

TAJIMA'S *D* CORRECTOR

ANNA RAMÍREZ-SORIANO AND RASMUS NIELSEN

FEBRUARY 2008

Most SNP data suffers from an ascertainment bias caused by the process of SNP discovery followed by SNP genotyping. SNP genotyping data have an excess of common alleles compared to directly sequenced data, making standard genetic methods of analysis inapplicable to this type of data. We have derived corrected estimators of the fundamental population genetic parameter $\theta = 4N_e\mu$ (N_e = effective population size, μ = mutation rate) based on the average number of pairwise differences and based on the number of segregating sites. We also derive the variances and co-variances of these estimators, and provide a corrected version of Tajima's *D* statistic. Tajima's *D* Corrector implements the corrected estimators we have derived and provides the corrected Tajima's *D* value of a sample, working both over simulation and empirical data, and over constant and changing *d* size.

An example of what Tajima's *D* Corrector can be used for is to find traces of selection on a gene through a genotyping study in which the SNPs have been selected from a discovery sample. That is, the SNP discovery protocol should be a) to resequence a subsample belonging to the genotyping sample and b) to genotype the SNPs found in the subsample *d* in the whole sample. Once the gene is genotyped, the value of the corrected Tajima's *D* can be calculated using this program. Moreover, its significance can also be easily computed by means of simulations, performed using any program aimed at doing coalescent simulations such as ms (Hudson, R. R. 2002. *Bioinformatics* 18: 337-338). The output of ms can be directly introduced to Tajima's *D* Corrector, where it is possible to simulate the ascertainment scheme done over the sample and to calculate the corrected Tajima's *D* over each sample in the simulations. Then, the significance of the Tajima's *D* of the gene can be easily found by looking where in the distribution obtained by simulations falls.

The program has been developed using Java version 6.0 and compiled under Windows, but it can also be used in Linux with graphic environment. To run Tajima's *D* Corrector, just double-click the .jar file or, if working from the command line, write `java -jar TajimaCorrection.jar`.

References: Ramírez-Soriano A, Nielsen R. Correcting Estimators of θ and Tajima's *D* for ascertainment biases caused by the SNP discovery process

INPUT FILES

GENERAL INFORMATION

Input files can be entered to the program typing them directly by hand or selecting them using the Browse button.

The only infile format accepted is the ms output format, both for simulations and for empirical data. This format is characterised by having a header and, below, the samples separated by a space and a double bar (//).

The header has the name of the program, the number of chromosomes per sample, the number of runs, the parameter for running ms and the seed used:

```
ms <chromosomes> <runs> -s/-t <various parameters>  
<seed>
```

```
ms 50 10 -s 111  
111
```

Of the header, the only information which will be actually used by Tajima's *D* Corrector is the number of chromosomes per sample and the number of runs.

The samples start with the double bar and next have a line with the number of segregating sites and another with their positions. Finally, there are the chromosomes, one per line:

```
//  
segsites: 10  
positions: 0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071  
101000000  
001000101  
0010001101  
001000010  
0010000101  
101010000  
0010000101  
001000000  
0010000101  
010100000  
0010000101  
0010000101  
0010000101  
0010010101  
001000000  
101010000  
0010001101  
0010000101  
001000000  
001000000
```

SIMULATION DATA

Samples generated using the ms program can be introduced directly into the program.

If some other program has been used to generate the samples, it has to be transformed into the ms output format before being used in this program. Some scripts which transform the outputs of some programs (such as Cosi or SelSim) into ms outputs can be downloaded from <http://www.snpator.com/public/downloads/aRamirez/>.

EMPIRICAL DATA

The samples obtained from empirical data must be transformed to an ms format. Thus, each position must be coded as 0 or 1; note that "0" does not necessarily denote ancestry:

```
ms 50 10 -s 111
111

//
segsites: 10
positions: 0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
```

If multiple population samples are to be analysed, they should be in separate files, and the program should be run once for each population file.

If the d sample size is not constant, each population file has to be accompanied by a second file with the same name that the one containing the sample with the extension .asc. This file must have two columns: one with the position of the SNP and the other with the size of the discovery sample for this SNP:

```
position dsample
21452 18
21662 17
22106 16
22328 3
22925 2
23393 1
24224 5
24685 17
24808 3
25062 2
25690 5
26249 1
```

CALCULATE TAJIMA'S D CORRECTED

CONSTANT D SIZE

When all the SNPs in the samples where the corrected Tajima's D needs to be calculated share the same d size, two parameters need to be specified: if the sample is already ascertained (which will usually be the case on empirical data but not on simulation data) and the discovery sample size.

This option does not allow data to have missings neither to work through windows along the region.

ASCERTAIN SAMPLES

"Ascertain simulations" specifies if the sample needs ascertainment or if it is already ascertained. If it is set to "No", the corrected Tajima's D will be calculated from the sample as it is introduced. This option should be used in empirical data and in simulation data if the ascertainment has been previously applied to the simulations.

If the sample needs to be ascertained, "ascertain simulations" should be set to "Yes". In this case, for each position a subsample of the size specified will be randomly selected. The SNP will only be considered to calculate Tajima's D if it is polymorphic in the subsample.

D SAMPLE SIZE

The discovery sample size should be specified here.

CHANGING D SIZE

This option should be set if the size of the d sample is not uniform over the SNPs. In this case the program accepts missings, which have to be coded as '?', as well as a different discovery sample size for each position, which has to be specified in a separated file (see infiles section). The formulas used to treat missings and different discovery sample sizes are explained at Ramírez-Soriano A, Nielsen R. Correcting Estimators of θ and Tajima's D for ascertainment biases caused by the SNP discovery process.

The corrected Tajima's D applied to non-uniform d sample size works through windows. The size of the windows and the step size between windows need to be specified at "windows size" and "step size" fields respectively. Both sizes have to be expressed in kilobases (kb) and must be integer number, as the program does not accept decimals.

OUTPUT FILES

The corrected Tajima's D will be given as an output file with the same name as the infile but with the extension .tcr. The output files are different depending if the input is simulation or empirical data.

CONSTANT D SIZE

Simulation data will provide an output file as follows:

Watterson's theta corrected	Tajima's theta corrected	Variance W_theta corrected	Variance T_theta corrected	Covariance corrected	Tajima's D corrected
17.000000	15.642857	70.940928	109.347628	84.654057	-0.409558
2.000000	0.542857	2.396624	2.952241	2.497456	-2.449229
13.000000	11.300000	43.936709	66.437652	51.801591	-0.653306
5.000000	4.771429	8.966245	12.336695	9.974563	-0.196445
4.000000	1.814286	6.379747	8.547731	6.984738	-2.233110
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.071429	1.000000	1.145714	1.000000	0.187122
5.000000	4.571429	8.966245	12.336695	9.974563	-0.368335
7.000000	6.142857	15.329114	21.897060	17.446582	-0.561171
1.000000	1.071429	1.000000	1.145714	1.000000	0.187122

This file has six columns, and the corrected Tajima's D value is the last. The other columns provide the corrected estimators of Watterson's and Tajima's theta, their variances and the covariance among them.

Moreover, if "Ascertain simulations" has been set to yes, another file will be generated with the sample obtained after ascertainment, that is, the sample over which the program will work. This file starts with the same name than the infile, but will have written at the end "_ascertained_x", where x will correspond to the size of the discovery sample. In that case, the output file with the corrected Tajima's D will be named as the file with the final sample. For example, if the infile was named Test.out and the discovery sample has been set to 5, two new files will be generated: Test_ascertained_5.out and Test_ascertained_5.tcr. They will contain, respectively, the ascertained sample and the corrected Tajima's D .

CHANGING *D* SIZE

Empirical data will provide an output file as follows:

Window	start_pos	end_pos	snp_num	average_n	W_theta	T_theta	average_Theta	Var_W_theta	VarT_theta	Cov	Tajima's D
1	0	49	49	139	23.432733	17.535322	23.432733	79.241026	163.471632	106.042844	-0.192556
2	20	49	29	140	12.906103	8.235925	12.906103	25.571425	51.268497	33.403851	-0.465518
3	28	49	21	139	8.681747	6.103672	8.681747	12.306826	24.077158	15.766162	-0.531380
4	35	49	14	139	5.739914	4.164504	5.739914	6.053543	11.337225	7.525796	-0.673489
5	41	49	8	139	2.749935	2.359691	2.749935	1.745456	3.019952	2.061302	-0.607096
6	42	49	7	141	2.284971	1.744746	2.284971	1.270244	2.165306	1.477794	-1.125558
7	44	49	5	141	1.670762	1.352576	1.670762	0.826446	1.330367	0.931516	-1.083072
8	44	49	5	141	1.670762	1.352576	1.670762	0.826446	1.330367	0.931516	-1.083072
9	46	49	3	142	1.028643	0.651729	1.028643	0.439866	0.651925	0.473517	-2.603770

This file has 11 columns.

Column 1: number of window.

Columns 2 and 3: absolute position of the start and end SNP. That is, 0 represents the first SNP in the sample.

Column 4: average number of SNPs per window.

Column 5: average number of valid chromosomes per window. For each position in the windows, the number of chromosomes corresponds to the number of chromosomes without missings.

Columns 6 and 7: corrected estimators of Watterson's and Tajima's thetas.

Columns 8 to 10: variances and covariance.

Column 11: corrected Tajima's *D*.