UQ-Brisbane SISG

Module 9: Gene Expression and Epigenetic Profiling

SISG Brisbane 2017

The University Of Queensland AUSTRALIA | IMB

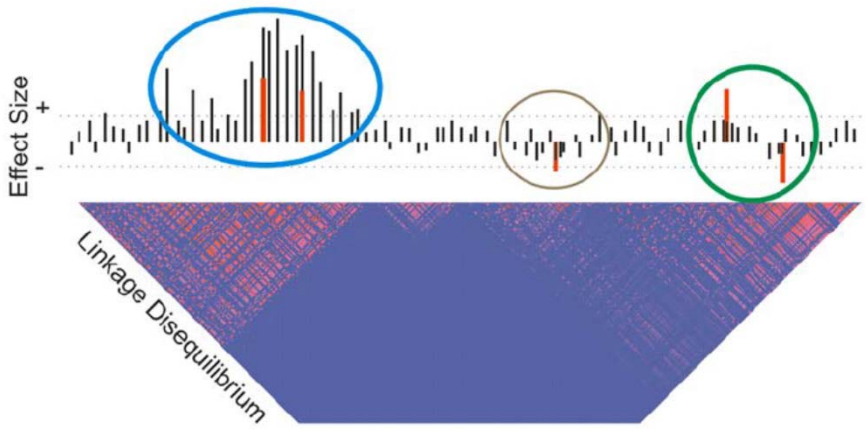
Tuesday February 14, 2017

"Colocalization"

Greg Gibson
Georgia Institute of Technology
greg.gibson@biology.gatech.com

Why Colocalized Signals do not alone imply Causation

Sampling variance means that we can only map "credible intervals"
Many genes harbor multiple eSNPs, and possibly multiple trait associated SNPs
LD means that multiple sites can interfere with one another in estimation of peak locations
The nearest gene is only sometimes the one affected by a SNP!

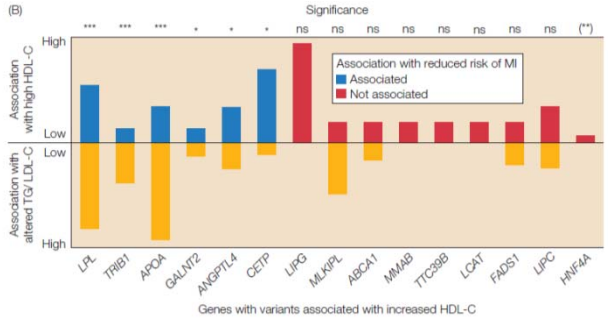


Mendelian Randomization

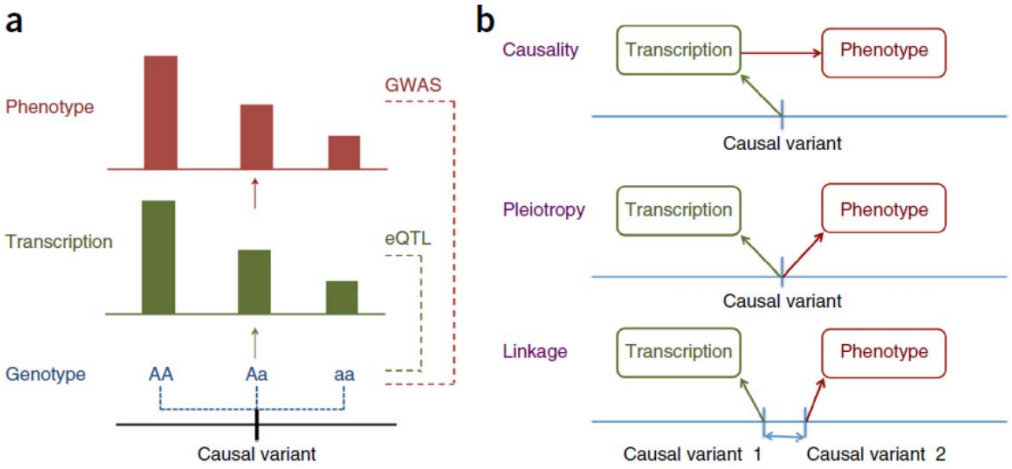
Often times the correlation between an exposure and an outcome (for example, smoking and lung cancer, or cholesterol and heart disease) may not be causal. If we are going to use drugs to reduce an exposure then it is important to know that to do so will be effective, namely that it is causal.

The classical way to do this is to conduct a randomized clinical trial.
Genetics provides a way to do it by assessing whether the genotypes that explain variation in the exposure also explain the relationship between the exposure and the phenotype.

That is, if a gene influences cholesterol levels, then it ought also to influence the risk of heart attacks, if the relationship is causal. It turns out that this condition is met for LDL, but not for HDL, which calls into question the notion that HDL is really protective.



Causality and Correlation for eQTL and GWAS



Statistical power of MR is proportional to: variance of SNP on transcript, variance of transcript on phenotype, N

Zhu et al (2016) *Nature Genetics* **48**: 481-487 "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets"

SMR theory

Let *y* be a phenotype
Let *x* be an exposure (transcript abundance)
Let *z* be a genotype which influences both the exposure and the phenotype. It is also called the "instrument".

Then b_{xy} is the slope (effect size) of the relationship between the transcript and the phenotype
And b_{zy} is the effect of the genotype on the phenotype (the GWAS association)
And b_{zx} is the effect of the genotype on the transcript (the eQTL)

$b_{xy} = b_{zy} / b_{zx}$ is interpreted as the effect free of non-genetic confounder. We do not need to have expression and trait data to estimate it if we have summary data for both the GWAS and the eQTL. Hence, the computation is called "Summary Mendelian Randomization", or "SMR"

Note though that SMR is not able to distinguish causation from pleiotropy.
In this it differs from MR, which is a test of causality – it takes genotypes that associate with an exposure and asks whether this explains why the exposure associates with the phenotype.
In SMR we know that the genotype associates with exposure (transcript) and phenotype and try to estimate whether this is consistent with the transcript and phenotype being correlated.

Zhu et al found 289 genes with highly significant P_{SMR} from blood eQTL with 5 traits (height, BMI, WHR, RA and SZ)

Pleiotropy and Linkage: The HEIDI Test

HEIDI = HEterogeneity In Dependent Instruments

It is a test designed to try to distinguish linkage from pleiotropy

Under pleiotropy, the b_{xy} (b_{zy} / b_{zx}) values for all SNPs should be the same, whereas they may vary for different SNPs if some SNPs in the LD block have variable associations with the transcript.

They found evidence for heterogeneity for 185 / 289 SMR loci, implying linkage over pleiotropy..
For the remaining 101 testable loci, two-thirds of the implicated genes are NOT the closest one to the SNP.

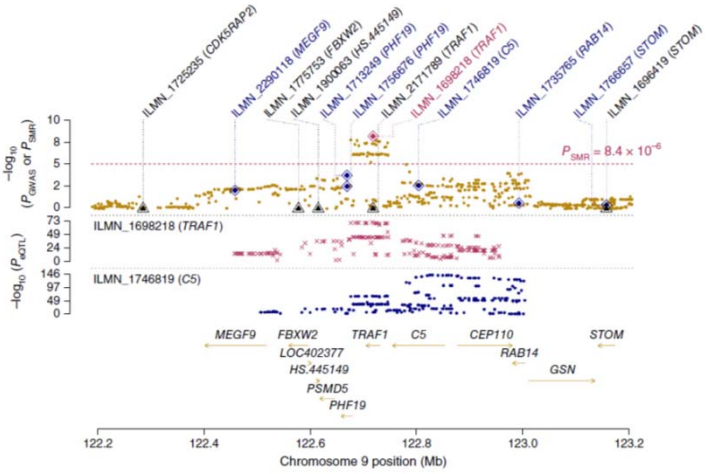
An Example: TRAF1 not C5 explains the RA association

The GWAS credible interval in the vicinity of TRAF1 contains the peak P_{SMR} with one of the *TRAF1* probes, not with the *C5* probe.

The eQTL peak in the region is actually stronger for C5 than TRAF1, but it is due to SNPs in variable LD, not to pleiotropy.

HEIDI confirms this since there is more heterogeneity for the blue than purple SNPs.

So, the TRAF1 region SNPs are GWAS significant and strong C5 eQTL, but probably mediate their effect through TRAF1.



Zhu et al (2016) *Nature Genetics* **48**: 481-487 "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets"

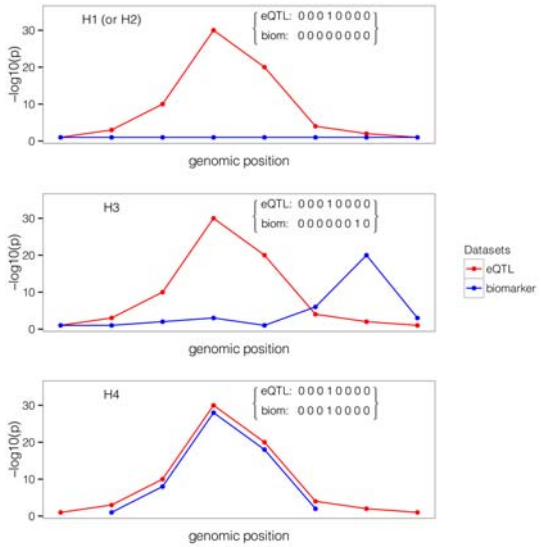
Coloc: A Bayesian test for colocalization of pairs of association signals

H1 is the hypothesis that there is only an eQTL signal at a locus

H2 is the hypothesis that there is only a GWAS signal at a locus.

H3 is the hypothesis that there are two independent eQTL and GWAS signals in linkage.

H4 is the strong hypothesis that the same SNP (not just the locus) is responsible for both the GWAS and eQTL.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383 "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics"

Examples of H3 and H4

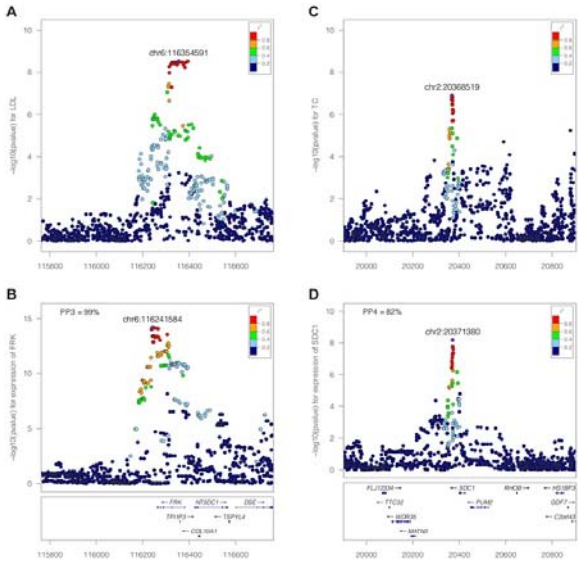
On the left, the profile of association at the FRK locus with LDL (top) is very different from that with *FRK* expression.

H3 is the supported hypothesis.

On the right, even though there are two different peak SNPs, they are in the same strong LD region and the profiles are almost the same for Total Cholesterol and *Soc1* expression.

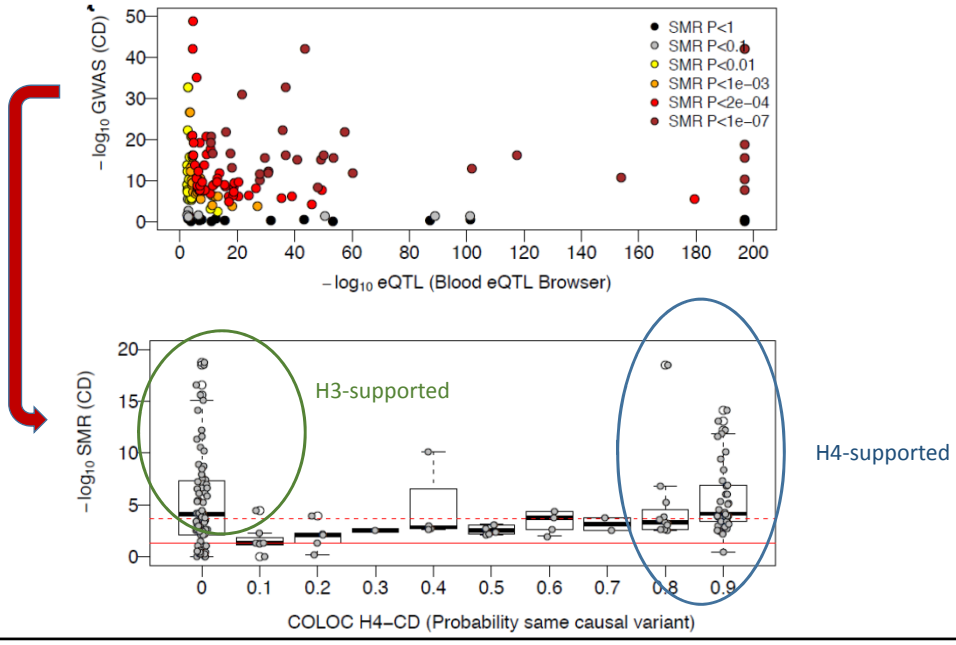
H4 is the supported hypothesis.

Bayesian analysis evaluate each H relative to the other four and generates a confidence level for the most likely one.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383 "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics"

SMR and coloc are complementary in our Crohns Disease study



Limitations of colocalization analyses

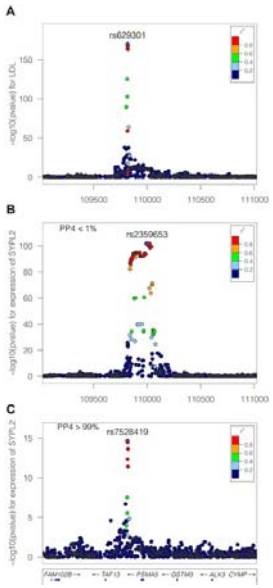
Heavily dependent on statistical power of the contributing analyses, which is generally relatively low

Depends on high quality imputation if the SNPs are not directly typed

Assumes that the GWAS and eQTL are evaluated on the same population (there is no stratification)

A negative result may arise if the incorrect tissue is being studied for the gene expression

Assumes there is a single causal variant at each locus for each effect (which is very unlikely) although this example shows that conditional analysis has the potential to resolve joint effects



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383 "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics"

Joint Mapping

A variety of open source methods are appearing that utilize Bayesian methods to perform joint mapping of eQTL

A statistical framework for joint eQTL analysis in multiple tissues.
Flutre T, Wen X, Pritchard J, Stephens M. *PLoS Genet*. 2013 **9**(5): e1003486.

This paper shows that combining signals across tissues increases power while also allowing assessment of whether the effect sizes are different in different cell types. Implemented in eQTLBMA software.

Cross-population joint analysis of eQTLs: Fine mapping and functional annotation.
Wen X, Luca F, Pique-Regi R. *PLoS Genet.* **11**(4): e1005176.

This paper shows that combining signals across populations increases power while also allowing assessment of how incorporating ENCODE data improves resolution. Implemented in FM QTL software.

Efficient integrative multi-SNP association analysis via Deterministic Approximation of Posteriors
Wen X, Lee Y, Luca F, Pique-Regi R. *AM J Hum. Genet.* **98**(6): 1114-1129.

This paper extends the framework for incorporating ENCODE data while allowing for multiple causal variants at each locus. Implemented in DAP software: http://github.com/xqwen/dap/