GGSB Prelim Q5 – Hang Chen

What is colocalization:

- Two association analyses (GWAS & eQTL)
- Associated SNPs are overlapped

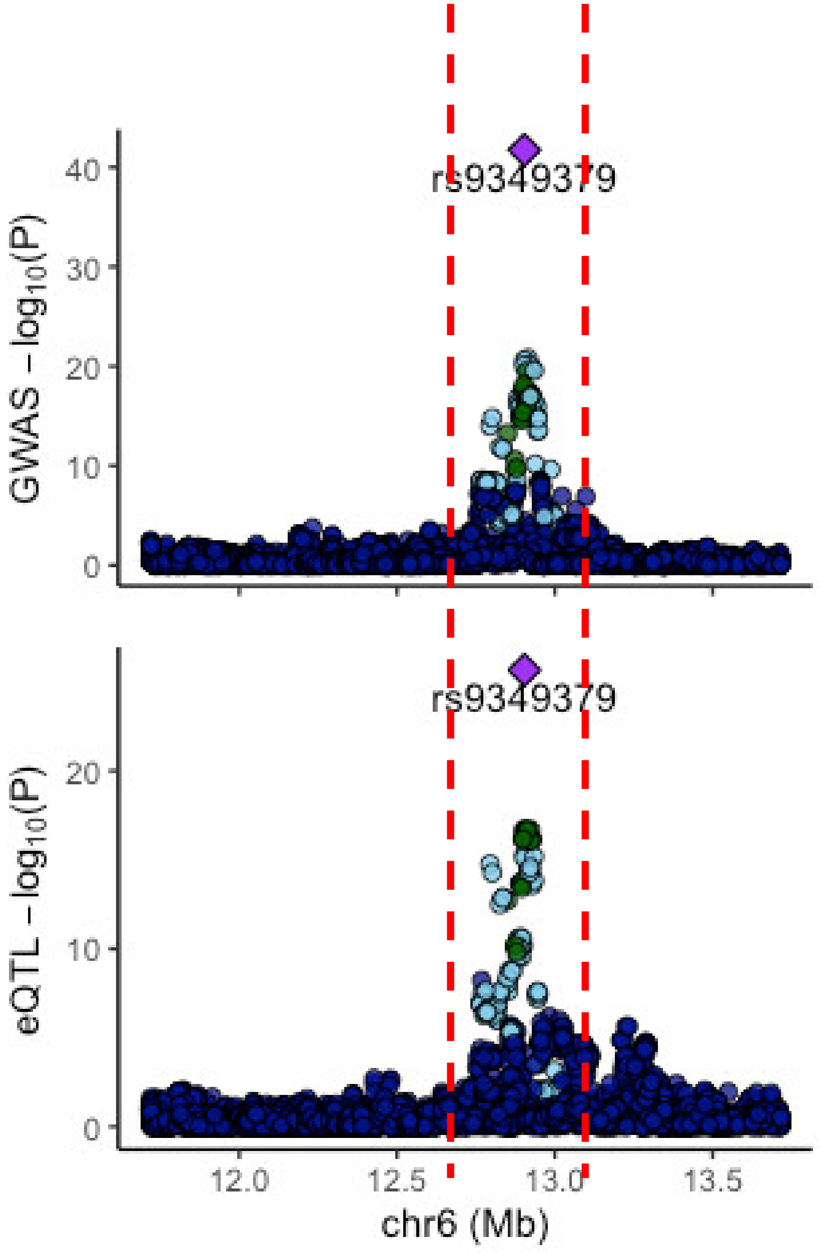
Why colocalization:

- GWAS lack molecular mechanisms
- GWAS hits are usually non-coding

How to do colocalization:

- Eyeballing?

We may not conclude a SNP-associated gene is causal for the same SNP-associated disease, but we can calculate and rank the probabilities.



(Zhang Lab@Columbia, 2019)

Question a)

1. Pleiotropy

```
SNP – X
 |
 | same SNP
 |
SNP – X' – Y
```

2. LD

```
SNP – X
 |
 | two SNPs in perfect LD
 |
SNP' – X' – Y
```

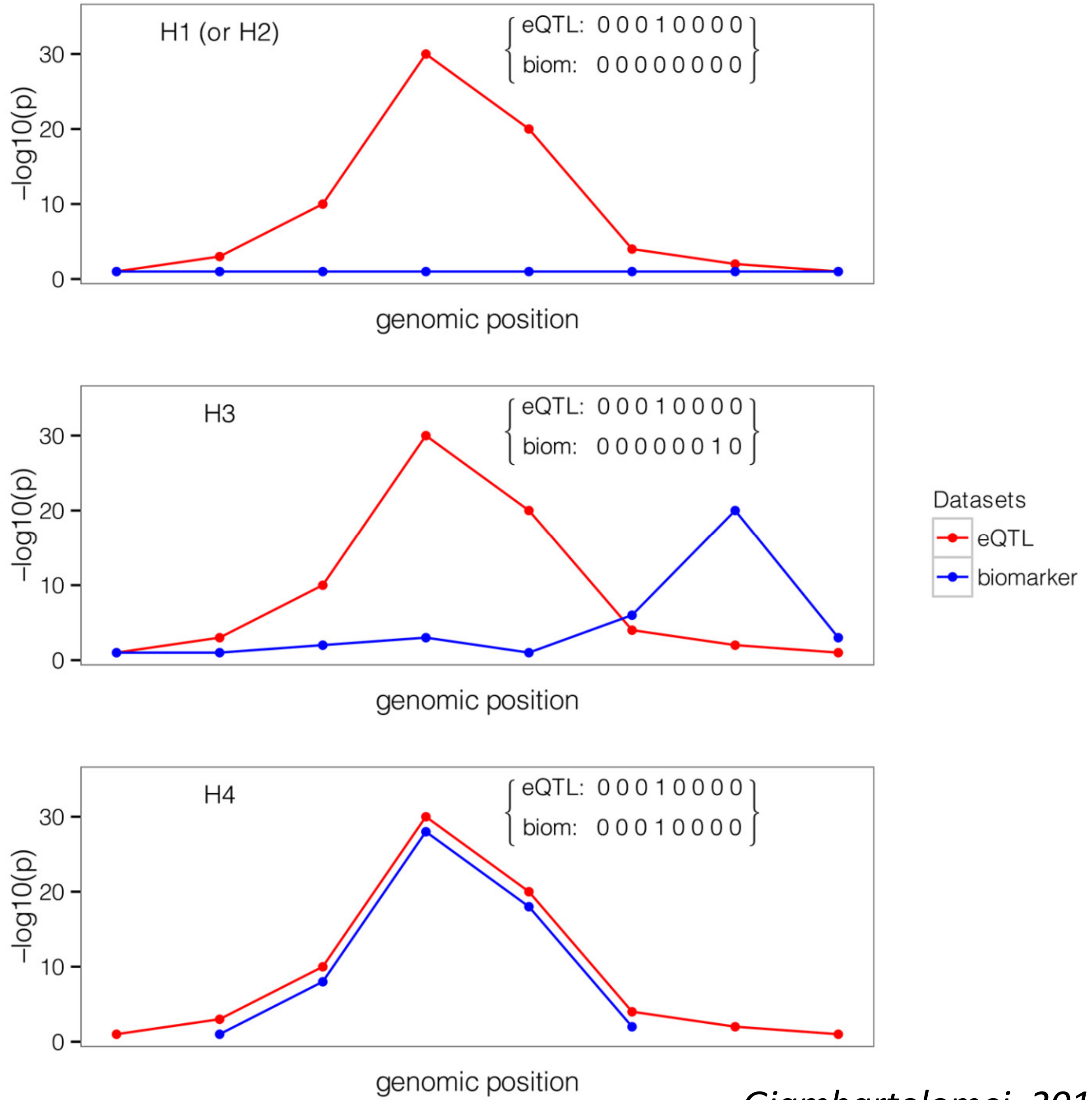
3. Intermediate

```
SNP – Int. – X
      |
      | same Int.
      |
      | – X' – Y
```

(X=gene, Y=disease)

Colocalization analysis

- Sequencing and routine QCs
- Two separate association analyses
- For each region, convert the association results to **binary** (each pair is a **configuration,** S)



Giambartolomei, 2014

- Then, test the five hypothesis:

  For a region,

  - H_0: No association with either trait
  - H_1: Association with trait 1, not with trait 2
  - H_2: Association with trait 2, not with trait 1
  - H_3: Association with trait 1 and trait 2, two independent SNPs
  - H_4: Association with trait 1 and trait 2, one shared SNP

- Assume there are Q SNPs in a region, and for each SNP, the probability of that it is associated with trait 1 is p_1, with trait 2 p_2, with both traits p_{12}, with no traits p_0.

  $(p_0 + p_1 + p_2 + p_{12} = 1)$

Question b)

Probabilities for each configurations:

- $P(S_0) = (p_0)^Q$
- $P(S_1) = (p_0)^{Q-1} \cdot p_1$
- $P(S_2) = (p_0)^{Q-1} \cdot p_2$
- $P(S_3) = (p_0)^{Q-2} \cdot p_1 \cdot p_2$
- $P(S_4) = (p_0)^{Q-1} \cdot p_{12}$

The *coloc* summarizes the results of the five hypotheses as posterior probabilities ($\mathrm{PP0}, \mathrm{PP1}, \mathrm{PP2}, \mathrm{PP3}, \mathrm{PP4}$).

Bayes’ theorem:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)}\cdot\mathrm{P(A)}}{\mathrm{P(B)}}$$

- posterior probability: $\mathrm{P(A|B)}$
- prior probability: $\mathrm{P(A)}$

- Bayes factor: $\frac{\mathrm{P(D|S_4)}}{\mathrm{P(D|S_0)}}$
  (the ratio of the likelihood of one particular hypothesis to the likelihood of another)

- Approximate Bayes factor: can be calculated from summary statistics (*p*-values) using Wakefield’s method.

Question c)

Deriving the PPs as a function of BFs:

$$\mathrm{PP4} = \mathrm{P(H_4|D)}$$

D: observed data

$$= \frac{\mathrm{P(D|S_4)}\cdot\mathrm{P(S_4)}}{\mathrm{P(D)}}$$

Reformatting P(D)

$$= \frac{\mathrm{P(D|S_4)}\cdot\mathrm{P(S_4)}}{\sum_{s\in s_h} \mathrm{P(D|S)}\cdot \mathrm{P(S)}}$$

Reformatting to BF

$$= \frac{\frac{\mathrm{P(D|S_4)}}{\mathrm{P(D|S_0)}}\cdot\mathrm{P(S_4)}}{\sum_{s\in s_h} \frac{\mathrm{P(D|S)}}{\mathrm{P(D|S_0)}}\cdot\mathrm{P(S)}}$$

$$= \frac{BF_4\cdot\mathrm{P(S_4)}}{\sum BF_h\cdot\mathrm{P(S)}}$$

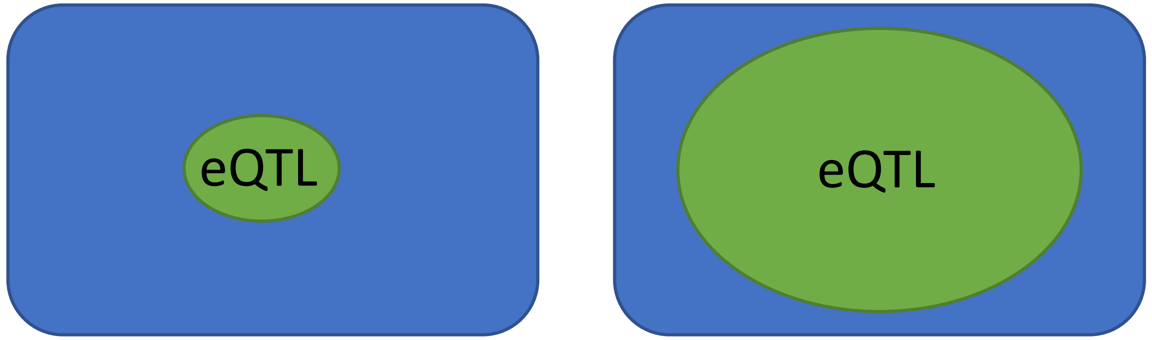
Slightly different from the equation in the *coloc* paper, but mathematically equivalent

$$\mathrm{PP4} = \frac{BF_4 \cdot \mathrm{P}(\mathrm{S}_4)}{\sum BF_h \cdot \mathrm{P}(\mathrm{S})}$$

About the prior:

- ***coloc***: $P_1 = 10^{-4}$, $P_2 = 10^{-4}$, $P_{12} = 10^{-6}$
  (meaning: 1/100 GWAS hits are also eQTL)
- ***eCAVIAR***: $P_{12} = P_1 \cdot P_2$ (two fine-mappings)
  (meaning: no eQTL enrichment in GWAS hits comparing to the whole genome)

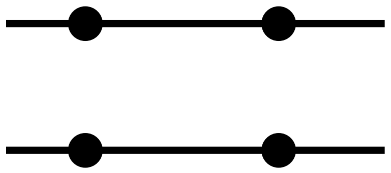
About eQTL enrichment in GWAS hits:



All SNPs

GWAS hits

Question e): an extreme example



One is GWAS hit

One is eQTL hit

- Assume the two SNPs are in perfect LD.
- We can only know that one of them is associated with trait-1 and one of them with trait-2.
- If there is no enrichment, the probability of colocalization in this region:

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

- If the enrichment level is extremely high, the probability of colocalization in this region:

$$\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1$$

- PPs are sensitive to priors. In the *eCAVIAR* case (two fine-mappings), fewer colocalization events will be called.

Question d/f/g) ***enloc***

- Main idea: evaluate eQTL enrichment level in GWAS hits from the original data.
- Method: Regress GWAS annotation odds ratio on eQTL annotation

$$\log\left[\frac{P(\gamma_i = 1)}{P(\gamma_i = 0)}\right] = \alpha_0 + \alpha_1 \cdot d_i$$

- γ_i: GWAS annotation; d_i: eQTL annotation
- α_1: indicates the enrichment level
  (an empirical way to assign the prior for PPs)

Relationship between *coloc* and *enloc*:

- *coloc* is a special case of *enloc.*
- *coloc* requires artificially assigned priors, or in other words, it bypasses the enrichment level assessing.
- *enloc* calculate the empirical priors from the data, which provides more accurate colocalization events calling.

Special thanks to Jing Gu, Charles Zhou, and Xuanyao Liu group.