Gene and pathway analysis of GWAS data

Christiaan de Leeuw

Miaoxin Li

Session

- **Theoretical overview**
  - Gene analysis
  - Pathway/gene-set analysis
- **MAGMA practical**
  - Gene analysis
  - (generalized) gene-set analysis
- **KGG practical**
  - Gene analysis
  - Gene-gene pair analysis in PPI networks

Gene analysis - overview

- **Joint analysis of (common) variants in a gene**
- **Goal: localization of associations on the genome**

- **Pro**
  - Reduce multiple testing (from up to 2.5M SNPs to around 20K genes)
  - Better power (potentially)
    - Aggregate weak individual effects to detectable signal
    - Detect haplotype effects

- **Con**
  - Disregards non-genic variation
  - Large sample sizes still required

Gene analysis - methods

- **Numerous methods: MAGMA, KGG, VEGAS, PLINK (/PLINK2), …**
- **SNP aggregation vs multi-marker model**

- **SNP aggregation approach**
  - Generate p-values per SNP, then aggregate into gene statistic
    - Gene statistic: mean SNP χ^2, top SNP χ^2
  - Only requires SNP p-values (and reference data, eg. 1,000 Genomes)
  - Often requires permutation to obtain gene p-value
    - Computationally expensive, uncertainty in p-value
  - Cannot detect haplotype effects

Gene analysis - methods

- **Numerous methods: MAGMA, KGG, VEGAS, PLINK (/PLINK2), …**
- **SNP aggregation vs multi-marker model**

- **Multi-marker model approach**
  - Linear or logistic regression with all SNPs in gene as predictors
    - Omnibus test: F-test, likelihood ratio test
  - Requires raw data, doesn't require permutation
  - Can detect haplotype effects
  - Easier to include covariates and interaction terms

- **Methods tend to converge as sample size increases**

Gene analysis - general issues

- **Annotation window**
  - Use nearby markers beyond transcription start/stop sites?
  - Recommendation: use very small or no window, or determine on gene-by-gene basis (eg. using eQTL data)

- **Population stratification**
  - More susceptible than single marker analysis
  - Recommendation: include (at least) 10 principal components

Gene-set analysis - overview

- **Joint analysis of functionally/biologically related genes**
  - Pathway = gene set, no internal structure

- **Pro**
  - Can aggregate multiple weakly associated genes into detectable signal
  - Leverage outside information to provide biological insight into phenotype
  - Shift in research question from “where are phenotype associations located on genome” to “which functional/biological properties and processes are relevant to the phenotype”

Gene-set analysis - overview

- **Joint analysis of functionally/biologically related genes**
  - Pathway = gene set, no internal structure

- **Con**
  - Difficult to determine which gene sets to test
    - Public database vs expert curated
    - Database-wide analysis vs specific hypotheses
  - Uncertainty in interpretation of results
  - Large sample sizes still required

Gene-set analysis - methods

- **Self-contained analysis**
  - PLINK, KGG-HYST, MAGMA, JAG, SETSCREEN, GenGen, …
- **Competitive analysis**
  - MAGMA, KGG-hyper, JAG, INRICH, MAGENTA, ALIGATOR, FORGE, GSEA, …
- **Different hypotheses**
  - Self-contained: are genes in the gene set (on average) associated with the phenotype?
  - Competitive: are genes in the gene set (on average) more strongly associated with the phenotype than other genes?

Gene-set analysis - methods

- **Compare eg. medical research**
  - Self-contained: does the health of patients taking drug X improve?
  - Competitive: does the health of patients taking drug X improve more than that of control patients?

- **Self-contained analysis does not correct for background association**
  - Gene-set p-values decrease as function of
    - Gene-set size
    - Degree of polygenicity and heritability
  - Strong association does not imply substantive relation to phenotype

Gene-set analysis - methods

- **Self-contained analysis**
  - P[illegible] KGG-HYST, MAGMA, JAG, SE[illegible]EN, GenGen, …

- **Competitive analysis**
  - MAGMA, KGG-hyper, JAG, INRICH, MAGENTA, ALIGATOR, FORGE, GSEA, …

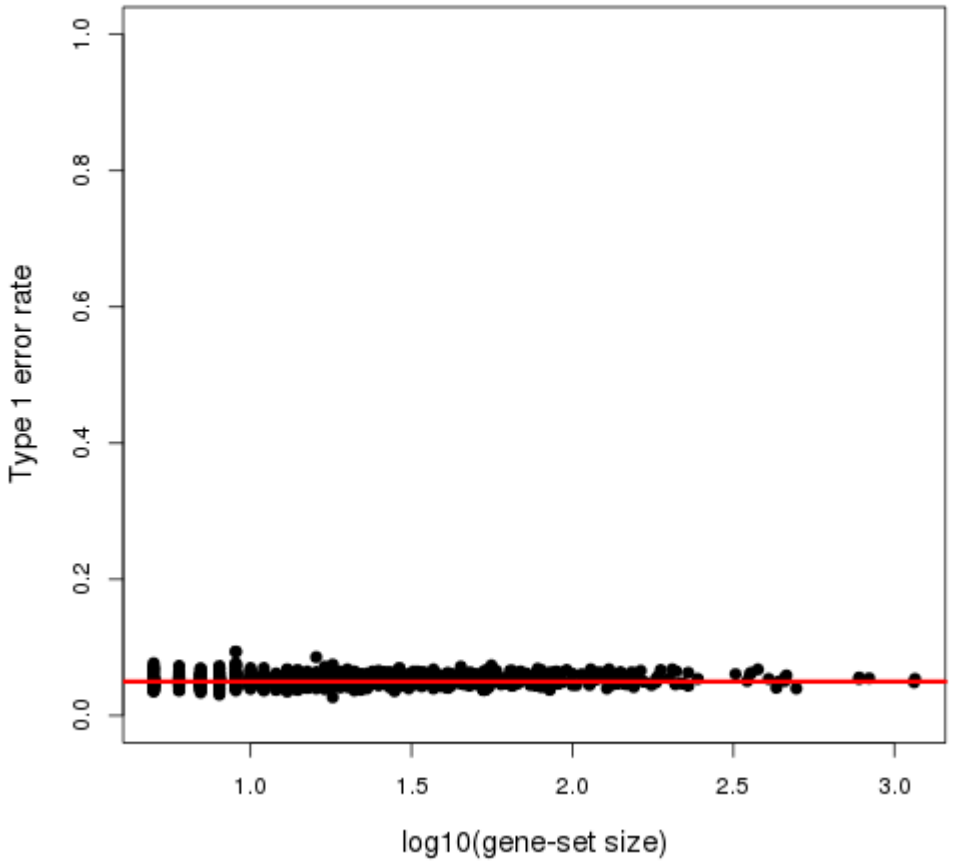
Gene-set analysis - methods

- **Competitive analysis**
  - MAGMA, KGG-hyper, JAG, INRICH, MAGENTA, ALIGATOR, FORGE, GSEA, …

Gene-set analysis - methods

- **Competitive analysis**
  - MAGMA, KGG-hyper, JAG, INRICH, MAGENTA, ALIGATOR, FORGE, GSEA, …
- **However…**
  - Type 1 error rates for many competitive methods not well controlled
    - Error rates are controlled averaged over many gene sets, but not always at individual gene set level

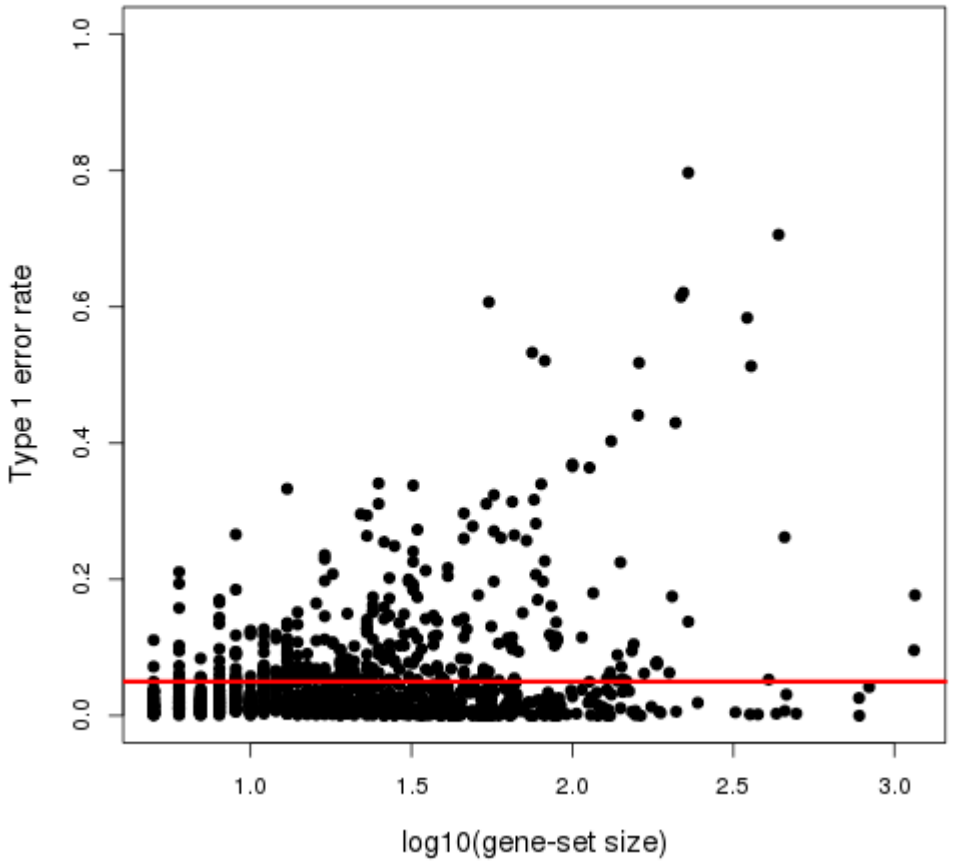
Individual gene-set type 1 error rates



MAGMA competitive analysis

“Well-controlled”

Individual gene-set type 1 error rates



Unnamed gene-set analysis method

“Not well-controlled”

Mean type 1 error rate at alpha of 5% is 0.051

Gene-set analysis - methods

- **Competitive analysis**
  - MAGMA, KGG-hyper, JAG, INRICH, MAGENTA, ALIGATOR, FORGE, GSEA, …
- **However…**
  - Type 1 error rates for many competitive methods not well controlled
    - Error rates are controlled averaged over many gene sets, but not always at individual gene set level
- **Recommendation**
  - MAGMA
  - INRICH (with higher SNP p-value cut-off, eg. 1st percentile of SNP p-values)

Gene-set analysis - general issues

- **Gene analysis issues**
  - Annotation window, population stratification
- **Interpretation**
  - Interpretation is relative to how association between genes and phenotype is defined
  - Interpretation is dependent on the quality of the gene set itself
    - ‘Measurement uncertainty’
  - Confounding

Gene-set analysis - confounding

- **AKA: correlation does not imply causation**

- **Suppose gene set A is truly associated with the phenotype:**
  - Any other gene set that has overlap with A may show a spurious association

- **Partial solution: conditional gene-set analysis (MAGMA only)**
  - Compute gene-set association conditioned on other gene sets or gene properties
  - Compare: principal components for population stratification
  - Only works if A is among the gene sets being tested

Practical

MAGMA

- **Goals**
  - Fast & user-friendly
  - Robust gene and gene-set analysis
  - Platform to include more biological information in genetic analysis

- **Program**
  - C++, Command-line interface
  - Binary PLINK data input

MAGMA

- **Gene analysis model**
  - Raw data: principal components regression + F-test
    - Rare variants: burden score
    - Optional GxE interaction component
  - SNP p-values: mean / top χ^2 statistic

- **Gene-set analysis model**
  - Based on gene analysis output
  - Linear regression model, genes as observations / rows in data matrix
    - Gene-sets as predictor of association in genes
  - Self contained analysis, (conditional) competitive analysis, gene property analysis

Practical

1. **Annotate SNPs to genes**
2. **Perform gene analysis (with 10 PCs as covariates)**
3. **Perform gene-set analysis**
4. **Perform conditional gene-set analysis with confounding gene-set**

- **Data**
  - Simulated GWAS data and phenotype; 400K SNPs, N = 2,500
  - 1011 Reactome gene sets

Practical

- **Open terminal window**
- **Copy practical files to local drive and move to new folder**
  - cp -R /faculty/christiaan/Boulder\ 2015/magma_practical .
  - cd magma_practical
- **Folder should contain**
  - GWAS data: boulder.bim, boulder.bed, boulder.fam
  - Covariate file: boulder_pca.cov
  - Gene-set file: reactome.sets
  - Gene definition file: NCBI37.3.gene.loc
  - Instructions: practical.pdf

Practical - key points

- **Step 2: gene analysis**
  - 3 genes are genome-wide significant (thresh. = 0.0000038)
  - Only 6.03% of genes have a p-value below 0.05
    - Modest polygenicity

- **Step 3: gene-set analysis**
  - Significant: 15 self-contained, 8 competitive
    - Despite only modest polygenicity in data
    - Gap is usually larger in practice
  - For first significant gene-set (SIGNALING_BY_NOTCH1_T)
    - Lowest gene p-value: 0.00034
    - 26.4% of genes have a p-value below 0.05

Practical - key points

- **Step 4**
  - 5 out of 7 gene-sets are no longer significant after correcting for the confounder gene-set
  - Conversely, the confounder set remains highly significant when conditioning on any of these 7 sets

	Self. P	Comp. P (step 3)	Comp. P (step 4)
Set 1	1.88e-12	3.79e-5	0.155
Set 2	1.12e-8	2.97e-5	2.6e-5
Set 3	4.02e-9	5.24e-8	0.077
Set 4	3.64e-14	2.07e-9	0.127
Set 5	7.94e-7	4.36e-6	0.736
Set 6	1.69e-9	3.57e-6	3.02e-6
Set 7	5.40e-15	6.87e-9	0.264