detect outlier loci

Arlequin ,Bayescan and PAML

Content

- Arlequin

-Fst

-Outlier loci

-manual

- Bayescan

F-dist

——use software Arlequin to analyze outlier

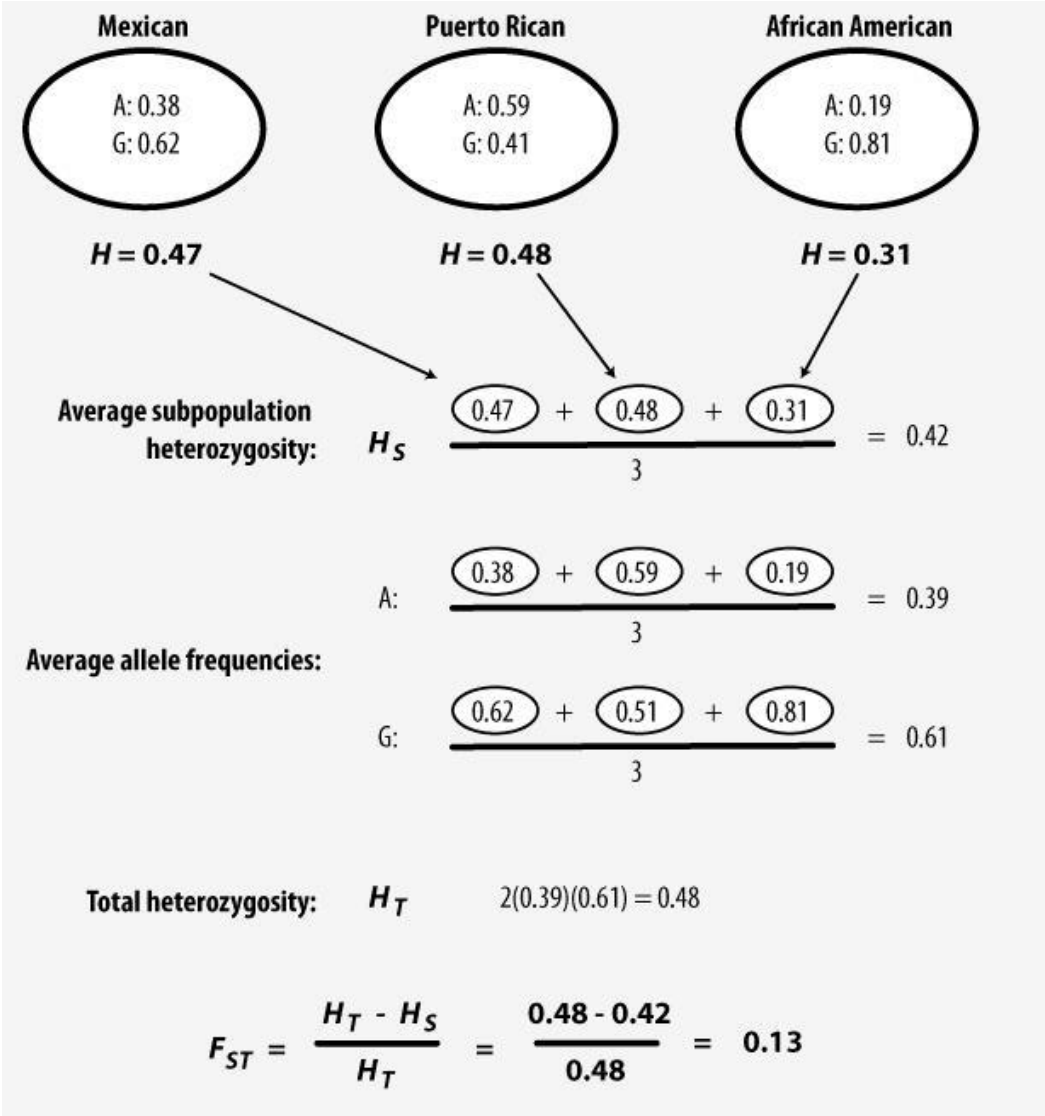
-Fst(Fixation index)

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

π_{Between} 表示不同群体间核苷酸多样性， π_{Within} 表示整个群体中核苷酸多样性

$$H_1 = \sum_{k=1}^{s} w_k h_k, \quad H_S = 1 - \sum_{i=1}^{2} \sum_{k=1}^{s} w_k q_{k(i)}^2, \quad H_T = 1 - \sum_{i=1}^{2} \bar{q}_i^2$$

$$F_{ST} = \frac{H_T - H_S}{H_T}$$



FST<0.05: no evident differentiation

0.05 ~ 0.15, slight differentiation;

0.15~0.25, evident differentiation; ;

>0.25 great differentiation。

Outlier loci

——use software Arlequin to analyze outlier

Definition of outlier loci（离群位点）

- definition： An outlier locus is one that has a distinct or significant allele frequency relative to assumption of neutrality (neutrality being the absence of directional selection).

- In software , some parameters used to evaluated outliers.

What caused appearance of outlier loci?

- Natural selection

  Natural selection is the differential survival and reproduction of individuals due to differences in phenotype.

Manuel

——use software Arlequin to analyze outlier

The birth of f-dist2

Beaumont and Nichols. Evaluating loci for use in the genetic analysis of population structure. (1996) Proc Roy. Soc. Lond. B. 263: 1619-1626

Evaluating loci for use in the genetic analysis of population structure

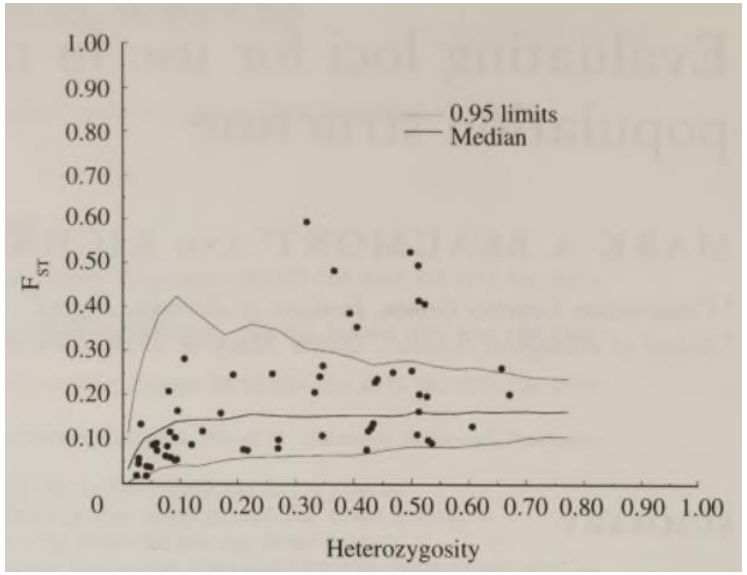
MARK A. BEAUMONT[1] AND RICHARD A. NICHOLS[2]

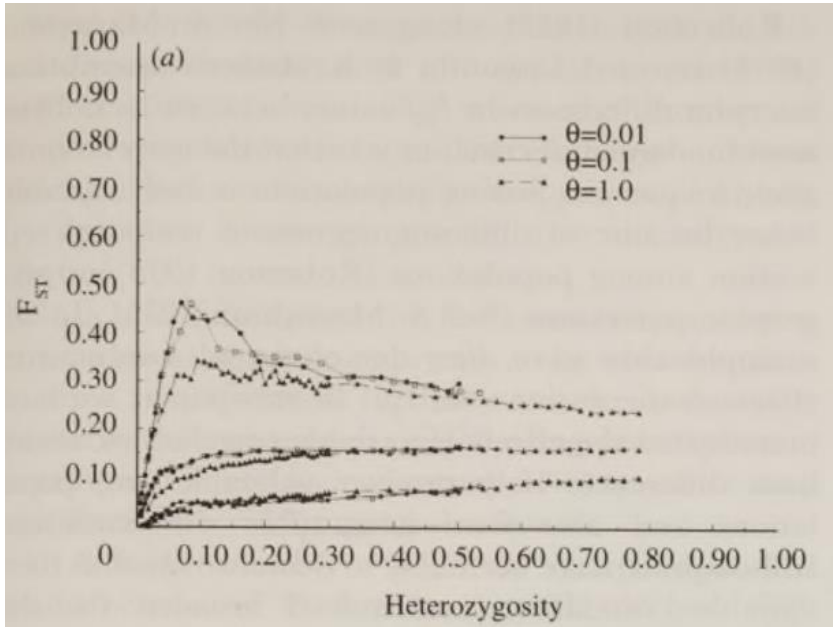
[1] *Conservation Genetics Group, Institute of Zoology, Regent's Park, London NW1 4RY, U.K.*
[2] *School of Biological Sciences, Queen Mary & Westfield College, Mile End Road, London E1 4NS, U.K.*

SUMMARY

Loci that show unusually low or high levels of genetic differentiation are often assumed to be subject to natural selection. We propose a method for the identification of loci showing such disparities. The differentiation can be quantified using the statistic F_{ST}. For a range of population structures and demographic histories, the distribution of F_{ST} is strongly related to the heterozygosity at a locus.

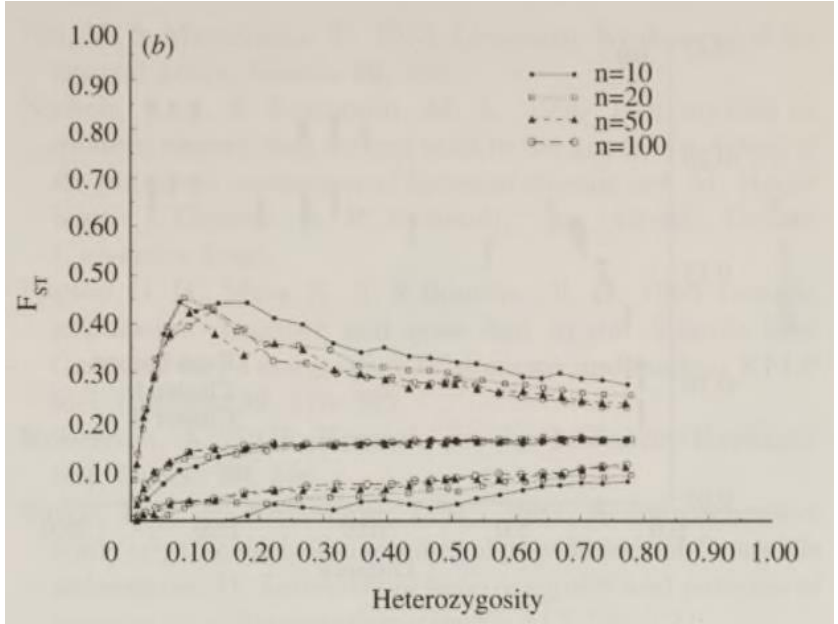
Outlying values of F_{ST} can be identified in a plot of F_{ST} *vs.* heterozygosity using a null distribution generated by a simple genetic model. We use published data-sets to illustrate the importance of the relationship with heterozygosity. We investigate a number of models of population structure, and demonstrate that the null distribution is robust to a wide range of conditions. In particular, the distribution is robust to differing mutation rates, and therefore different molecular markers, such as allozymes, restriction fragment length polymorphisms (RFLPS) and single strand conformation polymorphisms (SSCPS) can be compared together. We suggest that genetic variation at a discrepant locus, identified under these conditions, is likely to have been influenced by natural selection, either acting on the locus itself or at a closely linked locus.





The effect of different mutation rates on the expected distribution of FST.

the distribution of Fst depends quite strongly on the observed heterozygosity. The median value of FST drops off rapidly at heterozygosities less than 0.1. With higher mutation rates, at lower heterozygosities, the distribution is tighter and the median drops off faster.



The effect of equilibrium versus non-equilibrium population structure.

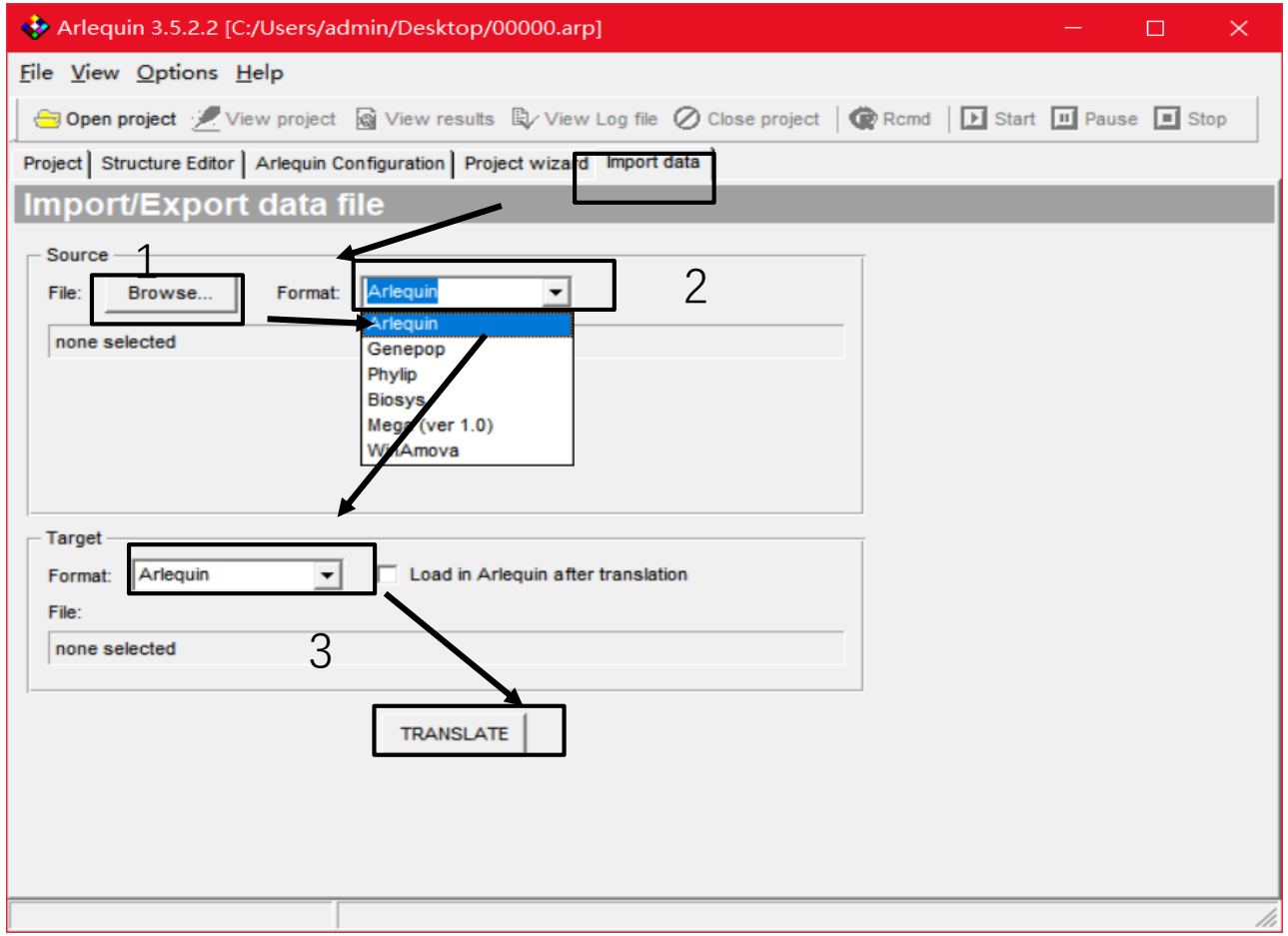
when very small samples are taken from each subpopulation, the distributions are broader. Even moderate sample sizes are surprisingly informative; the distribution for 50 is virtually indistinguishable from that of 100.

Arlequin

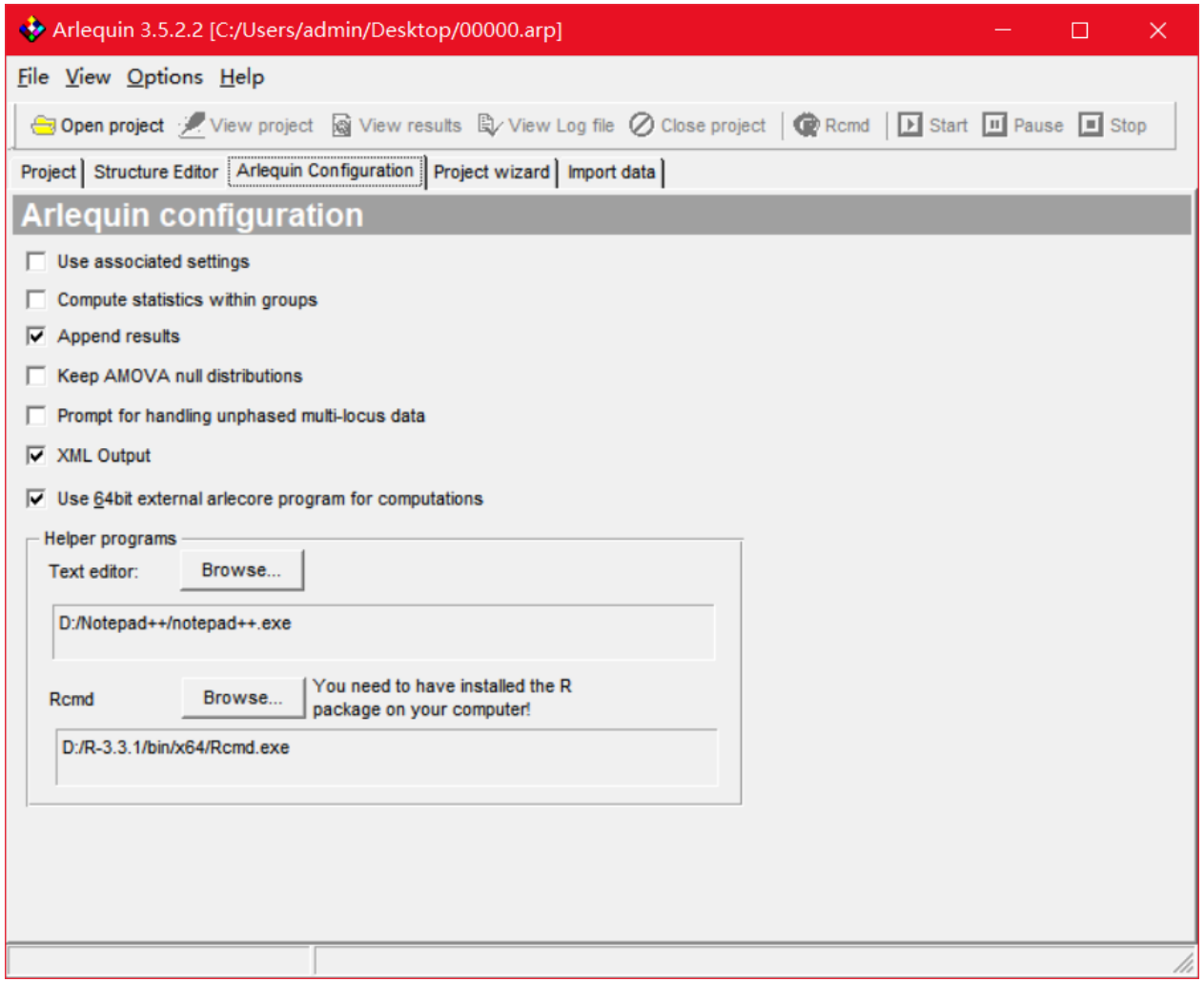
- **Arlequin** is a software that have a multi-function , we use one of its function(a **f-dist2** program) to calculate our data for indication of outliers loci.

Prepare

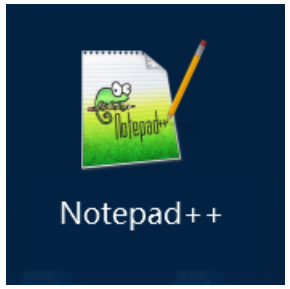
- Format conversion （must be extension of arp）



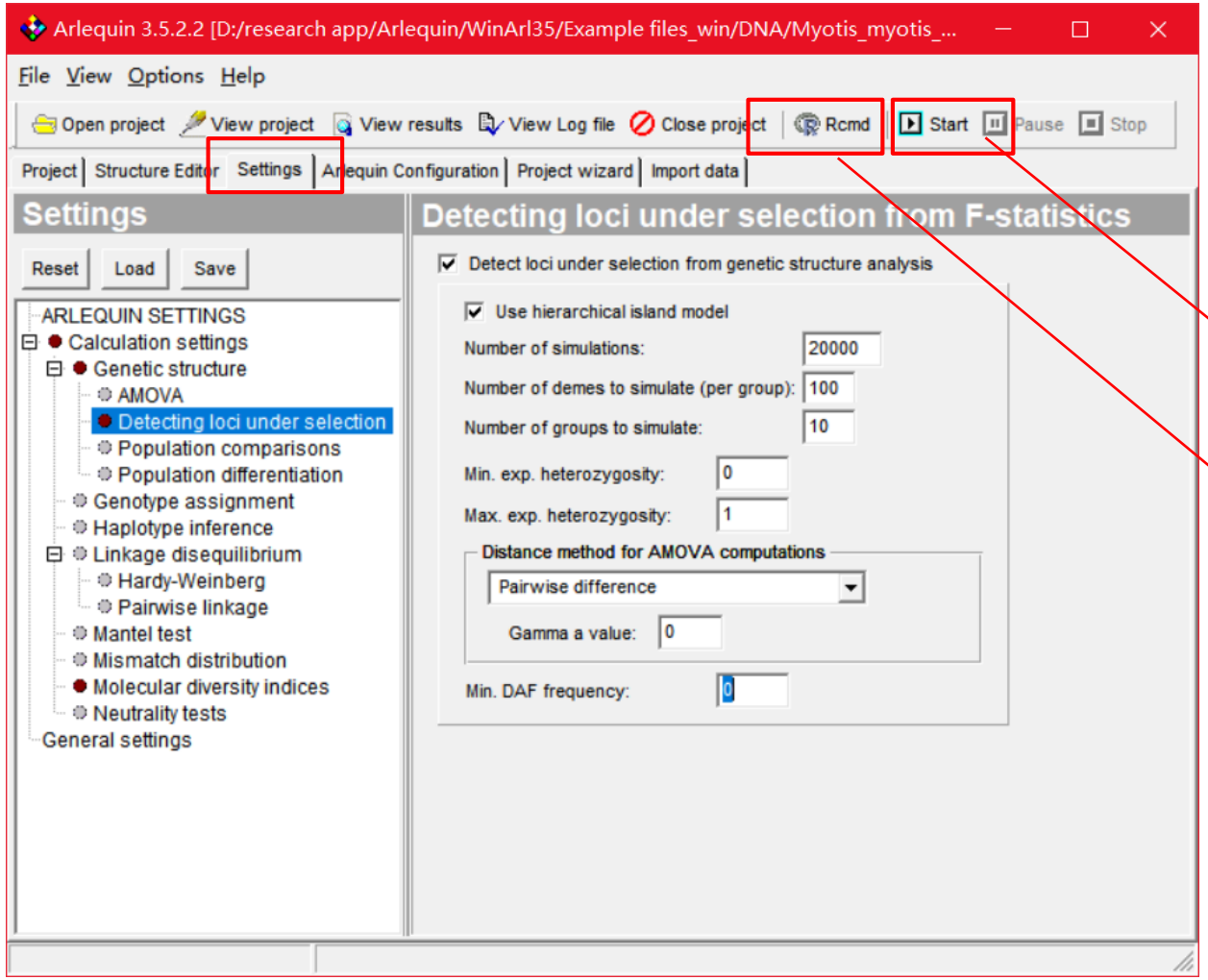
- 1.Choose your file.
- 2.Decide your file format , generally are DNA.
- 3.this format set as Arlequin.



select the path of text editor ,notepad++ for example .



path of R package,so you can get graph from your statistics with the help of Rcmd



checking the option "**Detect loci under selection from genetic structure analysis**" in the Detect loci under selection from F-statistic tab,

start Computation

report graphs

the p-values of each locus under neutrality and for a given genetic structure are output in a file called "**fdist2_ObsOut.txt**".

fdist2_ObsOut.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Locus	Obs. Het. BP	Obs FST	FST P-value	1-FST quantile	(if P-value=2 -> uncompu
1	0	0	-1	-1	
2	0	0	-1	-1	
3	0	0	-1	-1	
4	0	0	-1	-1	
5	0	0	-1	-1	
6	0	0	-1	-1	
7	0	0	-1	-1	
8	0	0	-1	-1	
9	0	0	-1	-1	
10	0.049365942	0.26257846	0.19363346	0.80636654	
11	0	0	-1	-1	
12	0	0	-1	-1	
13	1	1	1e-007	0	
14	1	1	1e-007	0	
15	0	0	-1	-1	
16	0	0	-1	-1	
17	0.21141304	0.58093176	0.039911138	0.96008886	
18	0	0	-1	-1	
19	0	0	-1	-1	
20	0	0	-1	-1	
21	0	0	-1	-1	
22	0	0	-1	-1	
23	0	0	-1	-1	
24	0	0	-1	-1	
25	0	0	-1	-1	
26	0	0	-1	-1	
27	0	0	-1	-1	
28	0	0	-1	-1	
29	0	0	-1	-1	
30	0	0	-1	-1	
31	0	0	-1	-1	
32	0	0	-1	-1	
33	0	0	-1	-1	
34	0	0	-1	-1	
35	0	0	-1	-1	
36	0	0	-1	-1	
37	0	0	-1	-1	
38	0	0	-1	-1	
39	0	0	-1	-1	

For each locus, we report：

i) the observed heterozygosity between population,

ii) the observed FST value

iii) the FST p-value

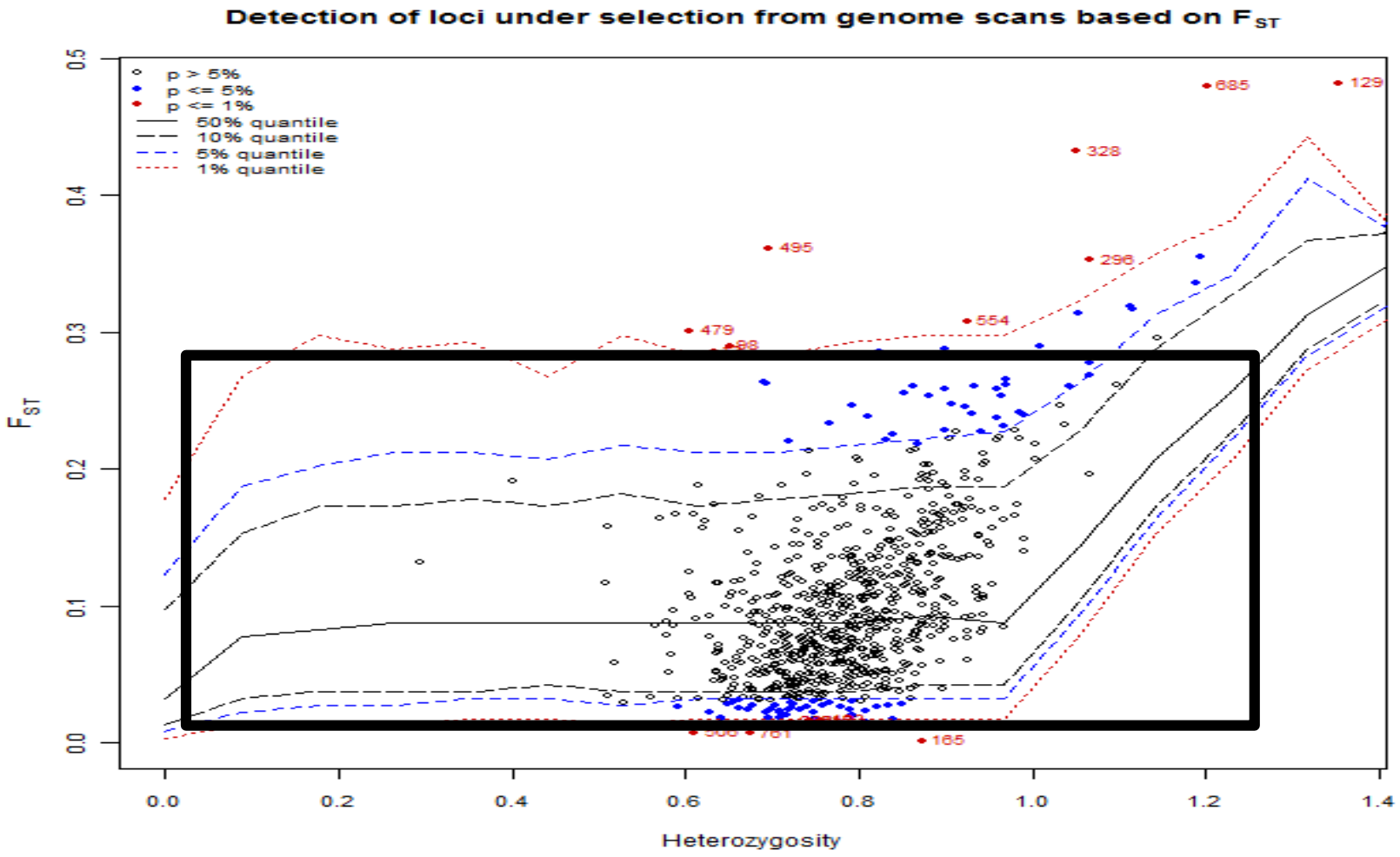
iv) 1 - the quantile of the observed FST in the distribution.

-P-value

the p-value or probability value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be greater than or equal to the actual observed results.

usually ,set p=0.05(5%) as the threshold

Fst-Heterozygosity figure



In data result , the points between two curve of the same confidence is consider as neutral.

Bayecsan

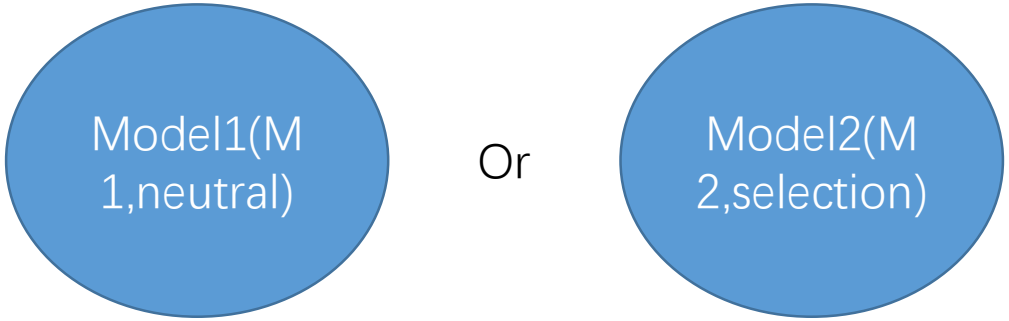
——A software

Introduction

BayeScan implements a reversible-jump MCMC algorithm for calculation.

A choice

- Model choice



- Which model locus prefer?

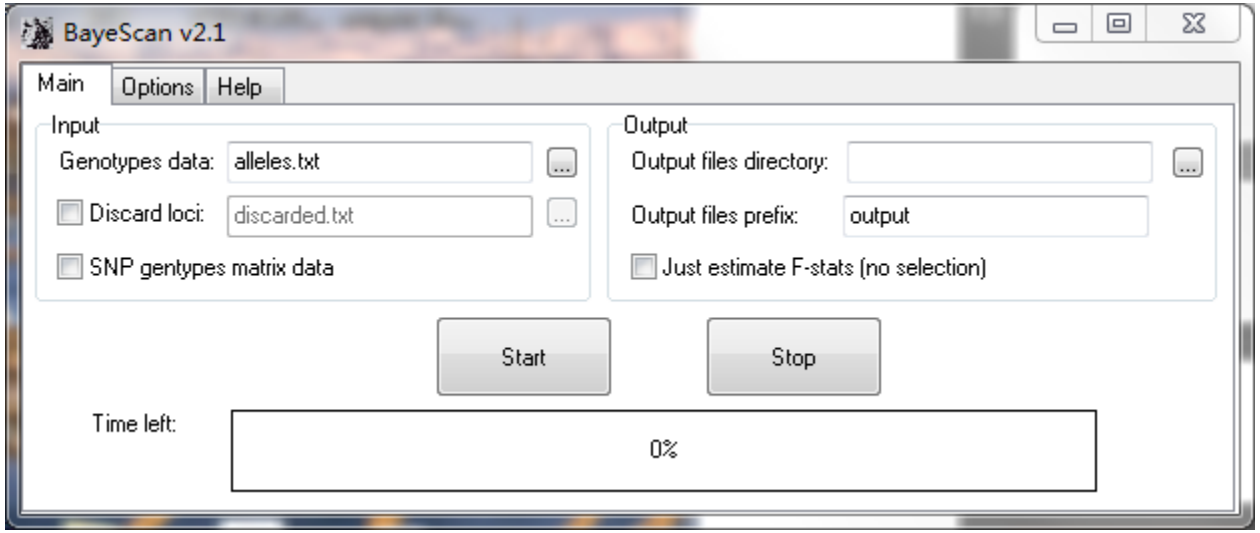
Jeffreys' scale of evidence for the choice

P(α≠0)	Bayes Factor (BF)	log10(BF)	Jeffreys' interpretation
0.50 → 0.76	1 → 3	0 → 0.5	Barely worth mentioning
0.76 → 0.91	3 → 10	0.5 → 1	Substantial
0.91 → 0.97	10 → 32	1 → 1.5	Strong
0.97 → 0.99	32 → 100	1.5 → 2	Very strong
0.99 → 1.00	100 → ∞	2 → ∞	Decisive

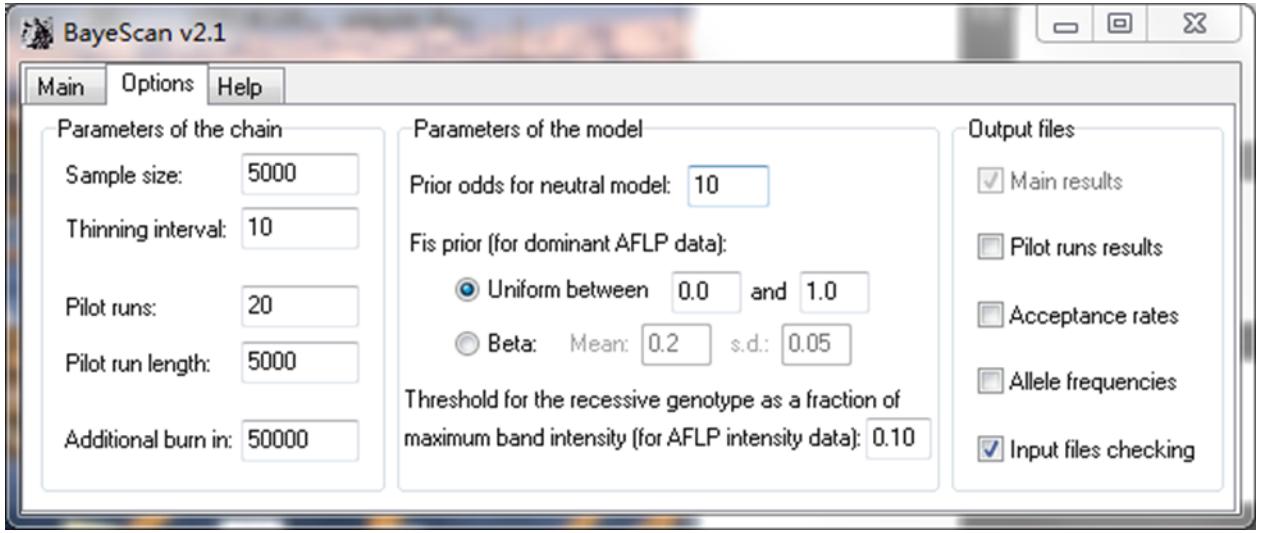
- Bayes factor(BF): The Bayes factor provides a scale of evidence in favor of one model versus another.
  BF=P(N|M2)/P(N|M1)
- Bayes factor also be a parameter for outliers evidence.
- For example, BF=2 indicates that the data favors model M2 over model M1.

How to use Bayescan?

- Download free in http://cmpg.unibe.ch/software/BayeScan/
- Process : Pilot runs(long time) ⟶ Calculation(relative short)
- Software interface

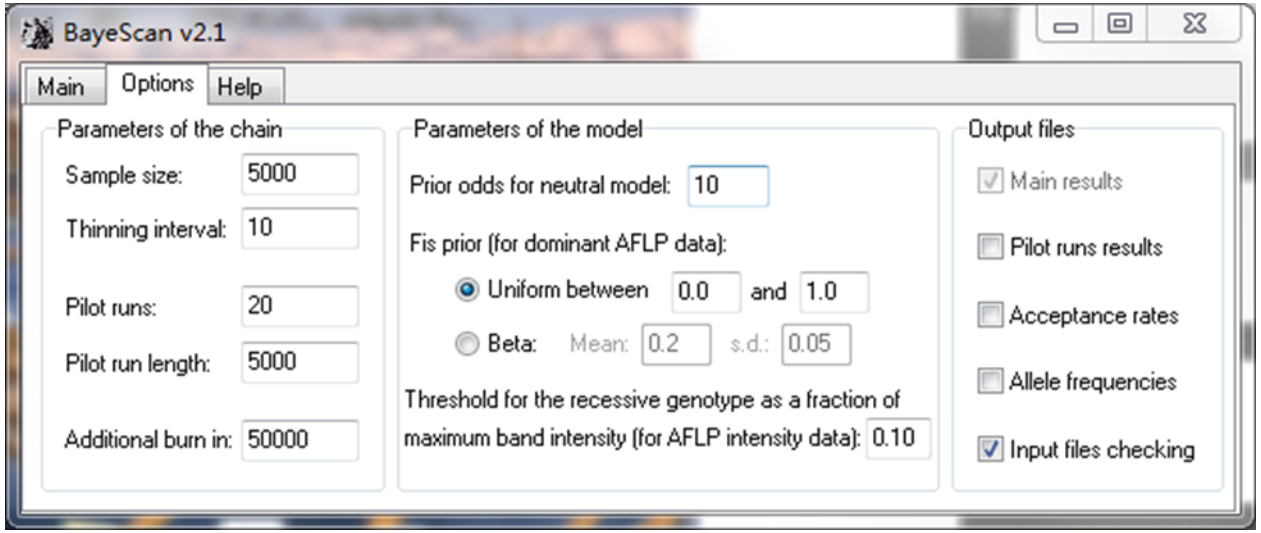


Setting parameters



- Samples size（样本大小）： We need set according to our number of samples.
- Thinning interval： The thinning interval is the number of iterations(迭代） between two samples.
- Burn in： A burn-in period can be necessary to attain convergence before starting the sampling ,default is 50000.
- Prior odds for neutral model(中性)： default is 10(you can use it),you can set this parameter based on your need. In some papers, author set four different prior odds to quantify how this parameter affect their result.

Setting parameters

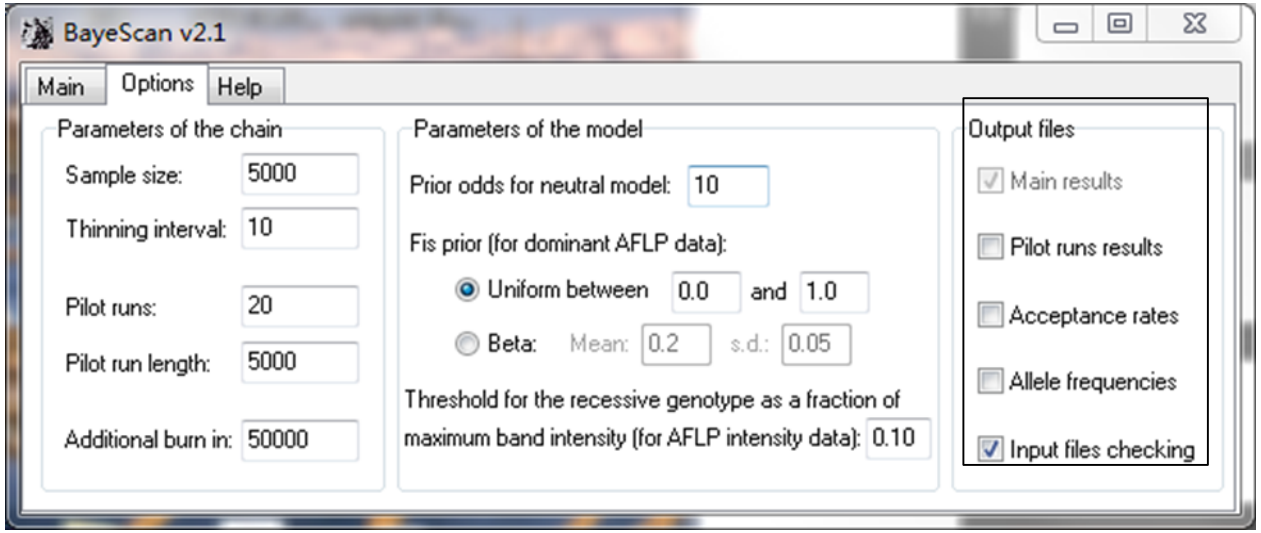


- Pilot runs（试运行）： We make by default 20. pilot run length : default 500.
- Function of pilot runs : Choose the **proposal distribution** for the reversible jump and adjusted the **acceptance rate** for each parameters.
- Proposal distribution（建议分布）: Proposal distributions have to be adjusted in order to have acceptance rates between 0.25 and 0.45. These values are automatically tuned on the basis of short successive pilot runs .

Setting suggestion

- Pilots runs : When calculation time is not a problem, increasing the number of pilot runs would be the first thing to do.
- Thinning interval : Increasing the sample size is generally useless, and one should rather increase thinning interval.

Output files choose



- Main results
- Pilot runs results
- Acceptance rates
- Allele frequencies
- Input files checking

Acceptable data types

- *Amplification intensity matrix for AFLP markers*
- *Dominant binary markers*
- *Codominant markers*
- *SNP genotype matrix*

SNP genotype(单核苷酸多态性基因分型)

- Source data example

The first 1 to 30 is meaning the samples of first population， and the second meaning second population， and maybe have many populations.

SNP genotype data

1	1	2	0	2	0	2	1	0	0	1	2	0	0	2	2	1	0	2	2	2	2
2	1	2	0	1	0	2	1	0	0	1	2	0	0	2	2	0	0	2	2	2	2
3	1	2	0	2	0	2	2	0	1	2	2	0	0	2	2	1	0	2	2	2	1
4	1	2	0	2	0	2	1	0	0	2	2	0	0	1	2	2	0	2	2	0	1
5	1	2	0	1	0	2	2	0	2	2	2	0	0	2	2	2	0	2	2	1	1
6	1	1	0	2	0	2	1	0	2	2	2	0	0	2	2	2	0	2	2	2	2
7	1	2	1	1	0	2	0	0	2	2	2	0	0	2	2	1	0	2	1	0	2
8	1	2	0	1	0	2	1	0	1	1	2	0	0	2	2	2	0	2	2	1	1
9	1	2	0	1	0	2	0	0	1	2	2	0	0	2	2	1	0	2	2	2	2
10	1	2	0	1	0	2	1	0	2	1	2	0	0	2	2	2	0	2	1	2	2
11	1	1	0	2	0	2	2	0	1	2	2	0	0	1	2	2	0	2	2	2	2
12	1	1	1	2	0	2	1	0	1	0	2	0	0	2	2	2	0	2	2	2	1
13	1	2	0	1	0	2	2	0	1	1	2	0	0	2	2	1	0	2	2	1	2
14	1	1	0	1	0	2	0	0	0	1	2	0	0	2	2	2	0	2	2	1	2
15	1	2	0	1	0	2	1	0	1	2	2	0	0	2	2	2	0	2	1	1	2
16	1	1	0	2	0	2	0	0	1	2	2	0	0	2	2	2	0	2	2	2	2
17	1	2	0	2	0	2	0	0	1	1	2	0	0	1	2	2	0	2	0	1	1
18	1	0	0	1	0	2	1	0	1	2	2	0	0	2	2	1	0	2	2	1	2
19	1	2	0	2	0	2	0	0	0	2	2	0	0	2	2	2	0	2	2	0	2
20	1	2	0	2	0	2	1	0	2	2	2	0	0	2	2	2	0	2	2	1	2
21	1	0	0	2	0	2	0	0	1	1	2	0	0	2	2	2	0	2	2	1	2
22	1	1	0	1	0	2	0	0	1	1	2	0	0	1	2	2	0	2	2	1	2
23	1	0	0	0	0	2	0	0	1	2	2	0	0	2	2	2	0	2	2	1	2
24	1	2	0	2	0	2	2	0	1	2	2	0	0	2	2	2	0	2	2	1	1
25	1	0	0	2	0	2	1	0	1	2	2	0	0	2	2	0	0	2	2	1	2
26	1	1	0	2	0	2	1	0	2	2	2	0	0	2	2	2	0	2	2	2	2
27	1	2	0	1	0	2	1	0	2	1	2	0	0	2	2	2	0	2	1	2	1
28	1	2	0	1	0	2	0	0	1	1	2	0	0	1	2	0	0	2	2	2	2
29	1	1	0	1	0	2	1	0	1	1	2	0	0	2	2	1	0	2	2	1	2
30	1	2	0	2	0	2	1	0	1	2	2	0	0	1	2	0	0	2	2	1	2
1	2	1	0	2	0	2	1	0	2	2	2	1	0	1	2	1	0	2	2	0	2
2	2	1	0	2	0	2	1	0	2	1	2	2	0	0	2	0	1	2	2	1	2
3	2	0	0	1	0	2	0	0	2	1	2	1	0	1	1	1	0	2	2	0	2
4	2	1	0	0	0	2	1	0	2	2	2	1	0	1	2	0	0	2	2	0	2
5	2	0	0	2	0	2	0	0	2	2	2	0	0	0	2	0	0	2	1	1	2
6	2	1	1	2	0	2	1	0	2	2	2	0	0	0	1	0	0	2	2	2	2
7	2	0	0	1	0	2	0	0	2	1	2	0	0	0	2	0	0	2	2	2	2
8	2	0	0	1	0	2	1	0	2	1	2	0	0	2	2	1	0	2	2	2	2
9	2	0	0	1	0	2	2	0	2	1	2	1	0	1	2	0	0	2	2	1	2
10	2	1	0	1	0	2	0	0	2	2	2	0	0	0	2	0	0	2	2	0	2

Output files format

Parameters

- Prob
- Log10(PO)
- Qval
- Alpha
- Fst

Locus

	prob	log10(PO)	qval	alpha	fst
1	0.0267559	-1.5608	0.871964	-0.0023432	0.19128
2	0.046823	-1.3087	0.85583	0.021966	0.19570
3	0.063545	-1.1684	0.82529	0.025038	0.19630
4	0.14381	-0.77477	0.74237	-0.13403	0.17785
5	0.17057	-0.68688	0.71224	-0.24207	0.17220
6	0.050167	-1.2772	0.84967	-0.0017856	0.19160
7	0.14047	-0.78668	0.74603	-0.11568	0.18107
8	0.053512	-1.2477	0.84143	0.015627	0.19509
9	0.21405	-0.56489	0.66346	-0.19938	0.17088
10	0.21405	-0.56489	0.66346	-0.26810	0.16958
11	0.25753	-0.45986	0.52365	0.20244	0.23109
12	0.18395	-0.64703	0.69597	-0.20868	0.17417
13	0.056856	-1.2198	0.83534	0.027735	0.19692
14	0.22742	-0.53110	0.62347	-0.30716	0.16422
15	0.046823	-1.3087	0.85583	0.023728	0.19619
16	0.19732	-0.60936	0.68997	-0.25624	0.16793
17	0.050167	-1.2772	0.84967	-0.0020146	0.19173
18	0.093645	-0.98581	0.79612	-0.046816	0.18675
19	0.046823	-1.3087	0.85583	-0.00012744	0.19173
20	0.21739	-0.55630	0.64596	0.21543	0.23544
21	0.063545	-1.1684	0.82529	0.0036964	0.19281
22	0.20067	-0.60025	0.68403	0.14675	0.22020
23	0.28094	-0.40816	0.48718	-0.39857	0.16018
24	0.073579	-1.1001	0.80955	0.021770	0.19548
25	0.060201	-1.1934	0.83213	0.0026949	0.19255
26	0.22074	-0.54781	0.63545	-0.28405	0.16673
27	0.040134	-1.3787	0.86575	0.0072718	0.19280
28	0.093645	-0.98581	0.79612	-0.040213	0.18765
29	0.16722	-0.69723	0.71706	0.11529	0.21396
30	0.040134	-1.3787	0.86575	-0.0057520	0.19103
31	0.066890	-1.1446	0.81973	-0.051574	0.18606
32	0.090301	-1.0032	0.79839	-0.038423	0.18727
33	0.10368	-0.93677	0.77831	-0.056977	0.18668
34	0.10368	-0.93677	0.77831	0.053242	0.20215

Evidence for outlier——log10（PO）

	prob	log10(PO)	qval	alpha	fst
1	0.0267559	-1.5608	0.871964	-0.0023432	0.19128
2	0.046823	-1.3087	0.85583	0.021966	0.19570
3	0.063545	-1.1684	0.82529	0.025038	0.19630
4	0.14381	-0.77477	0.74237	-0.13403	0.17785
5	0.17057	-0.68688	0.71224	-0.24207	0.17220
6	0.050167	-1.2772	0.84967	-0.0017856	0.19160
7	0.14047	-0.78668	0.74603	-0.11568	0.18107
8	0.053512	-1.2477	0.84143	0.015627	0.19509
9	0.21405	-0.56489	0.66346	-0.19938	0.17088
10	0.21405	-0.56489	0.66346	-0.26810	0.16958
11	0.25753	-0.45986	0.52365	0.20244	0.23109
12	0.18395	-0.64703	0.69597	-0.20868	0.17417
13	0.056856	-1.2198	0.83534	0.027735	0.19692
14	0.22742	-0.53110	0.62347	-0.30716	0.16422
15	0.046823	-1.3087	0.85583	0.023728	0.19619
16	0.19732	-0.60936	0.68997	-0.25624	0.16793
17	0.050167	-1.2772	0.84967	-0.0020146	0.19173
18	0.093645	-0.98581	0.79612	-0.046816	0.18675
19	0.046823	-1.3087	0.85583	-0.00012744	0.19173
20	0.21739	-0.55630	0.64596	0.21543	0.23544
21	0.063545	-1.1684	0.82529	0.0036964	0.19281
22	0.20067	-0.60025	0.68403	0.14675	0.22020
23	0.28094	-0.40816	0.48718	-0.39857	0.16018
24	0.073579	-1.1001	0.80955	0.021770	0.19548
25	0.060201	-1.1934	0.83213	0.0026949	0.19255
26	0.22074	-0.54781	0.63545	-0.28405	0.16673
27	0.040134	-1.3787	0.86575	0.0072718	0.19280
28	0.093645	-0.98581	0.79612	-0.040213	0.18765
29	0.16722	-0.69723	0.71706	0.11529	0.21396
30	0.040134	-1.3787	0.86575	-0.0057520	0.19103
31	0.066890	-1.1446	0.81973	-0.051574	0.18606
32	0.090301	-1.0032	0.79839	-0.038423	0.18727
33	0.10368	-0.93677	0.77831	-0.056977	0.18668
34	0.10368	-0.93677	0.77831	0.053242	0.20215

P(α≠0)	Bayes Factor (BF)	log10(BF)	Jeffreys' interpretation
0.50 → 0.76	1 → 3	0 → 0.5	Barely worth mentioning
0.76 → 0.91	3 → 10	0.5 → 1	Substantial
0.91 → 0.97	10 → 32	1 → 1.5	Strong
0.97 → 0.99	32 → 100	1.5 → 2	Very strong
0.99 → 1.00	100 → ∞	2 → ∞	Decisive

Jeffreys' scale of evidence

- Log10（PO），PO is meaning posterior odds（different from posterior probabilities）.
- As a result, a Bayes factor of 3 corresponding to a posterior probability of 0.76, is already considered as being a “substantial” evidence for selection, it was also considered evidence for **outlier behaviour** .
- **In the output data , the parameter of log10(PO) also regard as log10(BF)(said in manual).**

Other parameters

	prob	log10(P0)	qval	alpha	fst
1	0.0267559	-1.5608	0.871964	-0.0023432	0.19128
2	0.046823	-1.3087	0.85583	0.021966	0.19570
3	0.063545	-1.1684	0.82529	0.025038	0.19630
4	0.14381	-0.77477	0.74237	-0.13403	0.17785
5	0.17057	-0.68688	0.71224	-0.24207	0.17220
6	0.050167	-1.2772	0.84967	-0.0017856	0.19160
7	0.14047	-0.78668	0.74603	-0.11568	0.18107
8	0.053512	-1.2477	0.84143	0.015627	0.19509
9	0.21405	-0.56489	0.66346	-0.19938	0.17088
10	0.21405	-0.56489	0.66346	-0.26810	0.16958
11	0.25753	-0.45986	0.52365	0.20244	0.23109
12	0.18395	-0.64703	0.69597	-0.20868	0.17417
13	0.056856	-1.2198	0.83534	0.027735	0.19692
14	0.22742	-0.53110	0.62347	-0.30716	0.16422
15	0.046823	-1.3087	0.85583	0.023728	0.19619
16	0.19732	-0.60936	0.68997	-0.25624	0.16793
17	0.050167	-1.2772	0.84967	-0.0020146	0.19173
18	0.093645	-0.98581	0.79612	-0.046816	0.18675
19	0.046823	-1.3087	0.85583	-0.00012744	0.19173
20	0.21739	-0.55630	0.64596	0.21543	0.23544
21	0.063545	-1.1684	0.82529	0.0036964	0.19281
22	0.20067	-0.60025	0.68403	0.14675	0.22020
23	0.28094	-0.40816	0.48718	-0.39857	0.16018
24	0.073579	-1.1001	0.80955	0.021770	0.19548
25	0.060201	-1.1934	0.83213	0.0026949	0.19255
26	0.22074	-0.54781	0.63545	-0.28405	0.16673
27	0.040134	-1.3787	0.86575	0.0072718	0.19280
28	0.093645	-0.98581	0.79612	-0.040213	0.18765
29	0.16722	-0.69723	0.71706	0.11529	0.21396
30	0.040134	-1.3787	0.86575	-0.0057520	0.19103
31	0.066890	-1.1446	0.81973	-0.051574	0.18606
32	0.090301	-1.0032	0.79839	-0.038423	0.18727
33	0.10368	-0.93677	0.77831	-0.056977	0.18668
34	0.10368	-0.93677	0.77831	0.053242	0.20215

- Prob： posterior probability
- Alpha： A parameter.
- Fst： It is used to measure the degree of population differentiation. The value is from 0 to 1. 0 meaning didn’t differentiation,1 meaning total differentiation.

Another output file

	locus1	locus2	locus3	locus4	locus5	locus6	locus7	locus8	locus9	locus10	locus11	locus12	l
pop1	0.711884	0.0383577	0.731655	0.00381943	0.99629	0.417623	0.00426047	0.557452	0.76187				
pop2	0.255519	0.0577707	0.660167	0.0204624	0.996715	0.589003	0.00362909	0.995954	0.69720				
pop3	0.255208	0.0770812	0.398218	0.0395085	0.997712	0.378452	0.00514242	0.979112	0.46125				
pop4	0.245807	0.205442	0.662913	0.00596396	0.995036	0.44398	0.00331673	0.995231	0.727716				
pop5	0.706489	0.0109268	0.824454	0.0048077	0.994599	0.26173	0.0202165	0.995353	0.708924				
pop6	0.399883	0.148074	0.447008	0.0671368	0.996496	0.137655	0.0059532	0.994701	0.697022				
pop7	0.379191	0.132689	0.967713	0.0359968	0.996781	0.414805	0.00359171	0.996988	0.86373				
pop8	0.174903	0.0109931	0.950117	0.0382035	0.997245	0.165954	0.0509697	0.96492	0.784974				
pop9	0.0686699	0.598477	0.953177	0.00475556	0.995543	0.116255	0.00348044	0.996451	0.5559				
pop10	0.363761	0.168429	0.942653	0.00335977	0.996564	0.750001	0.00289939	0.995668	0.5582				

Also can output a “prefix-freq” document, including the allele frequencies of various locus in different populations.

Thank you for your watching