Probabilistic assignment of variants into infection susceptibility vs. disease severity effects
Matti Pirinen (matti.pirinen@helsinki.fi) 4-Dec-2021

We consider an assignment of variants reported in "covid_hgi_v6_topvariants_prob.txt" into two groups based on whether they seem to affect susceptibility to corona virus infection (INF) or severity of the infection (SEV). For this analysis, we meta-analyzed GWAS summary statistics from such versions of B2 (hospitalized for COVID vs. population) and C2 (infected vs. population) GWAS that only included those studies that had contributed some data to B2. Therefore, all studies included in this analysis had made some effort to distinguish from all the infected individuals those individuals who additionally were hospitalized. The sample sizes of these GWAS were 23,988 cases / 2,834,885 controls for B2 and 114,516 cases / 2,138,237 controls for C2.

We assume that all hospitalized cases (B2 cases) were included among the infected cases (C2 cases) of the corresponding study and that, for each study, the controls of B2 and C2 had a maximum overlap possible given the control counts in the two data sets.

Next, we explain how we defined the statistical models that represent INF and SEV, and how we compared these models at each SNP.

Intuitively, INF represents a variant that associates with susceptibility of infection but has no effect on the severity of infection. The allele frequency of such a variant is similar among the hospitalized cases (B2 cases) as it is among all infected (C2 cases). Thus, under the INF model, we assume similar effect size between C2 and B2, i.e., $\beta_{C2} \approx \beta_{B2}$.

Model SEV represents a variant that affects severity of infection ($|\beta_{B2}| > 0$) but not susceptibility to infection. If the cases of our susceptibility scan C2 were a random subset of infected individuals, then under the SEV model, $\beta_{C2} \approx 0$. However, since our C2 cases are strongly enriched for severe cases, we expect that in our data also $|\beta_{C2}| > 0$ even when a variant is affecting severity of infection but not susceptibility to infection. We expect that, for such variants, the effect size in C2 is proportional to its effect in B2, i.e., $\beta_{C2} \approx w_{SEV}\beta_{B2}$, where the constant of proportionality, $w_{SEV} < 1$, depends on the proportion of all C2 cases that are also B2 cases. If we imagine C2 analysis as a fixed-effect meta-analysis between B2 and an (imaginary) "non-severe infection vs. population" analysis that had no sample overlap with B2 analysis, then $w_{SEV} = n_{B2}^{(eff)}/n_{C2}^{(eff)}$, where $n_i^{(eff)} = 4R_iS_i/N_i$ is the effective sample size of study i with R_i the number of controls, S_i the number of cases and $N_i = R_i + S_i$. In our data, $n_{B2}^{(eff)}/n_{C2}^{(eff)} \approx 0.208$, where the effective sample sizes are computed by summing the effective sample sizes over individual studies of B2 and C2 analyses. Even when controls of B2 analysis and the imaginary "non-severe infection vs. population" analysis overlapped completely, the value of w_{SEV} would change little in these data. After accounting for the overlap in controls, we estimated a value of $w_{SEV} = 0.20$ that we used in the analyses described below.

To derive correlation r_{B2C2} between the effect size estimators of B2 and C2 analyses due to overlapping samples, we used formula

$$r_{B2C2} = \frac{\sum_{k=1}^{K} \sqrt{n_{B2,k}^{(eff)}\, n_{C2,k}^{(eff)}}\, r_{B2C2,k}}{\sqrt{\sum_{k=1}^{K} n_{B2,k}^{(eff)} \sum_{k=1}^{K} n_{C2,k}^{(eff)}}}$$

where subscript k refers to individual studies. The correlation $r_{B2C2,k}$ between B2 and C2 for study k is computed as in Bhattacharjee et al. (2012):

$$r_{B2C2,k} = \sqrt{n_{B2,k}^{(eff)}\, n_{C2,k}^{(eff)}} \left(\frac{S_{B2C2,k}}{S_{B2,k} S_{C2,k}} + \frac{R_{B2C2,k}}{R_{B2,k} R_{C2,k}} \right)$$

where $S_{B2C2,k}$ is the number of shared B2 and C2 cases in study k and similarly $R_{B2C2,k}$ is the number of shared controls. By applying this to the data, we estimated $r_{B2C2} = 0.45$.

With these estimates of $w_{SEV} = 0.2$ and $r_{B2C2} = 0.45$, and with the observed data at any one SNP containing effect estimates $(\hat{\beta}_{B2}, \hat{\beta}_{C2})$ and their standard errors (s_{B2}, s_{C2}), we derive the two models, INF and SEV, as follows.

Prior distribution for effects is zero-centered bivariate normal distribution

$$\begin{pmatrix} \beta_{B2} \\ \beta_{C2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Theta_i \right), \text{ with}$$

$$\Theta_{INF} = \tau^2 \begin{pmatrix} 1 & 1-\eta_{INF} \\ 1-\eta_{INF} & 1 \end{pmatrix} \text{ and}$$

$$\Theta_{SEV} = \tau^2 \begin{pmatrix} 1 & (1-\eta_{SEV}) w_{SEV} \\ (1-\eta_{SEV}) w_{SEV} & w_{SEV}^2 \end{pmatrix}.$$

We have used value $\tau = 0.1$ to define the expected effect sizes of the B2 analysis (implying roughly that 95% of the true effect sizes of the risk variants have odds-ratio below 1.2). By tuning the parameters η_{INF} and η_{SEV}, we can define how much deviation real effects can have from the theoretical relationships $\beta_{C2} = \beta_{B2}$ and $\beta_{C2} = w_{SEV}\beta_{B2}$ corresponding to models INF and SEV, respectively. We have set these values in such a way that, under both models, the mean (Euclidean) distance between the effect size and the corresponding line is 0.0025 and 95% of the effects are within 0.006 units from the line. This happens when $\eta_{INF} = 0.001$ and $\eta_{SEV} = 0.013$.

The likelihood for the observed data under both models is Gaussian

$$\begin{pmatrix} \hat{\beta}_{B2} \\ \hat{\beta}_{C2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_{B2} \\ \beta_{C2} \end{pmatrix}, \Sigma \right), \text{ where } \Sigma = \begin{pmatrix} s_{B2}^2 & s_{B2} s_{C2} r_{B2C2} \\ s_{B2} s_{C2} r_{B2C2} & s_{C2}^2 \end{pmatrix}.$$

It follows from Trochet et al. (2019) that we can analytically integrate the likelihood with respect to the prior distributions and the resulting marginal likelihood for each model is proportional to a Gaussian density function evaluated at the observed effect size estimates as

$$\Pr(\text{DATA} \mid \text{Model } i) \propto f_N\left(\begin{pmatrix} \hat{\beta}_{B2} \\ \hat{\beta}_{C2} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Theta_i + \Sigma \right).$$

We set equal prior probability on each model (i.e. 50% on INF and 50% on SEV), and consequently the posterior probabilities of the models will be proportional to their marginal likelihoods. These posterior probabilities are reported in "covid_hgi_v6_topvariants_prob.txt".

A limitation of this approach is that it classifies every variant between the two fixed models INF and SEV without considering a possibility that the variant might not fit either of these two models very well. We have chosen this approach since, based on the data shown in Figure "covid_hgi_v6_b2_c2_effects_plot.pdf", a large majority of the variants are well aligned with either INF or SEV model. Consequently, we kept this model comparison simple and only between INF and SEV rather than complicated it by inclusion of some additional models that would have little support from the observed data.

1. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P; GliomaScan Consortium, Yeager M, Chung CC, Chanock SJ, Chatterjee N. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet. 2012 May 4;90(5):821-35. doi: 10.1016/j.ajhg.2012.03.015. PMID: 22560090; PMCID: PMC3376551.

2. Trochet H, Pirinen M, Band G, Jostins L, McVean G, Spencer CCA. (2019): Bayesian meta-analysis across genome-wide association studies of diverse phenotypes. *Genetic Epidemiology*, 43: 532-547. https://doi.org/10.1002/gepi.22202