MASS TUTORIAL: Learn by Example

This example starts with a genotype file, a phenotype file and a mapping file that are used in the example of SCORE-Seq. We show the pipeline of meta-analysis for various rare variant tests.

Step 1: Run SCORE-Seq

Suppose that we are interested in T1, T5, MB, VT, and SKAT tests, as well as a maximum test over (T1, T5, MB), which is used to adjust for multiple testing with T1, T5, and MB burden scores. In the weight file for the maximum test, the first two columns are the gene ID and SNP ID, and the last three columns pertain to the T1, T5 and MB burden scores. For T1, the entry in each row indicates, by the values 1 vs 0, whether the MAF <= 1% or >1%. For T5, the entry in each row indicates, by the values 1 vs 0, whether the MAF <= 5% or >5%. For MB, the entry in each row is the Madsen-Browning weight based on the MAF of that SNP. To obtain the score statistics for MASS, we first run the following SCORE-Seq command:

```
$ SCORE-Seq -pfile phenoP.txt -gfile genoP.txt -mfile mappingP.txt -wfile
wfile.txt -ofile rareP.txt -vtlog vtP.log -snplog skatP.log -multilog
maxP.log -MAF 0.05 -MAC 5
```

The file `rareP.txt` contains scalar score statistics and variance estimates for T1, T5, and MB tests. The files `vtP.log`, `skatP.log`, and `maxP.log` contain the score vectors and information matrices for VT, SKAT, and maximum tests, respectively. The output for the first gene ACTBL2 is shown below. The types of statistics that will be used by MASS are boxed in different shapes, and different colors of the boxes represent different tests.

The file `rareP.txt` contains the following:

GENE_ID	T1_P	T5_P	MB_P	VT_P	SKAT_P	MAX_P	T1_U	T1_V	T1_Z	T5_U	T5_V	T5_Z	MB_U	MB_V	MB_Z	T1_MAC	T5_MAC	MB_MAC
ACTBL2	1.61E-01	1.27E-01	9.02E-02	3.34E-01	8.45E-01	1.56E-01	-3.25E+00	5.37E+00	-1.40E+00	-4.60E+00	9.08E+00	-1.53E+00	-6.48E+01	1.46E+03	-1.69E+00	26	43	43

The numbers inside the red rectangle are the score statistic and variance estimate based on the T1 burden score; the number inside the red round rectangle is the MAC of the T1 burden score. These numbers will be used to create the input files of the T1 test for MASS in Step 2.

The file `maxP.log` contains the following:

ACTBL2	26	-3.25E+00	5.37E+00	0	0
ACTBL2	43	-4.60E+00	5.40E+00	9.08E+00	0
ACTBL2	43	-6.48E+01	7.42E+01	1.09E+02	1.46E+03

The numbers inside the blue rectangle are the score vector (shown in the first column) and the lower triangular information matrix based on the three burden scores (T1, T5, MB). The three rows correspond to the three burden scores. You can compare, say, the score statistic (-3.25E+00) and the variance (5.37E+00) in the first row to the numbers inside the red rectangular of the file `rareP.txt` to reassure yourself that the first row indeed corresponds to the T1 burden score. The numbers inside the blue round rectangle are the MACs of the T1, T5, and MB burden

scores. These numbers will be used to create the input files of the maximum test for MASS in Step 2.

The file `vtP.log` contains the following:

ACTBL2	1.02E-02	43	-4.60E+00	9.08E+00	0	0	0	0
ACTBL2	1.13E-03	7	-1.76E+00	1.42E+00	1.55E+00	0	0	0
ACTBL2	9.07E-03	26	-3.25E+00	5.40E+00	1.48E+00	5.37E+00	0	0
ACTBL2	6.80E-03	18	-1.83E+00	3.83E+00	1.52E+00	3.74E+00	3.83E+00	0
ACTBL2	5.67E-03	12	-1.79E+00	2.66E+00	1.54E+00	2.53E+00	2.59E+00	2.62E+00

The numbers inside the green rectangle are the score vector (shown in the first column) and the lower triangular information matrix for a set of burden scores at different MAF thresholds. The column after the gene ID gives the specific MAF threshold. The numbers inside the green round rectangle are the MACs for the set of burden scores. These numbers will be used to create the input files of the VT test for MASS in Step 2.

The file `skatP.log` contains the following:

ACTBL2	rs3734173	1.02E-02	9	-2.65E-01	1.97E+00	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	1:173772298	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	9:5756326	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	6:47846556	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	rs75044156	9.07E-03	8	-1.42E+00	-3.08E-02	0	0	0	1.71E+00	0	0	0	0	0	0	0
ACTBL2	15:79749961	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	10:17089450	6.80E-03	6	-4.15E-02	-2.08E-02	0	0	0	-2.85E-02	0	1.27E+00	0	0	0	0	0
ACTBL2	9:116940613	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	22:29121015	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	10:16878375	1.02E-02	9	-1.09E+00	-4.52E-02	0	0	0	-2.84E-02	0	-1.82E-02	0	0	1.77E+00	0	0
ACTBL2	16:31087835	1.13E-03	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	rs61910679	5.67E-03	5	-2.82E-02	-2.42E-02	0	0	0	-1.88E-02	0	-1.37E-02	0	0	2.17E-01	0	1.10E+00

The numbers inside the purple rectangle are the score vector (shown in the first column) and the lower triangular information matrix for the set of SNPs in gene ACTBL2. The column after the gene ID gives the specific SNP ID. The numbers inside the purple dashed rectangle are the MAFs for the set of SNPs and the numbers inside the purple round rectangle are the MACs. These numbers will be used to create the input files of the SKAT test for MASS in Step 2.

Since we apply the MAC lower bound of 5, the corresponding components of the score vector and the rows and columns of the information matrix are zeroed out for the SNPs with MACs less than 5. We keep such SNPs instead of deleting them because it is possible that these SNPs have higher MACs in other studies, so that they can still contribute to the meta-analysis. Thus, we suggest to apply an MAC lower bound during the meta-analysis instead of during the analysis of the individual studies.

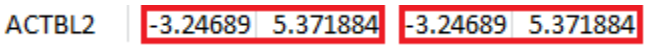
Step 2: Create input files for MASS

Suppose that we want to perform meta-analysis of two studies. For simplicity, the output from Step 1 is used as the output for both studies. In this way, we can demonstrate how to combine files from two studies by only using one set of files. A R script "createInput_MASS.r" shows

how to generate MASS input files from SCORE-Seq output files. At the end, we have a set of MAC input files "`MAC_<test>.txt`" and a set of score statistic input files "`score_<test>.txt`", where <test> is one of the tests: "T1", "T5", "MB", "VT", "SKAT", and "max". In addition, a weight file "`weight.txt`" is generated for SKAT.

(1) T1 test

The file `score_T1.txt` contains the following:



ACTBL2	-3.24689	5.371884	-3.24689	5.371884

The score statistics and variance estimates from study 1 and study 2 are inside the two red rectangles. These values are from the red rectangle in the file `rareP.txt` shown in the last step. In our example, the values from the two studies are identical because the output file of the 2nd study is also `rareP.txt`.

The file `MAC_T1.txt` contains the following:

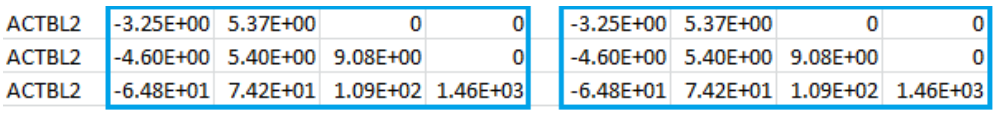
ACTBL2	52

The MAC inside the red round rectangle is the total MAC across two studies. Note that 52 is twice the value from the round rectangle in file `rareP.txt` since each of the two studies contributes 26 minor alleles.

The input files for the T5 and MB tests are generated in the same way as the T1 test.

(2) Maximum test

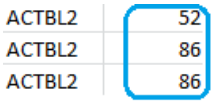
The file `score_max.txt` contains the following:



ACTBL2	-3.25E+00	5.37E+00	0	0	-3.25E+00	5.37E+00	0	0
ACTBL2	-4.60E+00	5.40E+00	9.08E+00	0	-4.60E+00	5.40E+00	9.08E+00	0
ACTBL2	-6.48E+01	7.42E+01	1.09E+02	1.46E+03	-6.48E+01	7.42E+01	1.09E+02	1.46E+03

The score vectors and information matrices from study 1 and study 2 are inside the two blue rectangles. These values are from the blue rectangle in the file `maxP.log` shown in Step 1.

The file `MAC_max.txt` contains the following:



ACTBL2	52
ACTBL2	86
ACTBL2	86

The MACs inside the blue round rectangle are the total MACs across the two studies. The value on each row is twice the value on the corresponding row inside the blue round rectangle of the file `maxP.log`.

(3) VT test

The file `score_VT.txt` contains the following:

ACTBL2	-4.60E+00	9.08E+00	0	0	0	0	-4.60E+00	9.08E+00	0	0	0	0
ACTBL2	-1.76E+00	1.42E+00	1.55E+00	0	0	0	-1.76E+00	1.42E+00	1.55E+00	0	0	0
ACTBL2	-3.25E+00	5.40E+00	1.48E+00	5.37E+00	0	0	-3.25E+00	5.40E+00	1.48E+00	5.37E+00	0	0
ACTBL2	-1.83E+00	3.83E+00	1.52E+00	3.74E+00	3.83E+00	0	-1.83E+00	3.83E+00	1.52E+00	3.74E+00	3.83E+00	0
ACTBL2	-1.79E+00	2.66E+00	1.54E+00	2.53E+00	2.59E+00	2.62E+00	-1.79E+00	2.66E+00	1.54E+00	2.53E+00	2.59E+00	2.62E+00

The score vectors and information matrices from study 1 and study 2 are inside the two green rectangles. These values are from inside the green rectangle of the file `vtP.log` shown in Step 1.

The file `MAC_VT.txt` contains the following:

ACTBL2	86
ACTBL2	14
ACTBL2	52
ACTBL2	36
ACTBL2	24

The MACs inside the green round rectangle are the total MACs across the two studies. The value on each row is twice the value on the corresponding row inside the green round rectangle of the file `vtP.log`.

(4) SKAT

The file `score_SKAT.txt` contains the following:

ACTBL2	-2.65E-01	1.97E+00	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	-1.42E+00	-3.08E-02	0	0	0	1.71E+00	0	0	0	0	0	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	-4.15E-02	-2.08E-02	0	0	0	-2.85E-02	0.00E+00	1.27E+00	0	0	0	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	0.00E+00	0.00E+00	0	0	0	0.00E+00	0.00E+00	0.00E+00	0	0	0	0	0
ACTBL2	-1.09E+00	-4.52E-02	0	0	0	-2.84E-02	0.00E+00	-1.82E-02	0	0	1.77E+00	0	0
ACTBL2	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTBL2	-2.82E-02	-2.42E-02	0	0	0	-1.88E-02	0.00E+00	-1.37E-02	0	0	2.17E-01	0	1.10E+00

-2.65E-01	1.97E+00	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
-1.42E+00	-3.08E-02	0	0	0	1.71E+00	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
-4.15E-02	-2.08E-02	0	0	0	-2.85E-02	0	1.27E+00	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0.00E+00	0.00E+00	0	0	0	0.00E+00	0	0.00E+00	0	0	0	0	0
-1.09E+00	-4.52E-02	0	0	0	-2.84E-02	0	-1.82E-02	0	0	1.77E+00	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
-2.82E-02	-2.42E-02	0	0	0	-1.88E-02	0	-1.37E-02	0	0	2.17E-01	0	1.10E+00

The score vectors and information matrices from study 1 and study 2 are inside the two purple rectangles. These values are from inside the purple rectangle in the file `skatP.log` shown in Step 1.

The file `MAC_SKAT.txt` contains the following:

ACTBL2	18
ACTBL2	2
ACTBL2	2
ACTBL2	2
ACTBL2	16
ACTBL2	2
ACTBL2	12
ACTBL2	2
ACTBL2	2
ACTBL2	18
ACTBL2	2
ACTBL2	10

The MACs inside the purple round rectangle are the total MACs across the two studies. The value on each row is twice the value on the corresponding row from inside the purple round rectangle of the file `skatP.log`.

The file `weight_SKAT.txt` contains the following:

ACTBL2	382.0073
ACTBL2	591.8771
ACTBL2	591.8771
ACTBL2	591.8771
ACTBL2	403.5866
ACTBL2	591.8771
ACTBL2	450.3862
ACTBL2	591.8771
ACTBL2	591.8771
ACTBL2	382.0073
ACTBL2	591.8771
ACTBL2	475.7386

The numbers inside the purple dashed rectangle are the diagonal entries of the weight matrix for the weighted quadratic statistic. In this case, the number on each row is calculated through a beta density function based on the MAF on the corresponding row from inside the purple dashed rectangle of the file `skatP.log`. This is the default weight in the SKAT test.

Step 3: Run MASS

Suppose that we want to apply the MAC lower bound of 2 to all tests. If the file for MAC is not provided when a positive MAC lower bound is specified, the software will output an error message "`Error: MAC information not provided.`" and exit. The meta-analysis results are given in the output files `meta_<test>.out`.

(1) T1, T5, and MB tests

```
   $ MASS -nstudy 2 -MAC_LB 2 -MAC MAC_T1.txt -ifile score_T1.txt -ofile
meta_T1.out
   $ MASS -nstudy 2 -MAC_LB 2 -MAC MAC_T5.txt -ifile score_T5.txt -ofile
meta_T5.out
   $ MASS -nstudy 2 -MAC_LB 2 -MAC MAC_MB.txt -ifile score_MB.txt -ofile
meta_MB.out
```

This is the expected software output from the first command:

```
@------------------------------------------------------------@
|      MASS       |      v3.0     |       19/March/2013      |
--------------------------------------------------------------
|               For documentation & citation                 |
|        http://www.bios.unc.edu/~lin/software/MASS          |
@------------------------------------------------------------@


MASS started: Sat Apr 20 23:53:03 2013


*----------------- Loading options...-----------------*
OPTIONS:
        -- The MAC lower bound is 2.
        -- MAC information is extracted from file 'MAC_T1.txt'.
        -- The method to combine score statistics is 'quadratic'.
        -- Meta results are written to file 'meta_T1.out'.


*----------------- Loading data...-----------------*

INPUT FILE: 'score_T1.txt'.........
successful

INPUT FILE: 'MAC_T1.txt'.........
successful

DATA SUMMARY:
#GENE = 86


*---------------- START MASS ----------------*
              Sat Apr 20 23:53:03 2013


*----------------- END MASS -----------------*
              Sat Apr 20 23:53:03 2013
```

There are three sections below the heading. The first section "Loading options" gives all the options (both user-specified and default) that are going to be applied to the meta-analysis. The second section "Loading data" reflects the data loading process and summarizes the total number of genes to be analyzed. The third section contains all possible messages during the analysis.

The results for the first 3 genes in the output files `meta_T1.out`, `meta_T5.out`, and `meta_MB.out` are shown in parallel below:

`meta_T1.out`

Gene_ID	Statistic	P-value
ACTBL2	3.92E+00	4.76E-02
ACTN4	5.40E-01	4.62E-01
ANK2	1.10E+01	8.96E-04

`meta_T5.out`

Gene_ID	Statistic	P-value
ACTBL2	4.67E+00	3.08E-02
ACTN4	5.40E-01	4.62E-01
ANK2	2.82E+00	9.33E-02

`meta_MB.out`

Gene_ID	Statistic	P-value
ACTBL2	5.74E+00	1.66E-02
ACTN4	1.47E+00	2.26E-01
ANK2	7.32E+00	6.82E-03

(2) Maximum test

```
$ MASS -nstudy 2  -MAC_LB 2 -MAC MAC_max.txt -method maximum -ifile
score_max.txt -ofile meta_max.out
```

The results for the first 3 genes are given in the MASS output file `meta_max.out` :

Gene_ID	Statistic	P-value	Flag
ACTBL2	5.74E+00	3.16E-02	0
ACTN4	1.47E+00	2.65E-01	0
ANK2	1.10E+01	2.00E-03	0

By comparing the output from the T1, T5, and MB tests, we can verify that the statistic of the maximum test is indeed the maximum over the three test statistics. The p-value is larger than the minimal p-value of the three tests due to the fact that the maximum test is adjusted for multiple testing. The values of the variable "Flag" are 0 for the three genes, which indicates that these p-values achieve the desired numerical accuracy in the multivariate normal integration.

(3) VT test

```
$ MASS -nstudy 2  -MAC_LB 2 -MAC MAC_VT.txt -method maximum -ifile
score_VT.txt -ofile meta_VT.out
```

The results for the first 3 genes are given in the output file `meta_VT.out`:

Gene_ID	Statistic	P-value	Flag
ACTBL2	4.67E+00	9.73E-02	0
ACTN4	5.40E-01	4.62E-01	0
ANK2	1.51E+01	6.79E-04	1

The value of the "Flag" variable for gene ANK2 indicates that the p-value may not achieve the desired numerical accuracy in the multivariate normal integration. This usually happens in a gene with a large number of thresholds or /and with large test statistics.

(4) SKAT

```
$ MASS -nstudy 2 -MAC_LB 2 -MAC MAC_SKAT.txt -method wquadratic -weight
weight.txt -ifile score_SKAT.txt -ofile meta_SKAT.out
```

The results for the first 3 genes are given in the output file `meta_SKAT.out`:

Gene_ID	Statistic	P-value	Flag
ACTBL2	5.18E+03	5.43E-01	0
ACTN4	1.75E+01	9.03E-01	0
ANK2	4.03E+04	3.19E-03	0

The p-value for the weighted quadratic test is based on a mixture chi-square distribution where numerical approximation is involved. The "Flag" variable is used to indicate whether the desired numerical accuracy is achieved.