Eco. Env. & Cons. 28 (January Suppl. Issue) : 2022; pp. (S1-S10)
Copyright@ EM International
ISSN 0971–765X

http://doi.org/10.53550/EEC.2022.v28i01s.001

Automation of the process of predicting the SSR as a phase variable within the entire *N. meningitidis* genome

Shaymaa Fouad Rasheed Al Khazraji and Mohammad Abdul Rahmman Al-Maeni

Biology Department, College of Science, University of Baghdad, Baghdad, Iraq

(Received 25 June, 2021; Accepted 27 August, 2021)

ABSTRACT

The abundance of repeat tracts along with genome of commensal and pathogenic *N. meningitidis* encourages us to think about a way of predicting the SSR that leads into phase variable mechanism. This prediction has to be automated using different language skills. Therefore our aim was to automate the process of predicting the SSR that leads into a phase variable mechanism relying on different criteria. These criteria were the length, polymorphic, instability and the value of Z score using a Markov model and synonymous shuffling model and the position of SSR within the gene or promoter. Perl script along with cgi and html was used for this purpose. Our automation program can detect three different categories for the SSR that leads into a phase variation mechanism which is weak, moderate and strong putative phase variable gene. We strongly recommended providing a good evidence for our model as it works correctly using experimental work.

Keywords : *Automationprocess, Simple sequence repeats, Phase variation, N. meningitidis*

Introduction

Generally, *N. meningitidis* colonies the upper respiratory tract with 10-30% being carriage isolates, however, in rare case some strains can evade blood vessels and cause septicemia and meningitis (Stephens, 2009; Caugant and Maiden, 2009). *N. meningitidis* causes septicemia and meningitis with very high rate and despite the presence of antibodies which are effective in clearance of causing disease agents, the commensal remain the main source of infection (Martin *et al.*, 2003; Saunders *et al.*, 2000). The emergence of some strains resist for different vaccines such as conjugate vaccines due to the presence of contingency loci has been proven (Bayliss *et al.*, 2001). The contingency loci are one of the most crucial defence mechanisms in the *N. meningitidis* which is triggers by the action of Localized hypermutation and play an important role in an adaptation of the bacterial cell in their host (Snyder *et al.* 2001; Orsi *et al.*, 2010). Therefore it is necessary to predict if the SSR has the ability to trigger phase variation and alter their repeats in a changeable manner and our goal was to automate the process of prediction for the putative phase variable genes relying on different criteria have been taken from these outhers (Saunders *et al.*, 2000; Martin *et al.*, 2003; Snyder *et al.*, 2001; Li *et al.*, 2004; Hsiang and Kussell, 2011; Orsi *et al.*, 2010; Passel and Ochman, 2007; Janulczyk *et al.*, 2010; ENDE *et al.*, 2000).

Materials and Methods

There were different characteristics that have been taken in the consideration for predicting SSR that leads to phase variation which was Z score calcu-

lated by Markov model, Z score calculated by synonymous shuffling model, the number of polymorphism of repeat tract more than cut off, the stability of repeat tract in 12 strains, the frameshift of repeat tract and position of repeat tract within gene (Saunders *et al.,* 2000; Martin *et al.,* 2003; Snyder *et al.,* 2001; Li *et al.,* 2004; Hsiang and Kussell, 2011; Orsi *et al.,* 2010; Passel and Ochman, 2007; Janulczyk *et al.,* 2010)

For the purpose of automation the process of determining the possibility of SSR that leads to phase variable genes we did the following steps;

Algorithm

All the criteria that have been used to evaluate if a particular SSR in particular gene could be considered as a phase variable gene, were taken to establish algorithm. The algorithm was written as follows

We have 6 vectors ;(A, B, C, D, E,F) ….where

A: is the value of Z score calculated by the Markov model

B: is the value of Z score calculated by a synonymous shuffling model

C: is the number of polymorphism of the repeat tract more than cut off

D: is the stability of the repeat tract in 12 strains

E: is the frameshift of the repeat tract

F: the position of the repeat tract within a gene

A =(a1, a2, a3 ……ai), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

in a1 :

1: over-represented repeat tract (positive value)

0: underrepresented repeat tract

B =(b1,b2,b3 ……bi), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

in b1 :

1: over-represented repeat tract (positive value)

0: underrepresented repeat tract

C=(c1,c2,c3 ……ci), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

Inc1:

1: there is more than one polymorphism (more than cut off) for repeat tract in 500 strains

0: there is no polymorphism (more than cut off) for repeat tract in 500 strains

D= (d1, d2,d3 ……di), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

Ind1:

1: repeat tract is stable within 500 strains

0: repeat tract is unstable within 500 strains

E= (e1,e2,e3 ……ie), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

Ine1:

1: there is frameshift in due to repeat tract

0: there is no frameshift in due to repeat tract

F=(f1,f2,f3 ……fi), n =327 where n is number of putative genes in 12 strains which are selected only on the length of the repeat tract

In f1 :

1: a position of the repeat tract at 5 end

0: a position of the repeat tract at 3 end

Condition

1. If ai,bi,ci,di,ei and fi = 0 then T =0 , i=1.....237, where I repeat tract with each gene
2. If one of them is equal to 1 and other zero then T=1
3. If two of them equal to 1 and other zero then T=2
4. If three of them equal to 1 and other zero then T=3
5. If four of them equal to 1 and other zero then T=4
6. If five of them equal to 1 and other zero then T=5
7. If all of them equal to 1 T=6 or if none of them equal to 1 T=0
8. Convert T to %

if T =< 30% then the gene is a none putative phase variable gene

if 30% < T > 50% then gene predicted as weak putative phase variable gene

if 50% < T > 60% then gene predicted as moderate putative phase variable gene

if T > 60% then gene predicted as strong putative phase variable gene

Statistics

The discriminant test was performed for all putative phase variable genes determined from 12 strains (genes with our cut off homopolymeric of G, C), known phase variable genes from literature as positive control and control gene (genes with SSR less than our cut off G, C less than 5) as negative control.

Programming

1. Perl script was written to count the number of polymorphism and stability for homopolymeric

repeat and other types of repeat tract (the script I, II) respectively (appendix).
2. Perl script was written to count Z score using Markov model for homopolymeric repeat and other types of repeat tract (script, III, V) respectively (appendix).
3. Perl script was written to count Z score using synonymous shaffling model for homopolymeric repeat and other types of repeat tract (script,VI,VII) respectively (appendix).
4. Perl script was written to identify if the change in fameshift due to SSR or indels for homopolymeric repeat and other types of repeat tract (script,VIII,VIIII) respectively (appendix).
5. Perl script was written to identify the location of SSR regarding with -10, -35 patterns for homopolymeric repeat and other types of repeat tract (script,VIIIII,VIIIIII) respectively (appendix).
6. genic.cgi and intergenic.cgi were written for the purpose of the final page loaded when genic and intergenic are selected respectively.
7. Rungenic.pm and Runintergenic.pm scripts run appropriate scripts when user selects genic and intergenic respectively.
7. test2.cgi was written for the purpose of start page. Loads fields for the user to input
8. upload.cgi was written to upload the cgi files into webpage.

Note

All the scripts and through mhogene79@yahoo.com

Results

The number of genes that fit with the cut off which was set for selecting length repeat tract was 327 from 12 strains. Our analysis showed that the genes which scored highly significant T values predicted as strong phase variable genes, were 65 (Table 1 appendix). Moreover, the genes that scored T values above the moderate threshold for phase variable genes were 50 (Table 2 appendix). The genes that scored T values which fit as weak phase variable genes were 15 (Table 3 appendix). Overall, genes were predicted as significantly phase variable genes (moderate and strong) were 133 out of 327. On the other hand, the T value for the control genes (negative control) and the experimental phase variable

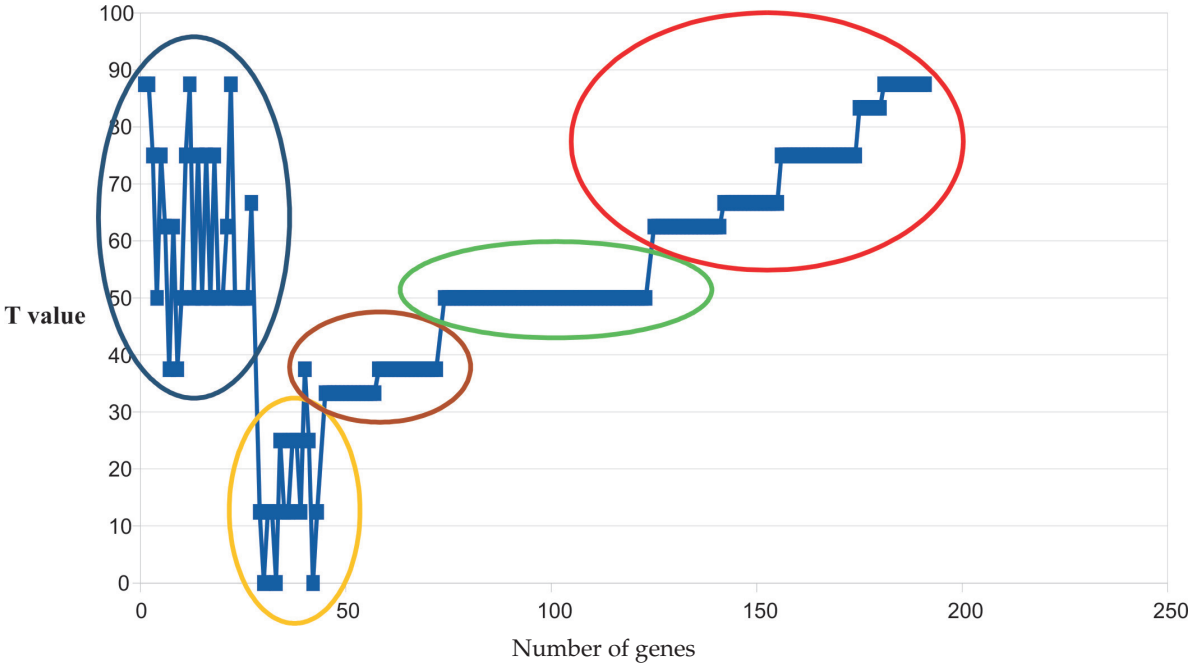

Fig. 1. Schematic representation of scattering T values of the putative phase variable genes that have been predicted from all the 12 strains , control genes (negative control) and experimentally phase variable genes (positive control) which have been selected from the literature. Blue circle: experimentally phase variable genes (positive control), Orange circle:control genes (negative control) , Brown circle: weak putative phase variable genes, Green circle: moderate putative phase variable genes and Red circle: strong putative phase variable genes

genes (positive control) was calculated (Table 4, 5 excel sheet appendix) respectively.

The discriminant test was achieved between all the putative phase variable genes that have been predicted from all the 12 strains and control genes (negative control) represent 65 genes that were selected not at the end of the countig and do not contain G or C repeats with tract lengths suitable as phase variation (less than 5 bps). In addition, experimentally phase variable genes (positive control) which have been selected from the literature. T value was calculated for each putative phase variable gene depending on a scoring of all the criteria that explained previously. T value was plotted for all the putative, control genes and experimentally known phase variable genes by which the control genes were scattered with T value under 30% and experimentally known phase variable genes were scattered with T value over 60 %. Meanwhile, the putative phase variable genes collected from 12 strains were accumulated with T values between (30-70)% as shown in Figure (1).

The result of prediction of putative phase variable genes showed that phase variations occur for all the genes that have an essential function to *N. meningitidis*. Phase variations occurred with genes that have different function as such metabolism, different enzymes, adhesion or synthesis different molecules, addition methyl group, outer membrane protein or process of ATP synthesis, restriction-modification system, production different types of protein such as global stress protein GspA, efflux pump component, haemoglobin receptor, hypothetical protein, pseudogenes, translation and replication process, biosynthesis different metabolic molecules, binding protein, component of pilin, component of antigen, and others.

The former webpage for the prediction program comprises five different inputs (Figure 2). The first input represents fasta of a single gene for scripts I, II and scripts VIII, VIIII while the second input represents single fasta of a whole -genome for scripts II, V. The third input enrolls with multifasta of strains related to the target genes for scripts I, II. The fourth input was designed to enter the type of repeat while the last input was formed for prediction -10 and -35 promoters for scripts VIII, VIII.

In the above example, all the required files were loaded and the genic region was selected, then the submit form icon was pressed; later the second webpage will appear (Figure 3). In this page, we will select all the type of analysis then press T value icon.

Finally, the result will be revealed and for the example above with (T,4) (Figure 4) the result stated the following;

Z score for the Markov model was 1 and a synonymous shuffling model was 1.3 there were two repeats the first one at position 11 and the second one at position 84. Both repeats are found one time, therefore, polymorphism and stability scored zero. The repeat at position 11 has OFF state, therefore, scores 1 while repeat at position 84 has ON the state,

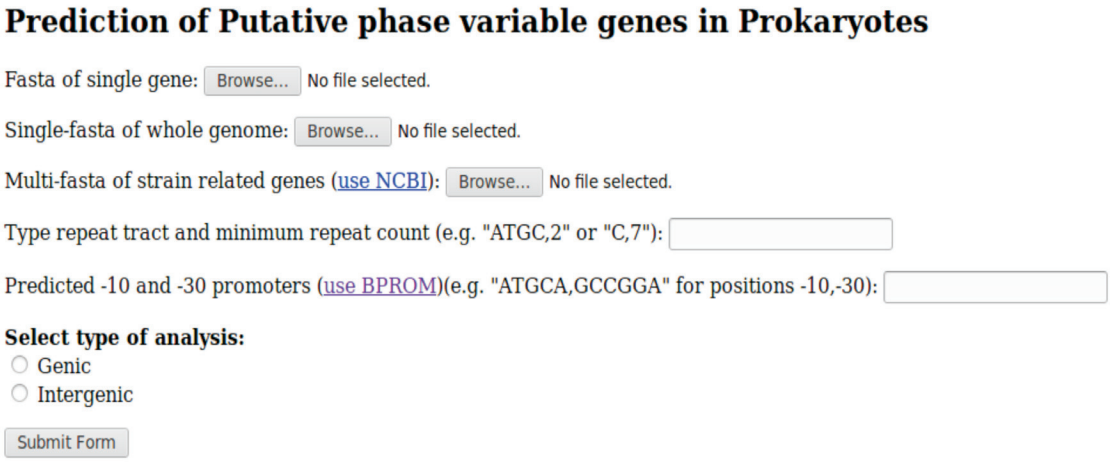

Fig. 2. Schematic representation of prediction of putative phase variable genes in prokaryotes. This page designed to be used to enter different inputs and type of repeat tract for the perdition weak, moderate and strong putative phase variable genes.

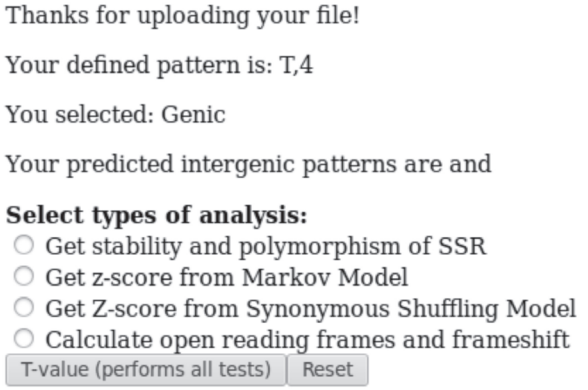

Fig. 3. Schematic representation of types of analysis for the prediction of putative phase variable genes in prokaryotes. This page designed to be used to select type of analysis for the perdition weak, moderate and strong putative phase variable genes.

therefore, scored zero. In summary, the repeat at position 11 has high possibility to generate phase variation than the repeat in position 84.

Discussion

The T value of 65 putative phase variable genes that collected from 12 strains was compatible with experimentally identified phase variable genes. These genes have a category as a strong putative phase variable gene. We recommended strongly doing experimental work for those genes to check whether they really are phase variable genes or there is some bias in our model. Anyway, the automation of this process is powerful to let other people interact and use the programs that enrolled with our model and we made life easy for them.

References

Bayliss, C.D., Field, D. and Moxon, E.R. 2001. The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J. Clin. Invest*. 107 : 657–662. http://dx.doi.org/10.1172/JCI12557.

Caugant, D.A. and Maiden, M.C.J. 2009. Meningococcal carriage and disease - population biology and evolution. *Vaccine*. 27 : B64–B70.

Cholon, D.M., Cutter, D., Richardson, S.K., Sethi, S., Murphy, T.F., Look, D.C. and St Geme, J.W., III. 2008. Serial isolates of persistent *Haemophilus influenzae* in patients with chronic obstructive pulmonary disease express diminish ing quantities of the HMW1 and HMW2 adhesins. *Infect. Immun*. 76 : 4463– 4468. http://dx.doi.org/10.1128/IAI.00499-08.

Granoff, D.M. 2010. Review of meningococcal group B vaccines. *Clin. Infect*. Dis. 50(Suppl 2):S54 –S65. http://dx.doi.org/10.1086/648966.

Janulczyk, R., Masignani, V., Maione, D., Tettelin, H., Grandi, G. and Telford, J.L. 2010. Simple sequence repeats and genome plasticity in *Streptococcus agalactiae*. J *Bacteriol*. 192 : 3990–4000.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. 2004. Microsatellites Within Genes: Structure, Function, and Evolution. *Mol. Biol. Evol*. 21(6) : 991–1007.

Lin, W.H. and Kussell, E. 2012. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Res*. 40 : 2399-2413.

Martin, P., van de Ven, T., Mouchel, N., Jeffries, A. C., Hood, D. W. and Moxon, E. R. 2003. Experimentally revised repertoire of putative contingency loci in *Neisseria meningitidis* strain MC58: evidence for a novel mechanism of phase variation. *Mol. Microbiol*. 50 : 245–257.

Orsi, R.H., Bowen, B.M. and Wiedmann, M. 2010. Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. *Genomics*. 11 : 102.

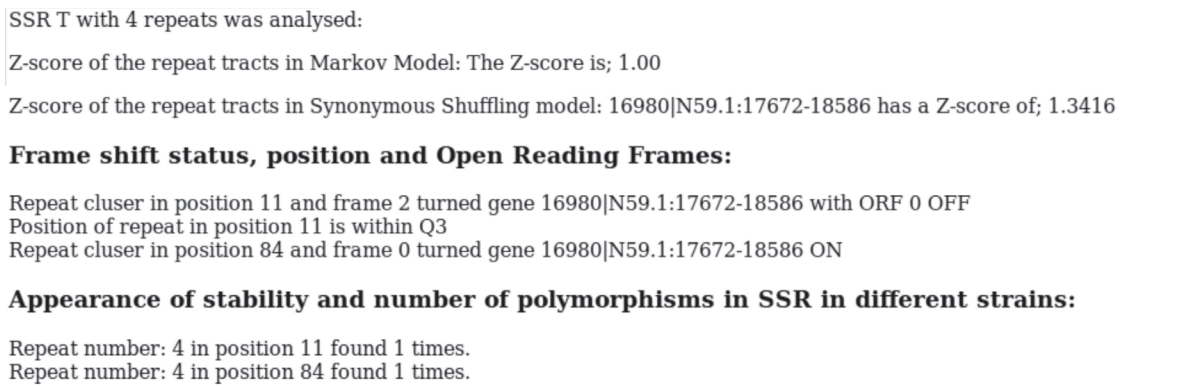

Fig. 4. Schematic representation of an example for the prediction of putative phase variable genes in prokaryotes. This page designed to show the output of different analysis for the perdition weak, moderate and strong putative phase variable genes.

Saunders, N. J., A. C. Jeffries, J. F. Peden, D. W. Hood, H. Tettelin, R. Rappuoli, and Moxon, E. R. 2000. Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol*. 37 : 207–215.

Snyder, L. A., Butcher, S. A. and Saunders, N. J. 2001. Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria spp*. *Microbiology*. 147 : 2321–2332.

Stephens, D.S. 2009. Biology and pathogenesis of the evolutionarily successful, obligate human bacterium *Neisseria meningitidis*. *Vaccine*. 27 : B71–77.

Truck, J. and Pollard, A.J. 2010. Challenges in immunization against bacterial infection in children. *Early Hum. Dev*. 86 : 695–701. http://dx.doi.org/10 .1016/ j.earlhumdev.2010.08.010.

Van der Ende, A., Hopman, C. T. and Dankert, J. 2000. Multiple mechanisms of phase variation of PorA in *Neisseria meningitidis. Infect Immun*. 68 : 6685–6690.

Van Passel, M.W. and Ochman, H. 2007. Selection on the genic location of disruptiveelements. *Trends Genet*. 23 : 601–604.

Appendix

Table 1 : genes that scored T value which can be predicted as strong putative phase varible Add these tables

gene	repeat	Zscore M	polymorphism of repeat tract above the cut off	Frame status	stability status	position	dead gene status	Z score SH	T value
NMB1595	ACGCGC3	0	0	11	1	1	0	1	5
NMB0841	28C7	1	1	11	1	0	0	0	5
ybiP	326G7	1	1	11	1	0	0	0	5
NMB0312	AAGC3	0	1	11	0	1	0	1	5
NMB1507	AAGC3	0	1	11	0	1	0	1	5
pglA	G(7-15)G8	1	1	11	1	0	0	0	5
NMB2032	C(8-15)61C11	0	1	11	1	1	0	0	5
NMB1255	C(7-11)G9	1	0	11	0	1	0	1	5
NMB0624	CAAACAA3	0	1	11	0	1	1	0	5
NMB1893	TTCC3	0	0	11	0	1	1	1	5
NMB1541-1	26C8	1	0	11	1	1	0	0	5
thiD	C(7-8)C7	1	1	0	1	1	1	0	5
NMB0195	GC6	1	0	11	1	1	0	0	5
NMB1818	cggg3	0	0	11	1	1	0	1	5
NMB0516	130A8	0	0	11	1	1	1	0	5
NMB0663 down -35	CTTCT3	0	1	1	1	0	0	1	4
NMB0444 bup -35	CTTCT3	0	1	1	1	0	0	1	4
NMB0993 down -35	236C7	1	1	1	1	0	0	0	4
NMB1052 up -35	153C18	1	1	1	1	0	0	0	4
NMB2142d own -35	59C8	1	1	1	1	0	0	0	4
NMB1543 down -35	30G7	1	1	1	1	0	0	0	4

Table 2 : genes that scored T value which can be predicted as moderate putative phase varible

gene	repeat	Zscore M	polymorphism of repeat tract above the cut off	Frame status	stability status	position	dead gene status	Z score SH	T value	T%
NMB1913	TTCC3	1	0	0	1	1	0	1	4	50
NMB0961	CAAAT3	1	0	0	0	1	1	1	4	50
NMB 2030	GGCGC 3	1	0	0	1	1	0	1	4	50
NMB1693	AC5	1	0	0	1	1	0	1	4	50
NMB0289	GCAG3	1	0	0	1	1	0	1	4	50
NMB0283	GCAG3	1	0	0	1	1	0	1	4	50
NMB1582	GCAG3	1	0	0	1	1	0	1	4	50
NMB1969-2	177C10	0	1	11	1	0	0	0	4	50
N114-01637	263G7	1	1	0	1	0	1	0	4	50
NMB1797	126C8	1	1	0	1	1	0	0	4	50
NMB1882-1	9C8	1	1	0	1	1	0	0	4	50
NMB1931	31G7	1	1	0	1	1	0	0	4	50
NMB0623	129C7	1	1	0	1	1	0	0	4	50
NMB2010-2	345G7	1	0	11	0	1	0	0	4	50
NMC1946	G(7-15)G10	0	1	0	1	1	0	1	4	50
NMB0039	C(7-11)C7	1	1	0	1	1	0	0	4	50
NMB0751	G(7-12)G8	1	1	0	1	0	1	0	4	50
NMB0970	C(7-10)C7	1	1	0	0	1	1	0	4	50
mbollM down -35	AGCC3	1	0	1	0	0	0	1	3	50
NMB1994-bet.(-10-135)	AAAT3	0	1	1	1	0	0	0	3	50
NMB1204 down -35	AC5	1	0	1	0	0	0	1	3	50
N59-01157 bet -10 -35	AT5	1	0	1	1	0	0	0	3	50
intg NMB1786 bet -10 -35	486T8	0	1	1	1	0	0	0	3	50

Table 3 : gene that scored T value which can be predicted as weak putative phase varible

gene	repeat	Zscore M	polymorphism of repeat tract above the cut off	Frame status	stability status	position	dead gene status	Z score SH	T value
N199-01208	GCCAA3	0	1	0	1	0	0	0	2
Intg NMB2114	C(8-10)C8	1	1	0	0	0	1	0	3
NMB0018 down -35	9C9	0	1	1	0	0	0	0	2
intg mfpsA down -35	162C10	0	1	1	0	0	0	0	2
NMB1053 down -35	248C9	0	1	1	0	0	0	0	2
NMB1988 down -35	333C9	0	1	1	0	0	0	0	2
NMB1969 down -35	228C10	0	1	1	0	0	0	0	2
lbpA | NMB1540	G8	1	0	1	0	0	0	0	2
N258.01007 up -35	GCAG3	1	0	1	0	0	0	0	2
intg 1719 up -35	340A8	0	1	1	0	0	0	0	2
NMB1508 up -35	AAGC3	0	1	1	0	0	0	0	2
N114-01031 down -35	GCCAA3	0	1	1	0	0	0	0	2
Intg NMB2036	C(7-14)C9	1	1	0	0	0	0	0	2
NMB0441 down -35	GAAC3	1	0	1	0	0	0	0	2
N73-00567	AGCC3	1	0	0	1	0	0	1	3
N114 01371	AAGC3	0	0	0	1	1	1	0	3
NMB111	GGCGC 3	1	0	0	0	1	0	1	3
NMB0468	TGTTT3	0	0	0	1	1	0	1	3
NMB0019	CGGTGG3	0	0	11	0	1	0	0	3
N199-00635	178C9	0	1	0	1	1	0	0	3
N73-01693	no matchC	0	1	0	1	0	1	0	3
NMB1836	241C9	0	1	0	1	1	0	0	3
NMB1443	no match	0	1	0	1	1	0	0	3
NMB0218	400C8	1	1	0	1	0	0	0	3
NMB0040	52C7	1	1	0	1	0	0	0	3

Table 4 : T value of control gene that selected with SSR less than our cut off

repeat tract	Zscore M	polymorphism of repeat tract above the cut off	Frame status	stability status	position	dead gene status	Zscore SH	T value	%100
C or G4	0	0	0	1	0	0	0	1	12.5
	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	1	12.5
	0	0	0	0	1	0	0	1	12.5
	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	0	0	2	25
	0	0	0	1	0	0	0	1	12.5
	0	0	0	0	1	0	0	1	12.5
	0	0	0	1	1	0	0	2	25
	0	0	0	1	1	0	0	2	25
	0	0	0	0	1	0	0	1	12.5
	0	1	0	1	1	0	0	3	37.5
	0	0	0	1	1	0	0	2	25
	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	1	12.5
	0	0	0	1	1	0	0	2	25
	0	0	0	0	1	0	0	1	12.5

Table 5 : T value of experimentally known phase variable genes that selected from literatures

genes	repeat tract	Zscore M	polymorphism of repeat tract above the cut off	Frame status	stability status	position	dead gene status	Zscore SH	T value	%100
pilC	G(7-15)G9	1	1	11	1	1	0	1	7	87.5
NMB1847	G(7-15)60G10	0	1	11	1	1	1	1	7	87.5
porA | NMB1429	G(13-17)G14	0	1	11	1	1	0	1	6	75
NMB1998-1	no match	1	0	11	0	1	0	0	4	50
NMB1668 -1	217C9	0	1	11	1	1	1	0	6	75
lbpA | NMB1540	G8	1	0	11	1	1	0	0	5	62.5
NMB1836	241C9	0	1	0	1	1	0	0	3	37.5
NMB2032	C(8-15)61C11	0	1	11	1	1	0	0	5	62.5
NMB-0218	400C8	1	1	0	1	0	0	0	3	37.5
NMB1969	177C10	0	1	0	1	1	1	0	4	50
NMB0831	G(7-11)G7	1	1	11	1	1	0	0	6	75
NMB0098	C(7-8)C8	1	1	11	1	1	1	0	7	87.5
nifS NMB1379	C8	1	0	11	0	1	0	0	4	50
NMB0415	G(7-12)G8	1	1	11	1	1	0	0	6	75
NMB0970	C(7-10)C7	1	1	0	0	1	1	0	4	50
NMB1892	(C7-8)C7	1	1	11	0	1	1	0	6	75
NMB0067	372C7	1	0	0	1	1	1	0	4	50
NMB1375	AGCC3	1	1	11	0	1	0	1	6	75
NMB1261	CCCAA3	0	1	0	0	1	1	1	4	50
NMB0961	CAAAT3	1	0	0	0	1	1	1	4	50
NMB1893	TTCC3	0	0	11	0	1	1	1	5	62.5
NMB1489	C(7)	1	1	11	1	1	1	0	7	87.5
NMB0368	213A8	0	0	11	1	1	0	0	4	50
NMB1931	31G7	1	1	0	1	1	0	0	4	50