**Meta-Analysis Practical - Instructions**

**Thursday 25th June 2015**

**Practical**

A consortium comprising five independent sites has been set up to (a) replicate findings at site 1 and (b) discover additional signals. In this practical you will calculate meta-analytic estimates of effect sizes across the five sites at the 3 top SNPs in your previous analysis and produce forest plots showing your results in order to establish whether your associations are replicated.

Tables 1-3 show total sample size (N), risk allele frequency, risk allelic odds ratios with associated standard errors and *P*-values for top GWAS SNPs rs112820984, rs376047 and rs57681667, respectively, in 5 separate studies: the GWAS we performed yesterday, a replication study, two Central African (CA) studies and a study from East Africa (CA).

We'll start by performing a *fixed-effect meta-analysis* using estimated effect sizes and standard errors estimated in each study. (Here, the analysis is a 'fixed-effect' analysis because it assumes that the true effect size is the same across studies - i.e there's no between-study heterogeneity in the genetic effect. This is the simplest meta-analysis to perform and the easiest to interpret.)

**Loading the data**

To get started, open RStudio and navigate to the practical folder, e.g. by using setwd()

setwd("/media/ubuntu/data/GEIA/Practicals/09\_Meta\_Analysis/")

We've stored the data from the studies in a file in the practical folder; let's load this and examine it:

data <- read.table("meta\_analysis\_practical\_data.txt", sep="\t", na.strings = "NA", header = T)

View(data)

The columns in the table give the name and alleles of the SNPs tested, case and control counts, the frequency of the non-reference allele in controls, the estimated odds ratio and standard error, and the p-value. Data for each SNP is also shown in the tables at the end of this document. (These are typical fields to report in a meta-analysis.) Before getting started it's worth doing some sanity checks on the data.

**Q. Check that the odds ratios and standard errors for the GWAS study match those in the GWAS scan.** If you completed yesterday's practical, you should be able to do this as follows:

gwas.data = read.table( "../07\_GWAS\_analysis/pccorrected-test.assoc.logistic", hea=T)

gwas.data[ gwas.data$SNP %in% data$snp & gwas.data$TEST == 'ADD', ]

**Q. In conducting a meta-analysis it's very important to make sure that all the effect sizes are reported consistently across studies. In particular, effects must be reported for the same alleles, otherwise the results will be nonsense. Look at the data and correct any problems with allele coding.** Note: if you have to swap the alleles for any studies, you should also fix the frequency value (replace freq with 1-freq) and the odds ratio (replace OR with 1/OR) for that row. The other fields can remain unchanged.

**Preparing the data**

The aim is to compute a meta-analysis effect size estimate, standard error and p-value by combining the per-study effect size estimates. Let’s do this for the top SNP in your GWA analysis. First create a subset of the data including output for the first SNP only.

snp <- "rs112820994"

data.1 <- data[which(data$snp==snp),]

Because effects were estimated from logistic regression, the appropriate scale to work on is the log-odds scale - i.e. we take log(OR). Let's make a vector of log ORs for the first SNP...

OR <- data.1$OR

beta <- log(OR)

…and a vector of standard errors:

se.beta <- data.1$std\_err

We'll use these values in the analysis, but for presentation we'll also translate these back into a confidence interval on the odds ratio scale:

CI.low <- exp(beta - qnorm(0.975)\*se.beta)

CI.high <- exp(beta + qnorm(0.975)\*se.beta)

CI <- paste(round(CI.low,2), "-", round(CI.high,2), sep="")

These values are currently missing in your output data. Now add these values to your data

 data.1$ci95 <- CI

To present the meta-analysis, we'll also compute the combined sample count and allele frequency:

ctls <- data.1$n\_ctls

cases <- data.1$n\_cases

fqs <- data.1$ctl\_freq

The combined case and control sample sizes are given by summing the study-specific sample sizes.

Nctls <- sum(ctls)

Ncases <- sum(cases)

To compute the combined control non reference allele frequency, we compute a weighted average frequency where the weights are the sample sizes.

combined.fq <- sum(fqs\*ctls)/Nctls

**Q. Use the top-right 'Environment' panel, or the print() function, to check these variables have sensible values.**

**Computing the meta-analysis**

Next we will perform a statistical meta-analysis. To do this we'll use the “mvmeta” package to compute a meta-analysis odds ratio and standard error, and then compute a P-value from these. mvmeta can perform both fixed and random-effects meta-analysis, and supports both uni- and multivariate traits. In this practical we'll only use the fixed-effect method. Load the library and look at its help page for more info:

library(mvmeta)

help(mvmeta)

As input the mvmeta() function requires a vector of estimated effect sizes (log odds ratios) at each site and a corresponding vector of estimated variances with method set to fixed. We've computed these values already; to meta-analyse the data for SNP rs112820994 :

m1 <- mvmeta(beta, S=se.beta^2, method="fixed")

summary(m1)

To obtain just information about the regression coefficients we could type:

coef = summary(m1)$coef

View(coef)

The output should look something like this:

Estimate Std. Error z Pr(>|z|) 95%ci.lb 95%ci.ub

0.4200257 0.03367056 12.47457 0 0.3540327 0.4860188

Here, Estimate and Std. Error are the meta-analysis estimate of the log odds ratio and its standard error under the fixed-effect model, z is the z-score (= the estimate divided by standard error), and Pr(>|z|) is the meta-analysis P-value. A 95% confidence interval on the log-odds scale is also given.

**Note:** In this output, the *P*-value is given as zero. In R, what this actually means is that the p-value is less than 2.2x10-16. We'll compute the P-value via a different method below which circumvents this problem.

To compute the *P*-value we compare the log odds ratio to a normal distribution. We compute a *2-sided P-value* because we tested for an effect in either direction.

pvalue2sided = 2\*pnorm(-abs(coef[3]))

Congratulations! You've completed meta-analysing the first SNP. Let's store the results in a new row in the data frame:

result = data.frame(

snp = snp,

ref\_allele = "T",

nonref\_allele = "C",

site = "Combined",

n\_cases = Ncases,

n\_ctls = Nctls,

ctl\_freq = round( combined.fq, 2 ),

OR = round( exp(coef[1]), 2 ),

std\_err = round( coef[2], 2 ),

ci95 = paste(round(exp(coef[5]),2), "-", round(exp(coef[6]),2),sep=""),

p\_value = pvalue2sided

)

data.1 = rbind( data.1, result )

And look at the result:

View(data.1)

If all has gone well, you should see a new row (with site = 'Combined') with the information from your meta-analysis.

**Creating a forest plot**

A table like the one we computed above is useful, but it's not very easy to spot trends on a table like that. To visualise the data, we'll create a forest plot that shows the effect sizes and their confidence intervals. We've put a script called create.snp.forestplot.R in the scripts/ directory. Load this into Rstudio (e.g. using File->Open menu) and run it.

Let's look at the arguments required for create.snp.forestplot

 args(create.snp.forestplot)

The arguments to be input are:

*beta* = vector of beta estimates of effect for selected SNPs

*se* = standard errors associated with beta estimates of effect

*fq* = allele frequency for selected SNPs

Optional arguments are:

*title* = title to appear on top of plot

*sitelabels* = vector of study labels to appear on the y-axis of the plot

Using the summary dataframe data.1 that you have created, you should now be able to produce a forest plot with this command:

create.snp.forestplot(

beta=log(data.1$OR),

se=data.1$std\_err,

fq=data.1$ctl\_freq,

title = snp,

sitelabels = data.1$site

)

Note: You may need to set the margins to allow the study labels to properly print.

 par(oma=c(4,8,2,2))

Note: You may also wish to reorder the data so that the meta-analysis data is at the bottom rather than the top of the plot.

data.1 <- data.1[6:1,]

Save the forest plot you've made for later reference. **(**E.g. you can do this by clicking 'Export' above the plot and then 'Save as image'. Make sure the directory is set to the Meta-analysis practical directory, and set the filename to the name of the SNP.)

**Q. Which allele is protective (i.e. reduces risk of disease) at this SNP and which is risk? Is the risk or the protective allele at a higher frequency?**

**Q. A fixed-effect analysis assumes that all the study-specific ORs are consistent with a single, true effect. Do all the study-specific odds ratios seem consistent with the meta-analysis odds ratio? Why do you say this?**

**Repeating the analysis for the remaining 2 SNPs**

Repeat the above analysis for the remaining 2 snps in the data. Save the completed data frame in a text file:

 completed\_data <- rbind(data.1, data.2, data.3)

write.table(completed\_data,"./solutions/completed\_data.txt",

sep="\t",row.names= F)

Print all your forest plots on one page. To do this adjust the mfrow setting and run each forest plot in succession.

 par(mfrow=c(3,1))

**Q. Do the forest plots for these SNPs look similar to the first one? Decide whether you think each SNP represents a real association.**

**Q. Are the studies well powered to replicate effects seen in your GWAS?**

**Q. Is there any significant heterogeneity of results across sites? Can you suggest the reasons for any heterogeneity?**

**Q. Hence, do you think that we should be using a fixed-effects model?**

A random-effects model can be fit by using method="reml" instead of method="fixed" in mvmeta(). If you have time, fit a random-effects model to one of the SNPs and inspect the output. We will not cover random-effects meta-analysis in detail here, but in brief it models the true effects as being possibly different but correlated between sites. The function will estimate the between-site correlation as well as giving an overall P-value.

**Congratulations! Your GWAS study is now complete.** (Your final task is to write it up and submit it to a major journal…good luck!)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Site** | **N****Cases** | **N****Ctls** | **Non-ref Allele Freq** | **Allelic OR****(Non-ref Allele)** | **Standard error** | **95%CI** | **P** |
| GWAS | 1297 | 1593 | 0.09 | 2.002 | 0.088 |   | 3.614x10-15 |
| Replication | 900 | 1080 | 0.17 | 1.480 | 0.080 |   | 1.140x10-6 |
| CA Study 1 | 500 | 600 | 0.10 | 1.620 | 0.131 |   | 2.110x10-4 |
| CA Study 2 | 700 | 840 | 0.20 | 1.540 | 0.085 |   | 3.420x10-7 |
| EA Study 1 | 2300 | 2760 | 0.17 | 1.390 | 0.050 |   | 4.620x10-11 |
| Combined |   |   |  |   |   |   | 1  |

Table 1 rs112820994. Reference allele = T; non-reference allele=C

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Site** | **N****Cases** | **N****Ctls** | **Non-ref Allele Freq** | **Allelic OR****(Non-ref Allele)** | **Standard error** | **95%CI** | **P** |
| GWAS | 1297 | 1593 | 0.81 | 0.71 | 0.069 | 0.62-0.81  | 3.83x10-7 |
| Replication | 900 | 1104 | 0.75 | 0.85 | 0.072 | 0.74-0.98  | 3.10x10-2 |
| CA Study 1 | 500 | 613 | 0.70 | 0.99 | 0.093 | 0.83-1.19  | 9.14x10-1 |
| CA Study 2 | 700 | 858 | 0.90 | 0.91 | 0.118 |  0.72-1.15 | 4.08x10-1 |
| EA Study 1 | 2300 | 2820 | 0.80 | 1.51 | 0.054 | 1.36-1.68  | 9.77x10-15 |
| Combined |  5697 |  6988 |  0.80 |  1.03 |  0.030 | 0.97-1.1  | 3.30x10-1  |

Table 2 rs376047. Reference allele = C; non-reference allele=T

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Site** | **N****Cases** | **N****Ctls** | **Non-ref Allele Freq** | **Allelic OR****(Non-ref Allele)** | **Standard error** | **95%CI** | **P** |
| GWAS | 1297 | 1592 | 0.04 | 1.974 | 0.124 |   | 4.41x10-8 |
| Replication | 900 | 1104 | 0.05 | 1.050 | 0.140 |   | 7.24x10-1 |
| CA Study 1 | 500 | 613 | 0.03 | 0.700 | 0.276 |   | 1.92x10-1 |
| CA Study 2 | 700 | 858 | 0.11 | 0.990 | 0.115 |   | 9.30x10-1 |
| EA Study 1 | 2100 | 2575 | 0.09 | 1.010 | 0.072 |   | 8.90x10-1 |
| Combined |   |   |   |   |   |   |   |

Table 2 rs57681667. Reference allele = A; non-reference allele=G