July 2021

Covariance of Interdependent Samples with Application to GWAS

Daniel Krefl[a,b] and Sven Bergmann[a,b,c]
[a] *Department of Computational Biology, University of Lausanne, Switzerland*
[b] *Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[c] *Dept. of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa*

We devise a significance test for covariance of samples not drawn independently, but with known inter-sample covariance structure. We propose a test distribution which is a linear combination of χ^2 distributions, with positive and negative coefficients. The corresponding cumulative distribution function can be efficiently calculated with Davies' algorithm at high precision. As an application, we suggest a test for dependence between SNP-wise effect sizes of two genome-wide association studies at the level of genes. This test can be extended to detect gene-wise causal links. We illustrate this method by uncovering potential shared genetic links between severity of COVID-19 and (1) being prescribed class M05B medication (drugs affecting bone structure and mineralization), (2) vitamin D (25OHD) and (3) serum calcium concentrations. Our method detects a potential role played by chemokine receptor genes linked to T_H1 versus T_H2 immune reaction, a gene related to integrin beta-1 cell surface expression, and other genes potentially impacting severity of COVID-19.

I. INTRODUCTION

Pearson's sample correlation is one of the most used techniques in data analysis, independent of the specific field of science. It is defined as the covariance divided by the standard deviations of two random variables and gives a measure of linear dependence. The core underlying requirement is that the observed samples are drawn independently from a joint bivariate distribution. However, not all applications possess independently drawn samples.

One particular application with dependent samples occurs in genome-wide association studies (GWAS). Specifically, GWAS correlate genotypes, most commonly single nucleotide polymorphisms (SNPs), with a phenotype of interest, both measured in the same study population. For human studies usually between 1 and 10 million SNPs are considered, and in most GWAS each of these SNPs is tested independently for correlation with the phenotype. By now, thousands of such GWAS have been conducted and identified a plethora of statistically significant associations of SNPs with complex traits. For most traits, in particular complex diseases or their risk factors, that have been assessed in very large populations (100k-1M subjects), usually hundreds of SNPs turn out to be significant, even after stringent correction for multiple hypotheses testing. Individual SNP-wise effect sizes are often very small, but add up to sizable narrow sense heritability, pointing to a polygenic genetic architecture [1]. Mapping SNP-wise effects to genes and annotated gene-sets (*pathways*) [2, 3] can yield valuable insights into the genetic underpinning and potential pathomechanisms of complex diseases and aid drug discovery and re-purposing.

However, due to Linkage Disequilibrium (LD) (*cf.*, [4]), many SNPs in close proximity are not independent, and this leads to dependencies between observed SNPs' effect sizes. This is of particular relevance when aggregating SNP-wise effects to genes or pathways. Specifically, gene-wise effects are computed by adding up the (squared) effects of all SNPs within the transcript region of a gene of interest, as well as sizable up- and downstream-regions that may contain regulatory elements of this gene. Pathway effects are computed from the gene-wise effects [2]. Thus, if not corrected for LD, the resulting gene and pathway scores may reflect the level of importance for the phenotype inaccurately. For this reason, techniques and tools have been developed to correct for LD in the aggregation process, like for example *Pascal* [2] and *MAGMA* [3]. These tools mainly differ in how they map SNP effect sizes to genes, how the LD structure is accounted for, and details of the numerical procedure to estimate significance.

Some SNPs are significantly associated with more than one trait. The phenomenon of a single genetic variant affecting two or more traits is called *pleiotropy* [5]. In the case of disease traits such a shared genetic component hints at the same functional pathology contributing to several diseases, *cf.*, [6]. At the gene level, usually a gene is considered to be relevant for two different traits if the respective gene effect sizes are significant in both traits. However, this may not be a good criterion under all circumstances. Protein coding genes often contain several independent LD blocks. Therefore, two traits may associate with genetic variation in two or more functionally different blocks of SNPs within the same gene, which may independently be significant. Hence, even though the two traits share the same significant gene, they may not share the same functional mechanism. To call a gene pleitropic, one should therefore move beyond comparing

single variants, and take all SNPs in the gene region into account.

Several methods have been proposed to uncover shared genetic origin of two traits from GWAS summary statistics: One early method is a test of co-localisation between GWAS pairs based on Bayesian statistics [7]. This method assumes that at most one association is present for each trait in the region of interest. The extension to the general case of multiple associations (which is usually the case) appears, however, to be non-trivial. A more recent method is cross-trait *LD score regression* [8], an extension of single-trait *LD score regression* (LDSR) [9], which is a method to estimate heritability and confounding biases from GWAS summary statistics. Like single-trait LDSR, cross-trait LDSR considers the effect sizes as random variables and uses *LD-scores* (*i.e.*, the sum of genetic correlations between a given SNP and all other SNPs) to estimate the genetic correlation between two traits. Yet this estimate is at the level of the entire genome, and it is difficult to use this approach to obtain estimates at the level of genes. The reason is that SNPs in the same genomic region are often in high LD, so the variables entering the linear correlation may be highly dependent, and this has to be corrected for. Similarly, the standard errors and p-values are estimated via resampling (jackknife). But this requires independent samples, which is not the case for strong LD. Another method to estimate local genetic covariance and correlation has been introduced in [10]. However, this method requires the computation of the inverse SNP-SNP correlation matrix, which is in general problematic as it often would need regularization.

Here, we propose to consider the second cross moment between the effects of SNPs for two traits within a gene region, corrected for LD, as a measure for pleiotropy. This gives us a simple, but systematic definition of co-significance of a gene for two traits, *i.e.*, if the SNP-wise cross-moment is significant. For a single SNP this reduces to testing against the product-normal distribution, and hence corresponds to a multiplicative meta-analysis, rather than an additive one like Fisher's. We show that for multiple SNPs with a non-trivial covariance structure, one can express the underlying test distribution in terms of a linear combination of χ^2 distributions, with a mixture of positive and negative coefficients. The corresponding cumulative distribution function can be efficiently calculated with Davies' algorithm at high precision. Thus, our method takes into account not only isolated significant SNPs, but all common SNPs within the gene region to call a gene co-significant for two traits. The assumption we have to make is that there is an underlying joint normal distribution, and the dependency pattern of the SNPs can be jointly inferred from a reference population. Furthermore, we assume that the study populations of the two traits do not have overlapping samples.

A limitation of our approach is that common genetic effects to different traits cannot be disentangled from potential other joint residual contributions. Yet, using GWAS data from different populations such contributions are likely to be negligible. Using the notion of Mendelian randomization, our statistic can be extended to test for a possible causal relation between the two traits, mediated by the tested gene. However, possible confounders have to be excluded by other means.

The outline of this paper is as follows: In sections II and III we introduce the probability distributions of relevance for this work. We briefly review the χ^2 distribution in section II, since it is essential for our work, as all the other distributions we consider can be expressed as a linear combination of χ^2 distributions. This holds in particular for the product-normal distribution, as we show in section III, as well as for the Variance-Gamma distribution discussed in appendix B. In section IV we explain how to perform a significance test of independence under the conditions described above. In section V we derive how to calculate the cumulative distribution function of a particular ratio distribution of relevance for a causality test. Simulated examples are discussed in section VI, followed by an application to real GWAS data in section VII. We demonstrate the utility of our method by a timely co-analysis of GWAS summary statistics on the severity of COVID-19, being prescribed class M05B medication, and vitamin D and calcium concentrations. In particular, we detect a potential role for severity of COVID-19 played by chemokine receptor genes linked to T_H1 versus T_H2 immune reaction, and a gene related to integrin beta-1 cell surface expression. Several other genes related to COVID-19, discussed before elsewhere in the literature, are replicated. In addition we uncover hints for a potential protective and/or therapeutic pathway related to being prescribed specific immune related medications (H03A, L04 and M01A).

II. LINEAR COMBINATION OF χ^2 DISTRIBUTIONS

We denote the χ^2-distribution with n degrees of freedom as $[\chi^2_n]$. It is well known that the sum of N independent $\chi^2_{n_i}$ distributed random variables v_i is again χ^2 distributed, *i.e.*,

$$\sum_i v_i \sim \left[\chi^2_{\sum_i n_i}\right] .$$

However, no closed analytic expression for the distribution Ξ of a general linear combination,

$$\sum_i a_i v_i \sim \sum_i v_i [\chi^2_{n_i}] = [\Xi]\,, \tag{1}$$

where a_i are real coefficients, is known. Nevertheless, a variety of numerical algorithms exist to compute the cumulative distribution function (cdf) of Ξ, denoted as F_Ξ, up to a desired precision. Perhaps most well known are Ruben's algorithm [11–13] and Davies' algorithm [14]. The latter is of most relevance for this work, as it allows for negative a_i.

With F_Ξ at hand, for a given real x a right tail probability p (p-value) can be calculated as

$$p = 1 - F_\Xi(x)\,. \tag{2}$$

A variety of approximations to the distribution Ξ exists, *cf.*, [15]. One of the more well known methods is the Satterthwaite-Welch approximation [16, 17], which matches the first two moments of Ξ to a Gamma distribution, denoted as Γ. In detail [18],

$$[\Xi] \approx g\,[\chi^2_h] = [\Gamma(h/2, 2g)]\,, \tag{3}$$

with

$$g := \frac{\sum_i d_i a_i^2}{\sum_i d_i a_i}\,, \quad h := \frac{\left(\sum_i d_i a_i\right)^2}{\sum_i d_i a_i^2}\,.$$

Here, d_i denotes the ith degree-of-freedom parameter of the ith χ^2 in equation (1) and $i \in \{1, ..., N\}$.

To our knowledge, the quality of second order approximations to Ξ at very high precision (below double precision) has not been evaluated in the literature, *cf.*, [15]. However, it is expected that in general the approximation breaks down far in the tail.

III. PRODUCT-NORMAL DISTRIBUTION

A central role in this work is played by the product-normal distribution, the distribution of the product of two normal-distributed random-variables w and z. The moment generating function for joint normal samples with correlation ϱ reads [19]

$$M_{w,z}(\nu) = \frac{1}{\sqrt{(1-(1+\varrho)\nu)(1+(1-\varrho)\nu)}}\,.$$

Our key observation is that the above moment generating function factorizes into moment generating functions of the gamma distribution, $M_\Gamma(\nu|\alpha,\beta) = \frac{1}{(1-\beta\nu)^\alpha}$, *i.e.*,

$$M_{w,z}(\nu) = M_\zeta(\nu|1/2, 1+\varrho)\, M_{-\xi}(\nu|1/2, 1-\varrho)\,.$$

Therefore,

$$zw \sim [\Gamma(1/2, 1+\varrho)] - [\Gamma(1/2, 1-\varrho)]\,. \tag{4}$$

For general parameters of the two gamma distributions the corresponding difference distribution is known as the bilateral gamma distribution [20]. (Note that the subtraction in equation (4) is in the distributional sense, so, even for $\varrho = 0$, the corresponding distribution does not vanish.)

Due to the well known relation between the gamma and χ^2 distributions, one can also express the product-normal distribution in terms of the χ^2 distribution introduced in the previous section. In detail,

$$zw \sim \frac{1+\varrho}{2}[\chi^2_1] - \frac{1-\varrho}{2}[\chi^2_1]\,. \tag{5}$$

The cdf of the product-normal can therefore be efficiently calculated using Davies algorithm, as the distribution (5) is simply a linear combination of χ^2_1 distributions. A similar relation can be derived for the product distribution of non-standardized Gaussian variables, albeit in terms of the non-central χ^2 distribution, *cf.*, appendix A. Note that the relation (4) allows for a simple analytic derivation of a closed form solution for the product-normal pdf, but not for the cdf. For completeness, details can be found in appendix B.

We conclude that the cdf of the product normal distribution can be calculated with Davies algorithm. In particular, note that for $\varrho = 0$ we can not make use of the Satterthwaite-Welch approximation to calculate the cdf, as the first cumulant vanishes and therefore (3) is not well defined.

IV. COVARIANCE SIGNIFICANCE TEST

Consider the index

$$I = \sum_i w_i z_i\,, \tag{6}$$

with w_i and z_i N independent samples of two random variables $z, w \sim \mathcal{N}(0,1)$. The index can also be written as

$$I = N\,\mathbb{E}(wz)\,,$$

with $\mathbb{E}$ denoting the expectation. Clearly, I is proportional to the standard empirical covariance of w and z.

For independent pairs of samples, the sampling distribution of I is simply a sum of independent product-normal distributions, hence we infer from the previous sections that for identically correlated random variables

$$I \sim \frac{1+\varrho}{2}[\chi^2_N] - \frac{1-\varrho}{2}[\chi^2_N]\,. \tag{7}$$

In particular, for $\varrho = 0$, we have that $I \sim X(N, 1)$, with X being the Variance-Gamma distribution discussed in more detail in appendix B. As mentioned already before, one should note that the difference is in the distributional sense and therefore generally non-vanishing. A null assumption of zero correlation (or some other fixed value) can therefore be tested via (2), as the cdf for I can be calculated explicitly and efficiently with Davies algorithm.

We have now all ingredients in place to discuss the problem we want to address with this paper. The main advantage of the above significance test is that it is straight-forward to relax the requirement of sample independence. That is, we can view the index I as a scalar product of random samples of $w \sim \mathcal{N}(0, \Sigma_w)$ and $z \sim \mathcal{N}(0, \Sigma_z)$, with $\mathcal{N}$ denoting here the multivariate Gaussian distribution and $\Sigma_.$ covariance matrices. For $\Sigma_. = 1$, the corresponding distribution of I is given by (7). In the general case, the inter-dependencies can be corrected for as follows.

We make use of the eigenvalue decompositions $U_w \Sigma_w U_w^T = \Lambda_w$ and $U_z \Sigma_z U_z^T = \Lambda_z$, with $\Lambda_.$ the diagonal matrix of eigenvalues of $\Sigma_.$, to decorrelate the elements of each set. The index I can then be written as

$$I = w^T z = w^T U_w^T U_w U_z^T U_z z = \hat{w}^T U_w U_z^T \hat{z} = \hat{w}^T K \hat{z}\,,$$

with $\hat{w} \sim \mathcal{N}(0, \Lambda_w)$, $\hat{z} \sim \mathcal{N}(0, \Lambda_z)$ and $K := U_w U_z^T$. In components, I reads

$$I = \sum_{i,j} K_{ij}\, \hat{w}_i \hat{z}_j\,.$$

The case of interest for this paper is $\Sigma_w = \Sigma_z =: \Sigma$ such that K is the identity matrix. In this case, making use of the moment generating function of the product-normal, as in section III, we can show that under the null of w and z being independent

$$I \sim \sum_i \frac{\lambda_i}{2}[\chi_1^2] - \sum_i \frac{\lambda_i}{2}[\chi_1^2]\,. \tag{8}$$

with λ_i the ith eigenvalue of Σ. Hence, I is distributed according to a linear combination of χ_1^2 distributions with positive and negative coefficients, and therefore the cdf and tail probability can be calculated with Davies algorithm.

Note that the above discussion can be extended to non-standardized variables (for $\varrho = 0$) via making use of the result of appendix A.

V. RATIO SIGNIFICANCE

Consider the normalized index

$$R = \frac{\sum_i w_i z_i}{\sum_j z_j^2}\,, \tag{9}$$

with w and z as in the previous section, in particular independent. The cdf for R can be calculated as follows (similarly for w and z interchanged). Clearly, for $\Sigma_w = \Sigma_z = \Sigma$ we have that

$$\Pr(R \le r) = \Pr(\hat{w}\hat{z} \le r\hat{z}^2) = \Pr((\hat{w} - r\hat{z})\hat{z} \le 0)\,.$$

We define $\hat{v} = \hat{w} - r\hat{z}$ such that $\hat{v} \sim \mathcal{N}(0, (1 + r^2)\Lambda)$. Note that the component-wise correlation coefficient ϱ between $\hat{v}$ and $\hat{z}$ reads

$$\varrho = -\frac{r}{\sqrt{1 + r^2}}\,.$$

Hence, from (5) and (A1) we deduce that

$$\begin{aligned} \hat{v}\hat{z} \sim &\sum_i \frac{\lambda_i \sqrt{1 + r^2}(1 + \varrho)}{2}[\chi_1^2] \\ &- \sum_i \frac{\lambda_i \sqrt{1 + r^2}(1 - \varrho)}{2}[\chi_1^2]\,. \end{aligned} \tag{10}$$

We conclude that

$$F_R(r) = \Pr(R \le r) = F_{\hat{v}\hat{z}}(0), \tag{11}$$

with $F_{\hat{v}\hat{z}}(0)$ the linear combination of χ_1^2 cdf evaluated at the origin. Hence, the cdf of the ratio (9) can be as well calculated with Davies algorithm. A consistency check can be performed as follows. In one-dimension, we have that R has to be Cauchy distributed. The corresponding cdf is given by $F_C = \frac{1}{2} + \frac{\arctan(r)}{\pi}$. Evaluation for various r shows agreement with values calculated from (10) in the one-dimensional case. Note that for non-standardized w_i and z_i similar expressions can be derived, albeit in terms of the non-central χ^2 distribution, *cf.*, appendix A.

VI. SIMULATIONS

In the following, several examples will be discussed, illustrating more specific aspects and applications of the theoretical material presented so far.

A. Single element

It is useful to discuss the single element case in more detail. The distribution (8) simplifies for $N = 1$ to

$$I \sim \frac{1}{2}[\chi_1^2] - \frac{1}{2}[\chi_1^2]\,,$$

which corresponds to the uncorrelated product normal distribution, *cf.*, (5). The index I for $N = 1$ is a measure of coherence (or anti-coherence) between z and w. The significance threshold curve for a fixed desired p-value,

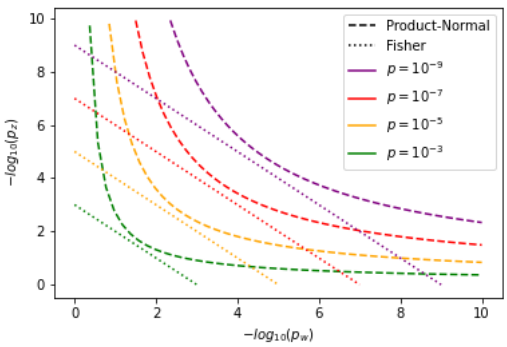

FIG. 1: Significance threshold curves in the one element case for the product-normal (dashed) and Fisher's method (dotted) for various p-values.

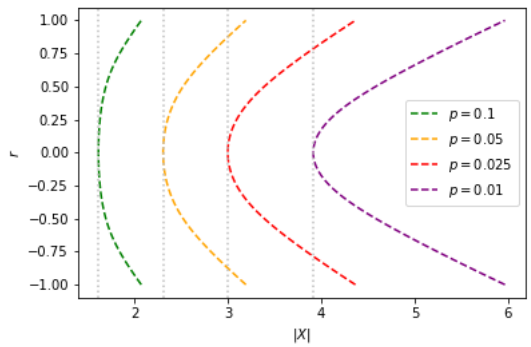

FIG. 2: Significance threshold curves of I in the two element case under variation of the inter-element correlation (y-axis). The x-axis corresponds to the argument of the tail probability defined in equation (2). The gray dotted lines mark the minimal value obtained for zero correlation.

say $p_I = 10^{-7}$, is illustrated in figure 1. The curve corresponding to a given p_I is unbounded. That is, for a given w, there is always a corresponding z such that the resulting product I is significant. This differs from Fisher's exact method which combines two p-values $p_{w,z}$ into a combined one (p_F) via $-2\log p_F = -2\log p_w - 2\log p_z \sim [\chi^2_2]$. Since Fisher's method combines significance by addition, the corresponding combined significance curve is bounded. Specifically, in the extreme case of one of the p-values being equal to one, the significance simply corresponds to that of the other p-value. In contrast, for the product-normal divergence between the p-values is penalized. If one of the p-values is large, say $p_w \simeq 1$, the other one has to be extremely small to achieve a given combined significance value, i.e. $p_z \ll p_w$ (*cf.*, figure 1). Therefore, one should see Fisher's method as being additive in the evidence, while the product-normal based method as being multiplicative.

B. Two elements

The importance of correcting for the inter-dependence between the elements of w and z can be seen easily in the $N = 2$ case. Consider the covariance matrices,

$$\Sigma = \Sigma_w = \Sigma_z = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} ,$$

such that $w \sim \mathcal{N}(0, \Sigma)$ and $z \sim \mathcal{N}(0, \Sigma)$, and with r varying. In figure 2 we show various significance threshold curves of I for varying r, as calculated from the distribution (8). Clearly, for increasing inter-element correlation, the significance threshold level rises. The magnitude of the effect increases with the desired level of significance.

C. Normal draws

Consider a correlation matrix Σ of dimension one hundred with off-diagonal elements identically set to 0.2. We draw 1000 pairs of independent samples of $\mathcal{N}(0, \Sigma)$ and calculate for each pair I defined in (6). A p-value is then obtained for each index value for the linear combination of χ^2 distributions (8) (also referred to as weighted χ^2), and for the gamma-variance distribution (B4). Recall that the latter does not correct for the off-diagonal correlations. We repeat the experiment with the off-diagonal elements set to 0.8, resulting in a stronger element-wise correlation. Resulting qq-plots for both cases are shown in figure 3.

We observe that the gamma-variance distribution (7) (with $\varrho = 0$) indeed becomes unsuitable for increasing element-wise correlation of the data sample elements. In detail, not correcting for the inter-sample correlation leads to more and more false positives with increasing correlation strength. In contrast, the weighted χ^2 distribution (8) yields stable results in both the weakly and strongly correlated regime, as is evident in figure 3.

VII. GENETIC COHERENCE

A. Generalities

As explained in the introduction, a prime example of very strong inter-element correlations are SNPs in LD.

Recall that the univariate least squares estimates of the effect sizes β reads

$$\beta_i = \frac{1}{n} x_i^T y \,, \tag{12}$$

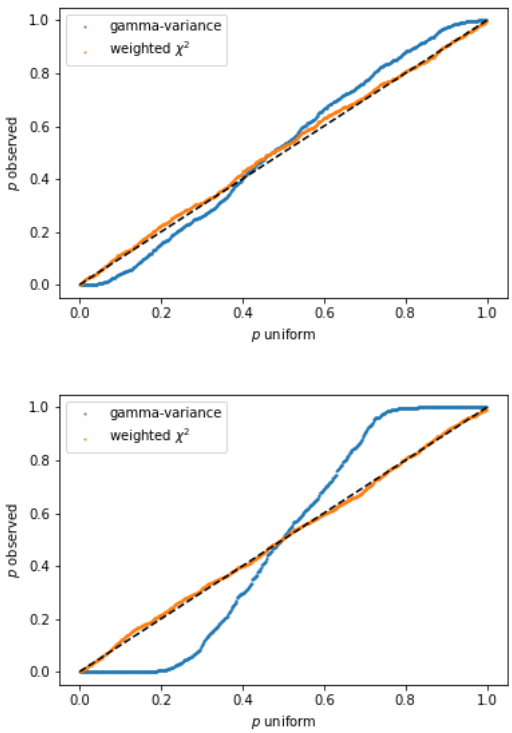

FIG. 3: QQ-plots of observed p-values resulting from the index I for 1000 pairs of samples of $\mathcal{N}(0,\Sigma)$ against uniform p-values. Top: Off-diagonal elements of Σ set to 0.2. Bottom: Off-diagonal set to 0.8. The blue curve is obtained using the gamma-variance distribution to perform the statistical test, while the orange curve is obtained via the weighted χ^2 distribution. The latter corrects for the correlation and therefore is well calibrated.

with x_i the ith column of the genotype matrix X of dimension (n,p) and y the phenotype vector of dimension n. Both x and y are mean centered and standardized. n is the number of samples and p the number of SNPs. The central limit theorem and standardization ensures that for n sufficiently large $z_i := \sqrt{n}\beta_i \sim \mathcal{N}(0,1)$.

As a multi-variate model, we have

$$y = X\alpha + \epsilon\,,$$

with α the vector of p true effect sizes and ϵ the n-dimensional vector of residuals with components assumed to be $\epsilon_i \sim \mathcal{N}(0,1)$ and independent. Substituting the multi-variate model into (12), yields

$$\beta_i = \frac{1}{n} x_i^T (X\alpha + \epsilon)\,.$$

a. Fixed effect size model: Under the null assumption that $\alpha = 0$ (no effects), we infer that

$$z_i = \frac{1}{\sqrt{n}} x_i^T \epsilon\,.$$

We can stack an arbitrary collection of such z_i to a vector z via stacking the x_i to a matrix x, such that

$$z = \frac{1}{\sqrt{n}} x^T \epsilon \sim \mathcal{N}(0,\Sigma)\,, \tag{13}$$

with $\Sigma := \frac{1}{n} x^T x$. Note that we made use of the affine transformation property of the multi-variate normal distribution.

It is important to be aware that z is only a component-wise univariate estimation, and hence the null model (13) is for a collection of SNPs with effect sizes estimated via independent regressions.

b. Effect sizes as random variables: We can also take the effects to be random variables themselves. Let us assume that independently $\alpha_i \sim \mathcal{N}(0, h^2/p)$, with h referred to as *heritability*. We then have that

$$z_i = \frac{1}{\sqrt{n}} \left(x_i^T X\alpha + x_i^T \epsilon \right)\,,$$

such that

$$z \sim \mathcal{N}(0, h^2 L) + \mathcal{N}(0,\Sigma) = \mathcal{N}(0, h^2 L + \Sigma)\,, \tag{14}$$

with $L := \frac{1}{np} x^T X X^T x$, and where we assumed that α and ϵ are independent. Two remarks are in order. As X runs over all SNPs, calculation of L usually requires an approximation, for instance, via a cutoff. Furthermore, the null model (14) requires an estimate of the heritability. Such an estimate can be obtained for via LD score regression [9].

B. Product normal test

For this paper, it is only of importance that in both cases *a.* and *b.*, the null model for z is a multi-variate Gaussian.

Therefore,

$$V := z^T z \sim \sum_i \lambda_i [\chi_1^2]\,, \tag{15}$$

with λ_i the ith eigenvalue of the covariance matrix of the Gaussian. As discussed in detail in [2], a GWAS gene enrichment test can be performed via testing against (15). This effectively tests against the expected variance of SNPs significances in the gene.

What we propose here, is to use (8) of section IV, *i.e.*,

$$w^T z \sim \sum_i \frac{\lambda_i}{2} [\chi_1^2] - \sum_i \frac{\lambda_i}{2} [\chi_1^2]\,, \tag{16}$$

for w and z resulting from two different GWAS phenotypes, to test for co-significance of a gene for two GWAS.

In more detail, a significance test can be performed either against the right tail of the null distribution (16) (*coherence*), or against the left tail (*anti-coherence*).

A clarifying remark is in order. Note that we do not centralize w and z over the gene SNPs. Hence, we do not test for covariance, but for a non-vanishing second cross-moment. After de-correlation, it is best to interpret this as testing independently each joint SNP in the gene for a coherent deviation from the null expectation. Therefore we refer to this test as testing for genetic coherence, or simply as cross-scoring. An example will be discussed below. For simplicity, in this work we only consider the fixed effect size model *a.* and assume that the correlation matrix Σ obtained from an external reference panel is a good approximation for both GWAS populations.

a. Direction of association: The direction of effect of the aggregated gene SNPs can be estimated via the index

$$D := \sum_i z_i \,. \tag{17}$$

In detail, making use of the Cholesky decomposition $\Sigma = CC^T$ (requires regularization of the estimated Σ), and the affine transform property of the multi-variate Gaussian, we have that as null

$$D \sim \mathcal{N}(0, |C|_F^2)\,, \tag{18}$$

with $|.|_F$ the Frobenius norm. Testing for deviations from D in the right or left tail, gives an indication of the direction of the aggregated effect size. Note that the (anti)-coherence test between pairs of GWAS introduced above is alone not sufficient to determine the direction for a GWAS pair, but requires in addition to test at least one GWAS via (17) to determine the base direction of the aggregated gene effect. Furthermore, at least one GWAS needs to carry sufficiently oriented signal in the gene such that (17) can succeed. We will also refer to testing for deviations from (18) as D-test.

C. Ratio test

Note that the ratio

$$R = \frac{w^T z}{z^T z} = \frac{\sum_i \lambda_i \bar{w}_i \bar{z}_i}{\sum_j \lambda_j \bar{z}_j^2}\,, \tag{19}$$

with $\bar{w}_i$ and $\bar{z}_i$ i.i.d. $\mathcal{N}(0,1)$, and λ_i the ith eigenvalue of Σ, can be interpreted as the weighted least squares solution in the case of heteroscedasticity for the regression coefficient of the linear regression between the de-correlated effect sizes of the two GWAS. Therefore, with the cdf for R derived in section V, we can test for a significant deviation from the null expectation of no relation. Note that in general R is not invariant under interchange of response and explanatory variable, and may be used under certain conditions to make inference about the causal direction, *cf.*, multi-instrument Mendelian randomization, in particular [21]. The main idea can be readily sketched.

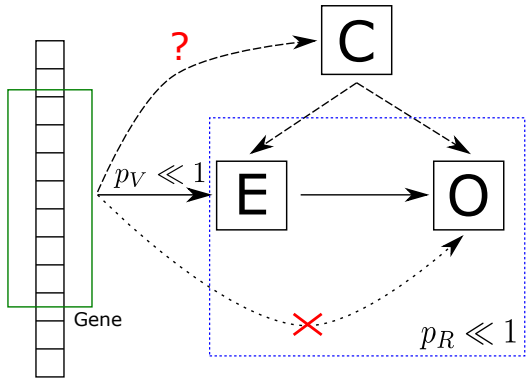

FIG. 4: Interpretation of the ratio test. The test detects potential gene-wise causal relations between a GWAS trait viewed as exposure (E) and a GWAS trait viewed as outcome (O). The association of the exposure has to be confirmed independently via (15). Potential confounders (C) have to be excluded by other means.

Note first that the ratio test tries to detect genes that are maximal (anti)-coherent over the SNP effects between two GWAS, but at the same time carry minimal variance in one of the GWAS, as is clear from (19). The purpose of this, at first sight somewhat non-intuitive, statistical test becomes more clear if one thinks in terms of one GWAS trait as being the exposure and the other as outcome. If the ratio tested gene does carry minimal SNP variance only over the outcome, but on the same time SNP coherence between the exposure and outcome, the causal direction from exposure to outcome is implied, if the exposure is confirmed to be associated to the gene via p_V obtainable from (15), and confounding factors can be excluded. The setup is illustrated in figure 4.

D. Pathway enrichment

The gene scores resulting from the above coherence or ratio test can be utilized instead of the usual gene scores resulting from (15) to perform a gene set (pathway) enrichment test. A pathway is thereby tested for enrichment in coherent or causal genes for a GWAS trait pair. We follow the pathway scoring methodology of [2]. In detail, genes of a pathway which are in close proximity are fused to so-called meta-genes and coherence or ratio test based gene scores are re-computed for the fused genes. The purpose of the fusion is to correct for dependencies between the gene scores due to LD. The result-

ing gene scores (p-values) are qq-normalized and inverse transformed to χ^2_1 distributed random variables. This is followed by testing against the χ^2_n distribution, with n the total number of (meta)-genes in the pathway.

E. SNP normalisation

As we have seen in section VI, the product normal combines evidence for coherent association in a multiplicative manner. A potential challenge to the method we propose arises when one of the two GWAS has associations with very low p-values. Such highly significant associations are common for GWAS with very large sample sizes. Without moderating these p-values, such associations may appear nominally co-significant as soon as the other GWAS provides a mild level of significance. We propose two possible strategies to mitigate this.

One strategy is to introduce a hard cutoff for very small SNP-wise p-values. The precise cutoff depends on the desired co-significance to achieve, and the amount of possible uplift of large p-values one finds acceptable. The dynamics is clear from figure 1. If we target a co-significance of $p = 10^{-8}$ and accept to consider SNPs with a p-value of 0.05 or less in one GWAS to be sufficiently significant, we have to cutoff p-values around 10^{-16}. While such a cutoff ensures that no SNPs with p-values above 0.05 in one GWAS can become co-significant due to very high significance (i.e. p-value below 10^{-16}) in the other GWAS, applying such a hard cutoff point hampers distinguishing differences in co-significance.

An alternative strategy is to transform all p-values, rather than only the most significant ones. For example, the so-called qq-normalisation, re-assigns uniformly distributed p-values according to the rank r, i.e. $p = (r+1)/(N+1)$, where N is the number of p-values. (This approach implies the strongest p-value moderation and therefore is very conservative.) The product-normal statistic in (16) is then computed with w and z according to the inverse χ^2 cdf of the respective p-values. Since for GWAS $N \simeq 10^6$ the most significant transformed p-value is $\sim 10^{-6}$, according to figure 1 the other p-value has to be smaller than $10^{-3} - 10^{-4}$ to achieve (genome-wide) co-significance.

Since the qq-normalisation allows for combining two GWAS with significantly different signal strengths without the need to introduce an adhoc cutoff, it is our approach of choice, and we will use it in the following example, and we also prefer to apply this transformation for the ratio test in order to compensate for different signal strengths between the GWAS.

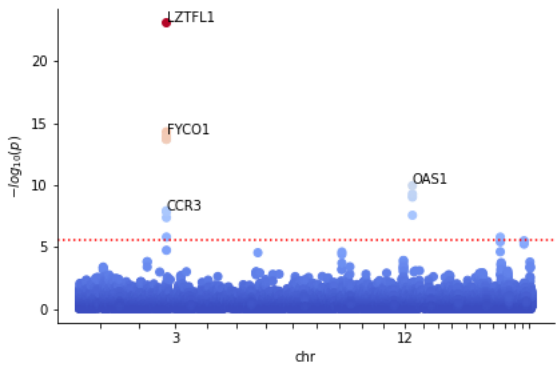

FIG. 5: Manhattan plot showing the strong gene enrichment on chromosomes 3 and 12 for the severe COVID-19 GWAS. The Bonferroni significance threshold (red dotted line) is taken to be 2.67×10^{-6} (0.05 divided by 18727, the number of tested genes). Only a selection of significant genes are labeled.

VIII. APPLICATION TO COVID-19 GWAS

A. Coherence test

To demonstrate the usefulness of our methods in the context of real and topical GWAS data we consider the recent meta-GWAS on very severe respiratory confirmed COVID-19 [22, 23] as main phenotype and co-analyse it with several other related traits. Specifically, we use the summary statistics resulting from European subjects excluding those from UK Biobank [22] (A2_ALL_eur_leave_ukbb_23andme, release 7. Jan. 2021), in order to avoid overlap with the secondary trait GWAS. This GWAS shows significant gene enrichment on chromosomes 3 and 12, as can be inferred via testing against the null model (15) with fixed effect sizes. The Manhattan plot for the resulting gene p-values is shown in figure 5.

We cross-scored this GWAS against a panel of GWAS on medication within the UK Biobank [24]. These GWAS on medication take prescription (self-reported intake) of 23 common types of medications as traits. Hence, in cross-scoring against the severe COVID-19 GWAS we hope to uncover whether there is a shared genetic architecture between severity and pre-disposition to being prescribed specific medications.

All calculations have been performed with the python package *PascalX* [25], which incorporates the methods described in section VII. Default settings and the European subpopulation of the 1000 Genome Project [26] as reference panel to estimate the SNP-SNP correlations Σ were used. We only considered protein coding genes and SNPs within a gene window extending $50kb$ beyond the

transcription start and end position for the cross-scoring. We transformed the raw GWAS p-values via *joint* rank transformation, for the reason discussed in section VII E. The directions of association are extracted from the signs of the raw effect sizes (β). All summary statistics data we made use of in this work have been made publicly available by the respective authors of the cited corresponding publications. Pathway enrichment has been tested against the gene sets included in MSigDB version 7.1 [27, 28].

a. Coherence: The cross-scoring results for the coherence test are shown in figure 6. Cross scoring of GWAS effects from prescription of medication group M05B (drugs affecting bone structure and mineralization) reveals joint signals with that for very severe COVID-19. We show the Manhattan plot resulting from the null model (13) and the index (16) for the medication group M05B in figure 7. We observe that SNP-wise effects in genes in the well-known COVID-19 peak locus on chromosome 3 appear to be coherent with those from being prescribed M05B medication, with strong significance for the chemokine receptor genes *CCR1*, *CCR3* and the gene *LZTFL1* in the region chr3p21. Note that the two chemokine receptors are Bonferroni significant under the number of genes and drugs tested. However, *LZTFL1* slightly missed the threshold. We tested the orientation of the aggregated associations of these genes via (17) over the COVID-19 GWAS and find a positive direction with $p_D \simeq 1.6 \times 10^{-4}, 7.5 \times 10^{-3}$ and 0.03, respectively.

For illustration, we show the spectrum of SNPs considered in the *CCR3* region and their SNP-SNP correlation in figure 8. One can see that there is a large block of SNPs in high LD, encompassing some of its 5'UTR and most of its gene body, all having positive associations with both COVID-19 and M05B drug prescription. Furthermore, a somewhat weaker signal of coherence can be seen for SNPs in the 3'UTR, some of which share negative associations. Using the method described in section IV, we calculate a p-value of $p \simeq 1.50 \times 10^{-8}$ for the significance of the coherence.

Note that the chemokine receptor of type 1 (*CCR1*) is involved in regulation of bone mineralization and immune/inflammatory response. In particular, chemokines and their receptors are critical for recruitment of effector immune cells to the location of inflammation. Mouse studies suggest that this gene plays a role in protection from inflammatory response and host defence [29, 30]. The chemokine receptor *CCR3* is of importance for regulation of eosinophils, a leukocyte involved in many inflammatory pathologies [31]. In particular, mouse models suggest a complex role of *CCR3* in allergic diseases [32]. In general, chemokines are suspected to be a direct cause of acute respiratory disease syndrome, a major cause of death in severe COVID-19. For a review of chemokines and their receptors in the COVID-19 context we refer to [33].

The SNP spectrum in the *LZTFL1* region is shown in figure 9. From the SNP-SNP correlation matrix on the right one can see that the region including the *LZTFL1* gene contains at least three sizable LD blocks. The p-value for the significance of the covariance is calculated to be given by $p \simeq 1.19 \times 10^{-7}$. The observed co-significance of *LZTFL1* for the two traits is in line with the fact that the gene *LZTFL1* modulates T-cell activation and enhances IL-5 production [34]. In particular, mouse models suggest that expression of IL-5 alters bone metabolism [35].

Direction of association and coherence suggests that genetic predispositions leading to M05B prescription may carry a higher risk for severe COVID-19, with shared functional pathways related to the genes discussed above. We list the most significantly enriched pathways for (anti)-coherent M05B and severe COVID-19 genes detected via the pathway enrichment test described in section VII D in table I. Note that we detect several Bonferroni significant pathways, and that most of the leading coherent pathways appear to be of interest in the COVID-19 context, as immune system related. It is also noteworthy that in looking at common pathways between M05B and COVID-19 we increase the signal strength, as only the lead pathway (*Roeth tert targets dn*) is also detected to be Bonferroni significant under the COVID-19 GWAS alone, *cf.*, table III in appendix C. The significant pathway of Roeth et al. [36] corresponds to genes that are significantly down regulated in T lymphocytes that overexpress human telomerase reverse transcriptase (hTERT). The also significant pathways of Kurozumi et al. [37] correspond to inflammatory cytokines and their receptors modulated in brain tumors after treatment with an oncocytic virus.

b. Anti-coherence: We perform a similar test between COVID-19 severity and drug classes as above for anti-coherence. The corresponding results for all medication classes are shown in figure 6. Despite the fact that no drug received a Bonferroni significant hit, we note that the drug classes H03A (Thyroid preparations), C10AA (HMG CoA reductase inhibitors), L04 (Immunosuppressants) and M01A (Anti-inflammatory and anti-rheumatic products) have gene hits with a p-value $< 1 \times 10^{-5}$. Interestingly, all of these drugs find applications in auto-immune diseases and allergies. The genes with a p-value below 1×10^{-5} for H03A are *HLA-DQB1*, for C10AA *SMARCA4* and *BCAT2*, for L04 *LZTFL1*, *TRIM10*, *TRIM15* and *TRIM26*, and for M01A *TRIM10* and *TRIM31*. Note that TRIM proteins are associated with innate immunity, and are in particular involved in

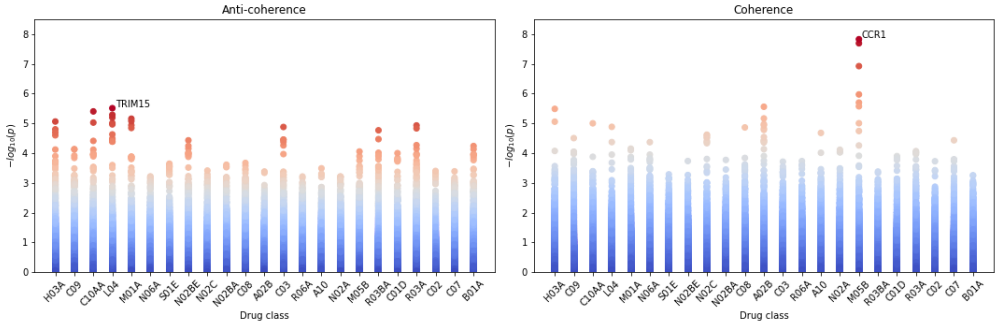

FIG. 6: Resulting p-values for cross scoring 23 drug classes GWAS against very severe COVID-19 GWAS for coherence, respectively, anti-coherence. Each data point corresponds to a gene and we marked the most significant gene in both cases. Left: Anti-coherence. Right: Coherence. Note that the drug class M05B shows significant enrichment in coherence with severe COVID-19.

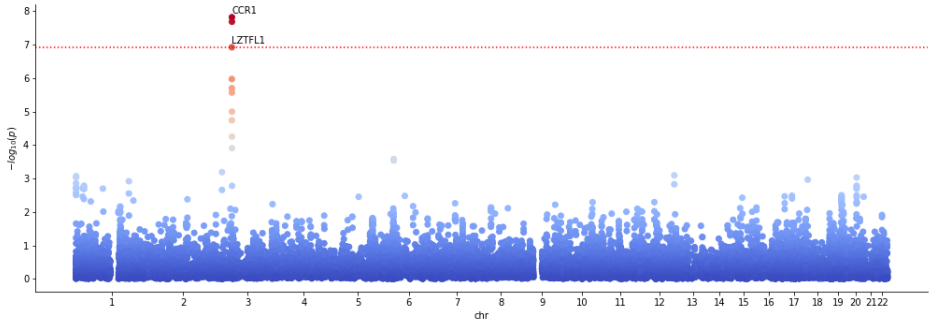

FIG. 7: Manhattan plot for cross scoring very severe confirmed COVID-19 with medication class M05B for coherence. Data points correspond to genes. The dotted red line marks a Bonferroni significance threshold of 1.16×10^{-7} (0.05 divided by the 18705 genes tested and 23 drug classes).

pathogen-recognition and host defence [38, 39]. For illustration of the anti-coherence case, we plot the SNP spectrum for *TRIM10* under L04 in figure 10. We tested the above genes for direction of aggregated effect sizes under the COVID-19 GWAS via (17) and found that the detected *TRIM* genes tend to be localized in the left tail. In particular, $p_D \simeq 0.04$ for *TRIM10* and *TRIM15* under L04 (right tail). Hence, a protective effect is suggested for these genes, and therefore conditions leading to prescription of these medications may imply a lower genetic risk for severe COVID-19. The pathway enrichment test for the anti-coherent L04 cross-scored severe COVID-19 genes show Bonferroni significant enrichment in interferon gamma related pathways, see table I. Note that *TRIM* proteins are expressed in response to Interferons, *cf.*, [38], and therefore the detected pathways are consistent with the above observed enrichment of TRIM genes. Remarkably, the individually non-Bonferroni significant genes aggregate to Bonferroni significant pathways. While two of the detected pathways are also Bonferroni significant under L04 alone, one pathway (*GO interferon gamma mediated signaling pathway*) is only Bonferroni significant under anti-coherence with COVID-19. Note that we detected for L04 individually 35 Bonferroni significant pathways, *cf.*, table III. Therefore, the coherence test singles out a small subset of potentially shared pathways with COVID-19.

c. M05B related traits: The main application of M05B medications is the treatment of osteoporosis. In order to investigate this potential link further, we cross-scored the COVID-19 GWAS against a selection of GWAS with phenotypes related to osteoporosis, namely, bone mineral density (BMD) estimated from quantitative heel ultrasounds and fractures [40], estrogene levels in men (estradiol and estrone) [41], calcium concentration [42], vitamin D (25OHD) concentration [43] and rheumatoid arthritis [44]. The inferred gene enrichment for coherence and anti-coherence with the COVID-19 GWAS is shown in figure 11. We found enrichment with gene p-values $< 10^{-5}$ for vitamin D and calcium. In particular, in the anti-coherent case the most significant genes for vitamin D are *OAS3*, *OAS2*, *OAS1*, *FYCO1*, *CXCR6* and *LZTFL1*. We show the corresponding Manhattan plot in figure 12. The D-test shows that all these genes

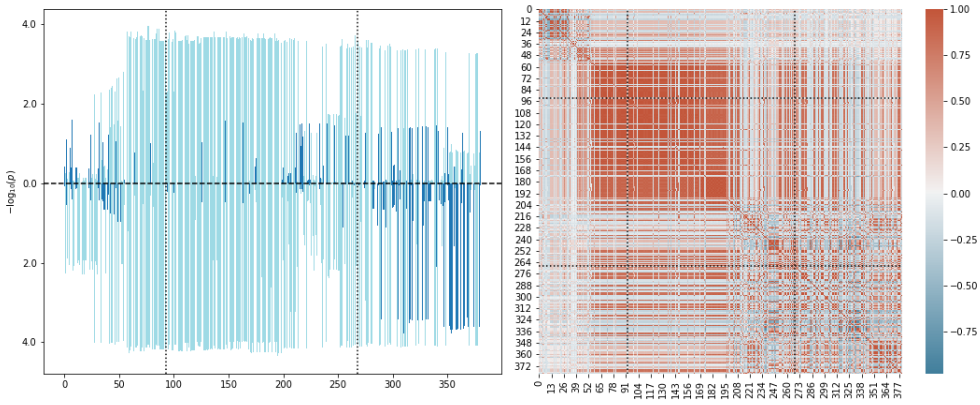

FIG. 8: Left: SNP p-values after rank transform for the *CCR3* gene. The x-axis is numbered according to the ith SNP in the gene window ordered by increasing position. The dotted black lines indicate the transcription start and end positions (first and last SNP). Up bars correspond to M05B and down bars to severe COVID-19. Light blue indicates positive and dark blue negative association. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel.

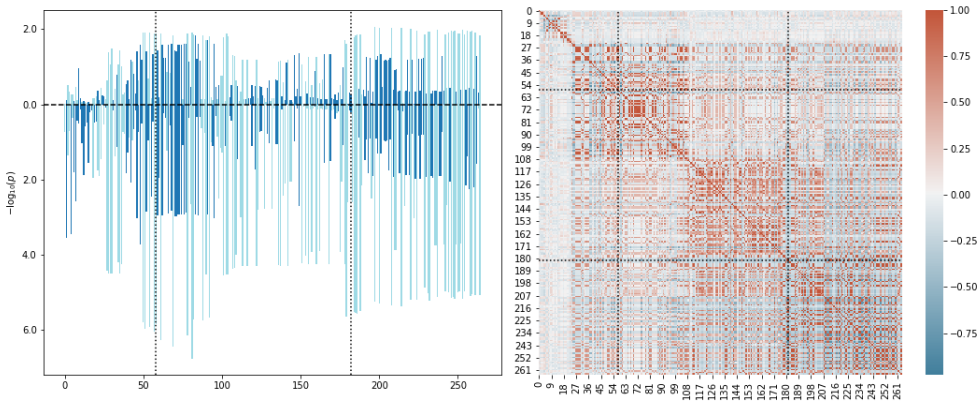

FIG. 9: Left: SNP p-values after rank transform for the *LZTFL1* gene. Annotation as in figure 8. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel. Note that the gene contains independent LD blocks.

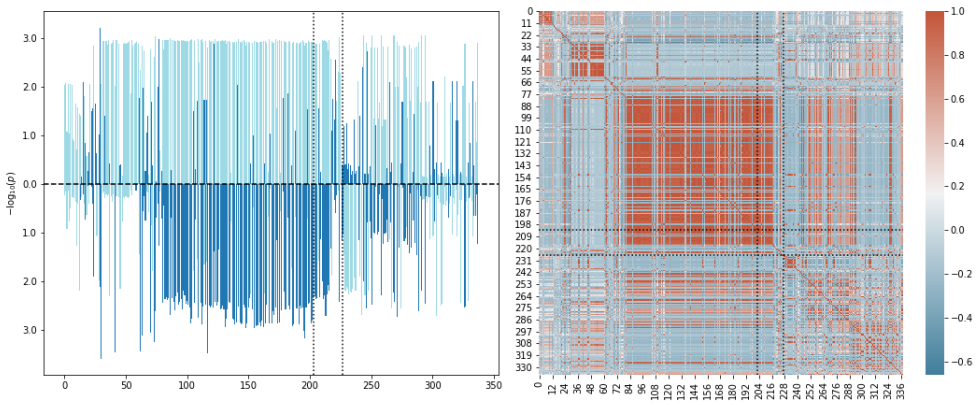

FIG. 10: Left: SNP p-values after rank transform for the *TRIM10* gene under medication L04. Annotation as in figure 8, but up bars corresponding to L04. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel.

trait	tail	enriched pathways	# genes	p-value
M05B	R	**Roeth tert targets dn** *	8	5.5×10^{-8}
		Kurozumi response to oncocytic virus and cyclic rgd	17	2.35×10^{-7}
		Kurozumi response to oncocytic virus	33	3.5×10^{-7}
		GO C-C chemokine binding	16	1.04×10^{-6}
		GO G-protein coupled chemoattractant receptor activity	18	2.54×10^{-6}
		chr3p21	6	4.82×10^{-6}
		GO chemokine binding	23	5.50×10^{-6}
		GO positive regulation of monocyte chemotaxis	17	5.83×10^{-6}
	L	-		
L04	R	GO inositol 1 4 5 trisphosphate binding	12	6.08×10^{-6}
		GSE13484 unstim vs yf17d vaccine stim pbmc dn	165	7.49×10^{-6}
	L	**GO interferon gamma mediated signaling pathway**	70	1.47×10^{-7}
		Gaurnier PSMD4 targets *	44	1.47×10^{-6}
		Reactome interferon gamma signaling *	67	1.65×10^{-6}
		GO negative regulation of cellular protein localization	110	2.68×10^{-6}
		Wilensky response to darapladib	25	5.6×10^{-6}

TABLE I: Enriched pathways for the M05B and L04 cross-scored severe COVID-19 genes. The coherent case corresponds to the right (R) tail, the anti-coherent case to the left (L) tail. We tested against the 25724 gene sets of MSigDB 7.1 and only list pathways with $p < 10^{-5}$. Bonferroni significant pathways ($p < 0.05/25724 \simeq 1.94 \times 10^{-6}$) are printed in bold. An asterix (*) indicates pathways which are also Bonferroni significant in one of the traits alone.

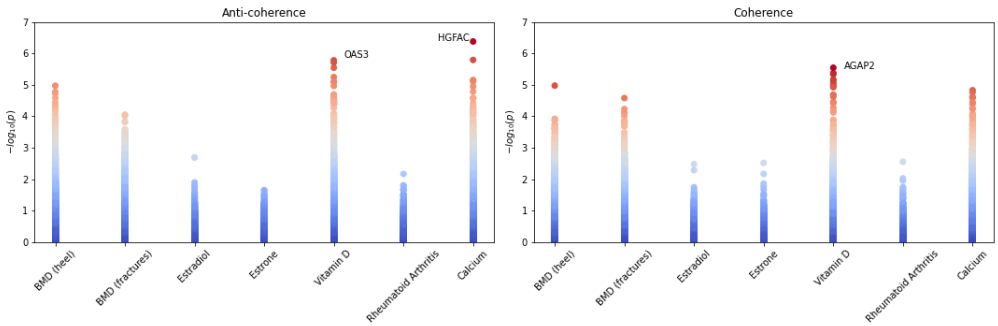

FIG. 11: Resulting p-values for cross scoring several GWAS related to osteoporosis against the severe COVID-19 GWAS. We observe enrichment in vitamin D and calcium with several p-values $< 10^{-5}$. The leading genes are indicated.

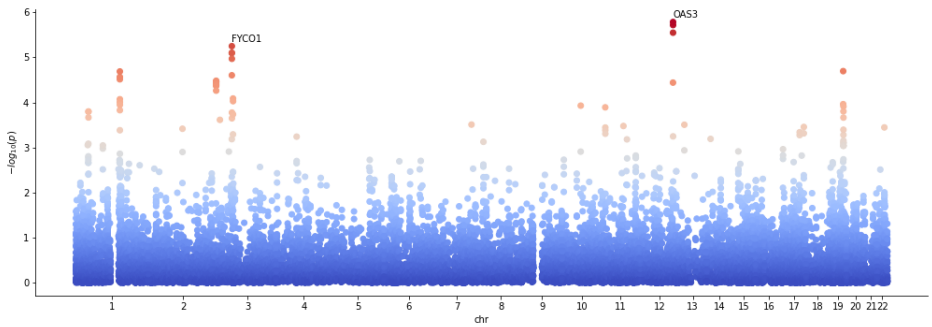

FIG. 12: Manhattan plot for cross scoring vitamin D concentration against severe COVID-19 in the anti-coherent case.

are in the right tail ($p_D < 3 \times 10^{-2}$), in particular *OAS3* with $p_D \simeq 7.21 \times 10^{-5}$, under the COVID-19 GWAS, and therefore suggests an increase in the risk for severe COVID-19 in the presence of effect alleles in these genes.

The *OAS* family are essential proteins involved in the innate immune response to viral infection. They are involved in viral RNA degradation and the inhibition of viral replication [45]. It has been observed that vitamin D can increase the expression of the *OAS* genes [46].

FYCO1 plays a role in microtubule plus end-directed transport of autophagic vesicles [47]. It has been demonstrated that SARS-CoV-2 inhibits autophagy activity [48]. A regulatory role of vitamin D on autophagy at different steps, including induction, nucleation and degradation has been suggested, *cf.*, [49].

The gene *CXCR6* ($p \simeq 7.96 \times 10^{-6}$) is expressed by subsets of T_H1 cells, but not by T_H2 cells, and may be important in trafficking of effector T cells that mediate type-1 inflammation [50]. It has been shown that the vitamin D analog TX527 promotes the surface expression of *CXCR6* on T-cells and inhibiting effector T cell reactivity while inducing regulatory T cell characteristics, promoting migration to sites of inflammation [51]. It is therefore suggestive that vitamin D concentration influences differences in immune system reaction, *i.e.*, between type-1 (inflammatory) or type-2 (anti-inflammatory), and thereby impacting the severity of COVID-19. Indeed, possible links between severity of COVID-19 and vitamin D are actively discussed in the current literature. See for instance [52–54] and references therein.

For calcium in the anti-coherent case, the top genes are *HGFAC* ($p \simeq 4.15 \times 10^{-7}$) and *DOK7* ($p \simeq 1.61 \times 10^{-6}$). Note that *HGFAC* is Bonferroni significant under the seven traits tested. Both genes sit in the right tail under the D-test (17) applied to the calcium GWAS ($p_D \simeq 5.87 \times 10^{-4}, 3.5 \times 10^{-5}$). Therefore, the aggregated variants in these genes imply that a predisposition for high calcium concentration may implicate a reduced risk for severe COVID-19, *i.e.*, a protective effect. Note that in general it is known that viruses appropriate or interrupt Ca^{2+} signaling pathways and dependent processes, *cf.*, [55].

The gene *HGFAC* plays a role in converting hepatocyte growth factor (HGF) to its active form. In particular, binding of HGF causes the up-regulation of *CXCR3*, which is primarily activated on T lymphocytes and NK cells. *CXCR3* is preferentially expressed on T_H1 cells, while *CCR3* on T_H2 cells [56]. In detail, *CXCR3* binds the chemokine receptor *CCR3* and prevents an activation of T_H2-lymphocytes. Thereby, a towards T_H1 biased inflammation immune reaction is triggered [57]. Note that *CXCR3* is able to increase intracellular Ca^{2+} levels [58].

The gene *DOK7* is of importance for neuromuscular synaptogenesis [59]. It activates *MuSK*, which is essential for maintenance of the neuromuscular junction as it is involved in concentrating *AChR* in the muscle membrane at the neuromuscular junction. The latter protein is critical for signaling between nerve and muscle cells, a necessity for movement, and is influenced by intra-cellular calcium [60]. Hence, it may be possible that *DOK7* could be involved in a genetic explanation of the muscle weakness impacting some of the severe COVID-19 patients [61].

Note that for the vitamin D and calcium GWAS, we observe cross enrichment in, both, the coherent and anti-coherent case with the severe COVID-19 GWAS, *cf.*, figure 11. For example, for vitamin D the leading gene hit is *AGAP2* ($p \simeq 2.84 \times 10^{-6}$). *AGAP2* modulates the transforming growth factor beta-1 (*TGF-β1*), one of the most potent pro-fibrotic cytokine known to date, currently accepted as the principal mediator of the fibrotic response in liver, lung, and kidney [62]. The D-test under the COVID-19 GWAS is however inconclusive (left-tail $p_D \simeq 0.15$).

Another implied coherent gene locus of potential interest is *OS9* ($p \simeq 4.2 \times 10^{-6}$). The corresponding protein binds to the hypoxia-inducible factor 1 (*HIF-1*). *HIF-1* is a key regulator of the hypoxic response [63, 64]. In particular, regulation of *HIF-1* interpolates between regeneration and scaring of injured tissue [65]. It is known that severe COVID-19 may lead to lung tissue fibrosis [66, 67]. The D-test applied to the COVID-19 GWAS shows that *OS9* is located in the left tail ($p_D \simeq 0.03$), and therefore a protective property of effect alleles in this gene are suggested.

B. Can coherence make a difference in identifying co-significant genes?

We next wanted to explore whether gene-wise significance based on the product normal statistic that tests for consistently coherent or incoherent SNP-wise associations within the gene-window gives different results than when first computing gene-wise significance for each trait (via the usual χ^2-test, *cf.*, eq. (15)), and then combining the corresponding scores, as if there was only a *single* genetic element affecting both traits (*cf.*, fig. 1 for the corresponding combined product normal p-values). As before, we use jointly qq-normalized GWAS p-values to avoid the risk of unintended uplift.

In general, we would expect that the simple approach described in this section would lead to more false positives, but would also lead to false negatives. The danger for false positives is clear, as the simple approach is blind to incoherent directions of association between the SNPs

of the two GWAS in the gene window under consideration. However, also false negatives may occur: In the proposed novel approach using relation (16), several independent weak signals may combine to a stronger signal, if the directions of association are consistent over the gene window. Such signals are invisible in the simple approach of comparing aggregated gene p-values.

Figure 13 (top plot) shows the scatter plot for gene-wise log-transformed p-values for the severe COVID-19 GWAS against the medication class M05B GWAS, and significance thresholds for the product-normal. We observe that in this example all of the genes which are significant ($p < 10^{-5}$) based on the SNP-wise product normal coherence test, are also significant under the simple single-element test. Yet, interestingly, two genes that pass the simple test, i.e. *CXCR6* and *CCR2*, are not significant under the coherence test. While these genes obtain significant association for both traits, their SNP-wise effects appear to be partly coherent and partly incoherent, explaining why the coherence test is not as significant, *cf.*, figure 14, which compares the SNP spectrum of *CCR2* against *CCR5*.

Nevertheless the corresponding coherence p-values are just above the 10^{-5} threshold. Since these genes are also cytokine receptors, it seems likely that their joint association with COVID-19 and M05B points to a common mechanism relevant for these traits that is modulated by variants in these genes.

A similar picture emerges for the genes with significant cross-scores of severe COVID-19 against vitamin D concentration, see figure 13 (middle plot): Their simple single-element statistic is also significant and conversely only very few genes whose cross-score is insignificant, *i.e.*, with $p < 10^{-5}$, have a significant single-element statistic, and if so only by a small margin.

A somewhat different pictures emerges from the co-analysis of COVID-19 and calcium GWAS signals: Here a number of genes with a significant simple single-element statistic do not achieve a significant cross-score, and there are several genes with the reverse signature of a significant cross-score, but insignificant single-element statistic. The latter include *LRPAP1* and *DOK7*, which both do not exhibit a strong association with COVID-19, yet the available annotation (*LRPAP1* codes for a LDL-receptor-related protein and *DOK7* has been assigned to the GO category "lipid binding") makes a role for these genes in severe COVID-19 plausible, as dislipidemia has been associated with the severity of COVID-19 [68].

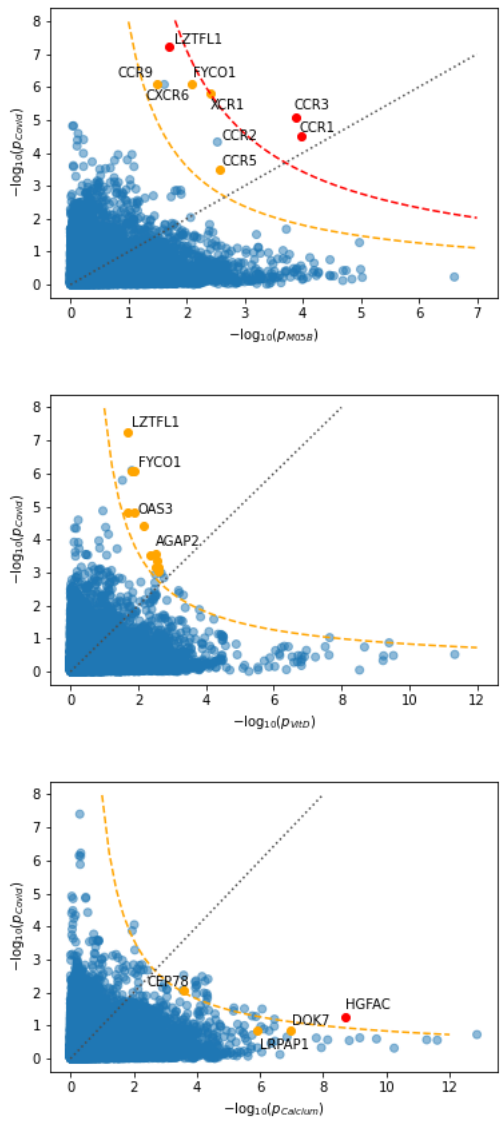

FIG. 13: GWAS gene scores, obtained separately for two GWAS using joint qq-normalization and (15), plotted against each other. Each point corresponds to a gene. Top: COVID-19 vs. M05B. Middle: COVID-19 vs. vitamin D. Bottom: COVID-19 vs. calcium. The gray dotted line marks the diagonal. The orange and red dashed curves mark the gene-wise product-normal threshold curves for significance of 10^{-5}, respectively, 10^{-7}. The color of a gene (red and orange) indicate the SNP-wise cross scored p-value ($< 10^{-6}$, respectively, $< 10^{-5}$).

C. Ratio test

Let us also briefly discuss an application of the ratio based causality test introduced in sections V and VII C. We ratio tested the COVID-19 GWAS against the vitamin D and calcium concentration GWAS in the coherent

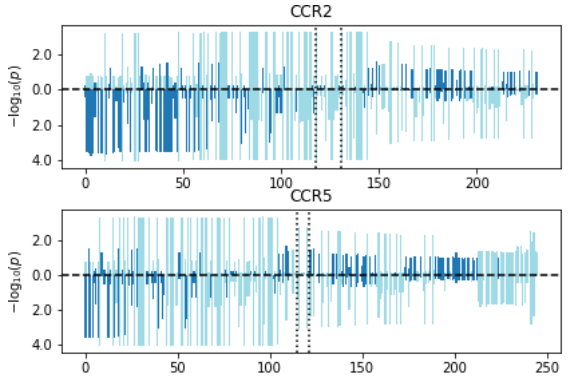

FIG. 14: SNP spectrum of the genes *CCR2* (top) and *CCR5* (bottom). Annotation as in figure 8. Note that the difference in coherence is difficult to call by eye as also LD has to be taken into account, but it appears also visually that *CCR2* is more incoherent.

case, making use of eq. (11) and (19), and found that Vitamin D carries two interesting hits which suggest causal pathways from genetic predisposition for Vitamin D concentration to severity of COVID-19. The ratio test Manhattan plot is shown in figure 15. The leading gene is *KLC1* ($p_R \simeq 9.18 \times 10^{-7}$). It is known that Kinesin-1 uncoats viral DNA and compromises nuclear pore complex integrity, allowing viral genomes nuclear access to promote infection [69]. Note that this gene also features in the COVID-19 virus-host protein interactions, belonging to the functional group of viral trafficking, as recently discussed in [70]. The gene carries under the outcome (COVID-19) a p-value of $p_V \simeq 0.95$ and under the exposure (vitamin D) a p-value of $p_V \simeq 3.23 \times 10^{-9}$. Under calcium we have $p_V \simeq 0.25$, hence calcium can be excluded as potential confounder. The gene is in the right tail ($p_D \simeq 0.07$) under the vitamin D GWAS. We conclude that the gene may mediate a causal relation.

Furthermore, besides the paralog gene *AL139300.1* of *KLC1*, the gene *ZFYVE21* ($p_R \simeq 3.9 \times 10^{-6}$) stands out. Under the COVID-19 GWAS (outcome) the gene carries a p-value of $p_V \simeq 0.99$ while under the vitamin D GWAS (exposure) a p-value of $p_V \simeq 6.5 \times 10^{-4}$. This implies a causal flow from genetic predisposition for vitamin D concentration to severity of COVID-19. We confirmed that calcium concentration is not a potential confounder for *ZFYVE21* ($p_V \simeq 0.26$). The D-test shows that the gene is in the right tail ($p_D \simeq 0.03$) and hence is associated with high vitamin D concentration. A possible explanation of this hit may go as follows. *ZFYVE21* regulates microtubule-induced PTK2/FAK1 dephosphorylation, which is important for integrin beta-1/ITGB1 cell surface expression. It has been discussed before in the literature that integrins in host cells may play the role of alternative receptors to ACE2 for SARS-CoV-2 [71, 72]. Hence, it is tempting to speculate that genetic predisposition for vitamin D concentration may also influence expression of the integrin receptors, and thereby the outcome of COVID-19 via an increased risk of cellular infection.

Another gene we observe to be of relevance in the coherent case is the Ring Finger Protein 217 (*RNF217*) with $p_R \simeq 1.83 \times 10^{-6}$ and calcium as exposure. We have $p_V \simeq 2.15 \times 10^{-5}$ under calcium and $p_V \simeq 0.82$ under COVID-19. Vitamin D has for this gene $p_V \simeq 0.26$, hence can be excluded as potential confounder. *RNF217* is a member of the E3 ubiquitin protein ligase family. Ubiquitination of proteins is a post-translational modification process with different cellular functions, including antiviral functions and virus replication [73]. A potential COVID-19 therapeutic pathway based on E3 ubiquitin ligases has been recently proposed in [74]. As the gene shows an illustrative coherence pattern (multiple LD blocks), we show the corresponding SNP correlation plot in figure 16.

We also tested the anti-coherent case. The gene *HOXC4* with $p_R \simeq 4.5 \times 10^{-6}$ is detected for M05B. Individual scored p-values for the gene read $p_V \simeq 2.04 \times 10^{-5}$ under M05B and $p_V \simeq 0.75$ under COVID-19. The D-test under M05B shows that the gene signal is located in the right tail, but not significantly. Modulo potential confounders, a causal relation from predisposition to take M05B to a protective property towards severity of COVID-19 via *HOXC4* is suggested. We have $p_V \simeq 0.27$ under calcium and $p_V \simeq 0.06$ under vitamin D, hence, a potential confounding role of vitamin D is possible. We note that *HOXC4* has been discussed before in the COVID-19 context [75]. The gene is related to an enhanced antibody response under the regulation of estrogens. This matches the observation of anti-coherence with severity of COVID-19.

Interestingly, also for the inverse causal relation, *i.e.*, a predisposition for severe COVID-19 implying a predisposition for the need to take M05B medication, a relevant gene is singled out. Namely, the gene *DPP9* with $p_R \simeq 3.88 \times 10^{-5}$, and $p_V \simeq 0.74$ under M05B, respectively, $p_V \simeq 2.57 \times 10^{-5}$ under COVID-19. For this gene p_V for calcium and vitamin D is not significant. This gene has been implicated before to be involved in the genetic mechanisms for critical illness due to COVID-19 [76]. Note that we also observe *DPP9* ($p_R \simeq 5.82 \times 10^{-7}$) as top hit for calcium concentration as outcome, *cf.*, the corresponding Manhattan plot shown in figure 17. Under both exposures the gene signal tends to be more in the right tail of the D-test, but is too weak for calling. Among the genes we detect in this ratio test, the gene *IFNAR2* ($p_R \simeq 9.70 \times 10^{-6}$) appears to be of particular

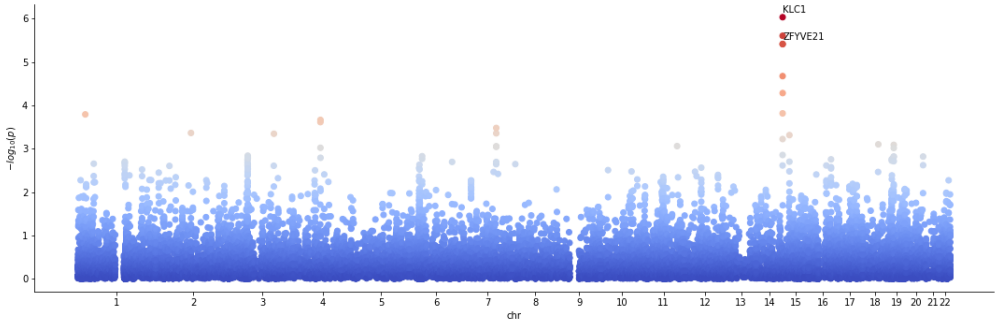

FIG. 15: Manhattan plot for ratio scoring Vitamin D concentration against severe COVID-19 in the coherent case, with ratio denominator given by COVID-19.

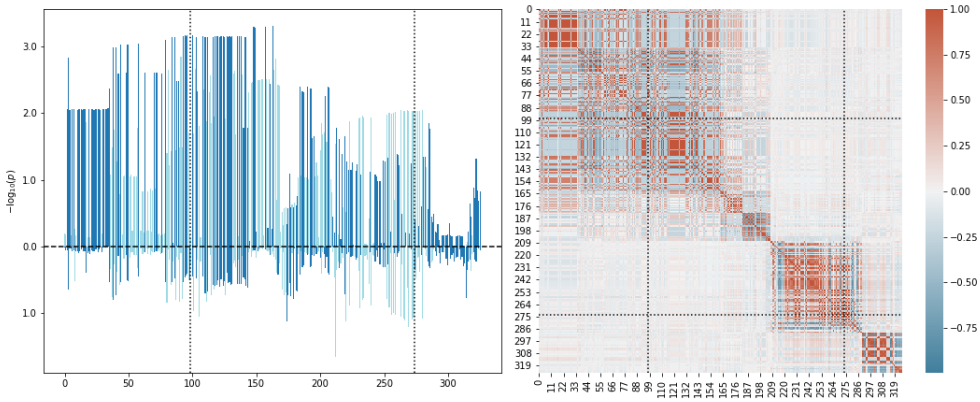

FIG. 16: Left: SNP p-values after rank transform for the *RNF217* gene under calcium concentration and severe COVID-19. Annotation as in figure 8, but up bars corresponding to calcium. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel.

relevance, as it is a known possible drug target against COVID-19 [76].

For vitamin D as exposure, we find in the anti-coherent case that the genes *MED23* ($p_R \simeq 5.22 \times 10^{-5}$ and *GATA4* ($p_R \simeq 7.88 \times 10^{-5}$) are the leading hits with a p-value $< 10^{-4}$. Both genes are not significant under calcium. The D-test shows that *MED23* is in the right tail under the exposure ($p_D \simeq 1.22 \times 10^{-3}$), whereas the test is inconclusive for *GATA4*. The Mediator subunit gene *MED23* has emerged before in SARS-CoV-2 screens, with a critical role of this complex during infection and death suggested [77]. *GATA4* has been observed previously in the context of the top significant biological processes likely associated with respiratory failure in COVID-19 patients [78].

Finally, note that the genes receiving very small p-values for our anti-coherent ratio test with calcium as exposure, i.e. *HGFAC* ($p_R \simeq 4.2 \times 10^{-5}$) and *DOK7* ($p_R \simeq 1.27 \times 10^{-5}$), were already detected above via the coherence test. Therefore, it is suggested that the corresponding pathways could be causal. However, for both genes we have that $p_V \simeq 0.01$ under vitamin D, and therefore vitamin D could be a potential confounder. In addition, we observe the leading gene *PPP2R3A* ($p_R \simeq 1.0 \times 10^{-5}$) with $p_V \simeq 1.9 \times 10^{-4}$ under calcium, but with a potential confounding role of vitamin D ($p_V \simeq 0.03$). Note that *PPP2R3A* interacts with *CDC6* [79], which is upregulated at the early stages of human coronavirus 229E infection [80], and initiates assembly of a pre-replication complex [81].

We also tested all cases against pathway enrichment as outlined in section VII D. The resulting enriched pathways with $p < 10^{-4}$ are listed in table IV of appendix C. Note that most of the detected enriched pathways are immune system related, and may be of interest for further investigations in the COVID-19 context.

IX. SUMMARY AND CONCLUSION

The work presented here relies on the novel mathematical insight that the null distribution of the normal product of two effect size vectors with a common known covariance structure, can be expressed in terms of

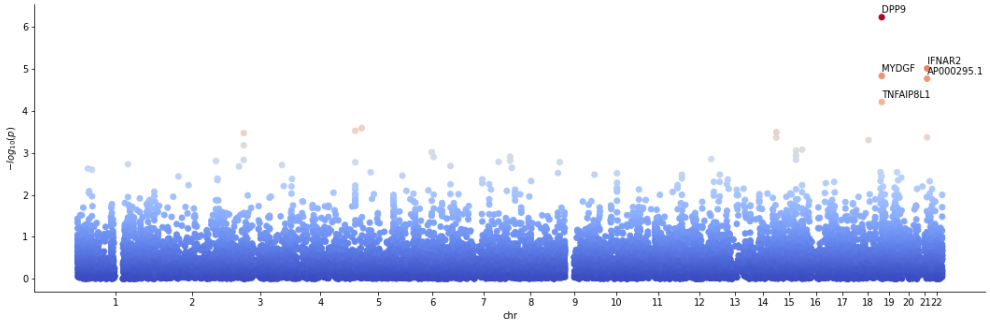

FIG. 17: Manhattan plot for ratio scoring calcium concentration against severe COVID-19 in the anti-coherent case, with ratio denominator (outcome) given by calcium concentration.

a weighted χ^2 difference distribution (*cf.*, eq. (8)). Moreover, we showed that the ratio between this product and the square of one of the vectors, which allows for testing causal relationships, also relates to this distribution. In both cases the corresponding cdf and tail probabilities can be computed efficiently with Davies' algorithm.

This insight has application for GWAS, because it enables testing for *coherent* SNP-wise effects within a gene window for two different traits, *cf.*, section VII. Importantly, our test is different from an additive test looking merely for co-significance. Indeed testing whether the signs of the effects sizes tend to be the same (or opposite) within the window, has potentially more power to identify genes that modulate both traits. Our test is statistically rigorous, properly taking into account the SNPs' correlation structure (i.e. LD), while still being able to be computed rapidly and accurately without any approximations.

We applied our method to identify genes with a role both in severe COVID-19 and medical conditions leading to the prescription of any of 23 medication classes. Our analysis revealed a particular strong signal of coherence between COVID-19 and M05B medications, with chemokine receptor genes *CCR1* and *CCR3* as lead hits. We then searched for coherence signals between COVID-19 and additional GWAS traits, such as vitamin D and calcium concentrations, known to be related to diseases treated with M05B drugs. This analysis implicated genes related to differentiation between type-1 and type-2 immune response. A possible explanation could be that many patients being prescribed class M05B medication suffer from osteoporosis. Vitamin D stimulates the absorption of calcium and therefore is often prescribed to these patients in order to increase their bone mineral density [82]. While data clearly support the function of vitamin D in bone growth and maintenance, evidence supporting the role of vitamin D in other health and disease processes, in particular in acute respiratory tract infections [83], often observed for patients with severe COVID-19, is less clear cut: Vitamin D is thought to reduce the risk of infection, mainly due to factors involving physical barriers, cellular natural immunity and adaptive immunity [84]. Furthermore, a lack of plasma vitamin D level has been associated with the risk of infection [85]. Yet, direct conclusive evidence for its proposed protective function specific to severe COVID-19 is still lacking [86]. In this work we found potential genetic hints that link the severity of COVID-19 to vitamin D concentration through specific genes operating in immune response pathways, including those coding for chemokine receptors. In addition, we detected a potential causal link from genetic predisposition for vitamin D deficiency to severity of COVID-19, mediated by the *ZFYVE21* gene, which is related to integrin beta-1/ITGB1 cell surface expression and thereby potentially impacting disease outcome by influencing alternative receptors to ACE2 for host cell entry.

Our causal analysis using the ratio test (*cf.*, eq. (19)) suggests that the genes *HOXC4* and *DPP9*, which have already been associated with severe COVID-19 [75, 76], may contribute to the protective role of conditions leading to M05B medication prescription, and that the genes *MED23* and *GATA4* may play a role in the causal link between vitamin D deficiencies and severe COVID-19. Note that the novel aspect of our method is its capacity to identify candidate genes mediating the causal effects, but we cannot exclude potential confounders in this causal analysis.

We also found that genetic variants in genes related to both the adaptive (HLA) and innate immune system (TRIM genes) that have a higher frequency in subjects being prescribed medications indicated for specific auto-immune disorders tend to reduce the risk of severe COVID-19. Indeed, it is well known that auto-immune disorders are more common in females [87] who also have a smaller risk of severe COVID-19 in comparison to men

[88]. While it is of course reasonable to expect that subjects with an increased risk for auto-immune disorders will tend to fight off infections more efficiently, the added value of our analysis is to pinpoint specific genes that are potentially involved in mediating this effect, and which may hint towards a protective pathway and/or therapeutic targets against severe COVID-19.

Our novel localized gene-centred approach was able to single out several plausible candidate causal genes, discussed in part already elsewhere in the literature in the COVID-19 context, and worthy of further investigations. Note that recent work using standard GWAS methods could not detect genetic evidence linking vitamin D to severity of COVID-19 [89]. This illustrates the power of the methods developed in this work as a discovery engine. We therefore believe that these methods are likely to be of utility for other studies trying to identify the genetic player mediating pleiotropic effects. An expected increase in the power of severe COVID GWAS and other relevant traits is likely to refine the picture starting to emerge in our analysis even further.

We consider it natural that the same genes, or even their domains, may exhibit effects with opposite signs for GWAS of different traits. Similarly pathways, or subsets thereof, may include sets of genes whose effects either coherently increase or reduce the effect on (or risk of) the trait. Coherence may occur at different levels of granularity (domain, gene, gene-pair, subunit or entire pathway), and these levels depend on which GWAS trait pair is co-analysed, specifically the complicated interplay between the common genetic components in each trait. Full genome based methods like genetic correlation or usual applications of Mendelian randomization can not resolve this fine structure, but rather yield only an aggregated trend for the traits. We therefore like to stress that statements made in this work regarding trait risk implications are not made for the traits in general, but rather for the respective gene under discussion. For a given pair of traits, the implication may differ between different genes.

In this work we assumed that there is no sample overlap between the GWAS under consideration and that the GWAS under consideration either have similar sample sizes, or that effects of sample sizes differences can be compensated by qq-normalization. Further work is needed to study the potential impact of sample overlap in more detail and to develop potential corrections along the lines of [90, 91]. Also our method would profit from better ways to deal with GWAS pairs whose traits exhibit very different effect size distributions or sample sizes.

Finally, we can envisage other, more general, applications of the methods discussed in this paper. For instance, our approach could be used to correct for the auto-correlation structures in estimating the significance of correlation between time-series. Hence, the technical results of this paper may be as well of interest for other domains.

Acknowledgments

We like to thank A. L. Button, A. Brümmer, Z. Kutalik and S. O. Vela for valuable comments on an earlier draft of the manuscript.

Appendix A: General product-normal

Let us briefly consider as well the distribution for xy with $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, *i.e.*, the product-normal for non-standardized Gaussian random variables. The moment generating function reads [19]

$$M_{x,y}(\nu) = \frac{e^{\frac{(\mu_x^2\sigma_y^2+\mu_y^2\sigma_x^2-2\varrho\mu_x\mu_y\sigma_x\sigma_y)\nu^2+2\mu_x\mu_y\nu}{2(1-\sigma_x\sigma_y(1-\varrho)\nu)(1+\sigma_x\sigma_y(1+\varrho)\nu)}}}{\sqrt{(1-\sigma_x\sigma_y(1+\varrho)\nu)(1+\sigma_x\sigma_y(1-\varrho)\nu)}} \,. \tag{A1}$$

Note that for, either, $\mu_x = 0$ or $\mu_y = 0$, the factorization of the main text still holds. Defining $\kappa_x = \frac{\mu_x}{\sigma_x}$ and $\kappa_y = \frac{\mu_y}{\sigma_y}$, the exponent of the exponential in $M_{x,y}\left(\frac{2\nu}{\sigma_x\sigma_y}\right)$ can be split for the general non-correlated case ($\varrho = 0$) into

$$\frac{(\kappa_x^2+\kappa_y^2+2\kappa_x\kappa_y)\nu}{2(1-2\nu)} - \frac{(\kappa_x^2+\kappa_y^2-2\kappa_x\kappa_y)\nu}{2(1+2\nu)} \,.$$

We deduce that we can still factorize the moment generating function, such that

$$\begin{aligned} xy \sim & \frac{\sigma_x\sigma_y}{2}\left[\chi_1^2\left(\frac{1}{2}(\kappa_x^2+\kappa_y^2+2\kappa_x\kappa_y)\right)\right] \\ & - \frac{\sigma_x\sigma_y}{2}\left[\chi_1^2\left(\frac{1}{2}(\kappa_x^2+\kappa_y^2-2\kappa_x\kappa_y)\right)\right] \,, \end{aligned} \tag{A2}$$

with $\chi_1^2(c)$ the non-central χ^2 distribution with one degree of freedom. Hence, the general product-normal can be expressed as a linear combination of non-central χ_1^2 distributions, for $\varrho = 0$.

Appendix B: Probability density functions

a. Product-Normal: Making use of the relation (4), the pdf, denoted as f, of the product-normal distribution can be calculated analytically via convolution (*cf.*, [92])

$$f_{\xi-\zeta}(x) = \frac{e^{\frac{x}{1-\varrho}}}{\pi\sqrt{1-\varrho^2}} \int_{\max(0,x)}^{\infty} dy (y^2-xy)^{-1/2} e^{-\frac{2y}{1-\varrho^2}} \,.$$

ϱ	0	0.3	0.6	0.9
MSE	3.3×10^{-15}	3.5×10^{-15}	4.6×10^{-15}	2.12×10^{-12}

TABLE II: Mean squared error (MSE) between numerical integration of (B1) and Davies algorithm for various ϱ and 1000 arguments evenly spaced in the range $[-4, 4]$.

Completing the square and invoking hyperbolic substitution, we arrive at

$$\begin{aligned} f_{\xi-\zeta}(x) &= \frac{e^{\frac{\varrho x}{1-\varrho^2}}}{\pi\sqrt{1-\varrho^2}} \int_0^\infty dt\, e^{-\frac{|x|}{1-\varrho^2}\cosh(t)} \\ &= \frac{e^{\frac{\varrho x}{1-\varrho^2}}}{\pi\sqrt{1-\varrho^2}} K_0\left(\frac{|x|}{1-\varrho^2}\right), \end{aligned} \tag{B1}$$

with K_0 the modified Bessel function of second kind at zero order. The result above for f is in agreement with the previous derivations of [93, 94]. Note that the analytic calculation of the corresponding cdf requires the solution of an integral of the type $\int_x^\infty dt\, e^{at} K_0(t)$. We are not aware of a known closed form solution.

For illustration, we plot the pdf for $\varrho = 1/2$ together with the corresponding histogram sampled from (4) in figure 18. We also show the cdf obtained via numerical integration of (B1). We verified that the numerical integration matches the results obtained via Davies algorithm for the cdf calculation for various ϱ, *cf.*, table II.

Note that since the pdf of the non-central χ^2 distribution includes a Bessel function, analytic calculation of the pdf of xy in the more general case of section A is more complicated than in (B1), and will not be discussed here.

b. Variance-Gamma: Consider a random variable X distributed according to

$$X(h, g) \sim [\Gamma(h/2, g)] - [\Gamma(h/2, g)]\,. \tag{B2}$$

The corresponding pdf can be calculated similar as above via convolution. We infer

$$f_{\xi-\zeta}(x) = \frac{e^{\frac{x}{g}}}{\Gamma(h/2)^2 g^h} \int_{\max(0,x)}^\infty dy\, (y^2 - xy)^{h/2-1} e^{-\frac{2y}{g}}\,.$$

Completing the square and using as before hyperbolic substitution, we arrive at

$$\begin{aligned} f_{\xi-\zeta}(x) &= \frac{x^{h-1}}{2^{h-1}\Gamma(h/2)^2 g^h} \int_0^\infty dt\, \sinh(t)^{h-1} e^{-\frac{|x|}{g}\cosh(t)} \\ &= \frac{K_{\frac{h-1}{2}}(|x|/g)}{g\sqrt{\pi}\,\Gamma(h/2)} \left(\frac{|x|}{2g}\right)^{\frac{h-1}{2}}, \end{aligned} \tag{B3}$$

with K_n a modified Bessel function of second kind at order n. Using the integral $\int_0^\infty dt\, t^{\mu-1} K_\nu(t) = 2^{\mu-2}\Gamma(\frac{\mu-\nu}{2})\Gamma(\frac{\mu+\nu}{2})$, we easily verify that $\int_0^\infty dx\, f_{\xi-\zeta}(x) = \frac{1}{2}$. Hence, due to symmetry the pdf is well normalized. However, we are not aware of closed form solutions for Bessel function integrals of the type $\int_x^\infty dt\, t^\nu K_\nu(t)$, which are needed to provide a closed form expression for the cdf.

The distribution given by (B3) occurred before in the finance domain as a special case of the *Variance-Gamma* distribution [95]. It can be traced back further to the distribution of the bivariate correlation. In detail, the gamma-variance corresponds to the off-diagonal marginal of a two-dimensional Wishart distribution, which models the covariance matrix [96]. However, what is new, to the best of our knowledge, is the expression in terms of the difference distribution in equation (B2).

For $h = n$ and $g = 1$, we have

$$X(n, 1) \sim [\Gamma(n/2, 1)] - [\Gamma(n/2, 1)] = \frac{1}{2}[\chi_n^2] - \frac{1}{2}[\chi_n^2]\,. \tag{B4}$$

Hence, for $n = 1$ we obtain the product-normal distribution with $\varrho = 0$ as discussed in the previous section. For general n we can view $X(n, 1)$ as the distribution of a sum of n independently distributed product-normal random variables. In particular, we can make use of Davies algorithm to calculate the cdf for $X(n, 1)$ exactly at a desired precision.

Appendix C: Enriched pathways

The enriched pathways ($p < 10^{-5}$) for the individual COVID-19, M05B and L04 GWAS, calculated as described in section VII D, are given in table III. Note that we detect for the COVID-19 trait only one Bonferroni significant pathway (*Roeth tert targets dn*).

We also tested for pathway enrichment of the genes under the ratio test for the combinations of traits investigated in section VIII C. We detected no Bonferroni significant hit. However, several of the lead pathways may be of potential interest in the COVID-19 context, as immune system related. Therefore, the top pathway scores ($p < 10^{-4}$) are listed in table IV.

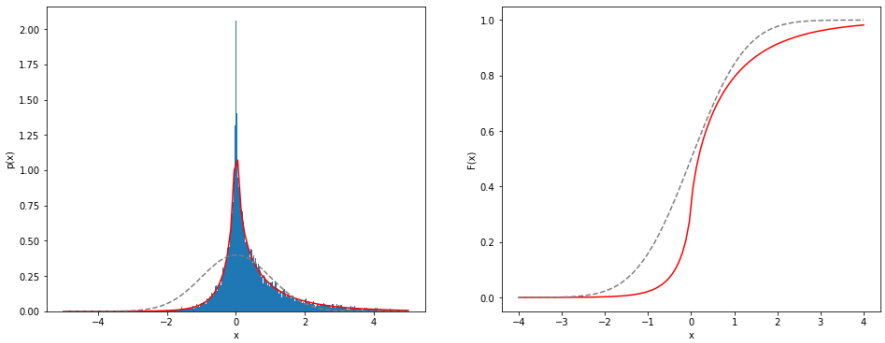

FIG. 18: Left: Histogram of the dependent product normal distribution with $\rho = 0.5$ obtained via subtracting 50.000 pairs of random samples of Gamma distributions following (4). The red line marks the pdf given in (B1). Right: The corresponding cdf obtained via Davies algorithm. For comparision, the dashed gray lines show the corresponding normal quantities.

trait	enriched pathways	# genes	p-value
COVID	**Roeth tert targets dn** *	8	6.9×10^{-7}
	chr3p21	6	5.54×10^{-6}
M05B	**GO skeletal system development**	439	6.51×10^{-8}
	GO ossification	358	3.52×10^{-7}
	GO biomineralization	145	1.35×10^{-6}
	GO tissue remodeling	158	3.04×10^{-6}
	GO MHC class II receptor activity	4	5.92×10^{-6}
	GO regulation of chondrocyte differentiation	50	9.08×10^{-6}
	KEGG cytokine cytokine receptor interaction	160	9.13×10^{-6}
L04	**Module 143**	9	2.35×10^{-12}
	KEGG allograft rejection	22	6.83×10^{-11}
	KEGG type I diabetes mellitus	27	1.08×10^{-10}
	Gaurnier PSDM4 targets *	44	1.49×10^{-10}
	Module 293	7	3.44×10^{-10}
	KEGG graft versus host disease	21	3.71×10^{-10}
	Reactome interferon gamma signaling *	67	4.11×10^{-9}
	GNF2 STAT6	72	1.75×10^{-8}
	PID IL12 2 pathway	55	1.87×10^{-8}
	Hallmark allograft rejection	173	2.01×10^{-8}
	KEGG autoimmune thyroid disease	25	2.46×10^{-8}
	GNF2 PTPN6	43	3.12×10^{-8}
	GO peptide antigen binding	12	3.35×10^{-8}
	KEGG antigen processing and presentation	40	6.47×10^{-8}
	Kim LRRC3B targets	28	9.17×10^{-8}
	Basso CD40 signaling up	93	9.27×10^{-8}

TABLE III: Enriched pathways for COVID-19, M05B and L04 GWAS. We tested against the 25724 gene sets of MSigDB 7.1 and only show pathways with $p < 10^{-5}$ for COVID-19 and M05B, respectively $p < 10^{-7}$ for L04. Bonferroni significant pathways ($p < 0.05/25724 \simeq 1.94 \times 10^{-6}$) are printed in bold. Pathways which are also Bonferroni significant under (anti)-coherence with the COVID-19 GWAS are marked with an asterix (*). Note that for L04 we detected 35 Bonferroni significant pathways and only a subset is listed.

exposure	outcome	tail	enriched pathways	# genes	p-value
M05B	COVID	L	Mori small pre BII lymphocyte dn	71	7.94×10^{-5}
		R	GO regulation of inflammatory response to wounding	4	3.65×10^{-5}
COVID	M05B	L	GO pos reg of transcr elongation from RNA polymerase II promoter	18	9.42×10^{-5}
		R	-		
calcium	COVID	L	GSE17721 LPS vs PAM3CSK4 16h bmdc dn	182	1.63×10^{-5}
			GSE31082 DP vs CD8 sp thymocyte up	187	2.93×10^{-5}
			GSE29618 LAIV vs TIV flu vaccine day 7 mdc dn	174	9.5×10^{-5}
		R	Wierenga STAT5A targets up	182	1.36×10^{-5}
			GO schwann cell migration	3	3.96×10^{-5}
			GSE21379 WT vs SAP ko tfh CD4 T-cell dn	180	4.14×10^{-5}
			Wierenga STAT5A targets group 1	117	7.37×10^{-5}
			Reactome RUNX1 reg estrogen receptor med trans	6	7.80×10^{-5}
			MIR3688 5P	128	8.53×10^{-5}
			MIR7113 3P	87	8.28×10^{-5}
COVID	calcium	L	GSE41176 WT vs TAK1 ko anti IGM stim B-cell 1h dn	181	1.37×10^{-5}
			GSE14350 IL2RB KO vs WT treg dn	185	5.06×10^{-5}
			GSE3720 VD1 vs VD2 gammadelta T-cell with PMA stim up	182	8.11×10^{-5}
		R	Farmer breast cancer cluster 5	11	2.66×10^{-5}
			Boylan multiple myeloma pca 3 up	73	5.17×10^{-5}
			Lee neural crest stem cell dn	107	9.93×10^{-5}
vitamin D	COVID	L	GO terminal bouton	50	2.98×10^{-5}
			GO molecular carrier activity	10	5.14×10^{-5}
		R	-		
COVID	vitamin D	L	Module 288	29	7.57×10^{-6}
			Wilensky response to darapladib	25	1.53×10^{-5}
			GO regulation of anion channel activity	9	4.20×10^{-5}
			Gross hypoxia via HIF1A only	8	4.99×10^{-5}
			GO neg reg of small GTPASE med signal transd	55	6.69×10^{-5}
			chr3p21	6	8.35×10^{-5}
		R	Vantveer breast cancer BRCA1 dn	42	2.46×10^{-5}
			GSE22886 unstim vs IL15 stim NK-cell up	178	4.38×10^{-5}

TABLE IV: Enriched pathways under the ratio test gene scores. We only show pathways with p-value $< 10^{-4}$. The left (L) tail corresponds to anti-coherence and, respectively, the right (R) to coherence. In total 25724 pathways of MSigDB 7.1 have been tested.

[1] Buniello, A. et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." Nucleic Acids Research, 2019, Vol. 47

[2] Lamparter D, Marbach D, Rueedi R, Kutalik Z, and Bergmann S. "Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics." PLoS Computational Biology 12, e1004714, 2016

[3] de Leeuw C, Mooij J, Heskes T, Posthuma D "MAGMA: Generalized gene-set analysis of GWAS data." PLoS Comput Biol 11(4): e1004219. doi.org/10.1371/journal.pcbi.1004219

[4] Slatkin M. "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future." Nat Rev Genet. 2008;9(6):477-485. doi.org/10.1038/nrg2361

[5] Solovieff, N., Cotsapas, C., Lee, P. et al. "Pleiotropy in complex traits: challenges and strategies." Nat Rev Genet 14, 483–495 (2013). doi.org/10.1038/nrg3461

[6] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, Albert-László Barabási, "The human disease network." Proceedings of the National Academy of Sciences May 2007, 104 (21) 8685-8690; doi.org/10.1073/pnas.0701361104

[7] Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, et al. "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics." PLOS Genetics 10(5): e1004383. doi.org/10.1371/journal.pgen.1004383

[8] Bulik-Sullivan B, Finucane HK, Anttila V, et al. "An atlas of genetic correlations across human diseases and traits." Nat Genet. 2015;47(11):1236-1241. doi.org/10.1038/ng.3406

[9] Bulik-Sullivan BK, Loh PR, Finucane HK, et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." Nat Genet. 2015;47(3):291-295. doi.org/10.1038/ng.3211

[10] Shi H, Mancuso N, Spendlove S, Pasaniuc B. "Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits." Am J Hum Genet. 2017;101(5):737-751. doi.org/10.1016/j.ajhg.2017.09.022

[11] H. Ruben, "Probability Content of Regions Under Spherical Normal Distributions, IV: The Distribution of Homogeneous and Non-Homogeneous Quadratic Functions of Normal Variables." Ann. Math. Statist. Volume 33, Number 2 (1962), 542-570.

[12] J. Sheil and I. O'Muircheartaigh, "Algorithm AS 106: The Distribution of Non-Negative Quadratic Forms in Normal Variables Journal of the Royal Statistical Society." Series C (Applied Statistics) Vol. 26, No. 1 (1977), pp. 92-98

[13] "Algorithm AS 204: The Distribution of a Positive Linear Combination of χ^2 Random Variables." Journal of the Royal Statistical Society. Series C (Applied Statistics) Vol. 33, No. 3 (1984), pp. 332-339

[14] R. B. Davies, "Numerical Inversion of a Characteristic Function." Biometrika, Vol. 60, No. 2 (Aug., 1973), pp. 415-417

[15] Dean A. Bodenham and Niall M. Adams, "A comparison of efficient approximations for a weighted sum of chi-squared random variables." Statistics and Computing volume 26, pages917–928(2016)

[16] B. L. Welch, "THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS WHEN THE POPULATION VARIANCES ARE UNEQUAL." Biometrika, Volume 29, Issue 3-4, February 1938, Pages 350–362, doi.org/10.1093/biomet/29.3-4.350

[17] F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components." Biometrics Bulletin, Vol. 2, No. 6 (Dec., 1946), pp. 110-114, International Biometric Society doi.org/10.2307/3002019

[18] G. E. P. Box, "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification." Ann. Math. Statist., Volume 25, Number 2 (1954), 290-302.

[19] C. C. Craig, "On the frequency function of xy." Ann. Math. Stat. 7 (1936) 1-15

[20] Uwe Küchler and Stefan Tappe, "Bilateral gamma distributions and processes in financial mathematics,", Stochastic Processes and their Applications, Volume 118, Issue 2, 2008, Pages 261-283, doi.org/10.1016/j.spa.2007.04.006

[21] Toby Johnson, "Efficient Calculation for Multi-SNP Genetic Risk Scores." Presented at the American Society of Human Genetics Annual Meeting, San Francisco, November 6–10, 2012

[22] The COVID-19 Host Genetics Initiative. "The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic." Eur. J. Hum. Genet. 28, 715–718 (2020). doi.org/10.1038/s41431-020-0636-6

[23] COVID-19 Host Genetics Initiative. "Mapping the human genetic architecture of COVID-19." Nature (2021). doi.org/10.1038/s41586-021-03767-x

[24] Wu, Y., Byrne, E.M., Zheng, Z. et al., "Genome-wide association study of medication-use and associated disease in the UK Biobank." Nat Commun 10, 1891 (2019). doi.org/10.1038/s41467-019-09572-5

[25] Krefl, D. and Bergmann, S., "PascalX". (2021). Zenodo. doi.org/10.5281/zenodo.4429921

[26] The 1000 Genomes Project Consortium "A global reference for human genetic variation," Nature 526, 68-74 (01 October 2015) doi.org/10.1038/nature15393

[27] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, Jill P. Mesirov "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of

the National Academy of Sciences Oct 2005, 102 (43) 15545-15550; doi.org/10.1073/pnas.0506580102

[28] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, Jill P. Mesirov, "Molecular signatures database (MSigDB) 3.0," Bioinformatics, Volume 27, Issue 12, 15 June 2011, Pages 1739–1740, doi.org/10.1093/bioinformatics/btr260

[29] Gao, JL., et al. "Impaired host defense, hematopoiesis, granulomatous inflammation and type 1-type 2 cytokine balance in mice lacking CC chemokine receptor 1." J Exp Med. 1997 Jun 2;185(11):1959-68. doi.org/10.1084/jem.185.11.1959

[30] Hickey MJ, Held KS, Baum E, Gao JL, Murphy PM, Lane TE. "CCR1 deficiency increases susceptibility to fatal coronavirus infection of the central nervous system." Viral Immunol. 2007 Dec;20(4):599-608. doi.org/10.1089/vim.2007.0056

[31] Daniel J. Dairaghi, Elizabeth R. Oldham, Kevin B. Bacon,Thomas J. Schall "Chemokine Receptor CCR3 Function Is Highly Dependent on Local pH and Ionic Strength." COMMUNICATIONS— VOLUME 272, ISSUE 45, P28206-28209, NOVEMBER 07, 1997 doi.org/10.1074/jbc.272.45.28206

[32] Alison A. Humbles, Bao Lu, Daniel S. Friend, Shoji Okinaga, Jose Lora, Amal Al-garawi, Thomas R. Martin, Norma P. Gerard, Craig Gerard "The murine CCR3 receptor regulates both the role of eosinophils and mast cells in allergen-induced airway inflammation and hyperresponsiveness." Proceedings of the National Academy of Sciences Feb 2002, 99 (3) 1479-1484; doi.org/10.1073/pnas.261462598

[33] Bariaa A. Khalil, Noha Mousaad Elemam, Azzam A. Maghazachi, "Chemokines and chemokine receptors during COVID-19 infection." Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 976-988, ISSN 2001-0370, doi.org/10.1016/j.csbj.2021.01.034

[34] Jiang, H., et al. "LZTFL1 upregulated by all-trans retinoic acid during CD4+ T cell activation enhances IL-5 production." J. Immunol. 196, 1081–1090. doi.org/10.4049/jimmunol.1500719

[35] Macias MP, et. al. "Expression of IL-5 alters bone metabolism and induces ossification of the spleen in transgenic mice." J Clin Invest. 2001;107(8):949-959. doi.org/10.1172/JCI11232

[36] Roeth A, Baerlocher GM, Schertzer M, Chavez E, Duehrsen U, Lansdorp PM. "Telomere loss, senescence, and genetic instability in CD4+ T lymphocytes overexpressing hTERT." Blood. 2005 Jul 1;106(1):43-50. doi.org/10.1182/blood-2004-10-4144

[37] Kurozumi K, Hardcastle J, Thakur R, Yang M, Christoforidis G, Fulci G, Hochberg FH, Weissleder R, Carson W, Chiocca EA, Kaur B. "Effect of tumor microenvironment modulation on the efficacy of oncolytic virus therapy." J Natl Cancer Inst. 2007 Dec 5;99(23):1768-81. doi.org/10.1093/jnci/djm229

[38] Ozato, K., Shin, DM., Chang, TH. et al. "TRIM family proteins and their emerging roles in innate immunity." Nat Rev Immunol 8, 849–860 (2008). doi.org/10.1038/nri2413

[39] Giraldo MI, Hage A, van Tol S, Rajsbaum R. "TRIM Proteins in Host Defense and Viral Pathogenesis." Curr Clin Microbiol Rep. 2020 Aug 8:1-14. doi.org/10.1007/s40588-020-00150-8

[40] Morris JA, et. al. "An atlas of genetic influences on osteoporosis in humans and mice." Nature Genetics volume 51, pages258–266(2019) doi.org/10.1038/s41588-018-0302-x

[41] Eriksson AL, et. al. "Genetic Determinants of Circulating Estrogen Levels and Evidence of a Causal Effect of Estradiol on Bone Density in Men." The Journal of Clinical Endocrinology & Metabolism, Volume 103, Issue 3, March 2018, Pages 991–1004, doi.org/10.1210/jc.2017-02060

[42] Neale lab, "UK Biobank calcium metabolite phenotype." Dataset ukb-d-30680_irnt from ieu open gwas project gwas.mrcieu.ac.uk/

[43] Revez JA, et. al. "Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration." Nat Commun 11, 1647 (2020). doi.org/10.1038/s41467-020-15421-7

[44] Okada Y, Wu D, Trynka G, et al. "Genetics of rheumatoid arthritis contributes to biology and drug discovery". Nature. 2014;506(7488):376-381. doi.org/10.1038/nature12873

[45] NCBI gene database www.ncbi.nlm.nih.gov/gene/4939

[46] Jadhav NJ, Gokhale S, Seervi M, Patil PS, Alagarasu K. "Immunomodulatory effect of 1, 25 dihydroxy vitamin D3 on the expression of RNA sensing pattern recognition receptor genes and cytokine response in dengue virus infected U937-DC-SIGN cells and THP-1 macrophages." Int Immunopharmacol. 2018 Sep;62:237-243. 10.1016/j.intimp.2018.07.019

[47] Pankiv S, Alemu EA, Brech A, Bruun JA, Lamark T, Overvatn A, Bjørkøy G, Johansen T. "FYCO1 is a Rab7 effector that binds to LC3 and PI3P to mediate microtubule plus end-directed vesicle transport." J Cell Biol. 2010 Jan 25;188(2):253-69. 10.1083/jcb.200907015

[48] Miao G, Zhao H, Li Y, Ji M, Chen Y, Shi Y, Bi Y, Wang P, Zhang H. "ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation." Dev Cell. 2021 Feb 22;56(4):427-442.e5. 10.1016/j.devcel.2020.12.010

[49] Wu S, Sun J. "Vitamin D, vitamin D receptor, and macroautophagy in inflammation and infection." Discov Med. 2011;11(59):325-335.

[50] Kim CH, Kunkel EJ, Boisvert J, Johnston B, Campbell JJ, Genovese MC, Greenberg HB, Butcher EC. "Bonzo/CXCR6 expression defines type 1-polarized T-cell subsets with extralymphoid tissue homing potential." J Clin Invest. 2001 Mar;107(5):595-601. 10.1172/JCI11902

[51] Baeke F, Korf H, Overbergh L, Verstuyf A, Thorrez L, Van Lommel L, Waer M, Schuit F, Gysemans C, Mathieu C. "The vitamin D analog, TX527, promotes a human CD4+CD25highCD127low regulatory T cell profile and induces a migratory signature specific for homing to sites

of inflammation." J Immunol. 2011 Jan 1;186(1):132-42. 10.4049/jimmunol.1000695

[52] Ali N. "Role of vitamin D in preventing of COVID-19 infection, progression and severity." J Infect Public Health. 2020 Oct;13(10):1373-1380. doi.org/10.1016/j.jiph.2020.06.021

[53] Pereira M., et. al. "Vitamin D deficiency aggravates COVID-19: systematic review and meta-analysis." Critical Reviews in Food Science and Nutrition (2020) doi.org/10.1080/10408398.2020.1841090

[54] Murdaca G., Pioggia G. and Negrini S. "Vitamin D and Covid-19: an update on evidence and potential therapeutic implications." Clin Mol Allergy. 2020; 18: 23. doi.org/10.1186/s12948-020-00139-0

[55] Yubin Zhou, Teryl K. Frey and Jenny J. Yanga "Viral calciomics: Interplays between Ca2+ and virus." Cell Calcium. 2009 Jul; 46(1): 1–17. doi.org/10.1016/j.ceca.2009.05.005

[56] Qin S, et. al. "The chemokine receptors CXCR3 and CCR5 mark subsets of T cells associated with certain inflammatory reactions". The Journal of Clinical Investigation. 101 (4): 746–54. doi.org/10.1172/JCI1422

[57] Loetscher P, et. al. "The ligands of CXC chemokine receptor 3, I-TAC, Mig, and IP10, are natural antagonists for CCR3." J Biol Chem. 2001 Feb 2;276(5):2986-91. doi.org/10.1074/jbc.M005652200

[58] Smit MJ, et. al. "CXCR3-mediated chemotaxis of human T cells is regulated by a Gi- and phospholipase C-dependent pathway and not via activation of MEK/p44/p42 MAPK nor Akt/PI-3 kinase". Blood. 102 (6): 1959–65. doi.org/10.1182/blood-2002-12-3945

[59] Kumiko Okada1, Akane Inoue1, Momoko Okada1, Yoji Murata1, et. al. "The Muscle Protein Dok-7 Is Essential for Neuromuscular Synaptogenesis." Science 23 Jun 2006: Vol. 312, Issue 5781, pp. 1802-1805 10.1126/science.1127142

[60] Laura J. Megeath and Justin R. Fallon "Intracellular Calcium Regulates Agrin-Induced Acetylcholine Receptor Clustering." Journal of Neuroscience 15 January 1998, 18 (2) 672-678; DOI: doi.org/10.1523/JNEUROSCI.18-02-00672.1998

[61] Chaolin Huang, Lixue Huang, Yeming Wang, Xia Li, Lili Ren, Xiaoying Gu, et al. "6-month consequences of COVID-19 in patients discharged from hospital: a cohort study" The Lancet, volume 397, issue 10270, p220-232, January 16, 2021 doi.org/10.1016/S0140-6736(20)32656-8

[62] Navarro-Corcuera A, Ansorena E, Montiel-Duarte C, Iraburu MJ. "AGAP2: Modulating TGFβ1-Signaling in the Regulation of Liver Fibrosis." Int J Mol Sci. 2020;21(4):1400. Published 2020 Feb 19. doi.org/10.3390/ijms21041400

[63] Lokmic Z, Musyoka J, Hewitson TD, Darby IA. "Hypoxia and hypoxia signaling in tissue repair and fibrosis." Int Rev Cell Mol Biol. 2012;296:139-85. doi.org/10.1016/B978-0-12-394307-1.00003-5

[64] Hong WX, Hu MS, Esquivel M, et al. "The Role of Hypoxia-Inducible Factor in Wound Healing." Adv Wound Care (New Rochelle). 2014;3(5):390-399. doi.org/10.1089/wound.2013.0520

[65] Lee, J.W., Ko, J., Ju, C. et al. "Hypoxia signaling in human diseases and therapeutic targets." Exp Mol Med 51, 1–13 (2019). doi.org/10.1038/s12276-019-0235-1

[66] Mirjam Kiener, Nuria Roldan, Carlos Machahua1, Arunima Sengupta, Thomas Geiser, Olivier Thierry Guenat, Manuela Funke-Chambour, Nina Hobi and Marianna Kruithof-de Julio "Human-Based Advanced in vitro Approaches to Investigate Lung Fibrosis and Pulmonary Effects of COVID-19." Front. Med., 07 May 2021 doi.org/10.3389/fmed.2021.644678

[67] Alison E. John, Chitra Joseph, Gisli Jenkins, Amanda L. Tatler "COVID-19 and pulmonary fibrosis: A potential role for lung epithelial cells and fibroblasts." Immunological Reviews, Wiley First published: 24 May 2021 doi.org/10.1111/imr.12977

[68] Wei X, Zeng W, Su J, Wan H, Yu X, Cao X, Tan W, Wang H. "Hypolipidemia is associated with the severity of COVID-19. J Clin Lipidol." 2020 May-Jun;14(3):297-304. doi.org/10.1016/j.jacl.2020.04.008

[69] Strunze S, Engelke MF, Wang IH, Puntener D, Boucke K, Schleich S, Way M, Schoenenberger P, Burckhardt CJ, Greber UF. "Kinesin-1-mediated capsid disassembly and disruption of the nuclear pore complex promote virus infection." Cell Host Microbe. 2011 Sep 15;10(3):210-23. doi.org/10.1016/j.chom.2011.08.010

[70] Dae-Kyum Kim, et. al., "A map of binary SARS-CoV-2 protein interactions implicates host immune regulation and ubiquitination." bioRxiv 2021.03.15.433877; doi.org/10.1101/2021.03.15.433877

[71] Sigrist CJ, Bridge A, Le Mercier P. "A potential role for integrins in host cell entry by SARS-CoV-2." Antiviral Res. 2020;177:104759. doi.org/10.1016/j.antiviral.2020.104759

[72] Dakal TC. "SARS-CoV-2 attachment to host cells is possibly mediated via RGD-integrin interaction in a calcium-dependent manner and suggests pulmonary EDTA chelation therapy as a novel treatment for COVID 19." Immunobiology. 2021;226(1):152021. doi.org/10.1016/j.imbio.2020.152021

[73] Valerdi KM, Hage A, van Tol S, Rajsbaum R, Giraldo MI. "The Role of the Host Ubiquitin System in Promoting Replication of Emergent Viruses." Viruses. 2021 Feb 26;13(3):369. doi.org/10.3390/v13030369

[74] Chatterjee, P., Ponnapati, M., Kramme, C. et al. "Targeted intracellular degradation of SARS-CoV-2 via computationally optimized peptide fusions." Commun Biol 3, 715 (2020). doi.org/10.1038/s42003-020-01470-7

[75] Picchiotti N., et. al. "Post-Mendelian genetic model in COVID-19," medRxiv 2021.01.27.21250593 doi.org/10.1101/2021.01.27.21250593

[76] Pairo-Castineira, E., Clohisey, S., Klaric, L. et al. "Genetic mechanisms of critical illness in COVID-19." Nature 591, 92–98 (2021). :doi.org/10.1038/s41586-020-03065-y

[77] Schneider W. M., et. al. "Genome-Scale Identification of SARS-CoV-2 and Pan-coronavirus Host Factor Net-

works Author links open overlay panel." Cell, Volume 184, Issue 1, 7 January 2021, Pages 120-132.e14 doi.org/10.1016/j.cell.2020.12.006

[78] Oh J.H., Tannenbaum A. and Deasy J.O., "Identification of biological correlates associated with respiratory failure in COVID-19." BMC Med Genomics. 2020; 13: 186. doi.org/10.1186/s12920-020-00839-1

[79] Davis AJ, Yan Z, Martinez B, Mumby MC. "Protein phosphatase 2A is targeted to cell division control protein 6 by a calcium-binding regulatory subunit." J Biol Chem. 2008;283(23):16104-16114. doi.org/10.1074/jbc.M710313200

[80] Friedman N, Jacob-Hirsch J, Drori Y, Eran E, Kol N, et al. (2021) "Transcriptomic profiling and genomic mutational analysis of Human coronavirus (HCoV)-229E -infected human cells." PLOS ONE 16(2): e0247128. doi.org/10.1371/journal.pone.0247128

[81] Speck C, Chen Z, Li H, Stillman B. "ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA." Nat Struct Mol Biol. 2005 Nov;12(11):965-71. doi.org/10.1038/nsmb1002

[82] Lips P, van Schoor NM. "The effect of vitamin D on bone and osteoporosis." Best Pract Res Clin Endocrinol Metab. 2011 Aug;25(4):585-91. doi.org/10.1016/j.beem.2011.05.002

[83] Jolliffe D, et. al., "Vitamin D supplementation to prevent acute respiratory infections: systematic review and meta-analysis of aggregate data from randomised controlled trials" medRxiv, Nov. 2020 (preprint) doi.org/10.1101/2020.07.14.20152728

[84] Rondanelli M, Miccono A, Lamburghini S, et al. "Self-Care for Common Colds: The Pivotal Role of Vitamin D, Vitamin C, Zinc, and Echinacea in Three Main Immune Interactive Clusters (Physical Barriers, Innate and Adaptive Immunity) Involved during an Episode of Common Colds-Practical Advice on Dosages and on the Time to Take These Nutrients/Botanicals in order to Prevent or Treat Common Colds." Evid Based Complement Alternat Med. 2018;2018:5813095. Published 2018 Apr 29. doi.org/10.1155/2018/5813095

[85] Merzon E, Tworowski D, Gorohovski A, Vinker S, Cohen A G, Green I, Frenkel-Morgenstern M, "Low plasma 25(OH) vitamin D level is associated with increased risk of COVID-19 infection: an Israeli population-based study." FEBS J. 2020 Sep;287(17):3693-3702. doi: 10.1111/febs.15495

[86] Vimaleswaran K S, Forouhi N G, Khunti K. "Vitamin D and covid-19" BMJ 2021; 372 :n544 doi.org/10.1136/bmj.n544

[87] Angum F, Khan T, Kaler J, Siddiqui L, Hussain A. "The Prevalence of Autoimmune Disorders in Women: A Narrative Review." Cureus. 2020;12(5):e8094. Published 2020 May 13. doi.org/10.7759/cureus.8094

[88] Peckham, H., de Gruijter, N.M., Raine, C. et al. "Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission." Nat Commun 11, 6317 (2020). doi.org/10.1038/s41467-020-19741-6

[89] Guillaume Butler-Laporte, et. al. "Vitamin D and COVID-19 susceptibility and severity in the COVID-19 Host Genetics Initiative: A Mendelian randomization study" PLOS Medicine doi.org/10.1371/journal.pmed.1003605

[90] Lin DY, Sullivan PF. "Meta-analysis of genome-wide association studies with overlapping subjects." Am J Hum Genet. 2009;85(6):862-872. doi.org/10.1016/j.ajhg.2009.11.001

[91] LeBlanc M, et. al. "A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework." BMC Genomics. 2018 Jun 25;19(1):494. doi.org/10.1186/s12864-018-4859-7

[92] G. L. Poe, E. K. Severance-Lossin and M. P. Welsh "Measuring the Difference (X - Y) of Simulated Distributions: A Convolutions Approach." American Journal of Agricultural Economics Vol. 76, No. 4 (Nov., 1994), pp. 904-915

[93] S. Nadarajah and T. K. Pogany, "On the distribution of the product of correlated normal random variables." C. R. Acad. Sci. Paris, Ser. I 354 (2016) 201-204

[94] Cui, G., Yu, X., Iommelli, S., and Kong, L. "Exact Distribution for the Product of Two Correlated Gaussian Random Variables." IEEE Signal Processing Letters, 23(11), 1662–1666. doi.org/10.1109/lsp.2016.2614539

[95] Dilip B. Madan and Eugene Seneta, "The Variance Gamma (V.G.) Model for Share Market Returns," The Journal of Business Vol. 63, No. 4 (Oct., 1990), pp. 511-524 (14 pages)

[96] K. Pearson, G. B. Jeffery and E. M. Elderton, "On the Distribution of the First Product Moment-Coefficient, in Samples Drawn from an Indefinitely Large Normal Population". (December 1929). Biometrika. Biometrika Trust. 21: 164–201. doi.org/10.2307/2332556