MOLOCO: A tool for detecting shared genetic variants across multiple phenotypes

Jimmy Z Liu and Joseph K Pickrell

June 7, 2016

1 Introduction

MOLOCO is an extension of coloc [Giambartolomei et al., 2014]. Our approach generalizes coloc to detect colocalization among any number of traits concurrently (rather than just two) and also takes into account sample overlap between the original studies.

2 Method

In this section, we describe the model used to estimate the probability that within a genomic locus across a set of traits, there is a single causal genetic variant that is shared across two or more of the traits. Our approach is a multiple-trait extension of the two-trait model described by Giambartolomei et al. [2014]. We wish to estimate the posterior probability that the observed data at a locus is consistent with a particular combination (or configuration) of shared and non-shared causal variants among the traits tested. We re-iterate the approach for two traits as described in Giambartolomei et al. [2014] for completeness, and also extend the approach beyond two traits. We also describe an approach to account for overlapping samples between the studied phenotypes.

In the simplest two-trait situation, we wish to estimate posterior probabilities for five possible configurations:

0. No variants associated with either trait
1. Causal variant associated with trait 1, not trait 2
2. Causal variant associated with trait 2, not trait 1
3. Causal variant associated with both traits 1 and 2 and are not shared
4. Causal variant associated with both traits 1 and 2 and are shared

For three traits, there are 15 possible configurations (e.g. there are two causal variants, the first variant is shared by two of the traits and the second variant is independent in the third); for four traits, there are 52 configurations and so on.

For each configuration S and observed data D, The likelihood of configuration h relative to the null (H_0) is given by (ref - Gianbartolomei):

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)} \tag{1}$$

Here, $\frac{P(D|S)}{P(D|S_0)}$ is the Bayes factor for each configuration, and $\frac{P(S)}{P(S_0)}$ is the prior probability of that configuration (see below).

2.1 Bayes factor computation

Bayes factors measure the relative support that a SNP is associated with a trait compared with the null model of no association. Approximate bayes factors were computed from summary association statistics using the Wakefield method [Wakefield, 2009]. For each SNP, let $\hat{\beta}$ be the maximum likelihood estimate of β (e.g. the genotypic log odds ratio for a case/control study or the regression coefficient for a quantitative trait) and $\sqrt{V}$ be the standard error of that estimate. Let the Z-score, $Z = \frac{\hat{\beta}}{\sqrt{V}}$. We set a normal prior on the true effect size $\beta_1 \sim N(0, W)$. The Wakefield approximate Bayes factor (WBF) for that SNP is then:

$$WBF = \sqrt{1-r} \times exp\Big[\frac{Z^2}{2}r\Big] \tag{2}$$

where $r = \frac{W}{W+V}$. We set $W = 0.1$ (ref paper here suggesting 0.1 ok?).

For each SNP and two traits, there are four possible models that can be considered:

0. M_0: The SNP is not associated with either trait
1. M_1: The SNP is associated with the first trait (but not the second)
2. M_2: The SNP is associated with the second trait (but not the first)
3. M_3: The SNP is associated with both traits

For the three alternative models, the Bayes factors are:

$$BF_1 = WBF_1 \tag{3}$$
$$BF_2 = WBF_2 \tag{4}$$
$$BF_3 = WBF_1 \times WBF_2 \tag{5}$$

BF_3 results from the assumption that the traits are independent and do not share overlapping controls. We relax this assumption later on. In general, across a set of traits $N = \{1, 2, 3...\}$, let $N_m \subseteq N$ be the subset of traits that share this SNP as a causal variant. The evidence to support the various combinations of N_M is:

$$BF^{(m)} = \prod_{i \in N_m} WBF_i \tag{6}$$

2.2 Prior probabilities

For each SNP, we assign a prior probability for each of the underlying models. In the two-trait example with four models:

0. π_0: Prior probability that the SNP is not associated with either trait
1. π_1: Prior probability that the SNP is associated with the first trait (but not the second)

2. π_2: Prior probability that the SNP is associated with the second trait (but not the first)

3. π_3: Prior probability that the SNP is associated with both traits

and $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$. In theory, we can set a different prior for each SNP and each combination N_m, though in practice, we assign prior probabilities according to how many traits that SNP is associated with (and constant across SNPs).

2.3 Configuration likelihoods

Using the BFs for each SNP and prior probabilities, we can compute the terms of equation 1 for a region with Q SNPs. In the two-trait example with five possible configurations:

$$\frac{P(H_0|D)}{P(H_0|D)} = 1 \tag{7}$$

$$\frac{P(H_1|D)}{P(H_0|D)} = \pi_1 \times \sum_{j=1}^{Q} BF_{1j} \tag{8}$$

$$\frac{P(H_2|D)}{P(H_0|D)} = \pi_2 \times \sum_{j=1}^{Q} BF_{2j} \tag{9}$$

$$\frac{P(H_3|D)}{P(H_0|D)} = \pi_1 \times \pi_2 \times \sum_{j}^{Q} \sum_{k}^{Q} BF_{1j} BF_{2j} I[j \neq k] \tag{10}$$

$$\frac{P(H_4|D)}{P(H_0|D)} = \pi_3 \times \sum_{j=1}^{Q} BF_{1j} BF_{2j} \tag{11}$$

where I is an indicator function equal to 1 if j and k are not the same, and 0 otherwise. Equation 10 can also be written as:

$$\frac{P(H_3|D)}{P(H_0|D)} = \pi_1 \times \pi_2 \times \sum_{j=1}^{Q} BF_{1j} \sum_{j=1}^{Q} BF_{2j} - \left[\frac{\pi_1 \times \pi_2}{\pi_3} \times \frac{P(H_4|D)}{P(H_0|D)} \right] \tag{12}$$

The likelihood of there being a set of m causal associations among M number of traits can be generalized as:

$$\frac{P(H_h|D)}{P(H_0|D)} = \prod_{i \in m} \pi^{(i)} \sum_{j=1}^{Q} BF_j^{(i)} - \frac{\prod_{i \in m} \pi^{(i)}}{\pi^{(1,2,\ldots M)}} \sum_{j=1}^{Q} \pi^{(1,2,\ldots M)} BF_j^{(1,2,\ldots m)} \tag{13}$$

Here, the superscripts (i) above π and BF indicate the posterior probability and Bayes factor respectively that the combination of traits i share a common causal variant.

For example, let there be three traits (1, 2 and 3) with two independent associations. The first signal colocalizes with traits 1 and 2, while the second is an independent association with trait 3. Hence, $M = 3$, $m = \big\{(1,2),(3)\big\}$ and:

$$\frac{P(H_h|D)}{P(H_o|D)} = \pi^{(1,2)} \pi^{(3)} \sum_{j=1}^{Q} BF_j^{(1,2)} \sum_{j=1}^{(Q)} BF_j^{(3)} - \frac{\pi^{(1,2)} \pi^{(3)}}{\pi^{(1,2,3)}} \sum_{j=1}^{Q} \pi^{(1,2,3)} BF_j^{(1,2,3)} \tag{14}$$

Similarly, let there be five traits (1, 2, 3, 4 and 5) with two independent associations. First one is common to traits 1, 2, 3, the second one common to traits 4 and 5. Hence, $M = 3$, $m = \big\{(1,2,3),(4,5)\big\}$ and:

$$\frac{P(H_h|D)}{P(H_o|D)} = \pi^{(1,2,3)}\pi^{(4,5)} \sum_{j=1}^{Q} BF_j^{(1,2,3)} \sum_{j=1}^{Q} BF_j^{(4,5)} - \frac{\pi^{(1,2,3)}\pi^{(4,5)}}{\pi^{(1,2,3,4,5)}} \sum_{j=1}^{Q} \pi^{(1,2,3,4,5)} BF_j^{(1,2,3,4,5)} \tag{15}$$

Then, the posterior probability supporting configuration h among H possible configurations, is:

$$PP_h = P(H_h|D) \tag{16}$$

$$= \frac{P(H_h|D)}{\sum_{i=0}^{H} P(H_i)} \tag{17}$$

$$= \frac{\frac{P(H_h|D)}{P(H_0|D)}}{1 + \sum_{i=1}^{H} \frac{P(H_i|D)}{P(H_0|D)}} \tag{18}$$

2.4 Accounting for overlapping samples

Thus far we have assumed that the phenotypes studied do not contain any overlapping individuals. In practice, individuals in some cohorts may partially or completely overlap, resulting in correlation among the SNP effect sizes. This may result in overestimates of the models that favor sharing of causal variants between the traits. To adjust SNP Bayes factors to account for this overlap, we extend the method described in Pickrell et al. [2016] to more than two traits.

For a given SNP and a set of N traits, let $\hat{\beta}_i$ and V_i be the estimated effect size and variance respectively for trait i, and $Z_i = \frac{\hat{\beta}_i}{\sqrt{V_i}}$. Let C_{ij} be the phenotypic correlation between traits i and j. The distributions of the effect sizes follows a multivariate normal distribution:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix}, \begin{bmatrix} V_1 & C_{1,2}\sqrt{V_1 V_2} & \cdots \\ C_{2,1}\sqrt{V_2 V_1} & V_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \right) \tag{19}$$

We place a multivariate prior on β_i with variance W_i, such that:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} W_1 & C_{1,2}\sqrt{(V_1+W_1)(V_2+W_2)} - C_{1,2}\sqrt{V_1 V_2} & \cdots \\ C_{2,1}\sqrt{(V_2+W_2)(V_1+W_1)} - C_{2,1}\sqrt{V_2 V_1} & W_1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \right) \tag{20}$$

Using this prior, we can estimate the posterior predictive distribution of the estimated effect sizes under alternate hypotheses:

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\end{bmatrix} | H_m \sim MVN\left(\begin{bmatrix}0\\0\\ \vdots\end{bmatrix}, \begin{bmatrix}V_1+W_1 & C_{1,2}\sqrt{(V_1+W_1)(V_2+W_2)} & \cdots\\ C_{2,1}\sqrt{(V_2+W_2)(V_1+W_1)} & V_2+W_2 & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}\right). \tag{21}$$

Here, the alternate hypotheses correspond to different sets of N_m traits that share an association at that SNP. In equation 21, we set $W_i = 0.1$ if trait $i \in N_m$ and 0 otherwise. Under the null hypothesis that there is no association with any traits at this SNP:

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\end{bmatrix} | H_0 \sim MVN\left(\begin{bmatrix}0\\0\\ \vdots\end{bmatrix}, \begin{bmatrix}V_1 & C_{1,2}\sqrt{V_1V_2} & \cdots\\ C_{2,1}\sqrt{V_2V_1} & V_2 & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}\right). \tag{22}$$

For each SNP, we can then calculate the Bayes factor for each subset of traits $N_m \in N$ that share this SNP as a causal variant:

$$BF^(m) = \frac{f(\vec{\beta}|H_m)}{f(\vec{\beta}|H_0)}, \tag{23}$$

where $f(\vec{\beta}|H_m)$ and $f(\vec{\beta}|H_0)$ are the probability density functions of the multivariate normal distributions described in equations 21 and 22 respectively. By default, all Bayes factors are calculated using equation 23 than those described in section 2.1.

References

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V., 2014. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet*, **10**(5):1–15.

Pickrell, J. K., Berisa, T., Liu, J. Z., Segurel, L., Tung, J. Y., and Hinds, D. A., 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, **advance online publication**:–.

Wakefield, J., 2009. Bayes factors for genome-wide association studies: comparison with P-values. *Genetic Epidemiology*, **33**(1):79–86.