SAS MACROs for SNP-phenotype association studies: implementations of the MAX test and MAX-maxT algorithms

User's manual

Dana Aeschliman, Marie-Pierre Dubé
Statistical Genetics Research Group
Montreal Heart Institute

October 19, 2007

1 Introduction

The purpose of this document is to give a brief introduction to our suite of macros for performing the MAX test and the MAX-maxT algorithm with SNP-phenotype case-control data.
Other information concerning this project may be viewed by clicking the "Downloads" tab at http://www.statgen.org.
We start off with a data set, d1_both, which is in the "TALL" format: four columns which are here called id_no, CC_vec, locus, A1 and A2. Although it doesn't matter for the purposes of running our macros, the data set is sorted by id_no, CC_vec, then locus.

```
                  Raw data in the TALL format                301
                                 17:03 Thursday, October 18, 2007

     id_no        CC_vec  locus               A1            A2
         1             1  vr_1                 0             0
         1             1  vr_10                0             0
         1             1  vr_11                0             0
         1             1  vr_12                0             0
         1             1  vr_13                1             0
         1             1  vr_14                1             0
         1             1  vr_15                0             0
         1             1  vr_16                0             1
         1             1  vr_17                1             0
         1             1  vr_18                0             0
```

2 The MAX test

Here is how the MAX test is performed from inside SAS using the data set d1_both:

```
%MAXTEST(DSN=d1_both, AFF_STAT=CC_vec, NUM_MVNS=10000,
use_lib=library, IDENT=id_no,
A1=A1, A2=A2, locus=locus, title1=Test run with synthetic data,
get_graphic=T, get_reports=T, top_SNPs=20);
```

Here is a list of all the macro variables used by the main macro max_test_macro.sas:

- dsn is the name of the input data set.
- aff_stat is the name of the variable on dsn which gives the affection status.
- locus is the name of the variable on dsn which gives the locus.
- A1 is the name of the variable on dsn which gives the first allele.
- A2 is the name of the variable on dsn which gives the second allele.
- NUM_MVNS is the number of Monte Carlo samples to be drawn to estimate the P-value of the association between a SNP and a phenotype.
- IDENT is the name of the variable on dsn which gives the subject ID.
- output_DSN is unimportant unless the main macro is being called by the wrapper macro implementing Westfall and Young's Algorithm 4.1.
- use_lib is used for several things. It's the library that holds the raw data. It's the library where the quality control (QC) files &use_lib..bad_ccs, &use_lib..miss_or_diff_als, and &use_lib..snp_warnings will be deposited. It's the library where the results file max_data will be deposited in the cases when use_lib≙WORK.
- skip_checks should be left equal to F.
- keep_details should be left equal to T.
- get_graphic is a flag telling if reports should be produced.
- ps_graphic and ls_graphic respectively control the output pagesize and linesize for graphics.
- get_reports is a flag telling if reports should be produced.
- ps_reports and ls_reports respectively control the output pagesize and linesize for reports.
- top_SNPs directs that the top top_SNPs SNPs be printed out in a separate report. Changing top_SNPs to a number less than 1 suppresses this output.

- ps_top_SNPs and ls_top_SNPs respectively control the output pagesize and linesize.

Here are some other notes on using the MAX test macros:

- The main macro includes %include statements to read the other macros. So all macros should be placed in the same directory.
- The macros assume that the data set is in the PROC CASECONTROL "long" format, like this data set, i.e. columns are locus name, subject number (an integer), Affection Status (0=Control, 1=Case), and binary alleles (can be coded either by the integers 0 and 1 or by letters).
- The values passed to the macro variables aff_stat, locus, A1, A2 and IDENT should **EXACTLY** match, including matching upper case and lower case letters, the names of the corresponding variables on the input data set. For example, if the variable on the input data set which codes for the locus name is SNP_ID then SNP_ID is acceptable while none of snp_ID, SNP_id or snp_id are acceptable.

3 The MAX-maxT algorithm

Here is some code that starts with a SAS data set in the "TALL" form, d1, and ends with maxT adjusted MAX test statistics.

```
libname library '';

%INCLUDE "prepare_4_step_down.sas";
%INCLUDE "do_maxT.sas";
%PREPARE4STEPDOWN(DSN=d1, AFF_STAT=CC_vec,
output_DSN=out_data, use_lib=library, IDENT=id_no,
num_resamps=420, overwrite_output_files=T, A1=A1, A2=A2,
locus=locus, keep_both=T,
skip_2_start2=F);


%do_maxT(use_lib=library, get_thin_ds=F, get_graphic=F, get_reports=T,
title1=MAX Test of Freidlin et al. (2002) followed by max T,
top_SNPs=40);
```

The last page of the output produced looks like this:

```
              MAX Test of Freidlin et al. (2002) followed by max T                20
                                  MAX-maxT results       15:54 Friday, October 19, 2007
       A total of 420 permutations of the CC Stat vector were performed.
        List of 40 SNPs most likely to be associated with the phenotype.

                     --------------Details of MAX test--------------
                                                                  max T
                     -------Components--------                  adjusted
           locus          Rec       Add       Dom       MAX    p-val(MAX)
           vr_38         2.86      2.42      1.76      2.86       0.37381
           vr_43        -2.57     -2.40     -1.87      2.57       0.70714
           vr_11        -1.86     -2.42     -2.12      2.42       0.84524
           vr_12         2.30      1.85      1.30      2.30       0.93571
           vr_14         0.50     -1.30     -2.29      2.29       0.93810
           vr_46         1.42      1.92      2.12      2.12       0.98333
           vr_17        -2.08     -0.90      0.12      2.08       0.98571
           vr_25         0.76      1.50      2.01      2.01       0.99048
           vr_57         1.93      1.84      1.38      1.93       0.99524
           vr_4          1.06      1.91      1.91      1.91       0.99524
           vr_50        -1.91     -1.85     -1.51      1.91       0.99524
           vr_5         -1.85     -0.65      0.00      1.85       1.00000
           vr_15         1.81     -0.70     -1.45      1.81       1.00000
           vr_59         1.80      1.45      0.99      1.80       1.00000
           vr_2          0.60      1.59      1.75      1.75       1.00000
           vr_47        -1.64     -1.45     -1.12      1.64       1.00000
           vr_1          0.81      1.61      1.63      1.63       1.00000
           vr_3          1.03      1.62      1.50      1.62       1.00000
           vr_23         1.56      1.62      1.39      1.62       1.00000
           vr_63         1.58      1.53      1.35      1.58       1.00000
           vr_65        -0.60      0.66      1.55      1.55       1.00000
```

Alternatively, if one is able to run multiple jobs simultaneously, one may wish to create the basic data set “both” which is row-wise resampled to produce the MAX statistics. Such a session in a Linux environment is sketched here.
We start as before, calling the macro PREPARE4STEPDOWN, but this time with less resampling:

```
libname library '';

%INCLUDE "prepare_4_step_down.sas";
%INCLUDE "do_maxT.sas";
%PREPARE4STEPDOWN(DSN=d1, AFF_STAT=CC_vec,
output_DSN=out_data, use_lib=library, IDENT=id_no,
num_resamps=20, overwrite_output_files=T, A1=A1, A2=A2,
locus=locus, keep_both=T,
skip_2_start2=F);
```

In the Linux environment, we copy the SAS files params.sas7bdat, both.sas7bdat, and max_data.sas7bdat along with all of the files containing the macros, into two subdirectories that we create, s2 and s3. Descending into s2, we start resampling again, outputting to out_data2 with the options keep_both=F and skip_2_start2=T:

```
libname library ’’;

%INCLUDE "prepare_4_step_down.sas";
%INCLUDE "do_maxT.sas";
%PREPARE4STEPDOWN(DSN=d1, AFF_STAT=CC_vec,
output_DSN=out_data2, use_lib=library, IDENT=id_no,
num_resamps=200, overwrite_output_files=T, A1=A1, A2=A2,
locus=locus, keep_both=F,
skip_2_start2=T);
```

We do the same in the subdirectory s3:

```
libname library ’’;

%INCLUDE "prepare_4_step_down.sas";
%INCLUDE "do_maxT.sas";
%PREPARE4STEPDOWN(DSN=d1, AFF_STAT=CC_vec,
output_DSN=out_data3, use_lib=library, IDENT=id_no,
num_resamps=200, overwrite_output_files=T, A1=A1, A2=A2,
locus=locus, keep_both=F,
skip_2_start2=T);
```

After these two processes finish, we return to the Linux environment. We copy the files produced, out_data2.sas7bdat and out_data3.sas7bdat to the original directory containing out_data.sas7bdat. We then build the file output_file_names and call the macro do_maxT:

```
libname library ’’;

data library.output_file_names;
output_file=‘‘LIBRARY.OUT_DATA ’’; num_resamps=20; output;
output_file=‘‘LIBRARY.OUT_DATA2’’; num_resamps=200; output;
output_file=‘‘LIBRARY.OUT_DATA3’’; num_resamps=200; output;
run;

%do_maxT(use_lib=library, get_thin_ds=F, get_graphic=F, get_reports=T,
title1=MAX Test of Freidlin et al. (2002) followed by max T,
top_SNPs=40);
```

We notice that the output produced varies somewhat from the previous output because of the relatively small number of resamples (420) done in both cases. The last page produced now looks like this:

```
             MAX Test of Freidlin et al. (2002) followed by max T               19
                                  MAX-maxT results       17:38 Friday, October 19, 2007
      A total of 420 permutations of the CC Stat vector were performed.
       List of 40 SNPs most likely to be associated with the phenotype.

                   --------------Details of MAX test--------------
                                                                  max T
                   -------Components--------                    adjusted
         locus          Rec       Add       Dom       MAX   p-val(MAX)
         vr_38         2.86      2.42      1.76      2.86      0.42619
         vr_43        -2.57     -2.40     -1.87      2.57      0.70238
         vr_11        -1.86     -2.42     -2.12      2.42      0.82619
         vr_12         2.30      1.85      1.30      2.30      0.91667
         vr_14         0.50     -1.30     -2.29      2.29      0.91667
         vr_46         1.42      1.92      2.12      2.12      0.97143
         vr_17        -2.08     -0.90      0.12      2.08      0.98333
         vr_25         0.76      1.50      2.01      2.01      0.98810
         vr_57         1.93      1.84      1.38      1.93      0.98810
         vr_4          1.06      1.91      1.91      1.91      0.99048
         vr_50        -1.91     -1.85     -1.51      1.91      0.99048
         vr_5         -1.85     -0.65      0.00      1.85      0.99524
         vr_15         1.81     -0.70     -1.45      1.81      0.99762
         vr_59         1.80      1.45      0.99      1.80      0.99762
         vr_2          0.60      1.59      1.75      1.75      0.99762
         vr_47        -1.64     -1.45     -1.12      1.64      1.00000
```

Here is a list of all the macro variables used by the macro PREPARE4STEPDOWN which is called at the start of the MAX-maxT algorithm:

- dsn is the name of the input data set.
- aff_stat is the name of the variable on dsn which gives the affection status.
- locus is the name of the variable on dsn which gives the locus.
- A1 is the name of the variable on dsn which gives the first allele.
- A2 is the name of the variable on dsn which gives the second allele.
- num_resamps is the number of permutations of the Affection Status vector and corresponding MAX statistics to be produced.
- IDENT is the name of the variable on dsn which gives the subject ID.
- output_DSN is the data set which will keep the num_resamps sets of MAX statistics which are produced by the current call to PREPARE4STEPDOWN.
- use_lib is used for several things. It's the library that holds the raw data. It's the library where the quality control (QC) files &use_lib..bad_ccs,

&use_lib..miss_or_diff_als, and &use_lib..snp_warnings will be deposited. It's the library where the results file max_data will be deposited in the cases when use_lib≙WORK.

- keep_both is a flag telling if a copy of the data set "both" should be saved in use_lib. If the user wants to produce the MAX statistics over several different time periods by repeated calls to PREPARE4STEPDOWN and this is the first time that PREPARE4STEPDOWN is being called, then this should be T.
- skip_2_start2 is a flag telling if the QC steps should be skipped. If the user wants to produce the MAX statistics over several different time periods by repeated calls to PREPARE4STEPDOWN and this is NOT the first time that PREPARE4STEPDOWN is being called, then this should be T. Otherwise, it should be F.
- overwrite_output_files is a flag telling if former results can be overwritten. As a precaution this should be left F.

Here is a list of all the MACRO variables used by the MACRO do_maxT contained in the file do_maxT.sas:

- use_lib is used for several things. It's the library that holds the raw data. It's the library where the quality control (QC) files &use_lib..bad_ccs, &use_lib..miss_or_diff_als, and &use_lib..snp_warnings will be deposited. It's the library where the results file max_data will be deposited in the cases when use_lib≙WORK.
- title1 is a title that can be changed by the user to output more informative graphs and reports.
- get_thin_ds is a flag telling if thin data set having only max_abs_zs, pval and locus should be output to text file "all".
- get_graphics is a flag telling if reports should be produced.
- ps_graphics and ls_graphics respectively control the output pagesize and linesize for graphics.
- get_reports is a flag telling if reports should be produced.
- ps_reports and ls_reports respectively control the output pagesize and linesize for reports.
- top_SNPs directs that the top top_SNPs SNPs be printed out in a separate report. Changing top_SNPs to a number less than 1 suppresses this output.
- ps_top_SNPs and ls_top_SNPs respectively control the output pagesize and linesize.

Here are some comments concerning the MAX-maxT algorithm:

- The MACROs assum e that the data set is in the PROC CASECONTROL "LONG" format, like this data set, i.e. columns are locus name, subject number (an integer), Affection Status (0=Control, 1=Case), and binary alleles (can be coded either by the integers 0 and 1 or by letters).

- The values passed to the MACRO variables aff_stat, locus, A1, A2 and IDENT should EXACTLY match, including matching upper case and lower case letters, the names of the corresponding variables on the input data set. For example, if the variable on the input data set which codes for the locus name is SNP_ID then SNP_ID is acceptable while none of snp_ID, SNP_id or snp_id are acceptable.

- The MACRO which will do the most work (and take the most time) is PREPARE4STEPDOWN which is contained in the file prepare_4_step_down.sas. This MACRO is a wrapper for MAXTEST which is contained in the file max_test_macro.sas. PREPARE4STEPDOWN includes %include statements to read the other MARCOs (except do_maxT, which is performed last). So all MACROs should be placed in the same directory.