Practical

CNCR CTGLAB

MAGMA

- **Tool for gene, gene set and gene property analysis**
  - Gene property = ‘continuous gene set’
    - Or: gene set = binary gene property
  - Command-line interface (Linux, Mac OS, Windows)

- **Today**
  - Gene analysis
  - Competitive gene-set analysis
  - Gene property analysis (tissue-specific expression)
  - Using raw genotype data as input

MAGMA

- **Other options**
  - Analysis on SNP p-value input
  - Rare variant (and rare + common) analysis features
  - Gene x environment analysis
  - Etc.

- **Upcoming release**
  - Gene-set level interaction analysis
  - Partitioning SNPs in genes
    - Eg. by functional SNP type

Practical

1. **Annotate SNPs to genes**
2. **Perform gene analysis (with 10 PCs as covariates)**
3. **Perform gene-set analysis**
4. **Perform tissue expression analysis**
5. **Perform joint gene-set / tissue expression analysis**

- **Data**
  - Simulated GWAS data and phenotype; 400K SNPs, N = 2,500
  - 1011 Reactome gene sets
  - Tissue-specific expression data for 11 tissues
    - Simulated, but based on real expression data

Practical

- **Open terminal window**
- **Copy practical files to local drive and move to new folder**
  - cp /faculty/christiaan/Boulder2017/magma_practical.zip .
  - unzip magma_practical.zip
  - cd magma_practical
- **Folder should contain**
  - GWAS data: boulder.bim, boulder.bed, boulder.fam
  - Covariate file: boulder_pca.cov
  - Gene-set files: reactome.sets, reactome_signif.sets
  - Tissue expression file: tissue_gex.cov
  - Gene definition file: NCBI37.3.gene.loc
  - Instructions: practical.pdf
    - Note: MAGMA commands in PDF may run over multiple lines, but should be entered in the terminal as a single command

Practical - key points

- **Step 1: annotation**
  - Out of 19,427 protein-coding genes in the gene location file, only 13,772 had any SNPs annotated to them
  - Restricts any conclusions to the annotated protein-coding genes, we cannot be sure whether the same relations hold in the other genes
    - Use genotype data with better coverage (eg. using imputation) to address this issue

- **Step 2: gene analysis**
  - 2 genes are genome-wide significant
    - Threshold = 0.05/13,772 = 3.63e-6
  - Only 6.22% of genes have a p-value below 0.05

Practical - key points

- **Step 3a: basic competitive gene-set analysis**
  - Out of 1011, there are 9 significant gene sets
    - Points to the underlying property (known pathway, cell function, biological process, etc.) playing a role in the phenotype
      - NOTE: only competitive gene-set analysis allows this kind of conclusion
    - Looking at the names, probably overlap between these gene sets
      - Use conditional gene-set analysis to improve specificity
  - For first significant gene-set (SIGNALING_BY_NOTCH1_T)
    - Lowest gene p-value: 0.00035
      - Genome-wide significance threshold = 3.7e-6
    - 28.3% of genes have a p-value below 0.05

Practical - key points

- **Step 3b: conditional competitive gene-set analysis**
  - 5 out of 8 gene-sets are no longer significant after conditioning on the Critical Pathway gene-set

Set	Comp. P (step 3a)	Comp. P (step 3b)
Signaling by Notch1 T	9.43e-7	8.04e-7
Constitutive Signaling by Notch1 HD + Pest Domain Mutants	8.89e-6	7.82e-6
Elastic Fibre Formation	6.46e-7	0.13
Activation of the Phototransduction Cascade	8.04e-6	0.051
The Phototransduction Cascade	4.55e-9	0.15
Notch1 Intracellular Domain Regulates Transcription	3.20e-5	2.85e-5
Inactivation Recovery And Regulation of the Phototransduction Cascade	1.26e-9	0.060
Molecules Associated with Elastic Fibres	4.80e-5	0.85
Critical Pathway	3.17e-12	-

Practical - key points

- **Step 3b: conditional competitive gene-set analysis**
  - 5 out of 8 gene-sets are no longer significant after conditioning on the Critical Pathway gene-set
  - Conversely, the Critical Pathway remains highly significant when conditioning on any of these 8 sets, suggesting that
    - Of these sets, the Critical Pathway set is most likely to be the true 'causal' gene set
    - The originally observed associations of the 5 sets that are no longer significant are driven entirely by their overlapping with this causal set
  - In practice, the true 'causal' set(s) may not be included in your analysis
    - And are unlikely to be helpfully labeled 'Critical Pathway' if they are

Practical - key points

- **Step 4a: basic tissue expression analysis**
  - All the tissue expression levels are significant, as is the mean expression level across tissues
    - In all likelihood, the associations per tissue are driven by the more general relation between gene expression and genetic association; not very informative

- **Step 4b: conditional tissue expression analysis**
  - Only the brain-specific expression level remains significant after conditioning on average gene expression level
    - More strongly (specifically) brain-expressed genes also tend to be more strongly associated with our phenotype; suggests that brain expression plays a role in (the genetics of) our phenotype

Practical - key points

- **Step 5**
  - The p-values remain effectively the same when conditioning on the average gene expression level, as well as when additionally conditioning brain-specific expression level
  - This suggests that the gene-set associations are not driven merely by gene expression effects (at least of the tissues we tested), which helps strengthen our interpretation of the gene-set associations

- **Any further questions?**
  - MAGMA program, manual and auxiliary files can be found on the MAGMA site: http://ctglab.nl/software/magma
  - Contact me for questions, suggestions, etc. at c.a.de.leeuw@vu.nl
  - These slides: /faculty/christiaan/Boulder2017/practical_slides.pdf