MTDT Webservice version

We present a webservice version of the M-TDT (the multi-locus transmission disequilibrium test) tool. This tool was developed to detect family-based multi-locus multi-allelic effects for qualitative or quantitative traits, extended from the original transmission disequilibrium test (TDT).

**Technology**

Our developers chose Node Js technology for Web and Desktop applications because of the performance benefits it offers such as fast execution, low resource reading and writing files, and ease to manage simultaneous customer requests. Also, the portability problems on Operating Systems such as Mac OS and Windows no longer arise compared to the first version developed with Python. Another advantage is that it offers more flexibility in setting up the web service and the desktop application with Electronjs.

**Interface**

The graphical interface is subdivided into several sections to facilitate the conduct of analyzes. The headings are as follows:

* User Space

This section allows you to create a user space in which the analysis will take place. The loaded files as well as the output files will be in this directory on the server. You can choose the name of this space at the Job title text input and submit to validate the creation of your space.





The tutorial, download and FAQ sections are not added yet.

To close your workspace, click on the button 

* Upload files and Summary Statistics

This is the section where the user uploads the files to be scanned.

file:///var/folders/01/f9lwwt853mq0r8vccnzqwnz40000gn/T/com.microsoft.Word/screenshot.pngsends you to your local files to select files whose formats must meet the following specifications:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

INPUT FILES FORMAT

\* No header in any of the 3 input files

\* Columns in all input files are tab separated

(i) the PED file: begins with 6 columns following by genotypes columns, respecting the following order

- family ID

- individual ID

- Father ID

- Mother ID

- sex

- affection (this column will not be taken into account, but its presence is required)

- genotypes columns (...)

!!! ADVICE: The 'ped' file can contain several markers, but we advice at the running step to include a limited number of markers, e.g. 3 to 5, (using the --markerset option describe below) to avoid both (i) facing a huge number of alternative hypotheses to test and (ii) having sparse count tables for transmitted versus non transmitted alleles combinations across markers. As discussed in the method paper, M-TDT has not been optimized for screen multiple markers effect within a large number of markers, but within a set predefined subgroup of markers of interest, e.g. markers with intermediate marginal effect.

(ii) the PHENOTYPE file: begins with 2 columns following by phenotypes columns, respecting the following order

- family ID

- individual ID

- phenotypes columns (...)

NOTA: Only individuals present in the phenotype file will be included in the analysis

(iii) MAP file: contains one column for genetic marker labels

- makers names (keeping the same order as in the 'ped' file)

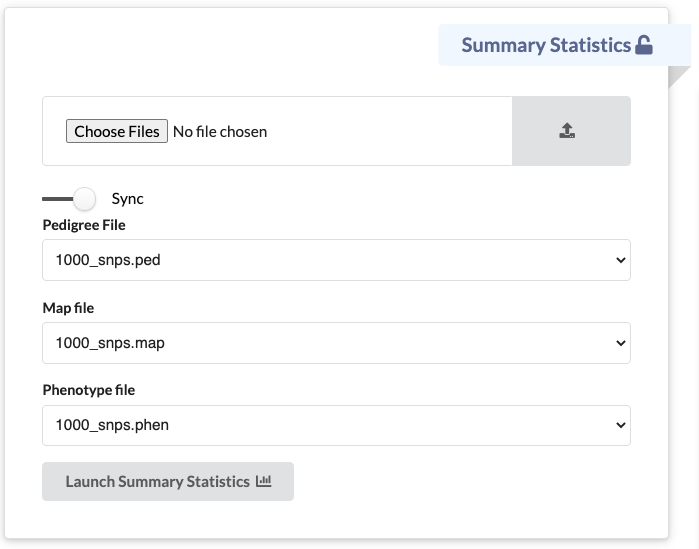
SOME RULES

- alleles should be coded in integer (1: major allele , 2: minor allele for a bi-allelic marker for instance)

- allele separator is SPACE (e.g., "1 2")

- missing genotypes should be coded "0 0"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

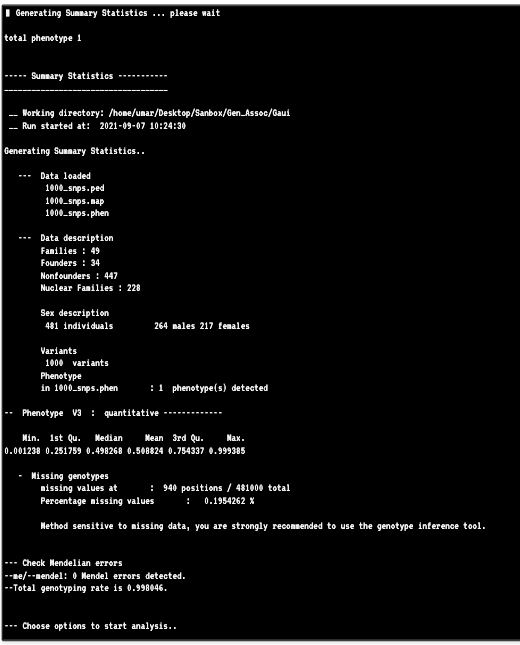


file:///var/folders/01/f9lwwt853mq0r8vccnzqwnz40000gn/T/com.microsoft.Word/screenshot.png Upload your files by clicking on this button.

You can now select your files in the dedicated fields. These three files must have the same base name.

file:///var/folders/01/f9lwwt853mq0r8vccnzqwnz40000gn/T/com.microsoft.Word/screenshot.png Activated, it allows you to pre-fill the other fields with the files corresponding to the same base name when a file is created for a control.

file:///var/folders/01/f9lwwt853mq0r8vccnzqwnz40000gn/T/com.microsoft.Word/screenshot.png generates a small description of your data. Here is a sample output of this feature.



You will be suggested to use the genotype inference option to try to find as many missing genotypes as possible since the method is very sensitive to missing data.

* Genotype Inference Option

Depending on how many missing markers you will have in your dataset, you can choose whether or not to use the genotype inference option.

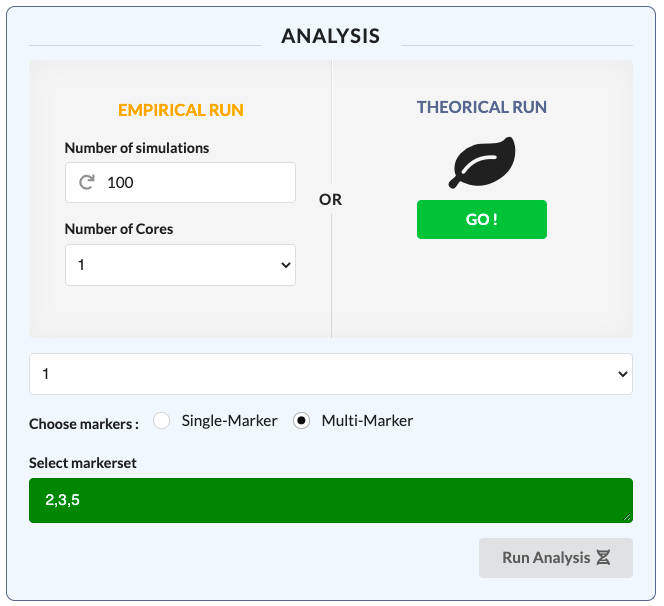


* Choose Options

It is in this section that the options are selected to launch the mtdt script. In the command line version of the tool, here is an example of a command that could be run:

Rscript mtdt.R --nbsim 10 --nbcores 4 --markerset 1,2,3 --pedfile 1000\_snps.ped --mapfile 1000\_snps.map --phenfile 1000\_snps.phen --phen 1

Here is what each of these options correspond to on the interface:



--nbsim

--nbcores

--markerset

--phen

--markerset

\* parameters

--pedfile: name of 'ped' file

--mapfile: name of 'map' file

--phenfile: name of phenotype file

--markerset: list indicating positions of the set of markers to analyze jointly (separated by "," without SPACE); position from the 'ped' file without counting the 6 first columns. If this "--markerset" option is not precise, the software will run like single-marker GWAS, screen all markers in the 'ped' file one-by-one in family-based association. A limited number of top SNPs from a GWAS screening can be subsequently used for multilocus analysis.

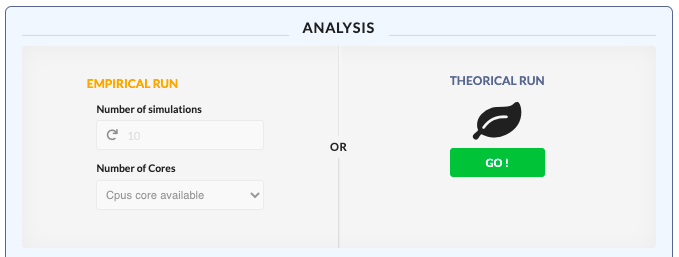
--phen: the position of the phenotype (only one) to analyze from the 'phenotype' file without counting the 2 first columns.

--nbsim: number of simulations for the computation of empirical P-values. The default value is 0 and will correspond to the case where only asymptotic P-values are provided. If sample size is limited or if there is LD (linkage disequilibrium) among markers analyzed, asymptotic theory is no more valid and then empirical P-values should be computed using simulations.

--nbcores: number of cores to use for the run, if the user wants to speed up the run by using multiple cores, as the program can run in parallelized at the simulation step if included to obtain empirical P-values. So, this option is useful only if —nbsim option is used.

Note: The distinction between empirical run and theoretical run is made at this level. By clicking on

 ,the parameters for an empirical run (number of simulations and number of cores) are disabled. To deactivate it, click on it again.



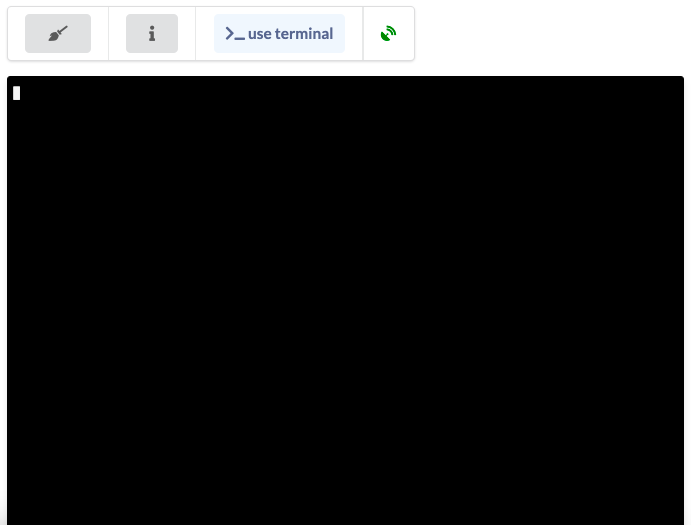
* Terminal

All outputs are displayed in the terminal. Some basic commands are available like “ls” or “clear”. More commands will be added so the user can have access to their directory.

System Info

Activate terminal

Clear



Example dataset:

Example datasets are available here: <https://github.com/avalanche-org/Gen_Assoc/tree/maindev/datasets>