Discovery at the interface of physical and biological sciences

open access

www.bioinformation.net Volume 9(10)

Software

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.

GI	MATo			fome Insert Pag	e Layout Po	Name	Review VI	ew -> 6	1997.0
Parametera			A	в	c	D	Ĕ		
Parameters.			1 Name	Seg Len	StartPos	EndPos	Repetitions	Motif	1
			2 >scatto	ld 1 42145699	2693	2704	6	TA	
Repeat-times: 5	Mini-leopth iz		2 weatte	10 1 02105699	8775	1776	6	0.0	
			d mentle	Id 1 42145695	10151	10160		TA	
ringnagnung. jo -	Krass tength: 10		a sealle	4 1 42245055	10101	10200		TO	
File Browne (Chgmatodist	ilica/litalica_164 fa		A sealle	LI 1 4314E600	30000	30073		TA	
			7 sealle	4 1 42145606	21001	22004		TA	
run	Limit		/ >scarro	10_1 42145695	31991	32004	2	IA	
1	Alter and the second se		o scarro	40_1 42143093	52520	52557		AG .	
			5 Ascarto	1d 1 42145695	00493	00502	5	OGT	
			10 /scano	42143033	01023	01037	5	TCA	
Morosatellite data were provided.			11 >scaffo	42145699	82963	82977	5	AD	
The Congression and a darkada and a second a		Fait	12 >scaffo	42145695	92711	92720	5	AT .	
Statistical Analysing			13 >scaffo	10_1 42145695	94436	94450	2	GGA	
statistic results were provided.	at2	Help	14 >scaffo	10_1 42145699	100221	100230	2	TA	
and contraction and a second second			15 >scaffo	Id_1 42145699	101677	101690	7	AT	
Dane.			10 >scaffo	10_1 42145695	101709	101726	9	AG	
				and the second se					
			17 >scaffo	ld_1 42145695	102643	102682	20	CT	
Genome-wide Microsof	ente Analyzing Test V1,2		17 >scaffo 18 >scaffo	ld_1 42145695 ld_1 42145695	102643	102682	20	TATG	
Genome wide Microsol China Tobacco B	ente Analyzing Toul V1,2 ens Research Deoler		17 >scaffo 18 >scaffo 19 >scaffo	Id_1 42145695 Id_1 42145695 Id_1 42145695 Id_1 42145695 Sitalica 164.6a	102643	102682 108960 109104	20	TATG	
Genome wide Microsof China tobacco di CNTC Zhengshou Tel	ellite Analyping Teal V1,2 ena Research Declay Sacco Research Institute		17 >scaffe 18 >scaffe 19 >scaffe Ready	ld_1 4214569 ld_1 4214569 ld 1 4214569 ld 1 4214569 Sitalica_164.fa	102643 108925 109093	102682 108960 109104	20 9 4 6 1 100% ①	TATG	
Genome wide Microsel China Tobacco G CNTC Zhengshou To (2) al 19 + cv - v Stellor,	ente Analyong Toli V1.2 ene Research Caoler sacco Research Institute		17 >scaffo 18 >scaffo 19 >scaffo 19 >scaffo Ready	Id_1 4214569 Id_1 4214569 Id_1 4214569 Sitalica_164.fa	102643 108925 109093	102682 108960 109104 109104	20 9 100% (-)		в
Genome wide Microsol Ghina Tobacco G CNTC Zhengshou Tol (2) La 49 + Co - V Stalica, The Pose Intel Page Lap	ente Analysing Teel V1,2 ene Research Daolor bacco Research Institute		17 >scaffo 18 >scaffo 19 >scaffo 19 >scaffo Ready Statics_184.fa Page Lagout 75	Id 1 4214569 Id 1 4214569 Sitalica 164.fa [2] a 9 + 0 + x 169 international form	102643 108925 109093 20109093	102682 108960 109104 109104 109104 109104 109104 109104 109104	20 9 4 6 100% (-) 		в
Ganome wide Microsol China Tobacco B CNTC Zhengshou Tol CNTC Zhengshou Tol (2) 109 + (n - (n Stalica, 100 -	entre Analysing Teel V1.2 ene Research Dector secco Research Institute	THE last of a constant fill be of a constant fill be been a constant fill be been a constant fill be constant fill be a constant fill be a constant fill be a constan	17 >scaffo 18 >scaffo 19 >caffo 19 + 11 Ready Statics JR4 to Page Lanut Fr	Id 1 4214569; Id 1 641, 62;	102643 108925 109093 109093 Statics 1844 or Page layord Free C & A	102682 108960 109104 	20 9 6 100% (-) 	CT TATG TA	в
Genome wide Microsel China Tobacco G CNTC Zhengshou Tol CNTC Zhengshou Tol (2) I W #9 + Cu - V Steller, China Fores Inium Page Lapo (2) China Fores Inium Page	ente Analyong Teil V1.2 ene Research Goder Saco Research Institute	The time base C1971 A 915 DCAGT	17 >Scaffc 18 >Scaffc 19 >scaffc 4 > 10 Ready Page Laevet Fo 5 8	Id 42145693 Id 42145693 Id 42145693 Sitalica 164.fa Id 194.fa	102643 108925 109093 109093 109093 Faalos 184.6x Frankow Frankow Frankow Frankow Frankow Frankow Frankow Frankow Frankow Frankow	102682 108960 109104 (1) (1)(1)(2)(2)(4) (1)(1)(2)(2)(4) (1)(1)(2)(4)(4) (1)(1)(2)(4)(4) (1)(1)(1)(1)(4)(4) (1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(1)(20 9 9 100% (-) 100% (-) 100% (-) 10 100% (-) 10 10 10 10 10 10 10 10 10 10 10 10 10	CT TATG TA	в
Genome wide Microsol Ghina Tobacco di CNTC Zhengshou Tol CNTC Zhengshou Tol (2) I al #9 + (% -) Stalica, The Pose Initit Page Ian Bose Initit Page Ian A B A B A B A B A B A B A B A B A B A B	ente Analysing Teel V1.2 erv Research Dactor Dacco Research Institute	IX Here terms been C1071 SIX GCAGT DIS AVGCA	17 >Scaffo 19 >scaffo 10 >sc	Id_1 4214569; Id_1 4214569; Id_1 4214569; Situlica 164.7a Id_1 4214569; Situlica 164.7a Id_1 4214569; Id_1 4214569; Id_1 4214569; Id_1 4214569; Id_2 4 Id_2 4 Id_2 4 Id_2 A Id_2 A Id_3 4 Id_4 4	102643 108925 20109093 2010900 2010900 20100000000000000000000	102652 105960 109104 1111623 U 1111623 U 1111623 U 111165 12 0	20 9 6 100% (-) 100% (-)		в
Genome wide Microsa China Tobacco G CNTC Zhengzhou Tol CNTC Zhengzhou Tol (2) 109 + 04 - 10 Steller, 100 Hone Intert Page Layo (1) 5 + 0 (1) 4 + 0 (2) 5 + 0	Analyzing Toell V1.2 and Research Dactor bacco Research Institute (2) I a 90 + Seal (1) France Interf Francis Fa	Image: Second	17 >scaffo 18 >scaffo 19 >scaffo 10 >sc	Id 1 4214569: id 1 4214569: Sitalica 104.fa Sitalica 104.fa 1047.casto 1047.casto 1047.casto 1047.casto 1047.rastrastrastrastrastrastrastrastrastrast	102643 108925 10993 10993 Salica Jilday Pagland Tan Tagla Jilday Tagla Tagla Jilday	102662 108960 109104 100105 100105 100 10010 Bate Table 6 E 0	20 9 9 100% (-) 100% (-)		в
Genome wide Microsol China Tobacco di CNITC Zhengshou Tol CNITC Zhengshou Tol Tobac Iniut Page Jan 1 gits pl was run and results w 2 Table 1 Mottif-merj Total 5 3 18509	ente Analysing Toil V1.2 ent Research Daole/ bacco Research Institute file for institute file for institute A B 15 Table 2 16 Motil total 37 TA 4478 16 QA 5723 19 QA 5723	IX III IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Chairs and the second s	Id_1 42145057 Id_1 42145505 Sitalica 164.fa Sitalica 164.fa Id 1 42145695 Sitalica 164.fa Id 1 42145695 Sitalica 164.fa Id 1 42145695 Id 1 42145695	Ladica, 164-16 v Ladica, 164-16 v Francisco V Francisco V California Californi California Californi California Califo	102662 105960 109104 I IIII C II IIII C III IIII C III IIII C III IIII C IIII IIII C IIII IIII IIII C IIII IIII IIII C IIII IIII IIII IIII IIII IIII IIII IIII IIII	20 9 4 6 100% (-)		в
Genome wide Microad China Tobacco di CNTC Zhengshou Tol CNTC Zhengshou Tol (2) ad #9 + 0 - V Stalica, 10 - 10 - 10 - 10 - 10 - 10 - 10 - 10 -	Analysing Toell V1.2 erv Research Daolor Daoco Research Institute Climer 2	X a true true true true true true true true	Statics J64 fa 17 >scaffo 18 >scaffo 19 >scaffo 19 >scaffo 14 + 1 + 1 Ready Statics J64 fa Page Laevet 1 1 1 1 1 1 1 1 1 1 1 1 1	Id_1 4214569; Id_1 4214569; Id_1 4214569; Situlica 164.7a Id_1 4214569; Situlica 164.7a Id_1 4214569; Id_1 4214569; Id_1 4214569; Id_1 4214569; Id_1 4214569; Id_2 A Id_2 A Id_2 A Id_2 A Id_2 A Id_2 A Id_3 A Id_3 A Id_3 A Id_4 A Id_4 A Id_5 A Id_4 A Id_4 A Id_4 A Id_6	102043 108925 109093 20 Partiana Partia	102682 108960 109104 IIII 2210 # Alexendi Iard Table 6 E 0	20 9 9 100% (-): 100% (-): 100% (-):		в
Centome wide Microsal China Tobacco G CNTC Zhengzhou Tol (2) I (9) + (9 - V Steller, 100 Hone Inter Papelays 615 + (100 Hone) 1 Table 1 1 Table 1	Image: Second	Image: State of the s	Statica_JS4_fa Page Larout Fe - 5 10 Statica_JS4_fa Page Larout Fe - 5 0 1 1 1 1 1 1 2 2 2 3 1 1 2 2 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1	Id_1 42145693 Id_1 42145693 Sitalica 1647,63 Sitalica 1647,63 Id_1 647,63 Sitalica 1647,63 Id_2 9,75 Id_1 647,63 Id_2 647,74 Id_2 647,74 Id_3 74,74 Id_3 74,74	L02043 108925 199093 199093 L09093 L09093 L09093 	102952 103960 109104 1111 col 10 1111 col 10 1111 col 10 1111 col 10 1111 col 10 10 10 10 10 10 10 10 10 10 10 10 10 1	20 9 100% (-) 100% (-)	CT TATG TA II II II II II II II II II II II II II	в
Genome wide Microad China Tobacco di CNITC Zhengshou Tol CNITC Zhengshou Tol Tobac inist Tess Lan Tobac inist Tess Lan Tess Lan T	Image: Second Control V1.2 Image: Second Control Second Control Second Research Institute Second Control Image: Second Control Second Con	IX I I I I I I I I I I I I I I I I I I	17 >Scaffo 18 >Scaffo 19 >Scaffo 19 >Scaffo Ready Statica 154 fa Page Lanut Fa 6 1 1014 - Fa 1 1014 - Fa 1 1014 - Fa 1 1014 - Fa 1 1014 - Fa 1014 -	Id_1 4214500; Id_2 A Id_2 A Id_2 A Id_1 A Id_2 A Id_3 A Id_3 A Id_3 A Id_3 A Id_3 A Id_3 A Id_3<	102043 108925 109093 200900 20090000000000	102662 102960 109104 1111623 U 1111623 U 1111623 U 111163 1111163 1111163 111163 111163 111	20 9 9 100% (-) 10% (-) 10% (-) 10% (-) 10%		в
Genome wide Microad China Tobacco di CNTC Zhengshou Tol CNTC Zhengshou Tol (2) ad #9 + 00 - V Stalica, 100 File Fore Inite Page Lapo 2 Table 1 3 Mattifumeri Total 4 2 31304 5 3 10509 1 4309 7 5 272 8 6 1327 8 6 1327 9 10 38	Analysing Teel V1.2 erv Research Dactor Dacor Research Institute Cline York Institute	X I I I I I I I I I I I I I I I I I I I	17 >Scaffc 18 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 14 + Page Lacout Page Lacout 1 1 1014/_above 0 46719 10 1014/_above 6333 4472 10	Id_1 42145693 Id_1 42145693 Id_1 42145693 Id_1 42145693 Situalica 164.7a Id_1 42145693 Situalica 164.7a Id_22 A Id_23 A Id_24 A Id_25 A Id_26 A Id_27 A Id_27 A Id_27 A Id_27 Sitabore Id_27 Sitafield	Lozo43 103925 1039555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 103955555 10395555555 103955555 103955555 1039555555 103955555 103955555555 1039555555 1039555555 10395555555555 1039555555555555555555555555555555555555	102952 103960 109104 11111 11111 11111 11111 111111	200 9 100% (-) 100% (-)	CT TATG TA M M M M M M M M M M M M M	в
Centome wide Microsol China Tobacco G CNTC Zhengzhou Tol (2) I (9) + (9 - V Steller, 1000 Inter Page lays 615 + (1000 Inter Page lays) 615 + (1000 Inter Page	Image: Second	Image: Second	17 >scaffc 18 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 10	Id_1 42145693 Id_1 42145693 Id_1 42145693 Sitzalica 1041,6a Id_1 1041,6a Sitzalica 1044,6a Id_1 1044,6a Id_1 1044,6a Id_1 1044,6a Id_1 1047,6a Id_2 1047,6a Id_2 1047,6a Id_2 1047,7a Id_2 1047,7a Id_2 5a Id_2 5a Id_3 5a Id_3 5a Id_3 5a Id_3 5a Id_3 5a Id_3 5a Id_4 1275,5a Id_4 1275,5a Id_4 1275,5a Id_4 1275,5a Id_4 1275,5a Id_6 1275,5a	L02043 1020925 109093 199093 199093 199093 199093 199093 10981,40009 7 60739 10981,40009 7 60739 10981,40009 7 60739 10993,40000 10993,40000 109093 10000000000	1020632 103960 109104 1111 cm 10 1111 cm 10 1111 cm 10 1111 cm 10 120	20 9 100% (-) 100% (-	CT TATG TA B B B B B B B B B B B B B B B B B B	B
Genome wide Microad Chris Tobaco di CNITC Zhengshou Tol CNITC Zhengshou Tol Tobac inist Page lan 1. gtts.pl was run and results w 2. Table J 3. Table J 3. Table J 4. gtts.pl was run and results w 2. Table J 3. doi:10.000 tobal 3. doi:10.0000 tobal 3. doi:10.00000 tobal 3. doi:10.0000000 tobal 3. doi:10.00000000000000000000000000000000000	Image: Second Control V1.2 envir Research Control Second Control bacco Research Institute Second Control Second Research Institute Second Contro <	IX and any	17 >scaffo 18 >scaffo 19 >scaffo 19 >scaffo 19 >scaffo 19 -scaffo 19 -scaffo 10 -sc	Id_1 42145693 Id_1 42145693 Id_1 42145693 Sitalica 164.fa Id_1 42145693 Sitalica 164.fa Id_1 42145693 Sitalica 164.fa Id_1 42145693 Id_1 42145693 Id_1 42145693 Id_1 42145693 Id_1 42145693 Id_2 A Id_1 4214746645 Id_1 42145693 Id_2 A Id_1 4214746645 Id_2 Jacoba Id_1 4214746645 Id_2 Jacoba	Lozo43 102925 109925 109093 Page Logod Page Logod Page Logod R 1 1098 _ 40719 7 1018 _ 40719 7 1018 _ 40719 7 1018 _ 40719 1029 3091 3092 3092 3092 3092 3092 3092 109093	102952 103960 109104 (IIII) Cli (III) (IIII) Cli (III) (IIII) Cli (III) (III) Cli (III) (IIII) Cli (III) (IIII) Cli (III) (III) Cli (20 9 9 100% (-) 10% (-) 10% (-) 10% 50 10% 50 10 10 50 50 51 51 51 51 51 51 51 51 51 51 51 51 51	CT TATG TA M M M M M M M M M M	в
Genome wide Microad China Tobacco di CNTC Zhengshou Tol CNTC Zhengshou Tol (2) ad 49 + 50 - 2 Stalics, 100 - 100 - 100 - 100 2 Stalics, 100 - 100 - 2 Stalics, 100 - 2 Stali	Analysing Toell V1.2 env Research Dactor bacco Research Institute Clip Processor Clip Processor To Tabler 2 16 16 17 18 19 10 10 11 12 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 20 21 22 24 25 26 27 28 29 20 21 22 24 25 26 27 28 29 20 210 210	X Image Ima	17 >Scaffc 18 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 14 + Page Lagout Fe - - - - - </td <td>Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_2 1435893 Id_3 42145893 Id_4 41472 Id_4 41472 Id_3 41472 Id_4 41474 Id_4 41474 Id_5 54104 Id_7 544164 Id_7 544164 Id_7 544164 Id_7 544164 Id_7 544164 Id_7 544164 Id_7 544164</td> <td>102043 103925 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 103955555 10395555 103955555 103955555 103955555 103955555 103955555 103955555 1039555555 1039555555 103955555 1039555555 1039555555 1039555555 1039555555 1039555555 1039555555 103955555555 10395555555 103955555555 10395555555555 1039555555555555555555555555555555555555</td> <td>102952 103960 109104 101105</td> <td>20 9 100% (-) 100% (-</td> <td>CT TATG TA M M M M M M M M M</td> <td>B</td>	Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_1 42145893 Id_2 1435893 Id_3 42145893 Id_4 41472 Id_4 41472 Id_3 41472 Id_4 41474 Id_4 41474 Id_5 54104 Id_7 544164	102043 103925 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 103955 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 103955555 10395555 103955555 103955555 103955555 103955555 103955555 103955555 1039555555 1039555555 103955555 1039555555 1039555555 1039555555 1039555555 1039555555 1039555555 103955555555 10395555555 103955555555 10395555555555 1039555555555555555555555555555555555555	102952 103960 109104 101105	20 9 100% (-) 100% (-	CT TATG TA M M M M M M M M M	B
Concorrected Microsol China Tobacco G CNTC Zhengzhou Tol CNTC Zhengzhou Tol (2) a 49 + 54 - 7 Steller, 100 - 1000 Inter Pape Iapo 615 + 6 2 a 8 8 3 about 1 3 about 1 4 about 1 3 about 1 4 about 1 5 a 1350 5 a 1350 5 a 1350 5 a 1350 7 5 327 8 a 1350 5 a 1350	Control Control Control Control Control <td< td=""><td>Image: Second Second</td><td>17 >Scaffc 18 >scaffc 19 >scaffc 10 1 <tr tr=""> <t< td=""><td>Id 42149607 Id 42143697 Id 42143697 Sitzalica 104/6 Sitzalica 104/6 Id 42143697 Sitzalica 104/6 Id 104/6 Id 104/6 Id 104/7 Id</td></t<></tr></td></td<> <td>Lozoda 102043 108925 199093 Logo93 Logo93 Logo94 109091 10909 10909 10909 10909 10909 10909 10909 109091 109093 10909</td> <td>1029632 103960 109104 (1) (1) (1) (1) (1) (1) (1) (1)</td> <td>200 9 9 100% (-) 9 100% (-) 10 10 10 10 10 10 10 10 10 10 10 10 10</td> <td>CT TATG TA M M M M M M M M M M M M M M M M M M</td> <td>в</td>	Image: Second	17 >Scaffc 18 >scaffc 19 >scaffc 10 1 <tr tr=""> <t< td=""><td>Id 42149607 Id 42143697 Id 42143697 Sitzalica 104/6 Sitzalica 104/6 Id 42143697 Sitzalica 104/6 Id 104/6 Id 104/6 Id 104/7 Id</td></t<></tr>	Id 42149607 Id 42143697 Id 42143697 Sitzalica 104/6 Sitzalica 104/6 Id 42143697 Sitzalica 104/6 Id 104/6 Id 104/6 Id 104/7 Id	Lozoda 102043 108925 199093 Logo93 Logo93 Logo94 109091 10909 10909 10909 10909 10909 10909 10909 109091 109093 10909	1029632 103960 109104 (1) (1) (1) (1) (1) (1) (1) (1)	200 9 9 100% (-) 9 100% (-) 10 10 10 10 10 10 10 10 10 10 10 10 10	CT TATG TA M M M M M M M M M M M M M M M M M M	в
Id 42149607 Id 42143697 Id 42143697 Sitzalica 104/6 Sitzalica 104/6 Id 42143697 Sitzalica 104/6 Id 104/6 Id 104/6 Id 104/7 Id									
Genome wide Microad China Tobacco di CATC Zhengshou Tol CATC Zhengshou Tol 2000 States 1 (2000 States) 2 (2000	State Analysing Toill V1.2 envi Research Daolo State bacco Research Institute State State File File File File File State File File File File File	X A SID ATGLA A SID ATGLA A SID CAAT A SID CAAT A SID CAAT B SID CAAT C SID CAAT C SID CAAT C SID Table 3 patred C SID Table 3 patred C SID CAAT C SID Table 3 patred C	17 >>scaffo 18 >>scaffo 19 >>scaffo 10 - 10 >>scaffo	Id_1 4214500; Id_2 407; Id_1 42147; Id_1 421450; Id_2 407; Id_2 5401; Id_1 5401; Id_2 5401; Id	102043 109925 109925 109093 109093 Page Layout Page Layout 8 1 10981_abusy 7 10981_abusy 7 10981_abusy 7 10981_abusy 7 129 9000 3791 8 9000 3791 8 9000 30925 109025 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 109925 1099555 1099555 1095555 10955555 1095555 10955555 10955555 109555555 10955555 10955555555 1095555555555	102952 103960 109104 (IIII) III II IIIII III III IIIII IIII IIIII III IIIII III IIIII III IIIIII IIIIII IIIIII IIIIII IIIIII	200 9 100% (-) 100% (-) 10% (-	CT TATG TATG TA B B B B B B B B B B B B B B B B B B	в
Genome wide Microad China Tobacco di CNTC Zhengshou Tol (2) (4) (2) (4) (3) (4) (4) (5) (5) (5) (4) (5) (5) (4) (5) (4) (5) (4) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (7) (5) (5) (7) (6) (5) (6) (5) (7) (6) (7) (7) (7) (7) (7) (7) (7) (7) (7) (7) (7) (7)	Analysing Toul V1.2 env Research Cacler bacco Research Institute Cline York Cline York France Cline York France Cline York France France <t< td=""><td>X Image Processing Hit Processing Image Processing C1071 Image Processing 0.15 GCAGT 0.15 GCAGT 0.15 GCAGT 0.15 Total_above 0.10 Table 3 paired 0.10 Table 3 paired 0.10 Table 3 paired 0.10 Table 3 paired 0.20 Table 3 paired 0.21 Cat/Tc 0.22 AT/AT 0.23 GC/OC 0.24 GC/OC 0.27 GT/AC 0.28 GC/OC 0.29 GC/OC</td><td>17 >Scaffc 18 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 14 4 Page Lagout Page 1 1 1014 above 0 46719 1 1 1014 6814 6333 4472 8600 3822 1993 1044 1815 1781 1993 1048</td><td>Id_1 42149507 Id_1 42149507 Id_1 42149507 Situatica 164.76 Situatica 164.76 Situatica 164.76 Situatica 164.76 Situatica 164.76 Loss cattoor/accasto 4002 Loss cattoor/accasto 4002</td><td>102043 103925 103925 103925 103925 103925 103925 103925 10392 10392 103925 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 10395555 1039555 1039555 1039555 1039555 1039555 10395555 1039555 1039555 1039555 1039555 10395555 10395555 10395555 103955555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 103955555 10395555 103955555 103955555 103955555 1039555555 103955555 103955555555 10395555555555 1039555555555555555555555555555555555555</td><td>1029632 103960 109104 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td><td>20 9 100% (-) 100% (-)</td><td>CT TATG TA B B B B F F E</td><td>в</td></t<>	X Image Processing Hit Processing Image Processing C1071 Image Processing 0.15 GCAGT 0.15 GCAGT 0.15 GCAGT 0.15 Total_above 0.10 Table 3 paired 0.10 Table 3 paired 0.10 Table 3 paired 0.10 Table 3 paired 0.20 Table 3 paired 0.21 Cat/Tc 0.22 AT/AT 0.23 GC/OC 0.24 GC/OC 0.27 GT/AC 0.28 GC/OC 0.29 GC/OC	17 >Scaffc 18 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 19 >scaffc 14 4 Page Lagout Page 1 1 1014 above 0 46719 1 1 1014 6814 6333 4472 8600 3822 1993 1044 1815 1781 1993 1048	Id_1 42149507 Id_1 42149507 Id_1 42149507 Situatica 164.76 Situatica 164.76 Situatica 164.76 Situatica 164.76 Situatica 164.76 Loss cattoor/accasto 4002	102043 103925 103925 103925 103925 103925 103925 103925 10392 10392 103925 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 1039555 10395555 1039555 1039555 1039555 1039555 1039555 10395555 1039555 1039555 1039555 1039555 10395555 10395555 10395555 103955555 10395555 10395555 10395555 10395555 10395555 10395555 10395555 103955555 10395555 103955555 103955555 103955555 1039555555 103955555 103955555555 10395555555555 1039555555555555555555555555555555555555	1029632 103960 109104 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 9 100% (-) 100% (-)	CT TATG TA B B B B F F E	в
Canone wide Microad Chris Tobaco di CNTC Zhengshou Tol CNTC Zhengshou Tol 1 199 + 01 - 12 Stalics, 1000 1 25.5 + 10 1 gts.pl was run and results with 2 13569 2 31364 - 2 3 13509 - 3 4 2 31369 5 - 2 31364 5 - 2 31369 4 3509 - 4 5 - 30 - 4 5 - 4 - 359 9 10 - 58 10 - 4 - 4 11 - 4 - 4 12 - 4 - 4 13 - 4 - 4 14 - 4 - 4 15 - 4 - 4 14 - 4 - 4 15 - 7 - 4 16 - 14 - 4 17 - 7 - 4 18 - 6728 - 4	Analysing Toill V1.2 mit Research Cacle/ bacco Research Institute Image: Construct Statute	IX IAI IAI IAI 1000 1000 1000 1000 1000 1000 1000 1000 0.00 1000 1000 1000 0.01 1000 1000 1000 0.01 Table 3 paced 000 1000 0.010 GC/GC 000 000 000 000 0.010 GC/GC 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000	17 >ScaffC 18 >scaffC 19 >scaffC 19 >scaffC 19 >scaffC 19 >scaffC 19 >scaffC 11 ScaffC 12 ScaffC 13 ScaffC 14 ScaffC 15 ScaffC 16 ScaffC 16 ScaffC 170 ScaffC 1872 ScaffC 1872 ScaffC 1872 ScaffC	Id_1 42145057 Id_1 42145057 Id_1 42145057 Id_1 42145057 Id_1 42145057 Id_1 4214507 Id_1 42147 Id_2 44 Id_1 42147 Id_2 44 Id_1 4214507 Id_1 42147 Id_2 54500 Id_1 54500 Id_2 54600 Id_2 54600 Id_1 54600 Id_1 54600 Id_1 54600 Id_2 54600 Id_2 54600 Id_1 54600 Id_2 54600 Id_2 54600 Id_3	Lanisa John San Ranisa John San Pagi Lavari Pagi Lavar	1029632 103960 109104 (IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	200 9 9 100% (-) 10% (в

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.

open access

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)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credite

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		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.