May 2021

Covariance of Interdependent Samples with Application to GWAS

Daniel Krefl and Sven Bergmann
Department of Computational Biology, University of Lausanne, Switzerland

We devise a significance test for covariance of samples not drawn independently, but with known inter-sample covariance structure. The test distribution we propose 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 propose a test for dependence between SNP-wise effect sizes of two genome-wide association studies at the level of genes. The 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, taking of class M05B medication (drugs affecting bone structure and mineralization), Vitamin D (25OHD) and Calcium concentrations. In particular, 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 can be found in genome-wide association studies (GWAS). In detail, GWAS correlate genomic single nucleotide polymorphisms (SNPs) with phenotypes in a study population, usually independently SNP by SNP. By now, thousands of such GWAS have been conducted and identified a plethora of statistically significant SNPs associated with disease risks and other relevant phenotypes, consistent with a polygenic genetic architecture [1]. These studies yield valuable information to better understand complex diseases and aid drug discovery and re-purposing.

However, due to Linkage Disequilibrium (LD) (*cf.*, [2]), many SNPs in close proximity are not independent, and this leads to dependencies between observed SNP effect sizes. This is of particular relevance in the aggregation to gene and pathway effect sizes. In this context, a pathway is understood as a functional set of genes and the SNP effect sizes in the gene regions are aggregated to a single effect. 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 [3] and MAGMA [4]. 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.

Often, a SNP is significant for more than one trait, *i.e.*, two or more traits may share the same genetic origin (pleiotropy) [5]. In particular, such a shared genetic component hints at that the same functional pathology may contribute to several diseases, *cf.*, [6]. At the gene level, usually a gene is considered to be relevant for two different traits if the gene carries a significant SNP effect size 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 genetic mechanism. In order to call pleiotropy of a gene, one should therefore move beyond comparing single variants, and take all SNPs in the gene region into account.

For a classic review on pleiotropy in complex traits we refer to [5]. In the following, some of the more novel methods proposed in the literature to uncover a fine-coarse shared genetic origin of two traits from GWAS summary statistics will be discussed.

One early method is the Bayesian based test of colocalisation between GWAS pairs of [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 for the usually available single SNP summary statistics appears, however, to be more involved.

A more recent method which attracted widespread attention and adoption to study the genetic component and correlation between traits is LD score regression [8, 9]. In this Ansatz, the effect sizes are considered as random variables and thereby naturally define a notation of heritability and (linear) genetic correlation. In partic-

ular, this leads to a simple and computational efficient method to calculate the correlation coefficient via regression. However, stratifying LD score regression onto the level of genes is in general difficult. Mainly because SNPs in the same genomic region are often in high LD, as already mentioned above. Hence, 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 demands 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 instead put forward to simply consider the raw second cross moment between the overlapping SNPs of two traits in a gene region, corrected for LD, to call 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 overlapping 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. Though our Ansatz does not intrinsically disentangle genetic and potential additional joint residual contributions, it nevertheless has utility in uncovering genes of relevance to more than one trait, as we will illustrate with an explicit example of current high interest in this work.

The technical novelty of the method we present here, which allows to perform such a significant test of covariance in the case of dependent samples, is the observation that 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. This allows one to take systematically into account not only isolated significant SNPs for calling a gene co-significant for two traits, but all common SNPs in the gene region. The assumption we have to make is that there is an underlying joint normal distribution, and the dependency pattern of the SNPs is known (for instance, the pattern can be approximated from a reference population). Furthermore, we assume that the study populations of the two traits do not have overlapping samples and that both populations can be approximated by an identical reference panel.

In the spirit of Mendelian randomization, the test 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 will introduce the probability distributions of relevance for this work. Most readers are probably familiar with the χ^2 distribution reviewed in section II. The role played in this work by the χ^2 is essential, 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 will see in section III, as well as for the Variance-Gamma distribution discussed in appendix B. The core sections of this paper constitute IV, in which we explain how we can perform a significance test of independence under the conditions described above, and section V, in which we derive a way 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. As illustrative example, we will use the developed technique to find computational predictions for shared genetic links between severity of Covid-19, class M05B medication taking, 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 taking of specific immune related medications (H03A, L04 and M01A).

II. LINEAR COMBINATION OF χ^2 DISTRIBUTIONS

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

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

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

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

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

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

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

One should note that the achievable precision of p calculated via (2) is limited by the available machine precision.

Note that also a varity 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].

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]

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

The core observation we will make use of 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]. One should note that the subtraction is in the distributional sense. Hence, even for $\varrho = 0$, the difference 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

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

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. EXAMPLE 1: 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 interesting 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). One should interpret the index I for $N = 1$ as a measure of coherence (or anti-coherence) between z and w. The significance threshold curve for a fixed desired p-value, say $p_I = 10^{-8}$, is illustrated in figure 1. One should note that the curve is unbounded. That is, for given w, there is always a corresponding z such that the resulting product I is significant. For instance, this differs from Fisher's exact method to combine p-values (meta-analysis). Fisher's method combines two p-values p_i via $-2\log p_1 - 2\log p_2 \sim [\chi_2^2]$. Due to its additive nature, Fisher's method results in a bounded significance curve. In the extreme case of one of the p-values being equal to one, the significance simply corresponds to 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, the other one has to be exponentially smaller in order to stay on the same significance curve,

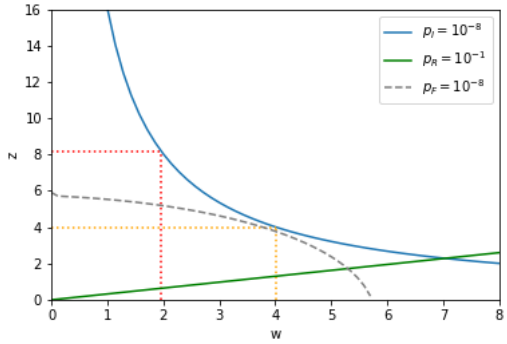

FIG. 1: Significance threshold curves of I (blue line) and R (green line) in the one element case for $p_I = 10^{-8}$, respectively, $p_R = 10^{-1}$, in the positive domain. The orange dotted lines mark the symmetric point with $p_z = p_w \sim 5 \times 10^{-5}$. The red dotted lines mark $p_w = 0.05$ and $p_z \sim 10^{-16}$. The gray dashed line illustrates the significance threshold based on Fisher's method with $p_F = 10^{-8}$.

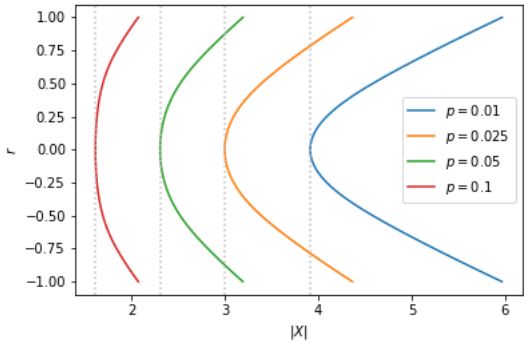

FIG. 2: Significance threshold curves of I in the two element case under variation of the inter-element correlation (y-axis). The x-axis corresponds to the argument of the tail probability defined in equation (2). The gray dotted lines mark the minimal value obtained for zero correlation.

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.

The linear threshold curve of R for $p_R = 10^{-1}$, defined in (9), is also shown in figure 1. Note the small slope, even for rather large p-values.

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. EXAMPLE 2: 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

$$\beta_i = \frac{1}{n} x_i^T y \,, \tag{11}$$

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

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

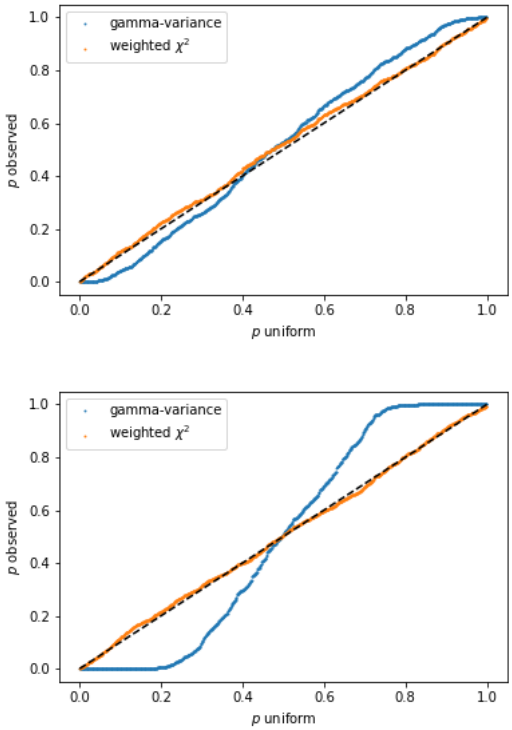

FIG. 3: QQ-plots of observed p-values resulting from the index I for 1000 pairs of samples of $\mathcal{N}(0,\Sigma)$ against uniform p-values. Top: Off-diagonal elements of Σ set to 0.2. Bottom: Off-diagonal set to 0.8. The blue curve is obtained using the gamma-variance distribution to perform the 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 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 (11), 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

$$z = \frac{1}{\sqrt{n}} x^T \epsilon \sim \mathcal{N}(0,\Sigma)\,, \tag{12}$$

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 (12) 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

$$z \sim \mathcal{N}(0, h^2 L) + \mathcal{N}(0,\Sigma) = \mathcal{N}(0, h^2 L + \Sigma)\,, \tag{13}$$

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 (13) requires an estimate of the heritability. Such an estimate can be obtained for via LD score regression [9].

c. Genetic coherence: 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,

$$V := z^T z \sim \sum_i \lambda_i [\chi_1^2]\,, \tag{14}$$

with λ_i the ith eigenvalue of the covariance matrix of the Gaussian. As discussed in detail in [3], a GWAS gene enrichment test can be performed via testing against (14). 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.*,

$$w^T z \sim \sum_i \frac{\lambda_i}{2} [\chi_1^2] - \sum_i \frac{\lambda_i}{2} [\chi_1^2]\,, \tag{15}$$

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 (15) (*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.

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

$$D := \sum_i z_i \,. \tag{16}$$

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

$$D \sim \mathcal{N}(0, |C|_F^2) \,, \tag{17}$$

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 (16) 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 (16) can succeed. We will also refer to testing for deviations from (17) as D-test.

e. Ratio test: Note that the ratio

$$R = \frac{w^T z}{z^T z} = \frac{\sum_i \lambda_i \bar{w}_i \bar{z}_i}{\sum_j \lambda_j \bar{z}_j^2} \,, \tag{18}$$

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.

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 (18). 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 (14), and confounding factors can be excluded. The setup is illustrated in figure 4.

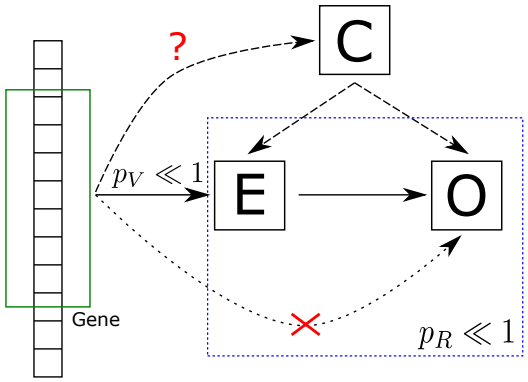

FIG. 4: Interpretation of the ratio test. The test detects potential gene-wise causal relations between a GWAS trait viewed as exposure (E) and a GWAS trait viewed as outcome (O). The association of the exposure has to be confirmed independently via (14). Potential confounders (C) have to be excluded by other means.

B. Single SNP

As we have seen in section VI, the product normal combines evidence in a multiplicative way. Obtaining a combined single SNP p-value from two different GWAS in this way comes however with a potential pitfall. Namely, if one of the SNPs p-values is small enough, we can always achieve co-significance. We see two possible strategies to mitigate this.

One way would be 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}, *cf.*, the red dotted lines in the figure. 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-significances.

Another possibility would be to first uniformize the p-values via ranking. That is, the p-values for all shared SNPs between the two GWAS are sorted for each GWAS and new p-values are calculated from the ranks. These ranked p-values (often also referred to as QQ-normalized) are then used as input for (15), after an appropriate inverse transform. Note that an unintended uplift of p-values, say above 0.05, could happen in the ranked case only for more than 10^{16} joined SNPs (*i.e.*, much more than in any realistic case).

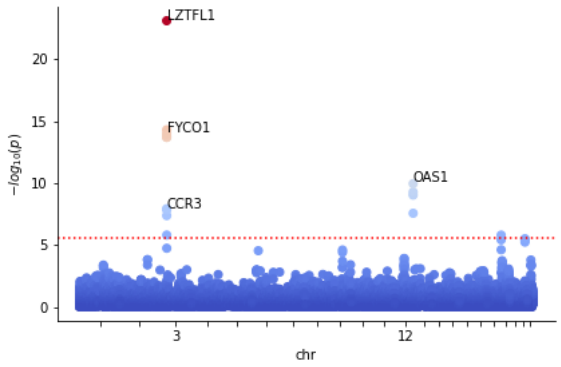

FIG. 5: Manhattan plot showing the strong gene enrichment on chromosomes 3 and 12 for the severe Covid-19 GWAS. The Bonferroni significance threshold is taken to be 2.67×10^{-6} (0.05 divided by number of tested genes, red dotted line). Only a selection of significant genes are labeled.

Since the ranking allows one to consider two GWAS with significantly different signal strength without the need to introduce an adhoc cutoff, the QQ-normalization is our approach of choice, and will be utilized in the following example. Note that also for the ratio test QQ-normalization is preferred in order to compensate for different signal strength between the GWAS.

C. Application to Covid-19 GWAS

1. *Coherence test*

We consider the meta-GWAS summary statistics on very severe respiratory confirmed Covid-19 cases versus the general population of European ancestry and without UK Biobank data contribution [22] (A2_ALL_eur_leave_ukbb_23andme, release 7. Jan. 2021). This GWAS shows significant gene enrichment on chromosomes 3 and 12, as can be inferred via testing against the null model (14) with fixed effect sizes. The Manhattan plot for the resulting gene p-values on these chromosomes is shown in figure 5. Note that even though the GWAS analysis implicates a set of disease relevant genes, the key causal drivers have not yet been identified.

We cross-scored this GWAS against a panel of GWAS on medication within the UK Biobank [23]. These GWAS on medication take 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 predisposition to taking specific medications. Note that we minimized possible effects of sample overlap via restricting to the Covid-19 meta-GWAS without UK Biobank contribution.

All calculations have been performed with the python package PascalX [24] with default settings, using the European subpopulation of the 1000 Genome Project [25] as reference panel to estimate the SNP-SNP correlations Σ. Only protein coding genes have been considered and the gene window has been extended by $50k$ at the transcription start and end position. The cross-scoring has been performed over the extended window. For the reason discussed in section VII B, we uniformize the raw GWAS p-values via joint rank transform. Note that we use the signs from the direction of association (β).

a. Coherence: The cross-scoring results for the coherence test are shown in figure 6. The medication group M05B (drugs affecting bone structure and mineralization) shows enrichment in very severe Covid-19. We show the Manhattan plot resulting from the null model (12) and the index (15) for the medication group M05B in figure 7. We observe that genes in the well-known Covid-19 peak locus on chromosome 3 appear to be coherent with M05B medication taking, with the strongest significance for the chemokine receptor genes *CCR1*, *CCR3* and *LZTFL1* in the region chr3p21. We test the orientation of the aggregated associations of these genes via (16) over the Covid-19 GWAS and find a positive direction (right tail).

For illustration, we show the spectrum of SNPs considered in the *CCR3* region and their SNP-SNP correlation in figure 8. The correlation between SNPs is clearly visible. Note that the SNPs are in high LD, as is evident from the heatmap shown on the right hand side of figure 8. Using the method described in section IV, we calculate a p-value of 1.50×10^{-8} for the significance of the coherence.

Note that the chemokine receptor of type 1 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 [26, 27].

The SNP spectrum in the *LZTFL1* region is shown in figure 9. In contrast to the gene discussed above, *LZTFL1* contains several independent LD blocks. The p-value for the significance of the covariance is calculated to be given by 1.19×10^{-7}. Additional indications for the observed co-signficance of *LZTFL1* for the two traits may be found in that it is known that the gene *LZTFL1* modulates T-cell activation and enhances IL-5 production [28]. In particular, mouse models suggest that expression of IL-5 alters bone metabolism [29].

Direction of association and coherence suggests that

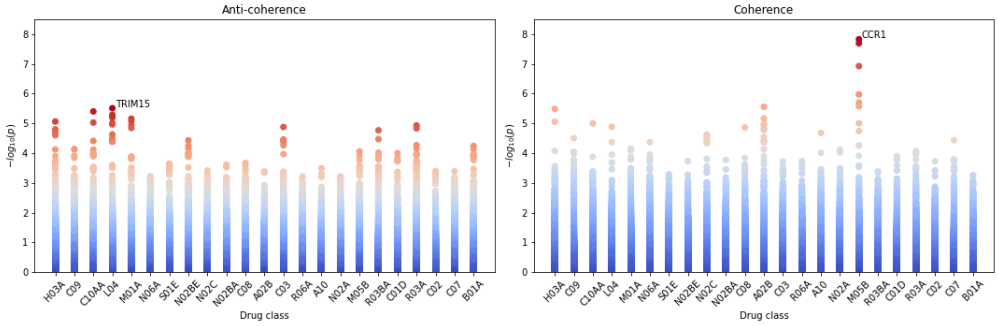

FIG. 6: Resulting p-values for cross scoring 23 drug classes GWAS against very severe Covid-19 GWAS for coherence, respectively, anti-coherence. Each data point corresponds to a gene and we marked the most significant gene in both cases. Left: Anti-coherence. Right: Coherence. Note that the drug class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

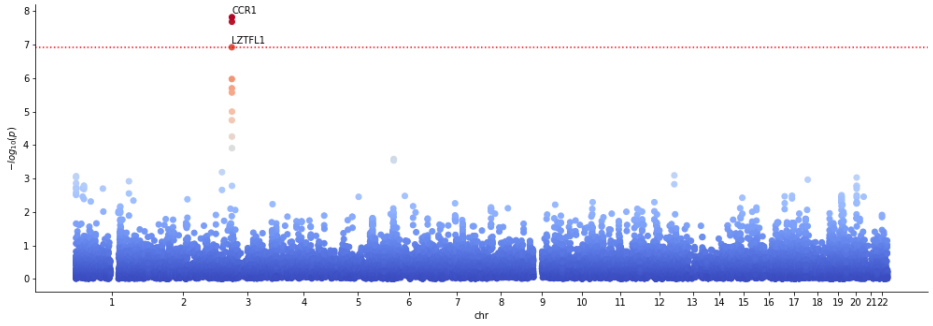

FIG. 7: Manhattan plot for cross scoring very severe confirmed Covid-19 with medication class M05B for coherence. Data points correspond to genes. The dotted red line marks a Bonferroni significance threshold of 1.16×10^{-7} (0.05 divided by number of genes tested and 23 drug classes.

genetic predispositions leading to M05B intake may carry a higher risk for severe Covid-19, with shared functional pathways related to the genes discussed above.

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 class possesses 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 $< 1 \times 10^{-5}$. 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 involved in pathogen-recognition and host defence [30]. 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 (16) and found that the detected genes for H03A, L04 and M01A tend to be localized in the left tail. Hence, a protective effect is suggested for these genes and therefore conditions leading to intake of these medications may imply a lower genetic risk for severe 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 score 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 [31], Estrogene levels in men (Estradiol and Estrone) [32], Calcium concentration [33], Vitamin D (25OHD) concentration [34] and Rheumatoid Arthritis [35]. The inferred gene enrichment for coherence and anti-coherence with the Covid-19 GWAS is shown in figure 11. We find 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

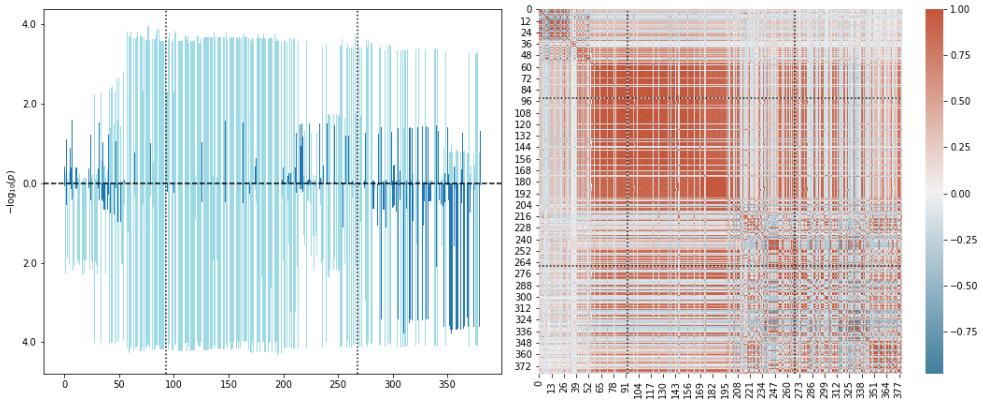

FIG. 8: Left: SNP p-values after rank transform for the *CCR3* gene. The x-axis is numbered according to the ith SNP in the gene window ordered by increasing position. The dotted black lines indicate the transcription start and end positions (first and last SNP). Up bars correspond to M05B and down bars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

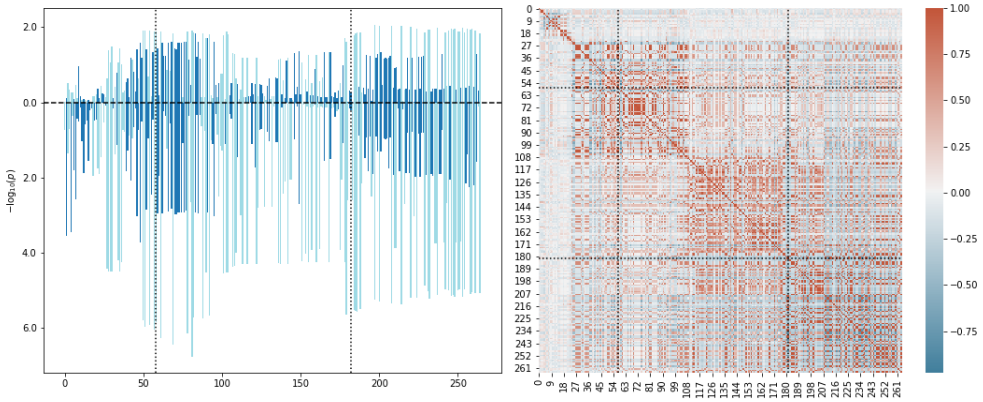

FIG. 9: Left: SNP p-values after rank transform for the *LZTFL1* gene. Annotation as in figure 8. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel. Note that the gene contains independent LD blocks.

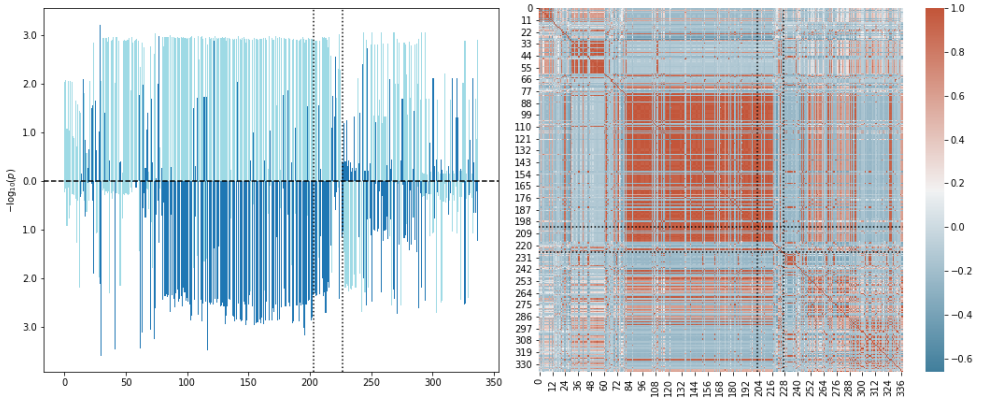

FIG. 10: Left: SNP p-values after rank transform for the *TRIM10* gene under medication L04. Annotation as in figure 8, but up bars corresponding to L04. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel.

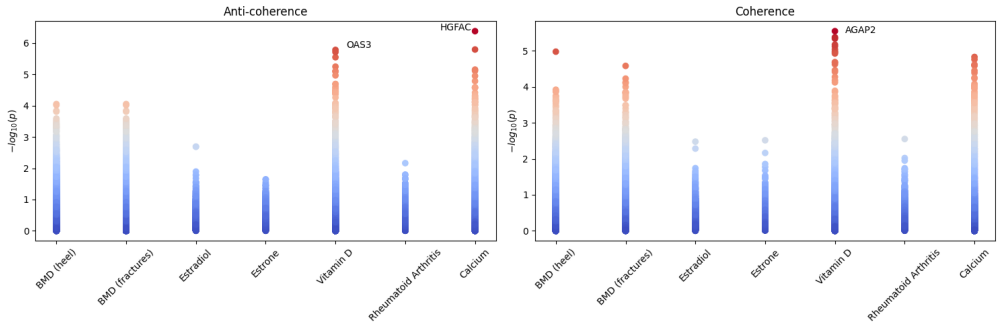

FIG. 11: Resulting p-values for cross scoring several GWAS related to Osteoporosis against the severe Covid-19 GWAS. We observe enrichment in Vitamin D and Calcium with several p-values $< 10^{-5}$. The leading genes are indicated.

plot in figure 12. 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 [36]. The gene *CXCR6* ($p = 7.96 \times 10^{-6}$) is expressed by subsets of T_H1 cells, but not by T_H2 cells. The D-test shows that all these genes are in the right tail and therefore suggest an increase in the risk for severe Covid-19 under variations in these genes. In particular, the anti-coherence of *CXCR6* towards Vitamin D concentration hints at that predisposition for Vitamin D concentration is linked to differences in immune system reaction, *i.e.*, between type 1 (inflammatory) or type 2 (anti-inflammatory). Indeed, possible links between severity of Covid-19 and Vitamin D are actively discussed in the current literature. See for instance [37–39] and references therein.

For Calcium the top genes are *HGFAC* ($p = 4.15 \times 10^{-7}$) and *DOK7* ($p = 1.61 \times 10^{-6}$). Note that *HGFAC* is Bonferroni significant under the seven traits tested. Both genes sit in the left tail under the D-test (16) applied to the Covid-19 GWAS. Therefore, for theses genes predisposition for high Calcium concentration suggest a reduced risk for severe Covid-19. Note that in general its is known that viruses appropriate or interrupt Ca^{2+} signaling pathways and dependent processes, *cf.*, [40].

The gene *HGFAC* plays a role in converting hepatocyte growth factor (HGF) to its active form. In particular, binding of HGF causes the upregulation 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 [41]. In detail, *CXCR3* binds the chemokine receptor *CCR3* and prevents an activation of T_H2-lymphocites. Thereby, a towards T_H1 biased inflammation immune reaction is triggered [42]. Note that *CXCR3* is able to increase intracellular Ca^{2+} levels [43].

The gene *DOK7* is of importance for neuromuscular synaptogenesis. 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. 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 [44].

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 one implied coherent gene of potential interest is *OS9* ($p = 4.2 \times 10^{-6}$). The corresponding protein binds to the hypoxia-inducible factor 1 (*HIF-1*). *HIF-1* is a key regulator of the hypoxic response. In particular, regulation of *HIF-1* interpolates between regeneration and scaring of injured tissue. It is known that severe Covid-19 may lead to lung tissue fibrosis. The D-test applied to the Covid-19 GWAS shows that *OS9* is located in the left tail and therefore a protective property of variations in this gene are suggested.

2. *Comparison to baseline*

It is illuminative to compare to a more simple way to detect potential co-significance of a gene for two GWAS. Namely, we may simply compute gene trait significance values via the original Pascal methodology, *i.e.*, making use of (14), and using gene-wise product-normal significance thresholds to detect genes of interest. Note that in doing so we ignore coherence of the SNPs inside the gene, and thereby compare only the total amount of variation. As before, we use jointly QQ-normalized GWAS p-values to suppress unintended uplift. 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 the simple approach is consistent with the results of the

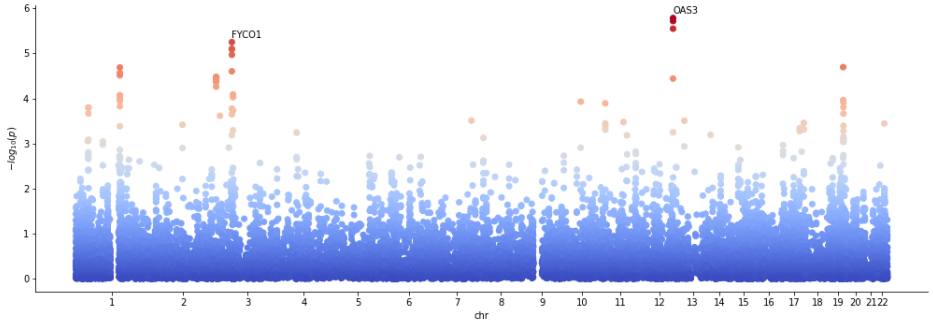

FIG. 12: Manhattan plot for cross scoring Vitamin D concentration against severe Covid-19 in the anti-coherent case.

more elaborate cross-scoring of the previous section. In particular, the genes *CCR1* and *CCR3* are confirmed to be of interest for both traits. Similar observations can be made for plotting the gene scores of severe Covid-19 against Vitamin D concentration, see figure 13 (middle plot).

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 (15), 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. That this is not only hypothetical may be seen at hand of figure 13 (bottom plot), there a gene p-value scatter plot is shown for severe Covid-19 against Calcium concentration. Remarkably, the detailed approach using individual SNP information marks genes to be of interest which are away from the gene-wise significant curve, and ignores other genes which are closer.

3. *Ratio test*

Let us also briefly discuss an application of the ratio based causality test introduced in sections V and VII Ae. For illustration, ratio test significance threshold curves for $p = 10^{-2}$ are indicated in figure 13. Note that the genes of non-interest are generally located between the two types of significance curves (in the figure between the black and orange dashed curves). Genes of relevance for the ratio test are instead engulfed by the coordinate axis and the ratio significance threshold curve (black dashed curves).

We ratio tested the Covid-19 GWAS against the Vitamin D and Calcium concentration GWAS in the coherent case and found that Vitamin D carries an interesting hit which suggests a causal pathway from genetic predisposition for Vitamin D concentration to severity of Covid-19. The ratio test Manhattan plot is shown in figure 14. In particular, besides the genes *KLC1* and *AL139300.1*, the gene *ZFYVE21* ($p_R = 3.9 \times 10^{-6}$) stands out. Under the Covid-19 GWAS (outcome) the gene carries a p-value of $p_V = 0.99$ while under the Vitamin D GWAS (exposure) a p-value of $p_V = 6.5 \times 10^{-4}$. 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* ($p_V = 0.26$). The D-test shows that the gene is in the right tail and hence is associated with high Vitamin D concentration. A possible explanation of the 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 [45, 46]. 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 $p_R = 1.83 \times 10^{-6}$ and Calcium as exposure. We have $p_V = 2.15 \times 10^{-5}$ under Calcium and $p_V = 0.82$ under Covid-19. The gene is associated with low Calcium concentration, as the D-test implies (left tail). The potential role played by this gene in the Covid-19 context is however unknown. As the gene shows an illustrative coherence pattern (multiple LD blocks), we show the corresponding SNP correlation plot in figure 15.

We also tested the anti-coherent case. The gene *HOXC4* with $p_R = 4.5 \times 10^{-6}$ is detected for M05B. Indi-

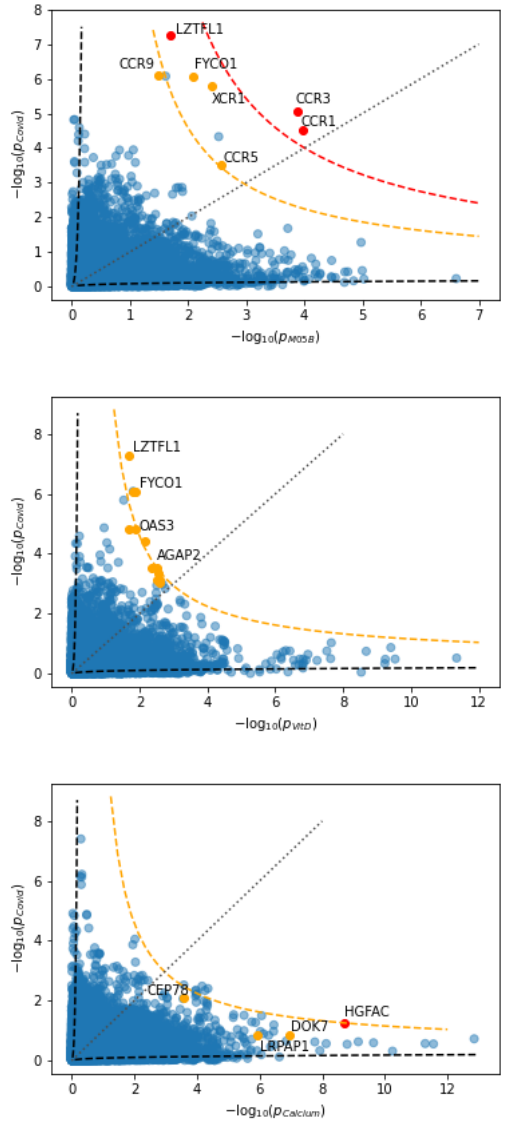

FIG. 13: GWAS gene scores, obtained separately for two GWAS using joint QQ-normalization and (14), plotted against each other. Each point corresponds to a gene. Top: Covid-19 vs. M05B. Middle: Covid-19 vs. Vitamin D. Bottom: Covid-19 vs. Calcium. The gray dotted line marks the diagonal. The orange and 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, $< 10^{-5}$). The black dashed curves indicate significance threshold curves of 10^{-2} for the ratio test.

vidual scored p-values for the gene read $p_V = 2.04 \times 10^{-5}$ under M05B and $p_V = 0.75$ under Covid-19. The D-test under M05B shows that the gene signal is located in the left tail. Hence, 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 note that *HOXC4* has been discussed before in the Covid-19 context [47]. 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 *DDP9* with $p_R = 3.88 \times 10^{-5}$, and $p_V = 0.74$ under M05B, respectively, $p_V = 2.57 \times 10^{-5}$ under Covid-19. This gene has been implicated before to be involved in the genetic mechanisms for critical illness due to Covid-19 [48]. Note that we also observe *DPP9* ($p_R = 5.82 \times 10^{-7}$) as top hit for Calcium concentration as exposure, *cf.*, the corresponding Manhattan plot shown in figure 16. Under both exposures the gene signal tends to be more in the right tail of the D-test. Another gene we detect in this ratio test is *IFNAR2* ($p_R = 9.70 \times 10^{-6}$), a known possible drug target against Covid-19 [48]. Note that this gene appears to be in the left tail of the D-test under Calcium as exposure.

For Vitamin D as exposure, we find in the anti-coherent case that the genes *MED23* ($p_R = 5.22 \times 10^{-5}$ and *GATA4* ($p_R = 7.88 \times 10^{-5}$) are the leading hits with a p-value $< 10^{-4}$. The D-test shows that *MED23* is in the right tail under the exposure, whereas the test is inconclusive for *GATA4*. The Mediator subunit *MED23* occurred before in SARS-CoV-2 screens, with a critical role of this complex during infection and death suggested [49]. *GATA4* has been observed previously in the context of the top significant biological processes likely associated with respiratory failure in Covid-19 patients [50].

Finally, note that we pick up the genes *HGFAC* ($p_R = 4.2 \times 10^{-5}$ and *DOK7* ($p_R = 1.27 \times 10^{-5}$), already detected before via the coherence test, as well via the anti-coherent ratio test with Calcium as exposure, suggesting a causal pathway. In addition, we observe for Calcium as exposure the leading gene *PPP2R3A* ($p_R = 1.0 \times 10^{-5}$), with unknown relation to Covid-19.

VIII. SUMMARY AND CONCLUSION

The core technical observation of this paper can be found in equations (5) and (B4), expressing the product-normal and variance-gamma distribution in terms of a χ^2 difference distribution. This allows for efficient calculation of the corresponding cdf and tail probabilities with Davies algorithm. This also allows one to perform a significance test of covariance for dependent samples, as long as the inter-sample correlation structure is known, see equation (8). Davies algorithm as well allows the cal-

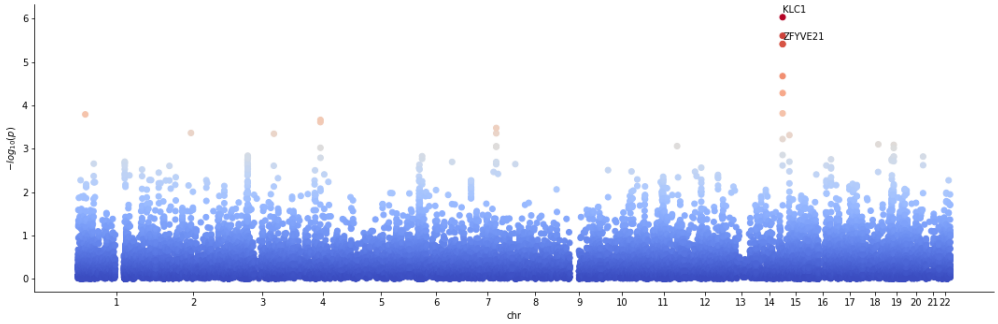

FIG. 14: Manhattan plot for ratio scoring Vitamin D concentration against severe Covid-19 in the coherent case, with ratio denominator given by Covid-19.

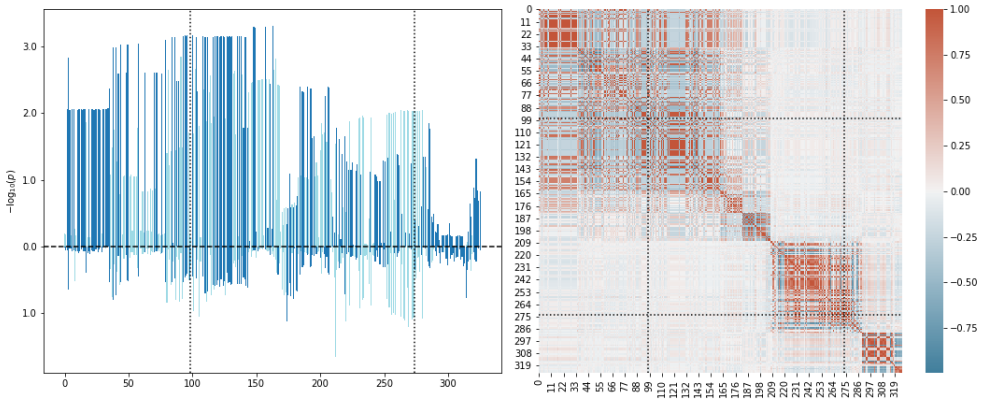

FIG. 15: Left: SNP p-values after rank transform for the *RNF217* gene under Calcium concentration and severe Covid-19. Annotation as in figure 8, but up bars corresponding to Calcium. Right: SNP-SNP correlation matrix inferred from the 1KG reference panel.

culation of the cdf of the ratio distribution (9), which finds application in testing for a causal relation.

We discussed a potential application for GWAS. Namely, to find coherent genes between different traits, *cf.*, section VII. We applied this method to uncover a potential link between severe Covid-19 and underlying conditions leading to usage of medications of class M05B. We found that chemokine receptors play a major role. Further investigations at hand of related GWAS phenotypes revealed genetic links to Vitamin D and Calcium concentrations, with Covid-19 cross implicated genes related to differentation between type 1 and type 2 immune response. A possible explanation of this emerging link being that patients taking class M05B medication usually suffer under Osteoporosis. Vitamin D stimulates the absorption of Calcium and therefore is related to bone mineral density [51]. However, it is also well known that Vitamin D contributes to reducing the risk of infection and death, mainly due to factors involving physical barriers, cellular natural immunity and adaptive immunity [52]. In this work we found hints that also the severity of Covid-19 may be impacted by Vitamin D concentration via influencing the immune response by pathways including chemokine receptors. In addition, we detected a potential causal link from genetic predisposition for Vitamin D concentration 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.

The previously in the severe Covid-19 context observed genes *HOXC4* and *DPP9* are also detected by our method, with a potential causal link to conditions leading to M05B medication intake. Similarly for *MED23* and *GATA4* with a potential causal link from Vitamin D concentration predisposition. Note that we can not exclude potential confounders.

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 taking 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, who also have a smaller risk

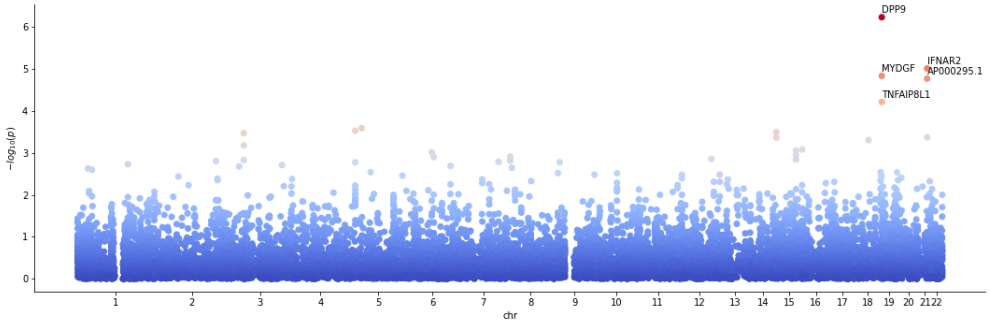

FIG. 16: Manhattan plot for ratio scoring Calcium concentration against severe Covid-19 in the anti-coherent case, with ratio denominator (outcome) given by Calcium concentration.

of severe Covid-19 in comparison to men. 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 and/or therapeutic pathway against severe Covid-19.

In general, we find it astonishing that the relatively simple methods presented in this work produce leading gene hits of apparently very high plausibility, and also re-discovers genes previously indicated in the literature. We therefore believe that the methods presented in this work may find wider utility.

Note that it is not a contradiction that we find the possibility for in effect diverging pathways between different genes for a pair of traits. It rather illustrates that in general the influence of one trait onto the other is a complicated interplay between in direction competing effects. More naive, 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. It would be interesting to study the potential impact of sample overlap in more detail and explicitly include the possibility for different sample sizes in the discussion of section VII. Artifacts of sample overlap could perhaps be corrected along the lines of [53, 54].

We can envisage other, more general, applications of the methods discussed in this paper. For instance, the methods might be of use 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

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

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

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.*, [55])

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

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 [56, 57]. 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 17. 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 I.

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 [58]. 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 [59]. 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.

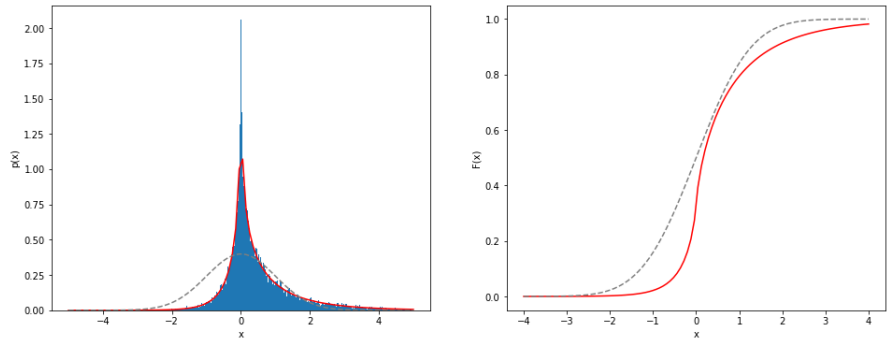

FIG. 17: Left: Histogram of the dependent product normal distribution with $\rho = 0.5$ obtained via subtracting 50.000 pairs of random samples of Gamma distributions following (4). The red line marks the pdf given in (B1). Right: The corresponding cdf obtained via Davies algorithm. For comparision, the dashed gray lines show the corresponding normal quantities.

[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] 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

[3] 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

[4] 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

[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] 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

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

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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

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

[34] 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

[35] 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

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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

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

[48] 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

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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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)

[59] 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