Genome-Wide Meta-analysis

Outline: Meta-Analysis for genetic studies

- Background
- Single-variant methods
- Gene-based test methods
- Issues specific to meta-analysis of genetic studies
- Using METAL software to perform meta-analysis of genome-wide association scans
- extra: Meta-analyzing gene-based tests using raremetalworker and raremetal

Background

- Have a collection of k primary studies ($i = 1\ to\ k$)
  - All studies tested the same null hypothesis, $\mathrm{H_O}$: $\beta = 0$
  - Each study i has estimate $\hat{\beta}_i$ and a p-value p_i for this hypothesis
- Want a single test of H_0 using the data from all k studies at significance level α
  - Usually: also want a pooled β estimate with a confidence interval
- Key assumption – independence of study results
  - There should be no subjects who are included in more than one study

Background

- First step: identify a consistent effect or association measure from each study in the collection to be pooled
  - Regression coefficient ($\hat{\beta}_i$) and its SE
  - p-value from test of association along with the direction of effect
- Next step: use the individual study effect estimates or association measures to arrive at a summary value

Background - typical scenario

- SNP of interest has two alleles, A and G
- Want to estimate the effect of the G allele on a particular phenotype
  - Fasting glucose (β coefficient from linear regression)
  - Diabetes risk (β coefficient from logistic regression)
- Use additive coding and a regression model to get an estimate of β in each study
- Use meta-analysis approaches to get a pooled estimate of association and effect (with SE and thus confidence interval) using estimates from all studies

Outline: Meta-Analysis for genetic studies

- Background
- Single-variant methods
- Gene-based test methods
- Issues specific to meta-analysis of genetic studies
- Using METAL software to perform meta-analysis of genome-wide association scans
- Meta-analyzing gene-based tests using raremetalworker and raremetal

Single variant meta-analysis methods

- Combining p-values (e.g., Fisher's method, Probit method)
  - Look to see if the set of p-values p_i are consistent with k random observations from a uniform distribution
  - Does not take direction of association into account → poor choice for genetic studies
- Fixed effects methods
  - Inverse-variance weights of regression estimates
  - Weighted Z-score (= signed p-value)
- Heterogeneity and random effects model

Combining effect sizes

- In medicine, combining effect sizes, or test statistics, rather than p-values, has been the main form of meta-analysis
- For case-control studies, the odds ratio (OR) or log(OR) (β estimate from regression) is the usual effect size
- For continuous traits, β estimate from regression is the usual effect size
- Anything with a standard error can be used as an effect size

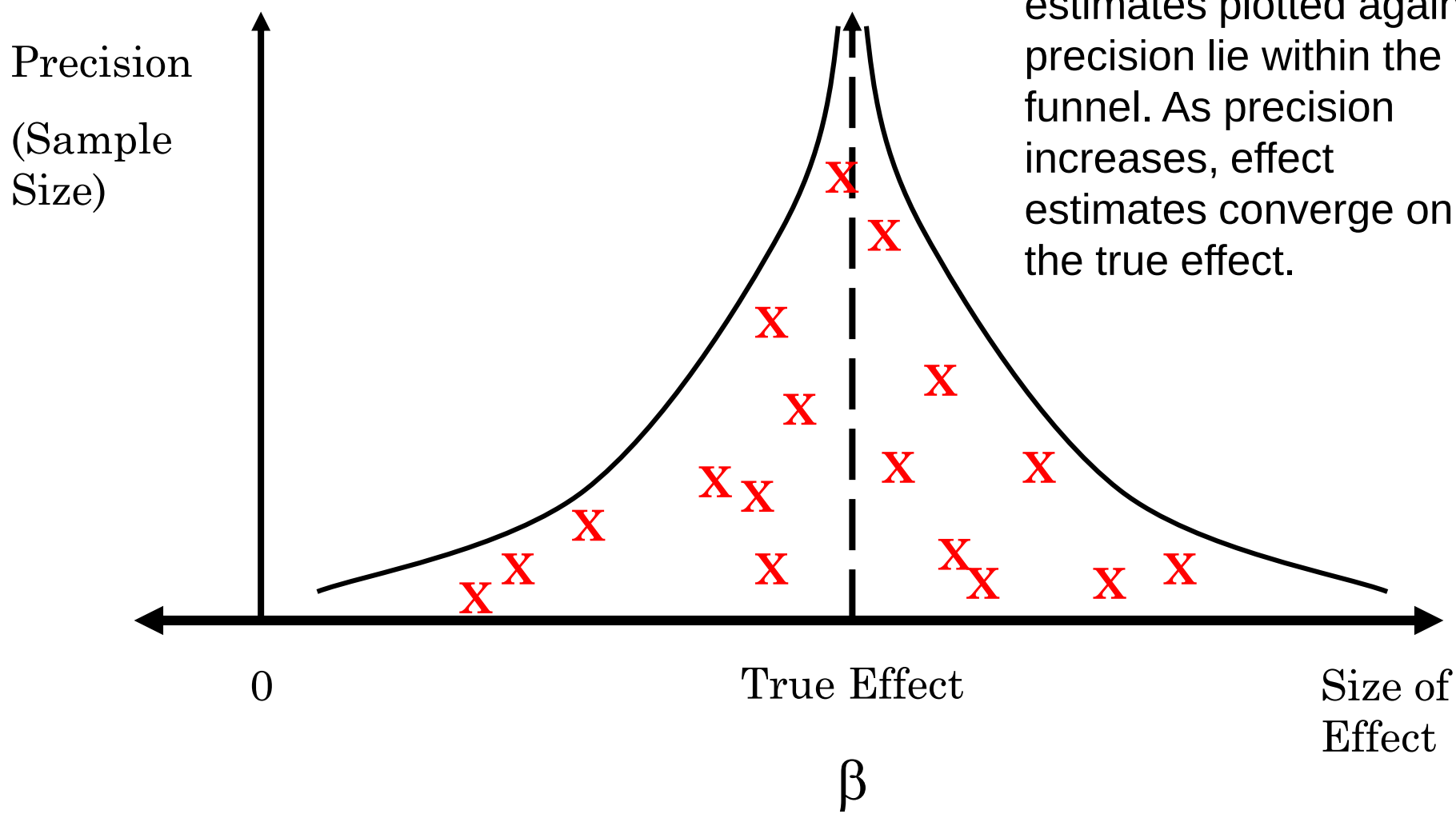
Fixed effects

- Combine effects from primary studies
  - Sampling error gives within-study error (due to sample size)
  - Is there also variability in effect between studies? (=Heterogeneity)
    - if no, use fixed effects analysis
    - If yes, use random effects analysis
  - Both analyses compute weighted averages of the study effects, but the weights are different

Fixed effects

- Assumes that all studies are estimating the same underlying population effect parameter β
- The expected value for each of the effect sizes in the set of primary studies is β
  - The only reason studies have different effect estimates is random sampling error
  - If any primary study had an infinitely large sample size, then the effect size estimate would exactly equal β

Fixed effects analyses



Fixed effects analysis assumes study effect estimates plotted against precision lie within the funnel. As precision increases, effect estimates converge on the true effect.

Inverse variance weighting

- Each study i provides an estimate $\widehat{\beta_i}$ of the population value β with a standard error estimate $SE(\widehat{\beta_i})$
- For the inverse variance weighting method, we assume that the $\widehat{\beta_i}$'s roughly follow a normal distribution
- Since all studies estimate β, it makes sense to estimate β by averaging the study estimates:
- $\hat{\beta} = w_1\hat{\beta}_1 + w_2\hat{\beta}_2 + \ldots + w_k\hat{\beta}_k = \sum_{i=1}^{k} w_i\hat{\beta}_i$
- w_{i} is the weight for study i
  - $0 \leq w_{\mathrm{i}} \leq 1$ for each i
  - $w_1 + \ldots + w_{\mathrm{k}} = 1$

Inverse variance weighting

- Studies are weighted in proportion to their precision (reciprocal of SE squared)

$$w_i \propto \frac{1}{SE\left(\hat{\beta}_i\right)^2}$$

  - Big study →smaller SE
  - Small study → bigger SE

Inverse variance weighting

- The optimal weight is:

$$w_i = \frac{\frac{1}{SE\left(\hat{\beta}_i\right)^2}}{\sum_{j=1}^{k} \frac{1}{SE\left(\hat{\beta}_\mathrm{j}\right)^2}}$$

- Denominator is the same for each study
- Ensures weights sum to 1

Inverse variance weighting

- The standard error of the combined estimate is

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{SE\left(\hat{\beta}_i\right)^2}}}$$

- The meta-analysis test statistic to test $\mathrm{H_O}$: $\beta = 0$ is $\mathrm{Z}_{\mathrm{meta}\beta} = \frac{\hat{\beta}}{SE(\hat{\beta})}$

Weighted Z-score approach

- Sometimes, the β estimates from each study do not measure exactly the same thing
- For example, the effect estimate from each study may not be on the same scale or comparable
  - Different assay used to measure trait
  - Different technologies used in each study
    - CES-D scores (60 vs 20 point scale)
    - Quantitative and semi-quantitative MRI
  - Different transformation applied to trait
    - rank normalized vs log transformed

→Instead of combining effect estimates, can combine test statistics

Weighted Z-score approach

- The weighted (signed) Z-score approach takes direction of association and sample size into consideration, without using effect estimates
- For each study with sample size n_i, the p-value for association is converted to a standard normal deviate (Z_i), with the sign of the normal deviate corresponding to the direction of the association
  - positive = effect allele increases risk or average trait value
  - negative = effect allele decreases risk or average trait value
- The null hypothesis is Ho: no association between SNP and phenotype
- The meta-analysis test statistic is $Z_{meta} = \frac{\sum \sqrt{n_i} Z_i}{\sqrt{\sum n_i}}$

Weighted Z-score approach

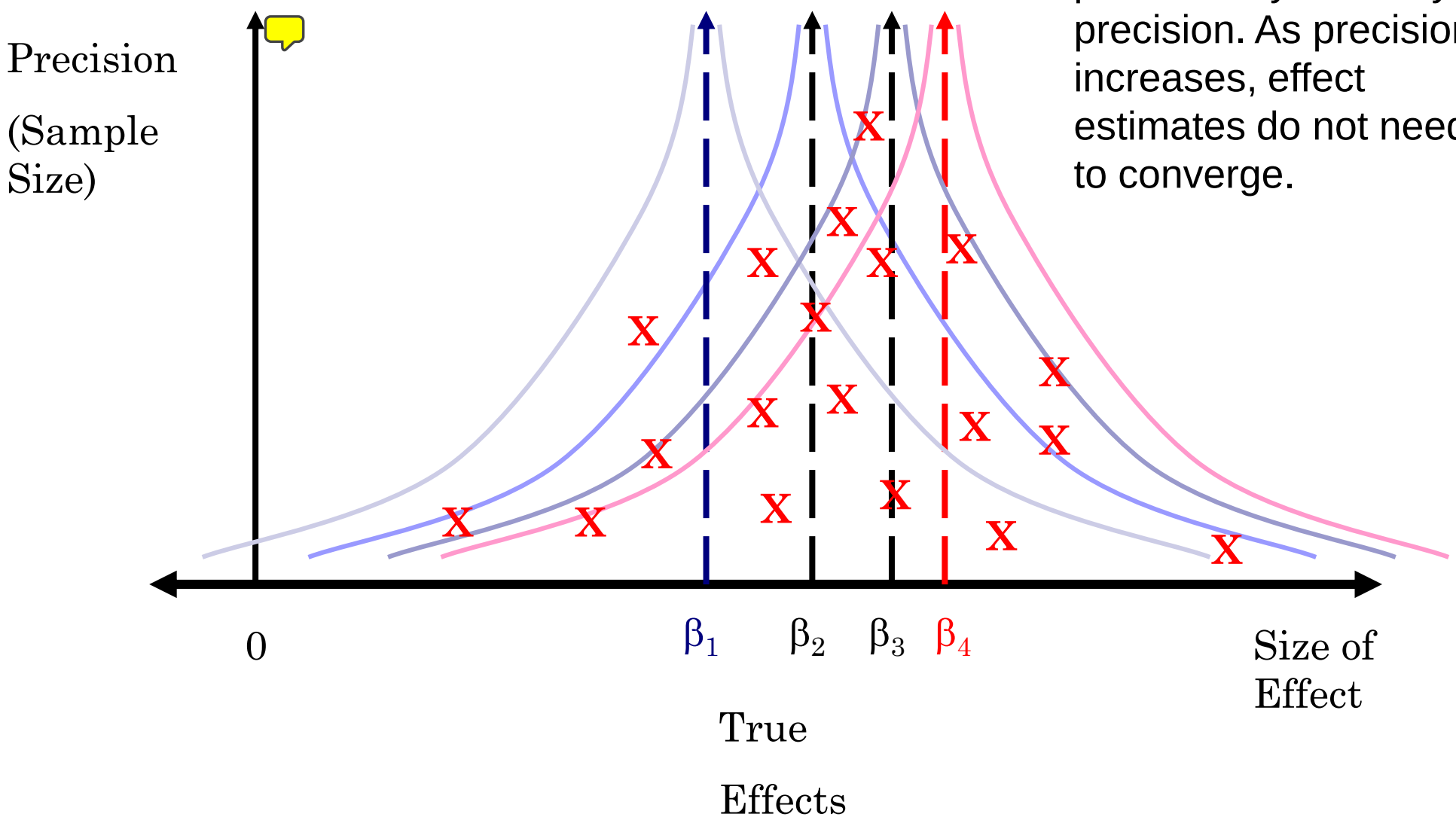
- Compare Z_{meta} to the standard normal distribution to determine significance
- No pooled effect size estimate produced

Heterogeneity

- With heterogeneity there is no single population effect size β
  - Each study effect size $\hat{\beta}_i$ is an estimate of the study-specific population effect β_i
  - The effect size for study i approaches β_i as the sample size increases
- As the individual study sample sizes increase, the effect sizes for all of the studies would ***not*** approach the same value
  - Some of the β_i might be the same, but not all of them

Heterogeneity

With heterogeneity there could be several funnels in the plot of study effect by precision. As precision increases, effect estimates do not need to converge.



Heterogeneity

- What are some possible causes of heterogeneity?

Testing for heterogeneity

- Most common test is Cochran's Q-test
- The formula for the Q-test statistic is

$$Q = \sum_{i=1}^{k} \frac{1}{SE\left(\hat{\beta}_i\right)^2} \left(\hat{\beta}_i - \hat{\beta}\right)^2$$

- $\hat{\beta}$ is the fixed effect combined effect size
- $\hat{\beta}_i$ is the effect size for study i
- $SE(\hat{\beta}_i)$ is the standard error of the estimate $\hat{\beta}_i$ from study i
- With homogeneous studies (H_0),
  - Q has a chi-square distribution with k-1 degrees of freedom
  - Q is compared to the upper 0.05 cutoff for a chi-square with $k - 1$ df to determine if the heterogeneity is significant

Testing for heterogeneity

- The form of the Q-test statistic is similar to the computation of variance
  - Differences between the study effect size and the fixed effects combined effect size are squared
  - The squared differences are weighted by the inverse of the effect size standard error squared and summed over all the studies

$$Q = \sum_{i=1}^{k} \frac{1}{SE\left(\hat{\beta}_i\right)^2} \left(\hat{\beta}_i - \hat{\beta}\right)^2$$

Testing for heterogeneity

- Testing for heterogeneity is somewhat problematic due to low power
- The test can still be useful to confirm the presence of heterogeneity
  - It may not be significant every time there is heterogeneity
  - It tends be significant too often if there are many studies (100+)
  - When significant with a small number of studies it is a strong indication of heterogeneity

Measuring Heterogeneity

- $I^2 = [Q - (k - 1)]/Q$

  where Q = Cochran's heterogeneity statistic,

  k = number of studies

- I^2 is the proportion of total variation across studies that is due to heterogeneity rather than chance (sampling error)

- Negative values of I^2 are set to to zero so that I^2 lies between 0 and 1

Higgins JP, Thompson SG. (2002) Quantifying heterogeneity in a meta-analysis. Stat Med. 21:1539-58.

Random effects model

- In the presence of heterogeneity, the fixed effect model is not appropriate
- As an alternative, in a random effects model, the β_i are assumed to follow a normal distribution with
  - mean = β
  - variance = τ^2
- These β_i are called random effects
- Incorporates heterogeneity in the meta-analysis model

Random effects model

A 2-level model

- Level 1: study effect size $\hat{\beta}_i$ provides an unbiased estimate of β_i, the population effect of study i

$$\hat{\beta}_i | \beta_i, s_i^2 \sim N(\beta_i, s_i^2)$$

where s_i^2 is the within study variance

- Level 2: population effect sizes β_i have a normal distribution around the central value β with variance

$$\beta_i | \beta, \tau^2 \sim N(\beta, \tau^2)$$

where τ^2 is the between-study variance

Random effects model

- The variance of an effect size estimate is more complicated than in the fixed effects model
- The variability of an effect size estimate $\hat{\beta}_i$ has two components in this model
  - Variability of $\hat{\beta}_i$ as an estimator of β_i: the *within study* variability s_i^2
    - Note that $SE(\hat{\beta}_i)$ provides an estimate of s_i
  - Variability of β_i around β: the *between study* variability (τ^2)

Random effects model

- With an estimate of τ^2
  - Make the study weights reflect both within and between study variation
  - For study i, the random effects weight is proportional to the reciprocal of the sum of the study standard error squared and the between study variability
  - $w_i^* \propto \frac{1}{SE(\widehat{\beta}_i)^2 + \tau^2}$

Random effects model

- Sum these values for all of the primary studies and divide each by the total

$$w_i^* = \frac{\dfrac{1}{SE\left(\hat{\beta}_i\right)^2 + \tau^2}}{\sum_{j=1}^{k} \dfrac{1}{SE\left(\hat{\beta}_j\right)^2 + \tau^2}}$$

- We use w^* to denote the random effects weights to keep them separate from the fixed effects weights

Random effects model

- So, in a random effects model

$$\hat{\beta}^* = \sum_{i=1}^{k} w_i^* \hat{\beta}_i$$

- And the standard error of the random effects combined estimate is

$$SE\left(\hat{\beta}^*\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{SE\left(\hat{\beta}_i\right)^2 + \tau^2}}}$$

How to estimate τ^2

- Usually use DerSimonian and Laird method of moments estimate of τ^2 that is based on the Q-test statistic

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum_{i=1}^{k}\frac{1}{SE\left(\hat{\beta}_i\right)^2} - \frac{\sum_{i=1}^{k}\frac{1}{SE\left(\hat{\beta}_i\right)^4}}{\sum_{i=1}^{k}\frac{1}{SE\left(\hat{\beta}_i\right)^2}}}$$

Single variant meta-analysis summary

- Two common fixed-effect approaches for meta analyses of association for individual variants in genetic studies
  - Weighted Z approach
  - Fixed Effects Regression Coefficients
- Both methods assume homogeneity of effects across studies
- Random effect models may be more appropriate when study effects are not homogeneous
  - Can measure and test for heterogeneity, but power is limited

Outline: Meta-Analysis for genetic studies

- Background
- Single-variant methods
- Gene-based test methods
- Issues specific to meta-analysis of genetic studies
- Using METAL software to perform meta-analysis of genome-wide association scans
- Meta-analyzing gene-based tests using raremetalworker and raremetal

Review: Rare variant tests – general framework

- Phenotype y_i has a distribution in the quasi-likelihood family
- Generalized linear model is:

$$h(\mu_i) = \alpha_0 + \alpha' \boldsymbol{X}_i + \beta' \boldsymbol{G}_i$$

$h(\mu_i) = \mu$ for a continuous trait

$h(\mu_i) = \mathrm{logit}(\mu)$ for a dichotomous trait

α_0: intercept

$\alpha' = (\alpha_1, \ldots, \alpha_q)'$: regression coefficients for covariates $\boldsymbol{X}_i$

$\beta' = (\beta_1, \ldots, \beta_m)'$: regression coefficients for allele counts $\boldsymbol{G}_i$

$i = 1, \ldots, n$ indexes individuals

$j = 1, \ldots, m$ indexes variants

Review: score statistic approach

- The score statistic for variant j is:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \widehat{\mu_i})$$

where $\widehat{\mu_i}$ is the estimated mean of y_i under H_0: $\boldsymbol{\beta} = 0$, i.e., it is estimated using the null model:

$$h(\mu_i) = \alpha_0 + \alpha' \boldsymbol{X}_i$$

- S_j is positive(negative) when variant j is associated with increased(decreased) disease risk or trait value

Burden Test

- The score test to test H_0: $\beta = 0$ is:

$$Q_{\mathrm{burden}} = \left(\sum_{j=1}^{m} w_j S_j \right)^2$$

- Q_{burden} has a chi-square distribution with 1 df
- How one defines the weights w_j will impact what types of variants are included in the score

Sequence Kernal Association Test (SKAT)

- In our model: $h(\mu_i) = \alpha_0 + \alpha' \boldsymbol{X}_i + \beta' \boldsymbol{G}_i$

we assume that the β_j are random, and have a distribution with mean 0 and variance $w_j^2 \tau$

- The null hypothesis is then $H_0: \tau = 0$, and is equivalent to testing

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

- The SKAT test statistic is:

$$Q_{\mathrm{SKAT}} = \sum_{j=1}^{m} w_j^2 S_j^2$$

SKAT Statistic

- Under the null hypothesis, $Q_{\mathrm{SKAT}} \sim$ mixture of χ^2 distributions
  - mixture depends on weights (W) and on LD between variants
  - p-values can be computed analytically -- without permutation
- Since SKAT is a function of S_j^2 rather than S_j, it is robust to groupings of variants that include positive and negative effects

Meta-analysis

- To perform a meta-analysis across $k = 1, \ldots, K$ studies, each study provides:
  - Allele frequency p_j for each variant $j = 1..m$
  - Score statistic S_{kj} for each variant
  - Between-variant relationship matrix

$$\boldsymbol{\Phi}_k = \boldsymbol{G}_k' \boldsymbol{P}_k \boldsymbol{G}_k$$

- $\boldsymbol{G}_k$ is the genotype matrix, and $\boldsymbol{P}_k$ is a projection matrix accounting for the fact that the effects of covariates are estimated
- Think of $\boldsymbol{\Phi}_k$ is a covariance matrix of the genotypes -- a way to measure the linkage disequilibrium

Meta-analysis

- Under assumption that that study cohorts share the same set of causal variants with the same effect size, the meta-analysis test statistics are

$$Q_{meta-SKAT} = \sum_{j=1}^{m} \left(\sum_{k=1}^{K} w_{kj} S_{kj} \right)^2$$

$$Q_{meta-burden} = \left(\sum_{j=1}^{m} \sum_{k=1}^{K} w_{kj} S_{kj} \right)^2$$

Meta-analysis

- w_{kj} is the weight for variant j in study k
- Distribution of the test statistics is determined as for single-study analysis:
  - $Q_{meta-burden}$ has a χ^2 distribution with 1 df
  - $Q_{meta-SKAT}$ is a mixture of χ^2 distributions that depends on the weights and on the correlation (LD) among the variants

Multi-variant gene-based tests

- Variant weights
  - Same w_j should be used by all studies if $\hat{\beta}$ of a variant score is meta-analyzed
  - w_j is often dependent on allele frequency – best to use combined study allele frequency; this can be easily managed with the sharing of scores, allele frequencies, and covariance matrix

Multi-variant gene-based tests

- SNP subsets
  - Need to pre-specify the variant subsets to get covariance matrices for the correct sets of variants
  - Different software deals with how to determine which variants to include together differently
    - Need to be aware of how this is done for the method you will be using for meta-analysis

Outline: Meta-Analysis for genetic studies

- Background
- Single-variant methods
- Gene-based test methods
- Issues specific to meta-analysis of genetic studies
- Using METAL software to perform meta-analysis of genome-wide association scans
- Meta-analyzing gene-based tests using raremetalworker and raremetal

Meta-analysis issues specific to genetic studies

- DNA Strand
- Choice of effect allele

Meta-analysis of genetic studies: Strand issue

- Allele A is complementary to T, C is complementary to G
  - A forward strand ⇒ T reverse strand
  - C forward strand ⇒ G reverse strand
  - G forward strand ⇒ C reverse strand
  - T forward strand ⇒ A reverse strand
- For a SNP with alleles A/C, A/G, C/T or G/T, it is easy to reverse the strand
  - Substitute T for A, G for C, C for G and A for T
  - Genotype AA becomes TT, AC becomes GT, CC becomes GG, etc.

Meta-analysis of genetic studies: Strand issue

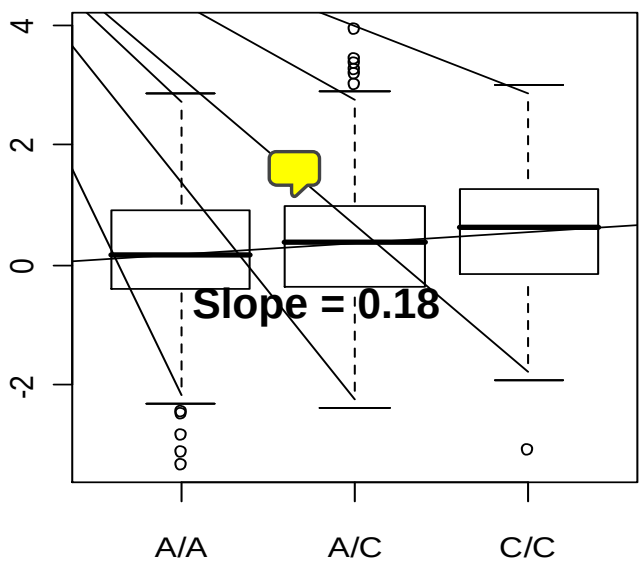
- For SNPs with alleles A/T or C/G, cannot make the correction without information about which DNA strand was typed
- Strand information is usually available from maker of genotyping platform used for genotyping
- For genome wide association study, need to clearly state what naming convention (forward/reverse, positive/negative, top/bottom) is requested
  - Will still need to double check A/T and G/C variants with frequencies near 0.50

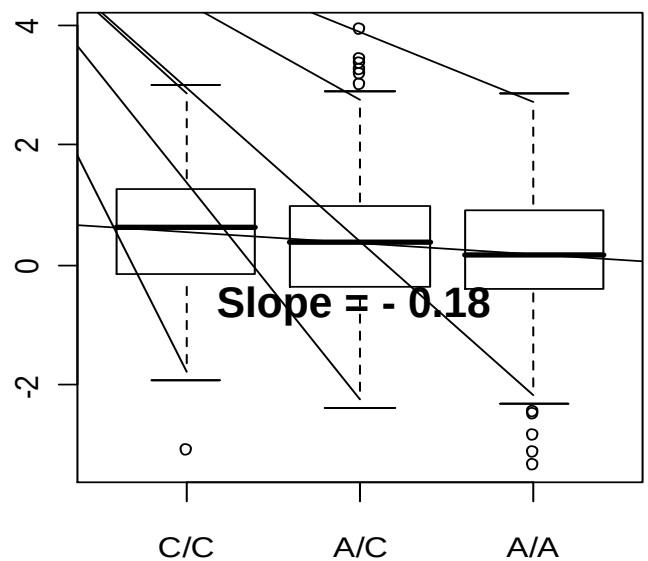
Meta-analysis of genetic studies: Choice of effect allele

- Most common genetic model used for meta-analysis: additive model
- Each study reports the association of trait with genotype coded as 0, 1 or 2 copies of the effect allele
- When studying genetic association, the choice of an effect allele may differ between studies

Meta-analysis of genetic studies: Choice of effect allele

- P-value not affected by choice of effect allele
- Estimate of effect size has opposite sign depending on choice of effect allele





- Want to take the direction of effect in consideration when combining results from multiple studies

Meta-analysis of genetic studies: Choice of effect allele

- Each study to be meta-analyzed needs to provide information about the effect and non-effect alleles
- For each SNP, an effect allele is selected
- For studies that coded the other effect allele, the sign of the effect estimate is "flipped"
  - Positive effect becomes negative effect estimates, and vice versa

Outline: Meta-Analysis for genetic studies

- Background
- Single-variant methods
- Gene-based test methods
- Issues specific to meta-analysis of genetic studies
- Using METAL software to perform meta-analysis of genome-wide association scans
- Meta-analyzing gene-based tests using raremetalworker and raremetal

Meta-analysis of genome-wide studies: software

- METAL is the most commonly used tool
  → we'll use this one today and in the homework
- Other meta-analysis software for genetic analysis will perform similar analyses
  - MetAble: http://www.genabel.org/packages/MetABEL
  - GWAMA: http://www.well.ox.ac.uk/gwama/
  - Metasoft: http://genetics.cs.ucla.edu/meta/
    - Fixed and random effects models
    - Has some unique models for heterogeneity

Using METAL for meta-analysis of genetic results

- URL: http://www.sph.umich.edu/csg/abecasis/Metal/
- Documentation: http://genome.sph.umich.edu/wiki/Metal_Documentation
- Change Log: http://www.sph.umich.edu/csg/abecasis/Metal/download/ChangeLog.html

METAL Documentation

https://genome.sph.umich.edu/wiki/METAL_Documentation

CENTER FOR STATISTICAL GENETICS

page | discussion | view source | history

quick links
- Abecasis Lab

teaching
- Biostatistics 666
- Biostatistics 615/815
- Biostatistics 602
- Short Workshops

navigation
- Main Page
- Community portal
- Current events
- Recent changes
- Random page
- Help

search

Go Search

toolbox
- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

METAL Documentation

(Redirected from Metal Documentation)

Contents [hide]

Useful Wiki Pages

There are a few pages in this Wiki that may be useful to METAL users. Here are links to key pages:

- The METAL Home Page

What does METAL do?

- Reads a command file that tells it:
  - Which files to open (one file per study, with all GWAS association statistics)
  - What columns to use for SNP name, allele coding, tests statistics, SEs, p-values, etc.
- Reads in all files, merges association information from each study
  - Checks coded alleles, flips alleles so all study test statistics are for the same coded allele
  - Computes the meta-analysis test statistics and p-values
  - Computes other statistics if requested (GC lambda, distribution of allele freqs, heterogeneity statistics etc)
- Usually, we perform meta-analysis genome-wide, so each input file will have millions of lines
  - Our example files are much smaller

Using METAL for meta-analysis of genetic results

- Copy the files in /project/bs859/class/class7 to your directory

```
cp project/bs859/class/class7  .
 commands.txt  covfiles  metal.txt  metal_GC.txt
raremetal_commands.txt  summaryfiles
```

metal.txt, metal_GC.txt: example command files that tells metal what to do
metal_commands.txt: commands for running metal

raremetal_commands.txt:
summaryfiles
covfiles

Using METAL for meta-analysis of genetic results

```
  ls $DATADIR
DGI_three_regions.txt          README.txt
three_region_map.txt
MAGIC_FUSION_Results.txt.gz  magic_SARDINIA.tbl
```

- Input files:
  - DGI_three_regions.txt
  - magic_SARDINIA.tbl
  - MAGIC_FUSION_Results.txt.gz
- Map file (not necessary for METAL run, useful for plotting):
  - three_region_map.txt

Example Data

- The example is drawn from a study of the genetics of glucose levels (Chen et al, Journal of Clinical Investigation, 2008; Prokopenko et al, Nature Genetics, 2009).
- Input files summarizing results for each of three studies (in a few interesting regions) :
  - DGI_three_regions.txt
  - MAGIC_FUSION_Results.txt.gz
  - magic_SARDINIA.tbl
- Each of the files uses a slightly different format to report results and this is accommodated in the metal.txt script

Meta-analysis of genetic studies: Software

- Weighted (signed) Z-score approach is default analysis implemented in software METAL
  - https://genome.sph.umich.edu/wiki/METAL_Documentation
  - Fixed effect inverse variance weighted approach also available in METAL
  - METAL has a heterogeneity testing option, but no random effect model option
- METAL will check for strand inconsistencies for SNPs with A/C, A/G, C/T and G/T alleles
  - Can handle A/T and C/G SNPs when strand information is provided, otherwise it uses given coding for all studies
- METAL will select a effect allele and change the sign of effect estimate for studies with opposite effect allele

METAL input files: DGI

more $DATADIR/DGI_three_regions.txt

CHR	POS	SNP	N	EFFECT_ALLELE	NON_EFFECT_ALLELE	EFFECT_ALLELE_FREQ	BETA	SE	r2	r2hat	P_VAL
2	169093837	rs2954939	1455	2	4	0.0621993	0.03369	0.07732	0.0001307	0.873	0.6698
2	169095689	rs12619614	1455	3	2	0.364948	0.01834	0.03898	0.0001523	0.947	0.6453
2	169095851	rs13415004	1455	4	2	0.189347	0.01141	0.04729	4.00E-05	0.955	0.8134
2	169095873	rs2724164	1455	3	2	0.0762887	-0.04199	0.06956	0.0002508	0.942	0.5547
2	169097055	rs11681374	1455	4	2	0.0343643	-0.0002721	0.1018	4.91E-09	0.928	0.9979
2	169097357	rs7584770	1455	2	3	0.0250859	0.1027	0.1188	0.0005143	0.863	0.3975
2	169097717	rs4399687	1455	2	3	0.0316151	0.02606	0.1069	4.09E-05	0.629	0.8114

Which columns do we need to do a meta-analysis?

Note: alleles here are 1,2,3,4
METAL assumes these correspond to A,C,G,T
so 1=A, 2=C, 3=G, 4=T

METAL input files: SARDINIA

$ more $DATADIR/magic_SARDINIA.tbl

SNP	CHR	POS	Rsq	AL1	AL2	FREQ1	TRAIT	EFFECT	SE	H2	LOD	PVALUE
rs1002666	2	169303525	0.9818	T	C	0.876	glucose_ND	0.029	0.041	0.025	0.112	0.4718
rs1002667	2	169303321	0.9794	A	G	0.876	glucose_ND	0.032	0.041	0.030	0.132	0.4362
rs1003456	2	169813878	0.9371	T	A	0.767	glucose_ND	0.000	0.032	0.000	0.000	0.9887
rs10167161	2	169187702	0.8657	G	A	0.987	glucose_ND	-0.061	0.092	0.012	0.096	0.5066
rs10169232	2	169879590	0.9994	C	G	0.684	glucose_ND	-0.006	0.030	0.002	0.009	0.8364

Which columns do we need for a meta-analysis?

METAL input files: FUSION

```
$ zcat $DATADIR/MAGIC_FUSION_Results.txt.gz |more
CHR POS SNP IMPUTATION STRAND EFFECT_ALLELE NON_EFFECT_ALLELE FREQ_EFFECT N BETA SE r2 CHI_SQ PVALUE
7 43511971 rs3807518 GEN + 1 2 0.857 1233 0.037 0.057 0.033 0.4068 0.5236
7 43512117 rs10256312 1.021 + 1 3 0.77 1233 0 0.049 0 0.0001 0.9924
7 43512161 rs3807517 0.698 + 2 4 0.138 1233 0.088 0.068 0.187 1.6961 0.1928
7 43515309 rs1181578 0.973 + 2 4 0.808 1233 0.039 0.051 0.047 0.5875 0.4434
7 43517495 rs1100405 0.975 + 2 4 0.665 1233 0.047 0.042 0.098 1.2295 0.2675
7 43517659 rs1181580 0.973 + 4 2 0.808 1233 0.039 0.051 0.047 0.5900 0.4424
7 43518135 rs13222617 1.022 + 3 1 0.771 1233 -0.002 0.049 0 0.0024 0.9607
7 43520278 rs861961 0.974 + 3 2 0.808 1233 0.039 0.051 0.048 0.5919 0.4417
7 43521623 rs849176 0.974 + 1 3 0.623 1233 -0.002 0.043 0 0.0016 0.9682
7 43523789 rs10225076 1.007 + 1 3 0.774 1233 -0.007 0.049 0.002 0.0178 0.8939
```

Which columns do we need for a meta-analysis?

Using METAL

- METAL requires a control file
- This file tells METAL what files have the individual study results, and what columns in those files are the effect size and SE, coded and reference allele designations, p-value, etc
- METAL will provide information about allele frequencies across studies if you tell it which column has the study-specific allele frequencies
  - Useful for QC/checking results

METAL command file

```
$ cat metal.txt
. . .
MARKER   SNP
WEIGHT   N
ALLELE   EFFECT_ALLELE NON_EFFECT_ALLELE
FREQ     EFFECT_ALLELE_FREQ
EFFECT   BETA
STDERR   SE
PVAL     P_VAL

PROCESS /projectnb/bs859/data/METAL_example/DGI_three_regions.txt
```

METAL command file

```
MARKER   SNP
ALLELE   EFFECT_ALLELE NON_EFFECT_ALLELE
FREQ     FREQ_EFFECT
WEIGHT   N
EFFECT   BETA
STDERR   SE
PVAL     PVALUE
PROCESS /projectnb/bs859/data/METAL_example/MAGIC_FUSION_Results.txt.gz

MARKER   SNP
DEFAULTWEIGHT  4108
ALLELE   AL1 AL2
FREQ     FREQ1
EFFECT   EFFECT
STDERR   SE
PVAL     PVALUE
PROCESS /projectnb/bs859/data/METAL_example/magic_SARDINIA.tbl

ANALYZE
```

Running METAL

- type: $metal metal.txt > metal.log
- Output files:
  - METAANALYSIS1.TBL
    - Meta-analysis results
  - METAANALYSIS1.TBL.info
    - Information about the cohorts
- First step: check log file for error messages and warnings

Log file

```
$ cat metal.log
MetaAnalysis Helper - (c) 2007 - 2009 Goncalo Abecasis
This version released on 2011-03-25
…
# Processing commands in metal.txt ...
## Set marker header to SNP ...
## Set weight header to N ...
## Set allele headers to EFFECT_ALLELE and NON_EFFECT_ALLELE ...
## Set frequency header to EFFECT_ALLELE_FREQ ...
## If you want frequencies to be averaged, issue the 'AVERAGEFREQ ON' command
## Set effect header to BETA ...
## Set standard error header to SE ...
## Set p-value header to P_VAL ...
################################################################################
## Processing file '/projectnb/bs859/data/METAL_example/DGI_three_regions.txt'
## Processed 2417 markers ...

## Set marker header to SNP ...
## Set allele headers to EFFECT_ALLELE and NON_EFFECT_ALLELE ...
## Set frequency header to FREQ_EFFECT ...
## If you want frequencies to be averaged, issue the 'AVERAGEFREQ ON' command
## Set weight header to N ...
## Set effect header to BETA ...
```

METAL output METAANALYSIS1.TBL.info

```
$ cat *.info
# This file contains a short description of the columns in the
# meta-analysis summary file, named 'METAANALYSIS1.TBL'

# Marker    - this is the marker name
# Allele1   - the first allele for this marker in the first file where it occurs
# Allele2   - the second allele for this marker in the first file where it occurs
# Weight    - the sum of the individual study weights (typically, N) for this marker
# Z-score   - the combined z-statistic for this marker
# P-value   - meta-analysis p-value
# Direction - summary of effect direction for each study, with one '+' or '-' per study

# Input for this meta-analysis was stored in the files:
# --> Input File 1 : /projectnb/data/meta/METAL_example/DGI_three_regions.txt
# --> Input File 2 : /projectnb/data/meta/METAL_example/MAGIC_FUSION_Results.txt.gz
# --> Input File 3 : /projectnb/data/meta/METAL_example/
```

- This file is important to interpret the “Direction” field in the output file (see next slide)

METAL output METAANALYSIS1.TBL

MarkerName	Allele1	Allele2	Weight	Zscore	P-value	Direction
rs217377	t	c	6796.00	0.813	0.4161	-++
rs4668077	a	g	6796.00	0.040	0.968	--+
rs16855496	a	g	4108.00	-0.092	0.9268	??-
rs217386	a	g	6796.00	0.789	0.4299	+++
rs2075070	a	g	6796.00	-1.018	0.3086	--+
rs10187002	a	t	4108.00	0.099	0.9208	??+
rs12785983	t	c	6796.00	-1.230	0.2188	+--
rs1100405	t	c	6796.00	-0.838	0.4018	---
rs12155014	t	c	6796.00	0.262	0.793	++-
rs2287619	t	c	6796.00	-2.139	0.03242	---
rs505899	a	t	6796.00	-0.572	0.5674	+--
rs4753424	t	c	6796.00	2.080	0.03752	+++
rs2724155	a	c	6796.00	0.461	0.6446	+++
rs6718042	a	c	6796.00	0.188	0.8508	--+
rs3845732	t	c	6796.00	1.032	0.3019	+-+
rs3755166	a	g	6796.00	-0.857	0.3914	---
rs768919	c	g	6796.00	1.050	0.2936	+-+
rs4753444	t	c	6796.00	-0.317	0.7516	--+
rs2433681	a	g	6796.00	-0.944	0.3449	-+-

- Default: sample-size weighted Z-score approach

METAL options

- Genomic control correction
  - When genome wide results are analyzed, may want to correct for possible inflation in type-I error
  - However, in the present example, only 3 regions are selected based on low p-values
    - Not appropriate to compute genomic control based on these 3 regions
  - May use genomic control lambda computed on full set of results
    - Available from supplementary Table 1 from original article:
      - Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, …, Dupuis J, Watanabe RM, Stefansson K, McCarthy MI, Wareham NJ, Meigs JB, Abecasis GR. Variants in MTNR1B influence fasting glucose levels. Nature Genetics. 2009;41:77-81.

METAL options: Genomic control correction

STUDY SAMPLE	TwinsUK	CoLaus	SardiNIA	Framingham	NTR/NESDA	NFBC1966	deCODE	Rotterdam Study	FUSION	DGI
DATA ANALYSIS										
Number of SNPs in analysis N imputed	2,434,545	2,557,249	2,252,558	2,540,223	425,052 (genotyped)	2,378,857	Genotyped 299,319 (imputed SNPs in 3 studied regions 2.385)	2,543,887	2,556,824	2,411,071
Trait transformation	Natural logarithm	Standardized, log10 transformed	Quantile normalization	Residuals	Natural logarithm	Natural logarithm	Natural log	Z-scaling of residuals of log-transformed trait	Inverse normalization of residuals	none
Adjustments	Age	Gender, age	BMI, age, age^2, sex	Gender specific residuals adjusted for age and age^2	Gender, age	Gender, 3 PCs based on GW data determining geographical differences	Gender, age	Gender, age	Age, age2, gender, birth province, study	Age, gender, log BMI, clinical site
Analysis method	Score test (FastAssoc)	Linear regression (additive model)	Score test (FastAssoc)	Linear Mixed Effect models	Regression, additive model, Wald test	Cochrane-Armitage test for additive genetic effect	Test for additive genetic effect	ML regression	Linear regression	Linear regression
Software for analysis	MERLIN	SNPtest	MERLIN	LMEKIN (R package)	Plink	SNPTEST	R / SNPTEST	ProbABEL	Merlin	PLINK
Genomic Control Lambda	1.002	1.009	1.061	1.013	1.014	1.017	1.094	n.a.	1.008	1.044

METAL options: Genomic control correction

- Modify metal.txt to add genomic control correction
  - Sardinia: 1.061
  - FUSION: 1.008
  - DGI: 1.044
- Need to write GENOMICCONTROL x before the “PROCESS filename” statement for each study
- Change output file name (so it doesn’t overwrite previous analysis)
  - OUTFILE METAANANLYSIS_GC_ .tbl
  - This statement should be before the “ANALYSE” command

Save the new metal input file with a new name! I called my edited file metal_GC.txt

METAL command file with GC correction

```
$ cat metal_GC.txt
…
MARKER   SNP
DEFAULTWEIGHT  4108
ALLELE   AL1 AL2
FREQ     FREQ1
EFFECT   EFFECT
STDERR   SE
PVAL     PVALUE
GENOMICCONTROL 1.061
PROCESS /projectnb/bs859/data/METAL_example/magic_SARDINIA.tbl

OUTFILE METAANALYSIS_GC_ .tbl
ANALYZE
```

Results with and without GC correction

- Zscore
  - Smallest p-value without GC correction (end of log file)

```
## Smallest p-value is 7.236e-12 at marker 'rs560887'
```

  - Smallest p-value with GC correction (end of log file)

```
## Smallest p-value is 2.219e-11 at marker 'rs560887'
```

- Results are less significant after GC correction
  - Why?

METAL options

- Inverse variance approach:
  - SCHEME STDERR
  - Need to specific name of standard error column for each study
    - E.g. STDERRLABEL SE
- Allele frequency options
  - AVERAGEFREQ ON
  - MINMAXFREQ ON
  - Provide average and SE of allele frequency for all cohorts
  - SE can be used to detect errors
    - Large SE may indicate allele coding errors
  - Need to specify allele frequency column for each study
    - E.g. FREQLABEL EFFECT_ALLELE_FREQ

Inverse Variance Approach in METAL

- Let's modify the file metal_GC.txt (keeping the genomic control correction) to perform Inverse Variance approach for pooling the beta coefficient estimates
- We will also add the option to compute the min, max and average allele frequency
- In input file, uncomment the lines (remove the "#"):
  - SCHEME STDERR
  - AVERAGEFREQ ON
  - MINMAXFREQ ON
- Required fields (STDERRLABLE and FREQLABEL) are already included in the example file
- Modify the output file name (so it doesn't overwrite previous GC results)
- Save the file with a new name and run the analysis

Inverse variance approach results

MarkerName	Allele1	Allele2	Freq1	FreqSE	MinFreq	MaxFreq	Effect	StdErr	P-value	Direction
rs217377	t	c	0.5963	0.0111	0.5770	0.6089	0.0149	0.0205	0.4685	-++
rs4668077	a	g	0.1552	0.0430	0.0990	0.1907	-0.0057	0.0281	0.8397	--+
rs16855496	a	g	0.9880	0.0000	0.9880	0.9880	-0.0110	0.1215	0.9279	??-
rs217386	a	g	0.4205	0.0609	0.3270	0.4740	0.0174	0.0214	0.4156	+++
rs2075070	a	g	0.5369	0.0353	0.5040	0.5873	-0.0241	0.0200	0.2285	--+
rs10187002	a	t	0.9890	0.0000	0.9890	0.9890	0.0120	0.1257	0.9239	??+
rs12785983	t	c	0.2910	0.0193	0.2780	0.3270	-0.0240	0.0220	0.2764	+--
rs1100405	t	c	0.3367	0.0040	0.3340	0.3433	-0.0185	0.0209	0.3749	---
rs12155014	t	c	0.9094	0.0259	0.8893	0.9450	0.0218	0.0357	0.5423	++-
rs2287619	t	c	0.9147	0.0137	0.9040	0.9344	-0.0738	0.0356	0.03842	---
rs505899	a	t	0.1791	0.0195	0.1570	0.1970	-0.0117	0.0258	0.6492	+--
rs4753424	t	c	0.6151	0.0141	0.6020	0.6306	0.0448	0.0212	0.03425	+++
rs2724155	a	c	0.9211	0.0264	0.9020	0.9850	0.0292	0.0451	0.5171	+++
rs6718042	a	c	0.8859	0.0406	0.8660	0.9730	0.0133	0.0370	0.7193	--+
rs3845732	t	c	0.2169	0.0650	0.1230	0.2640	0.0282	0.0258	0.2754	+-+
rs3755166	a	g	0.4104	0.0150	0.3938	0.4360	-0.0173	0.0204	0.3977	---
rs768919	c	g	0.0320	0.0046	0.0234	0.0350	0.0505	0.0512	0.3237	+-+
rs4753444	t	c	0.4653	0.0177	0.4488	0.4960	-0.0140	0.0200	0.4838	--+
rs2433681	a	g	0.0976	0.0120	0.0740	0.1060	-0.0356	0.0332	0.2833	-+-
rs13393173	a	g	0.2322	0.0130	0.2180	0.2450	0.0229	0.0237	0.3341	-++
rs6433109	a	c	0.5467	0.0194	0.5213	0.5750	0.0236	0.0206	0.2509	+++
rs3019218	c	g	0.4762	0.0304	0.4470	0.5190	0.0091	0.0196	0.6429	-++
rs512498	t	c	0.4069	0.0199	0.3780	0.4260	0.0052	0.0201	0.795	++-
rs2595650	a	g	0.5246	0.0218	0.4920	0.5450	-0.0192	0.0200	0.3376	-+-
rs12791593	t	c	0.2191	0.0027	0.2140	0.2210	0.0147	0.0237	0.5353	+++
rs3770636	t	g	0.9749	0.0074	0.9710	0.9890	0.0602	0.0757	0.4263	-++
rs605714	t	c	0.4631	0.0253	0.4380	0.4980	-0.0105	0.0199	0.5959	---
rs6483189	t	g	0.1322	0.0221	0.1120	0.1670	-0.0283	0.0302	0.3485	---
rs16855448	a	t	0.9890	0.0000	0.9890	0.9890	0.0120	0.1257	0.9239	??+
rs831019	t	g	0.5369	0.0176	0.5210	0.5570	0.0119	0.0214	0.5773	+++

Z-score versus pooling of effect estimates

- Zscore
  - Smallest p-value (with GC correction)

```
## Smallest p-value is 2.219e-11 at marker 'rs560887'
```

- Inverse variance pooling of effect estimate

```
## Smallest p-value is 6.1e-11 at marker 'rs560887'
```

- Z-score approach yields slightly more significant result

METAL options: Heterogeneity testing

- To include test for heterogeneity, change ANALYSE to
  - ANALYSE HETEROGENEITY

- Modify your latest command file to test for heterogeneity
- Don’t forget to modify the name of your output file!

Inverse variance approach results with Cochran Q-test

1	MarkerName	Allele1	Allele2	Freq1	FreqSE	MinFreq	MaxFreq	Effect	StdErr	P-value	Direction	HetISq	HetChiSq	HetDf	HetPVal
2	rs217377	t	c	0.5963	0.0111	0.577	0.6089	0.0149	0.0205	0.4685	-++	0	0.378	2	0.8279
3	rs4668077	a	g	0.1552	0.043	0.099	0.1907	-0.0057	0.0281	0.8397	--+	0	1.029	2	0.5977
4	rs16855496	a	g	0.988	0	0.988	0.988	-0.011	0.1215	0.9279	??-	0	0	0	1
5	rs217386	a	g	0.4205	0.0609	0.327	0.474	0.0174	0.0214	0.4156	+++	0	0.156	2	0.9252
6	rs2075070	a	g	0.5369	0.0353	0.504	0.5873	-0.0241	0.02	0.2285	--+	22.3	2.573	2	0.2763
7	rs10187002	a	t	0.989	0	0.989	0.989	0.012	0.1257	0.9239	??+	0	0	0	1
8	rs12785983	t	c	0.291	0.0193	0.278	0.327	-0.024	0.022	0.2764	+--	0	1.36	2	0.5065
9	rs1100405	t	c	0.3367	0.004	0.334	0.3433	-0.0185	0.0209	0.3749	---	0	0.613	2	0.7359
782	rs560887	t	c	0.3401	0.0344	0.2976	0.373	-0.1355	0.0207	6.10E-11	---	72.8	7.349	2	0.02536
783	rs1447351	a	g	0.5113	0.0541	0.428	0.56	-0.0734	0.0202	0.000281	---	21.3	2.541	2	0.2807

- rs560887: beta=0.1355 se=0.0207 95% CI: (0.094,0.176)
- Which allele increases glucose levels?
- Evidence for heterogeneity at rs560887
  - HetPVal = 0.02536; I^2 = 72.8
  - Should consider multiple testing correction for Heterogeneity test, but perhaps not as stringent as usual Bonferroni correction
- What do the individual study results look like at this SNP? (why do we see evidence of heterogeneity?)

Results from three cohorts for rs560887

FUSION

EFFECT_ ALLELE	NON_EFFECT_ALLELE	FREQ__ EFFECT	N	BETA	SE	λ_GC	SE_GC	PVALUE
4	2	0.314	1233	-0.139	0.044	1.008	0.0442	0.00169

DGI

EFFECT_ ALLELE	NON_EFFECT_ALLELE	EFFECT_ ALLELE_ FREQ	N	BETA	SE	λ_GC	SE_GC	P_VAL
4	2	0.297595	1455	-0.04571	0.03945	1.044	0.0403	0.257

Sardinia

AL1	AL2	FREQ1		EFFECT	SE	λ_GC	SE_GC	PVALUE
C	T	0.627		0.18	0.028	1.061	0.0288	1.36E-10

C is the glucose raising allele in all three cohorts

Homework

- Extension of the class example-- different results, and one additional study
  - Need to figure out how to specify comma-delimited file
- Need to merge meta-analysis results with one of the study-specific files to determine chromosome and position of the variants
- There are three loci on different chromosomes -- you need to report the 3 variants in each region with the smallest p-values

Outline: Meta-Analysis for genetic studies

- Extra: Meta-analyzing gene-based tests using raremetalworker and raremetal

Software options

- Two major choices:
  - seqmeta package in R
  - RAREMETALWORKER + RAREMETAL
- Significant drawbacks for each
- seqmeta
  - Cox, logistic, or linear model analyses
  - Does not keep track of coded alleles → relies on the analysts for each study to ensure common coded allele
  - Must predefine variant groups
- RAREMETAL
  - Linear model analysis only
  - Keeps track of coded allele, recodes when necessary
  - More flexible definition of variant groups
  - More flexible options for data input (MERLIN or vcf)

Using RAREMETALWORKER and RAREMETAL for grouped variant meta-analysis

Study $k = 1, \ldots, K$ analyzes data with RAREMETALWORKER, produces score statistics, allele frequencies, and covariance matrices

Sends results to meta-analyst

Meta-analyst uses RAREMETAL to produce meta-analysis statistics that combine the studies results

raremetal tutorial

- Raremetal example data sets:

```
$ ls $DATA2
exstudy1.dat        exstudy2.dat        group.file     study2.ped
exstudy1.vcf.gz     exstudy2.vcf.gz     region.groups
exstudy1.vcf.gz.tbi exstudy2.vcf.gz.tbi study1.ped
```

raremetal tutorial

- Get a quick idea of the data sets:

```
pedstats -p $DATA2/study1.ped -d $DATA2/exstudy1.dat --traitPDF -a
study1.pdf > study1.pedstats.log
pedstats -p $DATA2/study2.ped -d $DATA2/exstudy2.dat --traitPDF -a
study2.pdf> study2.pedstats.log

zcat $DATA2/exstudy1.vcf.gz |head -n 7
zcat  $DATA2/exstudy1.vcf.gz  |wc

zcat $DATA2/exstudy2.vcf.gz |head -n 7
zcat  $DATA2/exstudy2.vcf.gz  |wc

module load vcftools
vcftools --gzvcf $DATA2/exstudy1.vcf.gz --freq --out study1
vcftools --gzvcf $DATA2/exstudy2.vcf.gz --freq --out study2
```

raremetal tutorial

- Run raremetalworker on each data set, using trait QT1:

```
raremetalworker  --ped $DATA2/study1.ped --dat $DATA2/exstudy1.dat --vcf
$DATA2/exstudy1.vcf.gz --traitName QT1 --inverseNormal --makeResiduals --
kinSave --kinGeno --prefix STUDY1
```

```
raremetalworker  --ped $DATA2/study2.ped --dat $DATA2/exstudy2.dat --vcf
$DATA2/exstudy2.vcf.gz --traitName QT1 --inverseNormal --makeResiduals --
kinSave --kinGeno --prefix STUDY2
```

- For each study, this command transforms phenotype QT1 to normality, calculates trait residuals after adjusting for covariates, estimates relatedness between individuals and saves this for later use (--kinSave and --kinGeno), and generates score for each variant, and the covariance between pairs of markers, for use in meta-analysis
- It also creates some PDF files summarizing results

raremetal tutorial

- Several output files for each study:
- The critical results for doing the meta analysis are the *score.txt and *cov.txt files:

```
STUDY1.QT1.singlevar.score.txt      ## single variant statistics
STUDY1.QT1.singlevar.cov.txt        ## covariance matrices between score statistics
STUDY1.plots.pdf                    ## QQ plots and Manhattan plots
STUDY1.Empirical.Kinship.gz         ## Relatedness matrix
STUDY1.singlevar.log                ## Log file
```

raremetal tutorial

- covariance file:

```
$ head -n 4 STUDY1.QT1.singlevar.cov.txt
##ProgramName=RareMetalWorker
##Version=4.13.5
#CHROM  CURRENT_POS     MARKERS_IN_WINDOW       COV_MATRICES
9 44001280
44001280,44001379,44001983,44006342,44047825,44047839,44055726,44056400,44056412,44057574,4
4058842,44058893,44079591,44081288,44090195,44097343,44098979,44116609,44116956,44117052,44
117769,44117961,44117985,44118188,44118191,44118353,44130960,44131324,44153100,44153248,441
56378,44156472,44159722,44187577,44223113,44223180,----   0.0159717,0.014753,0.00024573,-
4.03085e-06,0.000188971,-3.26979e-05,0.00187045,-3.94089e-06,-0.00104282,-4.14834e-05,-4.05452e-
06,-4.20392e-06,-4.19761e-06,0.00630938,-0.00613228,-7.98346e-06,-0.000332018,-8.11202e-06,-
0.00641209,-0.00170944,-4.00865e-06,-1.2361e-05,-0.000114491,0.0141212,-0.00639423,0.0141212,-
3.88244e-06,-7.90183e-06,
```

- format is: location, list of markers in the window (comma separated), and the covariances of those markers with the first marker (comma separated)

raremetal tutorial

```
$ more  STUDY1.QT1.singlevar.score.txt
##ProgramName=RareMetalWorker
##Version=4.13.5
##Samples=4000
##AnalyzedSamples=4000
##Families=4000
##AnalyzedFamilies=4000
##Founders=4000
##AnalyzedFounders=4000
##Covariates=AGE,sex
##CovariateSummaries    min     25th    median  75th    max     mean    variance
##AGE   40      44      48      51      55      47.5355 18.7259
##sex   1       1       1       2       2       1.40525 0.241083
##InverseNormal=ON
##TraitSummaries        min     25th    median  75th    max     mean    variance
##QT1   118     121.8   125     126.8   140.4   124.701 11.3288
## - NullModelEstimates
## - Name       BetaHat SE(BetaHat)
## - Intercept  1.58353e-05     0.0158077
##AnalyzedTrait -3.66226        -0.676064       -0.000626657    0.674883        3.6
```

raremetal tutorial

```
$ more  STUDY1.QT1.singlevar.score.txt
---
## - NullModelEstimates
## - Name       BetaHat SE(BetaHat)
## - Intercept  1.58353e-05     0.0158077
##AnalyzedTrait -3.66226       -0.676064       -0.000626657    0.674883       3.6
6226    1.682e-05       0.999814
##Sigma_g2_Hat  4.53866e-05
##Sigma_e2_Hat  0.999533
##Heritability=-nan
#CHROM  POS     REF     ALT     N_INFORMATIVE   FOUNDER_AF      ALL_AF  INFORMATIVE
_ALT_AC CALL_RATE       HWE_PVALUE      N_REF   N_HET   N_ALT   U_STAT
SQRT_V_STAT
       ALT_EFFSIZE     PVALUE
9       44001280        G       A       4000    0.008125        0.008125        65-1       1       3935    65
0       10.3998 7.99291 0.162785        0.193216
9       44001379        T       C       4000    0.106125        0.106125        849
       1       0.73873 3198    755     47      29.3132 27.5629 0.0385844       0.2
```

raremetal tutorial

- Once we have run RAREMETALWORKER on all studies, we can run RAREMETAL to do the Meta-Analysis
  - first index the result files generated by RAREMETAL
  - This step relies on bgzip and tabix, two tools that allow rapid indexing and retrieval of results from compressed text files

```
bgzip STUDY1.QT1.singlevar.score.txt
tabix -c "#" -s 1 -b 2 -e 2  STUDY1.QT1.singlevar.score.txt
bgzip STUDY1.QT1.singlevar.cov.txt
tabix -c "#" -s 1 -b 2 -e 2  STUDY1.QT1.singlevar.cov.txt

bgzip STUDY2.QT1.singlevar.score.txt
tabix -c "#" -s 1 -b 2 -e 2  STUDY2.QT1.singlevar.score.txt.gz
bgzip STUDY2.QT1.singlevar.cov.txt
tabix -c "#" -s 1 -b 2 -e 2  STUDY2.QT1.singlevar.cov.txt.gz
```

raremetal tutorial

- Need a file that defines variant groups – just like EPACTS

```
$ cat  $DATA2/group.file
GENE1   9:45368740:G:A 9:45375164:C:T 9:45375295:C:T 9:45377254:G:A 9:45377290:C:T 9:45377654:A:G 9:45381530:G:A 9:45381836:C:A 9:45381860:C:T 9:45385488:G:A 9:45389198:G:C
GENE2  9:45404058:C:T
GENE3    9:45411110:T:C 9:45412040:C:T 9:45412056:G:A 9:45412079:C:T 9:45412097:G:T
GENE4   9:45419555:A:G 9:45422446:A:T
GENE5   9:45445541:C:T 9:45445586:G:A 9:45448036:T:C 9:45448070:G:A 9:45448465:T:G 9:45448507:A:C
GENE6   9:45451743:C:T 9:45451745:C:G 9:45451769:G:A 9:45451987:G:A 9:45452080:G:A 9:45452429:A:C
```

raremetal tutorial

raremetal --summaryFiles summaryfiles --covFiles covfiles --groupFile $DATA2/group.file --SKAT --burden --hwe 1.0e-05 --callRate 0.95 --longOutput --tabulateHits --hitsCutoff 1e-05 --prefix COMBINED.QT1 --labelHits --geneMap $SCC_RAREMETAL_DIR/src/raremetal_4.13.5/raremetal/data/refFlat_hg19.txt.gz

- Filters summary statistics based on HWE p-value and variant call rate
- Generates single variant meta-analysis results
- Generates gene-level meta-analysis results using simple burden test (all variants equal weight) and SKAT using the group definitions in group.file
- Tabulates significant genes with detailed single variant results included
- Generates a PDF file summarizing results
- Also can specify maximum MAF with –maf
  - The default is maf<0.05; since we did not specify –maf, only variants with frequency <0.05 will be included in the SKAT and burden tests

raremetal tutorial

```
$ ls COMBINED.QT1.*
COMBINED.QT1.meta.SKAT.results
COMBINED.QT1.meta.tophits.SKAT.tbl
COMBINED.QT1.meta.burden.results
COMBINED.QT1.meta.tophits.burden.tbl
COMBINED.QT1.meta.plots.pdf
COMBINED.QT1.raremetal.log
COMBINED.QT1.meta.singlevar.results
```

raremetal tutorial

- Our groupfile has only a 6 genes defined
- 2 of the genes have 1 or 2 SNPs
- Results file has results only for GENE1, GENE3 and GENE5

rare variant meta-analysis comments

- Best to include ALL variants in the study-specific analyses – share all variants, allele frequencies, covariance matrices
- This allows flexibility at meta-analysis stage
  - Different subsets of SNPs (e.g., all MAF<0.01, MAF<0.01 AND nonsynonymous, etc)
  - Different subsets based on MAF (where MAF can be computed across all studies)