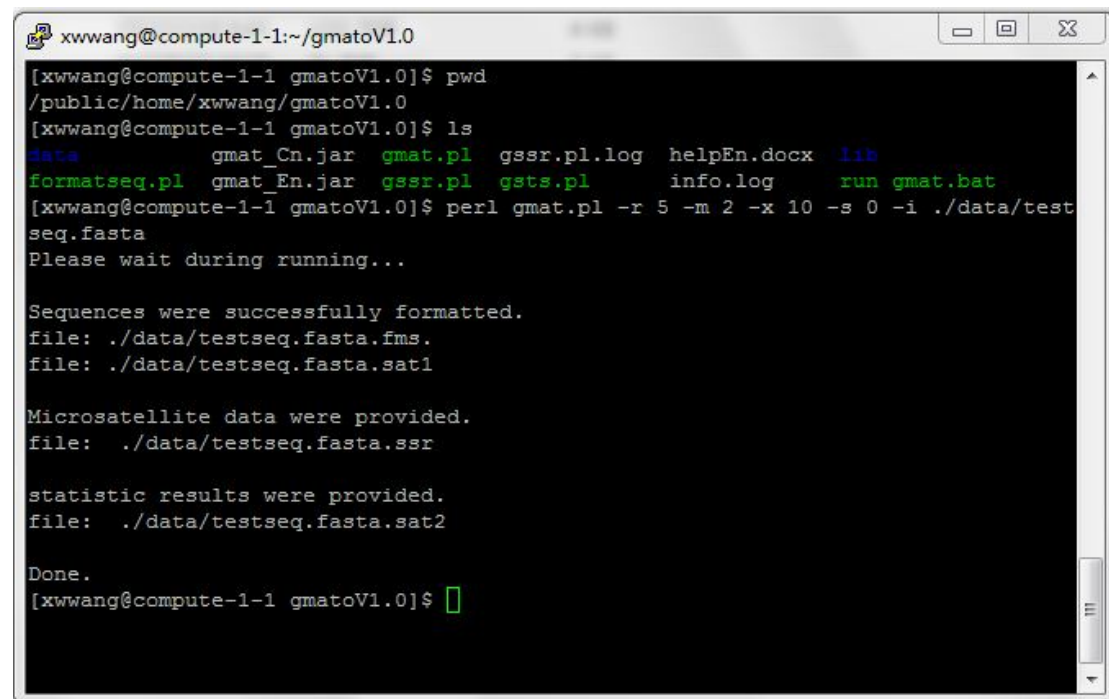
example:

Install the program GMATo in folder "gmatov1.0". Copy the sequence file to a folder called "data"

(any name you can set) in the directory "gmatoV1.0". Then go to the directory of "gmatoV1.0". Type the following command to run program GMATo. Set up the parameters to -r 5 -m 2 -x 10 -s 0 . Note there is a space between letters. This running screen captured in terminal was shown in Figure 3.

Command:

```
perl gmat.pl -r 5 -m 2 -x 10 -s 0 -i ./data/testseq.fasta
```



```
xwwang@compute-1-1:~/gmatoV1.0
[xwwang@compute-1-1 gmatoV1.0]$ pwd
/public/home/xwwang/gmatoV1.0
[xwwang@compute-1-1 gmatoV1.0]$ ls
data          gmat_Cn.jar  gmat.pl      gssr.pl.log  helpEn.docx  lib
formatseq.pl  gmat_En.jar  gssr.pl      gsts.pl      info.log     run gmat.bat
[xwwang@compute-1-1 gmatoV1.0]$ perl gmat.pl -r 5 -m 2 -x 10 -s 0 -i ./data/testseq.fasta
Please wait during running...

Sequences were successfully formatted.
file: ./data/testseq.fasta.fms.
file: ./data/testseq.fasta.sat1

Microsatellite data were provided.
file: ./data/testseq.fasta.ssr

statistic results were provided.
file: ./data/testseq.fasta.sat2

Done.
[xwwang@compute-1-1 gmatoV1.0]$
```

Figure 3 GMATo running screen at Linux system

Windows

In the command terminal, type the following command followed by return to run GMATo. There is a single space between words. For the meaning of the parameters, please read related information in section 0 at page 3.

Command in Windows:

```
perl [PATH]\gmat.pl parameters -i [PATH]\SequenceFileName
```

If the working directory is same as the soft GMATo and DNA Sequence File is also in the same

directory, the command will be as simple as the following command.

```
perl gmat.pl parameters -i SequenceFileName
```

Usually the working directory is same as the GMATo and DNA Sequence File is in another folder such as c:\data directory, the command will be this listed in the following line.

```
perl gmat.pl parameters -i c:\data \SequenceFileName
```

example:

e.g. install GMATo in directory "H:\gmatoV1.0" and put the sequence file in its subfolder called "data".

Set up the parameters to -r 5 -m 2 -x 10 -s 0 . Note there is a space between letters. The order of parameter can be changed. The order of -m 2 -x 10 -s 0 -r 5 is the same as to -r 5 -m 2 -x 10 -s 0.

Go to GMATo soft directory using command:

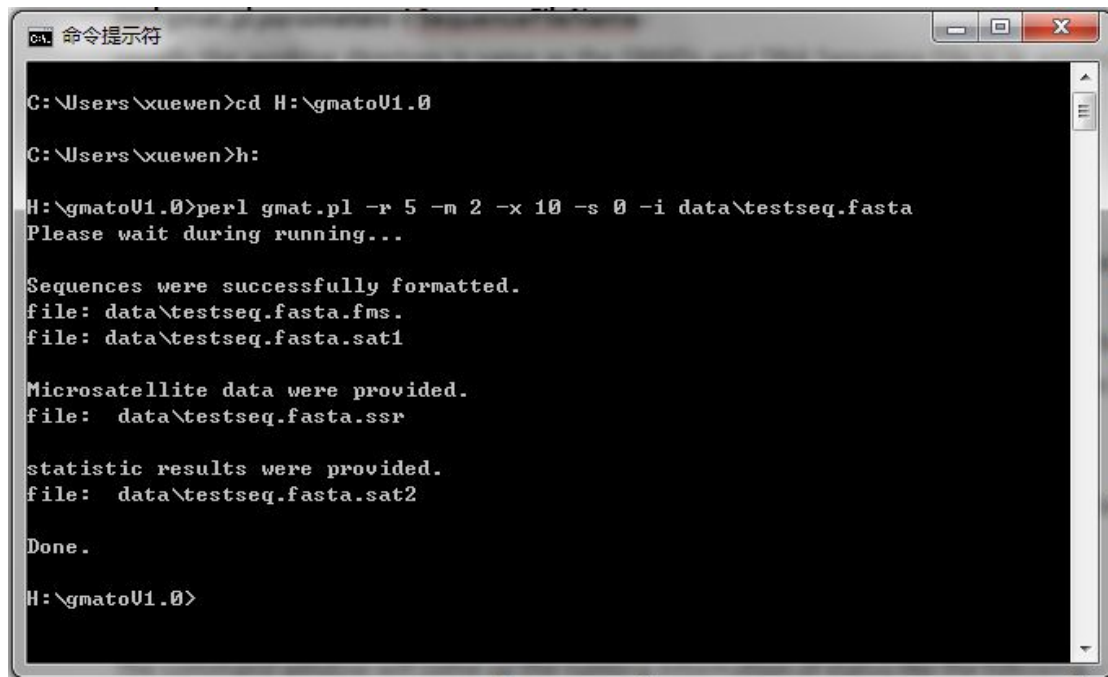
```
cd H:\gmatoV1.0
```

```
h:
```

In command window, type the following command in MS DOS command shell followed by return.

```
perl gmat.pl -r 5 -m 2 -x 10 -s 0 -i data\testseq.fasta
```

The command window will come up the running information of status like Figure 4.



```
C:\Users\xuewen>cd H:\gmatoU1.0
C:\Users\xuewen>h:
H:\gmatoU1.0>perl gmat.pl -r 5 -m 2 -x 10 -s 0 -i data\testseq.fasta
Please wait during running...

Sequences were successfully formatted.
file: data\testseq.fasta.fms.
file: data\testseq.fasta.sat1

Microsatellite data were provided.
file: data\testseq.fasta.ssr

statistic results were provided.
file: data\testseq.fasta.sat2

Done.
H:\gmatoU1.0>
```

Figure 4 GMATo running screen at DOS shell

5. Input file

The input file can be prepared using any text processing soft and then saved in plain text format. If the sequences were downloaded from the NCBI or from a genome sequence database, the file can be directly used as an input file. If you prepared the sequences from your experiment, you should create the file in the described below. The format of input file is similar to (raw) fasta format, beginning with a > sign for the sequence name in one line identified by line return sign and then followed by one or multiple lines of DNA sequence. For soft GMATo V1.0 or above, the DNA sequence may be ATGC or any nucleotide letter, number, space, tab and empty lines, and line return, which is very useful for some DNA copied from web page such as Arabidopsis web TAIR.

There is no length limitation for the number of sequences and length of each sequence if its length is less than the computer maximum memory.

```
>testseq1
```

```
GCTA123456GGGTTGGGGCTAATCTCAAGTGTGATTACTGTCTCTCTCTCTCTCTNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNATATTATTACAATTAGAACAAATCTACCATACCTATTTTTTTTATATTCATTTCTTCCG
TAAAAAAAAGAAGGTATGGTAGGTTTGTCTAATTGTAATAATATTGATATATATATATATATATATATATA
TATATAATATATATTAGGTATTATCCTATATTCATATACTAATCTGTTTTTGTCAATTACGACAAAACCTTCCTAA
GCACTCTGACGCTAGAT    CCAG
```

```
TATATTCATTTCTTCCGTAAAAAAAAGAAGGTATGGTAGGTTTGTCTAATTGTAATAATATTGATATATATAT
ATATATATATATATATATATAATATATATTAGGTATTATCCTATATTCATATACTAATCTGTTTTTGTCAATTA
CGACAAAACCTTCCTAAGCACTCTGACGCTAGATCCAG
```

```
>MethylFiltered.Decon.masked_contig_79|1992:2197|206/189-189
```

```
GTCCTTGACAGCGAACTTGAAGAACTGGTAATGCT
TGTACATCTTAAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTATCTAATA
TACATTACATAAGTACACTGGTTGATTGTCATATGCTCTTTGTTTAAACTGTTTATTGTTTGAATGCAGCTATTC
CAGTTGGCCAGTTATTGTTCTGTTGTACTCTGTTGTCGGGTGGC
```

```
>AT2G18790.1 from TAIR http://arabidopsis.org/servlets/TairObject?type=sequence&id=25602
```

```
1 CTTCAATTTA TTTTATTGGT TTCTCCACTT ATCTCCGATC TCAATTCTCC
51 CCATTTTCTT CTCCTCAAG TTCAAAATTC TTGAGAATTT AGCTCTACCA
```

101 GAATTCGTCT CCGATAACTA GTGGATGATG ATTCACCCTA AATCCTTCCT

.....

6. Output files

There are two result files generated by GMATo: SSR location file with suffix .ssr and SSR distribution statistical file suffix .sat2 in the file name. The location of the files will be the same directory as the input sequence file.

6. 1. Formatting report

The formatting report with suffix .sat1 in the file name is a summary of cleaned input sequences which is processed by formatting module. The report summarizes the total number of input sequences, total length (bp) after cleaning during formatting and total chunking sites.

An example of the formatting report of *Setaria Italica* whole genome sequences:

formatchunk.pl was run and results were yielded at Mon Dec 24 03:09:09 2012

statistic summary of input sequence(s)

total input sequences #: 336

total length (bp) after formatted: 405737341

total chunking sites after formatted: 138

6. 2. SSR location file

By default, the SSR loci file has the same name as the input file plus suffix .ssr in the end of file name.

There are 6 columns of data with one head line of clear self-explained titles including source sequence name, source sequence length, repeated sequence starting position, repeated sequence ending position, repetitions (repeated times) and motif of repeated sequence. Each line lists one repeated locus. If there are more loci in the same input source sequence, the repeated loci will be present in lines sequentially. The format of this output file is in tabular plain text.

An example of data in SSR location file:

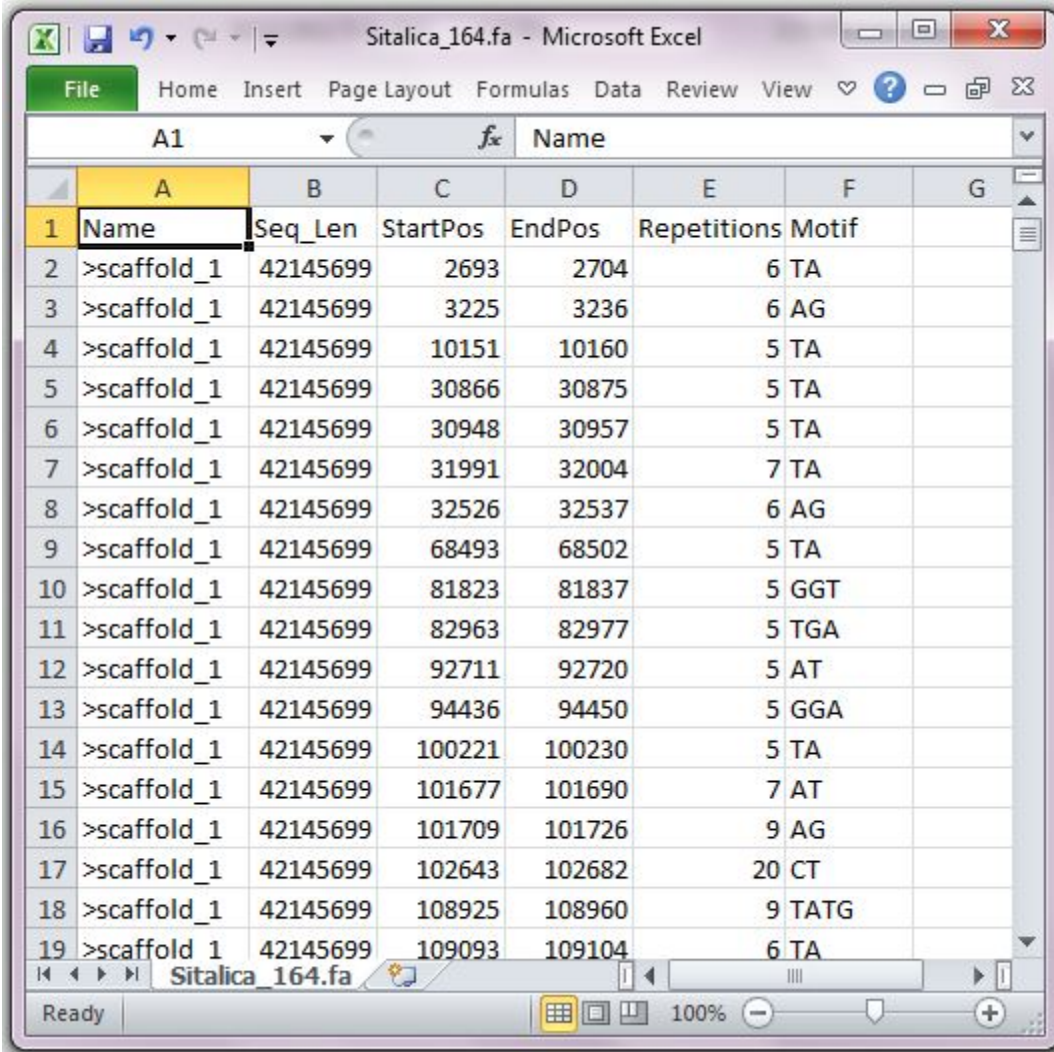
Name	Seq_Len	StartPos	EndPos	Repetitions	Motif
>testimperfect	83	36	53	9	TC

```

>testseq1      506      36      51      8      TC
>testseq1      506     191     222     16     AT
>testseq1      506     382     413     16     AT
>testseq2      184      24      41      6     ACC
.....

```

Another example of viewing the data in SSR location file in Microsoft Excel (Figure 1):



	A	B	C	D	E	F	G
	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	30948	30957	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	

Figure 5 Viewing SSR loci information file in MS Excel

6. 3. SSR distribution statistical file

By default, the SSR distribution statistical file has the same name as the input file plus suffix .sat2 in

the end of file name. The format of this output file is also in tabular plain text.

The statistical distribution file provides four statistical results at different types of classification at genome aspect. A summary is generated in the end of each classification. For the classifications described below, please refer to the **Example of SSR distribution statistic file** given page 16.

Classification I (table 1): The motif length statistics provides overview information for the type, abundance, and rank of each length-based motif, also providing total types and total motifs.

Classification II (table 2): the motif statistics provides distribution for each motif based on sequence, i.e. detailed motif sequence, occurrence, total motif number and total occurrence. This result reveals the motif composition, distribution by abundance rank in a genome.

Classification III (table 3): grouped complementary motifs statistics provides distribution for complementary motifs such as TC/GA. The results list grouped motif types and corresponding occurrence in ranked order, also providing the total of grouped types and total occurrence. Classification III resolves the distribution of grouped motif among all chromosomes/scaffolds. Classification II can provide insight to the distribution of a certain motif pair in an investigated chromosome/scaffold. e.g. The distribution of motif TC may be different from its complementary motif GA in a chromosome.

Classification IV (table 4): Chromosome level distribution statistics provides the total occurrence of motif(s) and SSR frequency (loci/Mbp) at each chromosome or super-scaffold. This statistic data provides insight to comparison of motif(s) frequency among chromosomes. Furthermore, altering motif length used in mining, the corresponding statistical data also provide comparison information of frequency for a given motif among chromosomes for evolution investigation.

Example of SSR distribution statistic file of Zea Mays (>2Gb genome):

The following data is part of SSR distribution data from SSR distribution statistic file. The genome

sequence was Zea Mays (corn) sequence downloaded from NCBI. Parameters setting are -m 2 -x 10 -r 5 and -m1 -x 1 -r 10.

gsts.pl was run and results were yielded at Fri Sep 21 06:09:05 2012

Table 1

Motif(-mer)	total
2	158848
1	85905
3	27279
4	1086
5	156
6	59
7	3
8	1
total_above	total_above
8	273337

Table 2

Motif	total
AT	24589
TA	23164
G	23093
C	22616
GA	21064
CT	20246
T	20241
A	19955
TC	16676

AG	15536
CG	7918
GC	7716
TG	7033
CA	5840
AC	5116
GT	3950
CAG	3695
...	
CTGCTC	1
CACTA	1
GCAGT	1
CTCGAC	1
total_above	total_above
256	273337

Table 3 paired

Grouped_Motif	total
G/C	45709
T/A	40196
GA/TC	37740
AG/CT	35782
AT/AT	24589
TA/TA	23164
TG/CA	12873
GT/AC	9066
CG/CG	7918
GC/GC	7716
CAG/CTG	6100

...

TACGA/TCGTA 1

CTGAC/GTCAG 1

TACTGC/GCAGTA 1

total_above total_above

172 273337

An example of classification 4 for Parameters setting -m 2 -x 10 -r 5

Table 4

SeqID	Total_Motifs	SeqSize	Frequency(Motifs/Mb)
>1 chromosome:AGPv2:1:1:301354135:1 chromosome 1	27585	301354135	91.536822615691
>4 chromosome:AGPv2:4:1:241473504:1 chromosome 4	22002	241473504	91.1155867436288
...			
>UNKNOWN chromosome:AGPv2:UNKNOWN:1:7140151:1 chromosome UNKNOWN	419	7140151	58.6822323505483
>Mt chromosome:AGPv2:mitochondrion:1:569630:1 chromosome mitochondrion	36	569630	63.1989185962818
>Pt chromosome:AGPv2:chloroplast:1:140384:1 chromosome chloroplast	7	140384	49.8632322771826
total_above	total_above	total_above	average_frequency
13	187432	2066432718	90.703170912531

7. Support and Contact

Any suggestion and question are welcome. We try to answer your question(s). However, it is not guaranteed to always answer every question. I believe with your help will speed improving this soft GMATo.

Xuewen Wang, PhD

Email: xwwang@ymail.com