

Meta-analysis of association studies: Practical session

Kalliope Panoutsopoulou

Much of the material below (within boxes) is taken from the tutorial on the GWAMA website (<http://www.well.ox.ac.uk/gwama/>).

GWAMA: software tool for meta analysis of whole genome association data

Magi R, Morris AP: GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010, 11:288.

INTRODUCTION

GWAMA (Genome-Wide Association Meta Analysis) software has been developed to perform meta-analysis of the results of GWA studies of binary or quantitative phenotypes. The software incorporates error trapping facilities to identify strand alignment errors and allele flipping, and performs tests of heterogeneity of effects between studies.

GWAMA file format

Each GWA study file has mandatory column headers:

- 1) MARKER – snp name
- 2) EA – effect allele
- 3) NEA – non effect allele
- 4) OR - odds ratio
- 5) OR_95L - lower confidence interval of OR
- 6) OR_95U - upper confidence interval of OR

In case of quantitative trait:

- 4) BETA – beta
- 5) SE – std. error

Study files might also contain columns:

- 7) N - number of samples
- 8) EAF – effect allele frequency
- 9) STRAND – marker strand (if the column is missing then program expects all markers being on positive strand)
- 10) IMPUTED – if marker is imputed or not (if the column is missing then all markers are counted as directly genotyped ones)

Input files must be either tab or space delimited. Files must not have empty columns as multiple separators are treated as one.

RUNNING GWAMA

Command line options:

gwama

--filelist {filename} or -i {filename} Specify studies' result files. Default = gwama.in

--output {fileroot} or -o {fileroot} Specify file root for output of analysis. Default = gwama (gwama.out, gwama.gc.out)

--random or -r Use random effect correction. Default = disabled

--genomic-control or -gc Use genomic control for adjusting studies' result files. Default = disabled

--genomic-control-output or -gco Use genomic control on meta-analysis summary (i.e. results of meta- analysis are corrected for gc). Default = disabled

--quantitative or -qt Select quantitative trait version (BETA and SE columns). Default = binary trait

--map {filename} or -m {filename} Select file name for marker map.

--threshold {0-1} or -t {0-1} The p-value threshold for showing direction in summary effect directions. Default = 1

--help or -h Print this help

--version or -v Print GWAMA version number

OUTPUT FILES

GWAMA generates following output files:

gwama.out (or 'fileroot'.out if --output option is used)

This file contains results of meta-analysis. Output file has following columns:

chromosome - Marker chromosome

position - Marker position (bp)

rs_number - Marker ID

reference_allele - Effect allele

other_allele - Non effect allele

OR - Overall odds ratio for meta-analysis

OR_95L - Lower 95% CI for OR

OR_95U - Upper 95% CI for OR

z - Z-score

p-value - Meta-analysis p-value

-log10_p-value - Absolute value of logarithm of meta-analysis p-value to the base of 10.

q_statistic - Cochran's heterogeneity statistic

q_p-value - Cochran's heterogeneity statistic's p-value

i2 - Heterogeneity index I2 by Higgins et al 2003

n_studies - Number of studies with marker present

n_samples - Number of samples with marker present (will be NA if marker is present in any input file where N column is not present)

effects - Summary of effect directions ('+' - positive effect of reference allele, '-' - negative effect of reference allele, '0' - no effect (or non-significant) effect of reference allele, '?' - missing data)

gwama.gc.out (or 'fileroot'.gc.out if --output option is used)

This file contains *lambda* values for GC correction. The file is only generated, if **-gc** option is used.

gwama.log.out

This file contains all log information about current GWAMA run. Each error and warning has unique error code. More information for them can be found from gwama.err.out file.

gwama.err.out

This file contains all errors and warning generated during GWAMA run. Information about any error can be searched according to error code. For example in UNIX:

grep E00000001 gwama.err.out

gives information about error E00000001

You have conducted a GWAS for your disease of interest. You have prioritised 3 SNPs for follow up (rs3, rs4, rs5) which are directly typed in your study. You also want to examine 2 SNPs in candidate genes for your disease of interest for which some evidence for association has been reported in the literature (rs1 and rs2). You will perform a meta-analysis by combining your selected SNPs with *in silico* replication set 1. You will then select SNPs for *de novo* replication (replication set 2).

1/ Start by examining your association results file study_1.assoc. This is the output of a basic association analysis performed in PLINK. Familiarise yourself with the information given in this file. Note that A1 (which stands for allele 1) is the effect allele i.e. the allele to which the odds ratio (OR) has been estimated in PLINK. 95L and 95U are the lower and upper confidence interval of OR respectively. F_A is the frequency of A1 in affected and F_U is the frequency of A1 in unaffected.

1a/ Which SNP shows the strongest evidence for association?

1b/ Which of these columns are mandatory for the meta-analysis? See GWAMA file format.

2/ You have identified another group which has GWAS data on your disease of interest and has agreed to provide *in silico* replication for the 5 SNPs requested.

2a/ Run a fixed-effects meta-analysis model using your study file (study_1.gwama) and *in silico* replication set1 (repl_1.gwama). You will need to create a file that specifies the filenames of the studies to be meta-analysed, see meta_1.gwama.in.

```
GWAMA -i meta_1.gwama.in -o meta_1_results
```

Check the log and error files. Is there anything wrong in the 2 results files? How has the program coped with it?

2b/ Look at the repl_1.gwama file more carefully. Are any of the SNPs imputed? Is there a need for eliminating any of the SNPs based on the information that has been provided?

2c/ Can you think of which other QC metrics should have been included in the replication file?

3a/ Re-run the meta-analysis using the updated replication set 1 results file (repl_1_updated.gwama).

```
GWAMA -i meta_1_updated.gwama.in -o meta_1_updated_results
```

Is there more/less evidence for association for your SNPs of interest?

3b/ What does ? mean in the effects column?

3c/ If you had funds to prioritise 3 of the 5 SNPs for *de novo* replication which SNPs would you prioritise and why?

4/ You decided to genotype all 5 SNPs in an independent case/control study sample and have got the results of the analysis (repl_2.gwama).

4a/ Run the meta-analysis using all three studies (study_1.gwama, repl_1_updated.gwama, repl_2.gwama). Is there more/less evidence for association with your SNPs of interest?

4b/ Which of the SNPs (if any) shows significant association with the disease? Can you rule out association for any of the SNPs examined?

4c/ What can you tell about the between-study heterogeneity for the SNPs examined in the three studies?

5/ Re-run the meta-analysis for the three studies using random effects correction. Do the results change, why?

6/ If you were examining association with a quantitative trait how would the gwama study files differ? Which command line option would you use?

7/ You examined 5 SNPs which you prioritised based either on statistical significance from your study (rs3, rs4, rs5) or biological candidacy (rs1, rs2). Based on the evidence from the combined analysis what can you conclude about these two approaches?