Multiple-testing Correction in Metabolome-wide Association Studies

Alina Peluso

Oak Ridge National Laboratory   https://orcid.org/0000-0003-2895-0406

Robert Glen

Imperial College London

Timothy M D Ebbels (✉ t.ebbels@imperial.ac.uk)

Imperial College London

Multiple-testing correction in metabolome-wide association studies

Alina Peluso[1], Robert Glen[1,2] and Timothy M D Ebbels[1*]

Abstract

The search for statistically significant relationships between molecular markers and outcomes is challenging when dealing with high-dimensional, noisy and collinear multivariate omics data. Permutation procedures allow for the estimation of adjusted significance levels without assuming independence among the metabolomic variables. Nevertheless, the complex non-normal structure of the metabolic profiles and outcomes may bias the permutation results leading to overly conservative threshold estimates i.e. lower than a Bonferroni or Sidak correction. Within a univariate permutation procedure we employ parametric simulation methods based on the multivariate (log-)Normal distribution to obtain adjusted significance levels which are consistent across phenotypes while effectively controlling the type I error rate at the α level. Next, we derive an alternative closed-form expression for the estimation of the number of non-redundant metabolic variates based on the spectral decomposition of their correlation matrix. The efficacy of our methods is tested for different model parametrizations and across a wide range of correlation levels of the variates using both synthetic and real data sets. Both the permutation-based formulation and the more practical closed form expression are found to give an effective indication of the number of indipendent metabolic effects exhibited by the system.

Keywords: FWER; MWAS; MWSL; Multiple testing; Permutation; Correlated tests

1 Introduction

In omics studies many hundreds to tens of thousands of molecular variables are collected for each individual, leading to high-dimensional multivariate data which are highly collinear. When analysing these data, thousands of hypothesis tests are conducted simultaneously, thus effective methods to adjust for multiple testing are a central topic, especially in the context of Metabolome-Wide Association Studies (MWAS) [1]. The aim is the detection of statistically significant relationships between molecular concentrations and disease phenotypes while minimising the risk of false positive associations. A widely used approach for multiple testing is the false discovery rate (FDR) [2] which controls the expected proportion of falsely rejected hypotheses among all those rejected. This approach is effective in the case of independent or positive dependent tests. While there have been some attempts to deal with correlated tests such as [3] that proposed a simple but highly conservative procedure, in general correlation among tests is still a problem for FDR methods. Besides FDR corrections, family-wise error rate

(FWER) procedures control the probability of making at least one false conclusion (i.e. at least one Type I error). The FWER provides a more stringent control of Type I error compared to the FDR. Nevertheless, FWER conventional methods such as the Bonferroni [4] or Sidak [5] adjustment are known to be overly conservative when the tests are correlated. On the other hand, resampling-based methods such as the permutation test are a standard tool to simultaneously assess the association of different correlated molecular quantities with an outcome of interest. These procedures can be conducted in both a parametric or non-parametric fashion. Parametric approaches are the preferred methods as they have relatively high power if the assumptions (e.g. normality of the data) hold. Nevertheless, in the context of MWAS the metabolic profiles are very rarely normally distributed nor present a symmetric distribution, and this may bias the result of the chosen significance test. Thus, a first aim of this study is to overcome this issue and derive a valid yet stable metabolome-wide significance level (MWSL) across phenotypes with diverse distributional properties. The proposed approach is based on a permutation procedure built from parametric approximation methods via the multivariate Normal and log-Normal distributions to describe, at least approximately, the

---

[*]Correspondence: t.ebbels@imperial.ac.uk
[1]Faculty of Medicine, Department of Surgery & Cancer, Imperial College London, South Kensington Campus, SW7 2AZ London, UK
Full list of author information is available at the end of the article

set of metabolic profiles while retaining their complex correlated structure up to the 2nd order moments, and while effectively controlling the expected overall type I error rate at the α level. While the proposed re-sampling-based method is accurate and asymptotically consistent it demands intensive computation. In the context of genomic studies there have been several attempts to formulate the problem in terms of estimating the number of non-redundant molecular quantities as a closed-form eigenvalue-based measure from the spectral decomposition of the empirical correlation matrix of the molecular variables. The available measures proposed by [6], [7], [8], [9], and [10] are proven to be not sufficiently accurate as a valid substitute for the proposed permutation procedure. Therefore, a second aim of this study is to derive a permutation-free closed-form estimate of the MWSL to express the number of non-redundant molecular quantities. Both the permutation-based MWSL formulation and the more practical closed form expression are tested on synthetic and real data.

2 Permutation-based MWSL estimation

2.1 Permutation algorithm

Suppose the data consists of n observations, and let Y be the response variable or phenotype, $X = (X_1, \ldots, X_M)^T$ the vector of M predictors or features, and $Z = (Z_1, \ldots, Z_P)^T$ the vector of P fixed effects covariates. The permutation-based MWSL estimation can be described as follows.

- Step (1): Shuffle i.e. re-sample without replacement, the outcome variable Y together with the set of fixed effects confounders Z if any. In this way, the n subjects are re-sampled under the null hypothesis of no association.
- Step (2): To estimate the relationship between the outcome and the set of features while accounting for possible confounding effects, compute M regression models in a univariate approach, that is by using one feature at a time. From each model store the p-value associated with the feature of interest. When appropriate, other approaches than regression methods can be used for testing of association e.g. correlation or t-test.
- Step (3): Extract the minimum of the set of M p-values as this indicates the highest threshold value which would reject all M null hypotheses.
- Step (4): Repeat Step (1)-(3) for K times, where K is at least $n/2$ times [11]. The K minimum p-values are the elements of the new vector q.
- Step (5): Sort the elements of q, and take the (αK)-value of this vector. This value is the MWSL estimate. An approximate confidence interval can be obtained by treating the true po-

sition of the MWSL estimate as a Binomial random variable with parameters K and α. Then, using the Normal approximation to the Binomial, we obtain the $z_{(1-\alpha)\%}$ confidence limits by extracting the elements of q in positions $(\alpha K) \pm \{(1-\alpha)\sqrt{\alpha K(1-\alpha)}\}$.

- Step (6): Compute the effective number of tests (ENT) defined as the number of independent tests that would be required to obtain the same significance level using the Bonferroni correction ENT$=\frac{\alpha}{\text{MWSL}}$. The ENT estimate measures the extent that the M markers are non-redundant. Therefore, the ratio R$=\frac{\text{ENT}}{M}\%$ of the effective and the actual number of tests (ANT or M) is a measure of the dependence among features, which it is expected to be closer to 0% when highly correlated features are considered.

Similar versions of this procedure have been previously applied in different studies e.g. by [12] to approximate the genome-wide significance threshold for dense SNP and resequencing data, or by [13] for urinary metabolic profiles. Recently in the context of NMR metabolic profiling studies [14] employed the permutation algorithm to perform a series of MWAS for serum levels of glucose. Counterintuitively ENT estimates greater than the ANT were found, with an R ratio for glucose over 400%.

2.2 Parametric simulation methods

The underlying assumption of the permutation procedure is that the p-values are properly calibrated, that is, every metabolite-specific p-value is uniformly distributed, i.e. p-value$_m \sim U(0, 1)$ where $m = 1, \ldots, M$, when the null hypothesis is true. Because the MWSL is the minimum p-value across the metabolite specific tests, all it takes is one poorly calibrated test with an erroneous small p-value to bias the MWSL estimation. In metabolomics studies very often the features are not normally distributed. Nevertheless, normality matters but only sometimes. It matters when both the feature and the outcome have a skew distribution [15], while it has very little effect when either the feature or the outcome is normally distributed. In this context, we investigate the properties of the permutation approach for significance level estimation by employing the multivariate Normal and the multivariate log-Normal distributions to describe, at least approximately, the set of correlated features and to obtain stable estimates of the MWSL while effectively controlling the maximum overall type I error rate at the α level. We assume that the data are already centred so that the means equal zero. Therefore, $X \sim \mathcal{N}_M(\mu, \Sigma^*)$ is the multivariate Normal distribution employed to simulate the set of features where $\mu = \text{E}[X] = (\text{E}[X_1], \ldots, \text{E}[X_M])^T = \mathbf{0}$

is the M-dimensional mean vector of zero means, and Σ^* is the $(M \times M)$ shrinkage estimator of the covariance matrix as described by [16]. The shrinkage estimator is always positive definite, well-conditioned, more efficient and therefore preferred to the unbiased estimator Σ, or to the related maximum likelihood estimator Σ_{ML}. Where the probability density of a feature is right skewed, we use the multivariate log-Normal approximation. In such cases, the features are first transformed i.e. the absolute value of their minimum, plus one unit, is added to their original value. The algorithm is applied to real-data and simulated scenarios to illustrate the results for different model parametrizations and distributional features of the outcome, as well as to investigate different correlation levels across features and between the features and the phenotype.

3 Practical approximation of the ENT

The empirical method of computing the permutation test p-value is hampered by the fact that a very large number of permutation is required to correctly estimate small, and therefore interesting p-values. Thus, we now present a more efficient alternative to a standard permutation test to derive the MWSL. To distinguish from the effective number of non-redundant variates from the permutation procedure which has been defined as ENT in Section 2, here we refer to the estimate from this practical approximation approach as Meff. It has previously been shown that the collective correlation among a set of variables can be measured by the variance of the eigenvalues (λs) derived from a correlation matrix ([6], [17]). In particular, high correlation among variables leads to high λs, that is, when all variables are completely correlated, the first λ equals the number of variables in the correlation matrix (M) and the rest of the λs are zero. Vice versa in the case of no correlation among variables, all the λs will be equal to 1 with zero variance. Hence, the variance of the λs will range between zero, and M. Based on this concept, within the genomics field several methods have been proposed for estimating the ENT from the correlations between variates. Among the first, [7] proposed to use the variance of the λs to estimate the ENT for the limiting cases of none/ fully correlated variables, and a ratio of the eigenvalue's variance to its maximum M for intermediate situations. [8] suggested summing the λs, after substituting 1 for the λs that are greater than 1. [9] suggested defining ENT as the number of λs which can explain a certain percentage of the variation within the data. However, it is unclear how the percentage should be chosen as overly large or small value of it would result in an FWER that is overly conservative or liberal. [10] proposed a measure of ENT based on

a λs ratio function. In the context of our analyses, the Meff measures proposed by these authors were not sufficiently accurate as a valid substitute of the permutation procedure, hence we propose an empirical closed-form expression directly related to the correlation among metabolomics variates as follows

$$\mathrm{Meff}_{\mathrm{MWSL}} = \left(\frac{\sum_{m=1}^{M} \sqrt{\lambda_m}}{\log(\lambda_1)} \right)^2 \Big/ \left(\frac{\sum_{m=1}^{M} \lambda_m}{\lambda_1} + \sqrt{\lambda_1} \right).$$

This formulation balances the information from the λ_m with $m = 1, \ldots, M$ estimated eigenvalues from the correlation matrix of the metabolites concentrations, with the contribution of the first eigenvalue λ_1 which measures the primary cluster in the matrix, its number of variables, and the average correlation among the features [18]. This formulation is of interest in the context of correlated variates, that is when at least all variates but one are independent, i.e. for $\lambda_1 > 1$, and therefore $\log(\lambda_1) > 0$.

To summarize the full algorithm as an alternative to the permutation procedure, we propose the following.

- Step (1): Compute the $\mathrm{Meff}_{\mathrm{MWSL}}$ with the proposed formulation.
- Step (2): Under the null hypothesis the p-value of each metabolite follows a Uniform distribution, i.e. p-value$_m \sim U(0, 1)$, where $m = 1, \ldots, M$. The distribution of minimum p-values can be approximated by the minimum order statistics (r=1), that is $U_{(1)} \sim \mathrm{Beta}(1, M)$ in the case of not correlated molecular variates, and $\mathrm{Beta}(1, M')$ with $M' \leqslant M$ in the case of correlated features. The limit case of very highly correlated features with $M' = 1 (<< M)$ reduces to sampling from a $\mathrm{Beta}(1, 1)$ which equals a $U(0, 1)$. It follows that the $\mathrm{Meff}_{\mathrm{MWSL}}$ can be used to approximate the distribution of minimum p-values sampling from a $\mathrm{Beta}(1, \mathrm{Meff}_{\mathrm{MWSL}})$.
- Step (3): The MWSL and its respective $z_{(1-\alpha)}\%$ confidence limits can be derived as described in Section 2.1, Step (5)-(6) of the permutation procedure.

4 Results

4.1 Study of experimental metabolomics data

The MWAS approach was employed to investigate the association between human serum ^{1}H NMR metabolic profiles and various clinical outcomes in the Multi-Ethnic Study of Atherosclerosis (MESA) [19]. The data have been extensively described in [14]. Briefly, the cohort includes participants (51% females, 49% males), aged 44-84 years, (mean=63 years) from four different ethnic groups: Chinese-American, African-American, Hispanic, and Caucasian, all recruited between 2000-2002 at clinical centres in the United States

and free of symptomatic cardiovascular disease at baseline. Demographic, medical history, anthropometric, and lifestyle data, as well as serum samples were collected, together with information on diabetes, and lipid and blood pressure treatment. Metabolic profiles were obtained using ^{1}H NMR at 600 MHz and processed as detailed in [20]. The outcomes of interest are glucose concentrations and the body mass index (BMI). Table 1 presents the descriptive statistics for the clinical outcome measures, while Table 2 reports the descriptive statistics for the fixed effects covariates used in the study. Three sets of NMR spectra are considered: (1) a standard water-suppressed one-dimensional spectrum (NOESY), (2) a Carr-Purcell-Meiboom-Gill spectrum (CPMG), and (3) a lower resolution version of the CPMG data (BINNED). The BINNED version consists of M=655 features, while the NOESY and CPMG contain M=30,590 features. The BINNED data sample comprises of n=3,500 individuals, while the NOESY and CPMG data have n=3,867 individuals. All MWSL calculations are performed for $\alpha = 0.05$.

From the conventional permutation procedure applied to the BINNED data shown in Figure 1, when the real features are considered, there is instability in the estimation of the ENT across the different outcomes, and in particular the ENT estimate for glucose is above the ANT. When the data are simulated from a multivariate log-Normal or Normal as described in Section 2.2 the ENT estimates are stable across the different outcomes and remain bounded below the total number of features with an average ENT around 350 and an R ratio around 50%. To assess the validity of this result in terms of redundancy of the set of features we considered principal component analysis (PCA) as an alternative method for estimating the ENT [6–8]. The cumulative proportion of variance explained by the first 350 PCs is around 99%. This is consistent with the interpretation that there are effectively 350 uncorrelated features in the data.

Figure 2 reports the ENT estimates for CPMG data. Without any transformations applied, there is a very large variation across the ENT estimates for the different outcomes, and in particular a very high and meaningless estimate for glucose levels which goes beyond R=400%. On the other hand, when the set of features is simulated from the multivariate Normal and from the multivariate log-Normal distribution the corresponding ENT estimate is below the total number of features, and stable across different outcomes with an average ENT of around 16,000 features and an R ratio around 50%. In this case the usefulness of the proposed permutation method to estimate the ENT is

clear in a comparison with the PCA-based ENT estimate would be constrained to the maximum number of PCs ($n = 3,866$ i.e. max no. PC is n-1).

Figure 3 reports the ENT estimates for the NOESY data which are below R=100% but vary across outcomes when the original set of features is considered. When simulated features from the multivariate (log-) Normal distributions are considered we obtain lower ENT values than the ones from the CPMG data, with an average ENT of around 2,700 features and an R ratio around 9%. This result was expected due to the reduced influence of broad signals in CPMG spectra compared to NOESY, which is linked to a weaker covariance structure of the data. By applying a PCA to the NOESY data the cumulative proportion of variance explained by the first 2,700 PCs is around 99%, and this is in line with our findings.

Next, by exploiting the approximation method described in Section 3, we derive the proposed $\text{Meff}_{\text{MWSL}}$. Table 3 provides this estimate compared to the available alternative methods proposed by [7], [8], [9], and [10], and the ENT estimate from the permutation procedure which is the averaged estimate of the results obtained via the multivariate and log-multivariate Normal transformations. Considering the complexity of the eigenvalue structure in cases of very large data sets, the proposed $\text{Meff}_{\text{MWSL}}$ in most cases seems to be able to consistently quantify, at least approximately, the correlation structure of the metabolomic variates. Based on this Meff estimate, to derive the MWSL and its confidence limits we simulate from a $\text{Beta}(1, \text{Meff}_{\text{MWSL}})$ which let us obtain MWSL estimate of the same order of magnitude as these from the permutation procedure as shown in Table 4.

4.2 Simulation study

We now broaden the investigation by considering various correlation levels of the set of molecular variates as well as cases of correlation between the outcome and the variates. At first we generate various sets of variates each of these with a specific and well bounded correlation level. This has be done following the algorithm described in Section 7.1. Specifically, we generated nine sets of variates covering the whole range of positive correlation levels. Next, we generate outcomes both correlated and uncorrelated to the variates which we will employ within the permutation procedure to estimate the ENT across the various sets of correlated molecular variates. *Uncorrelated outcomes* of different shapes are easily simulated via parametric distributions such as the Normal distribution for a symmetric outcome, the Skew-Normal distribution for a left skewed outcome, and a Weibull distribution for a right skewed outcome. Figure 4 shows the ENT

estimates in the case of correlated variates and uncorrelated outcomes. Simulated *correlated outcomes* can be obtained as a linear combination of few randomly chosen molecular variates with an added noise, or via procedures based on Cholesky decomposition as done when simulating correlated features following the algorithm detailed in Section 7.1. Figure 5 show the ENT estimates from the permutation procedure for the various sets of synthetic molecular variates and the correlated simulated phenotypes. We conclude that correlation to the outcome makes no discernible difference to relationship between ENT and variate-variate correlation. Lastly, we apply the Meff$_{\text{MWSL}}$ approximation to derive the results in Table 5 and Table 6. The ENT from the permutation procedure is averaged from the results in Figure 4 and Figure 5. In this simulated environment, the Meff$_{\text{MWSL}}$ approximation outperforms other available methods and well describes the permutation-based ENT estimates.

5 Validation of the approach

A type I error (false-positive) occurs when a true null hypothesis is being rejected. To check whether the permutation procedure accounts for the FWER at the α level, for each metabolic variate and across the permutation replicates, we measure the type I error rate as the number of occurrences of having a p-value less or equal than the MWSL. Rather than the original phenotypes we test the multivariate (log-)Normal permutation procedure to calculate the MWSL using various synthetic outcomes which represent a wide range of real-data phenotypes. In particular, we employ a continuous outcome from a Normal distribution, a discrete-binary outcome from a Binomial distribution, a discrete-count outcome from a Poisson distribution, and a time-to-event survival outcome from the Cox proportional hazards model as in [21]. We benchmark our result on the MESA BINNED data but also on a set of synthetic variates obtained via a nonparametric approach using PCA (see Section 7.2). We divide the data into test and non-test sets, compute a PCA model of the non-test data, and predict the test data based on this model. This approach allowed us to generate synthetic data based on the structure of the real data without involving bootstrap/permutation methods [22] which we already employ to estimate the MWSL. Following the algorithm of Section 7.2 applied to the MESA BINNED data, we define the *test* and the *nontest* set by randomly sampling $n_t = 1,500$ and $n_{\bar{t}} = 3,500 - 1,500 = 2,000$ observations, respectively. From the PCA on the *nontest* set we select 350 PCs to be used to build the simulated *test* set of molecular variates $\hat{X}_t$. Table 7 and Table 8 confirm that the MWSL procedure effectively controls the FWER at around the (default) α-level of 5%.

6 Conclusions

In this paper we focus on assessing univariate test significance in multi collinear omics data by estimating a significance level threshold controlling the family wise error rate. The proposed procedure is based on an iterative permutation approach via univariate regression models while other measure of association may be used when appropriate. The molecular variates are simulated via parametric methods such as multivariate Normal and multivariate log-Normal distributions to retain the correlation structure in the data, while controlling the false positive rate at the desired level. When the permutation procedure is applied to the approximated data the MWSL is stable across outcome measures with diverse properties. In MWAS, the metabolic profiles often exhibit a high degree of collinearity, and this is supported by our finding that in all scenarios considered, when parametric methods are applied to approximate the structure of the data, the MWSL estimated through the permutation procedure is larger than the threshold obtained via a metabolome-wide Bonferroni or Sidak corrections. Therefore, the corresponding ENT is always less than the actual number of tests as it mainly depends on the extent of correlation within the data. The extent of collinearity is summarized by the R ratio (%) of effective to actual number of tests. For the examples in this paper, R was found to be around 50% for the CPMG data (high-resolution and and BINNED version), and around 9% for the NOESY resolution. This is consistent with the expected higher degree of correlations between spectral variables in the NOESY data. As with other approaches, the proposed closed-form Meff approximation to the permutation-based ENT could be tentatively interpreted as the number of independent metabolic processes exhibited by the system. Both the MWSL or the Meff estimate can be employed downstream of the analysis to identify differentially regulated metabolites.

7 Appendix

7.1 Parametric algorithm to generate synthetic variates

- Step (1): Generate a square $(M \times M)$ correlation matrix A assuming all variables have unit variance, i.e. the M elements on the diagonal are 1s. The $[M(M-1)]/2$ elements of the upper triangular matrix are sampled from Uniform distributions bounded by a certain interval, e.g. high correlation level within the interval $[0.75, 0.85]$, medium correlation in $[0.45, 0.55]$, or low correlations in $[0.25, 0.35]$. The lower triangle elements are copied from the upper triangle.
- Step (2): As the λs of A are required to be greater than zero, compute S as the nearest pos-

itive definite to the correlation matrix A achieving $\{\min \|A - S\|_F : S$ is a correlation matrix$\}$, where $\|A\|_F^2 = \sum_{i,j} a_{ij}^2$ as described by [23].

- Step (3): Derive the lower triangular L via Cholesky decomposition of matrix S such that $S = LL'$.

- Step (4): M multivariate Normal features with zero means result from the product ZL between the $(n \times M)$ matrix Z of M random N(0,1) i.i.d. features, and the $(M \times M)$ lower triangular matrix L. The correlations of the simulated features are very close to those assigned in matrix A.

7.2 Nonparametric algorithm to generate synthetic variates

- Step (1): By randomly sampling n_t observations from the original data matrix of variates X, construct the $(n_t \times M)$ *test* set of variates X_t, and the $(n_{\bar{t}} \times M)$ *nontest* set $X_{\bar{t}}$, with $n_t < n$ and $n_{\bar{t}} = n - n_t$.

- Step (2): Standardise the *test* and the *nontest* set by subtracting its respective vector of column means i.e. μ_t and $\mu_{\bar{t}}$, and dividing by its standard deviations i.e. σ_t and $\sigma_{\bar{t}}$, to respectively obtain Z_t and $Z_{\bar{t}}$.

- Step (3): Compute PCA over the *nontest* set by applying singular value decomposition (SVD) such that $Z_{\bar{t}} = U_{\bar{t}}\Sigma_{\bar{t}}V_{\bar{t}}^T$, where $V_{\bar{t}}^T$ is the $(M \times M)$ matrix of loadings, while the PC scores are obtained as the product between the $(n_{\bar{t}} \times n_{\bar{t}})$ matrix $U_{\bar{t}}$ of eigenvectors of $Z_{\bar{t}}Z_{\bar{t}}^T$, and the $(n_{\bar{t}} \times M)$ diagonal matrix $\Sigma_{\bar{t}}$.

- Step (4): Use the *nontest* loadings $V_{\bar{t}}$ combined with the test X_t to compute the $(n_t \times M)$ matrix $\hat{U}_t\hat{\Sigma}_t$ of PC predicted scores for the *test* set.

- Step (5): Build the $(n_t \times M)$ simulated *test* set of variates $\hat{Z}_t$ as the product of the predicted scores from Step (4), and the matrix of loadings $V_{\bar{t}}^T$ from Step (3) such that $\hat{Z}_t = \hat{U}_t\hat{\Sigma}_t V_{\bar{t}}^T$. We note that S PCs, with $S \leq M$, can be selected to be used in the predictions, thus $\hat{Z}_t$ would result from the product of the $(n_t \times S)$ matrix of PCs and the $(S \times M)$ matrix of loadings.

- Step (6): From the simulated *test* set of standardised features $\hat{Z}_t$ compute the $(M \times M)$ set of simulated features as $\hat{X}_t = \hat{Z}_t\sigma_t + \mu_t$.

To simulate the set of variates in such way the sample size of the data should be large enough for the data to be split between the test and the nontest set, and no missing values are allowed. Nevertheless, a possible extension of this method would consider the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm as a modified PCA to accommodate missing values [24].

Author details

[1]Faculty of Medicine, Department of Surgery & Cancer, Imperial College London, South Kensington Campus, SW7 2AZ London, UK. [2]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, CB2 1EW London, UK.

References

1. Holmes, E., Loo, R.L., Stamler, J., Bictash, M., Yap, I.K., Chan, Q., Ebbels, T., De Iorio, M., Brown, I.J., Veselkov, K.A., *et al.*: Human metabolic phenotype diversity and its association with diet and blood pressure. Nature **453**(7193), 396 (2008)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
3. Benjamini, Y., Yekutieli, D., *et al.*: The control of the false discovery rate in multiple testing under dependency. The annals of statistics **29**(4), 1165–1188 (2001)
4. Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze **8**, 3–62 (1936)
5. Šidák, Z.: Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association **62**(318), 626–633 (1967)
6. Cheverud, J.M.: A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**(1), 52 (2001)
7. Nyholt, D.R.: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics **74**(4), 765–769 (2004)
8. Li, J., Ji, L.: Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity **95**(3), 221 (2005)
9. Gao, X., Starmer, J., Martin, E.R.: A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic epidemiology **32**(4), 361–369 (2008)
10. Galwey, N.W.: A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society **33**(7), 559–568 (2009)
11. Paparoditis, E., Politis, D.N.: The local bootstrap for periodogram statistics. Journal of Time Series Analysis **20**(2), 193–222 (1999)
12. Hoggart, C.J., Clark, T.G., De Iorio, M., Whittaker, J.C., Balding, D.J.: Genome-wide significance for dense snp and resequencing data. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society **32**(2), 179–185 (2008)
13. Chadeau-Hyam, M., Ebbels, T.M., Brown, I.J., Chan, Q., Stamler, J., Huang, C.C., Daviglus, M.L., Ueshima, H., Zhao, L., Holmes, E., *et al.*: Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. Journal of proteome research **9**(9), 4620–4627 (2010)
14. Castagné, R., Boulangé, C.L., Karaman, I., Campanella, G., Santos Ferreira, D.L., Kaluarachchi, M.R., Lehne, B., Moayyeri, A., Lewis, M.R., Spagou, K., *et al.*: Improving visualization and interpretation of metabolome-wide association studies: An application in a population-based cohort using untargeted 1H-NMR metabolic profiling. Journal of proteome research **16**(10), 3623–3633 (2017)
15. Box, G.E., Watson, G.S.: Robustness to non-normality of regression tests. Biometrika **49**(1-2), 93–106 (1962)
16. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology **4**(1) (2005)
17. Cheverud, J.M., Rutledge, J., Atchley, W.R.: Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. Evolution **37**(5), 895–905 (1983)
18. Friedman, S., Weisberg, H.F.: Interpreting the first eigenvalue of a correlation matrix. Educational and Psychological Measurement **41**(1), 11–21 (1981)
19. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., JacobsJr, D.R., Kronmal, R., Liu, K., *et al.*: Multi-ethnic study of atherosclerosis: objectives and design. American journal of epidemiology **156**(9), 871–881 (2002)
20. Karaman, I., Ferreira, D.L., Boulangé, C.L., Kaluarachchi, M.R., Herrington, D., Dona, A.C., Castagné, R., Moayyeri, A., Lehne, B., Loh, M., *et al.*: Workflow for integrated processing of multicohort untargeted 1h nmr metabolomics data in large-scale metabolic epidemiology. Journal of proteome research **15**(12), 4188–4194 (2016)
21. Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate cox proportional hazards models. Statistics in medicine **24**(11), 1713–1723 (2005)
22. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications (1970)
23. Higham, N.J.: Computing the nearest correlation matrix—a problem from finance. IMA journal of Numerical Analysis **22**(3), 329–343 (2002)
24. Martens, H., Martens, M.: Multivariate analysis of quality. An introduction. IOP Publishing (2001)
25. Horizon2020, E.C.: PhenoMeNal (Phenome and Metabolome aNalysis): Large-scale Computing for Medical Metabolomics (2015-2018). https://phenomenal-h2020.eu/

Tables

Table 1 Descriptive statistics for the clinical outcome measures.

Table 2 Descriptive statistics for the fixed effects covariates.

Table 3 Real data: Comparison of estimation of the number of non-redundant variates from the permutation method (ENT, obtained as the average of the ENT estimates for all the clinical outcomes measures considered via the multivariate Normal and the multivariate log-Normal methods) and via approximation procedures based on the eigenvalues of the correlation matrix of the metabolite concentrations (Meff). R=ENT/ANT(%) ratio in brackets. ANT=655 for the MESA BINNED data, and ANT=30590 for the NOESY and CPMG data. Meff estimates closest to ENT estimates in bold.

Table 4 Real data: MWSL estimation comparison between the permutation method and the approximation procedure generating the distribution of the minimum p-value as a Beta(1,$\mathrm{Meff}_{\mathrm{MWSL}}$).

Table 5 Simulated data: Comparison of estimation of the number of non-redundant variates from the permutation method (ENT, obtained as the average of the ENT estimates for all the simulated uncorrelated and correlated outcomes measures considered via the multivariate Normal and the multivariate log-Normal methods) and via the approximation procedure based on the eigenvalues of the correlation matrix of the metabolite concentrations (Meff). R=ENT/ANT(%) ratio in brackets. ANT=650. Meff estimates closest to ENT estimates in bold.

Table 6 Simulated data: MWSL estimation comparison between the permutation method and the approximation procedure generating the distribution of the minimum p-value as a Beta(1,$\mathrm{Meff}_{\mathrm{MWSL}}$).

Table 7 BINNED data: ENT estimates with 95% confidence intervals in brackets, and type I error estimation from the permutation procedure for dvarious simulated outcome measures: continuous, discrete-binary, discrete-count, time-to-event survival. 5,000 permutations. ANT=655.

Table 8 PCA simulated data (ANT=655, n_t=1,500, PCs=350): ENT estimates with 95% confidence intervals in brackets, and type I error estimation from the permutation procedure for various simulated outcome measures: continuous, discrete-binary, discrete-count, time-to-event survival. K=5,000 permutations.

Figures

> **Figure 1** BINNED data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

> **Figure 2** CPMG data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

> **Figure 3** NOESY data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

> **Figure 4** ENT for uncorrelated outcomes across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.

> **Figure 5** ENT for correlated outcome across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.

Additional Files
Additional file 1 — Supplementary results including Figures, Tables, and the description of the simulation algorithms used to generate the synthetic variates.
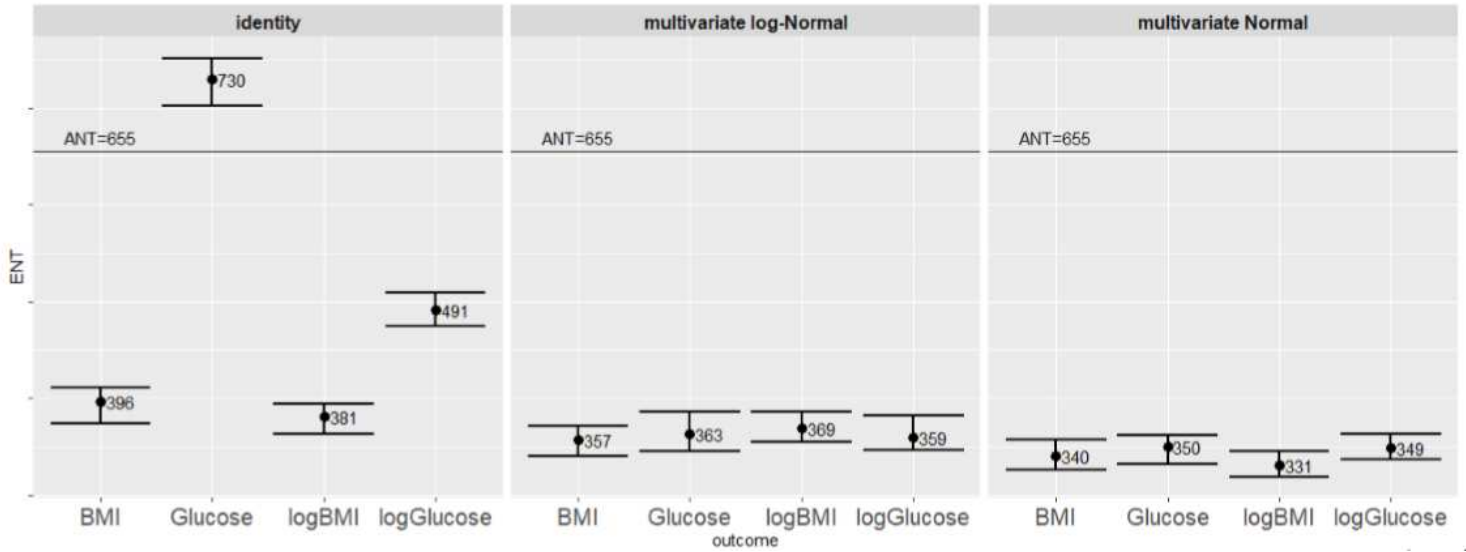Additional file 2 — R tutorial to enhance the application of the MWSL R package.

Figures

Figure 1

BINNED data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.
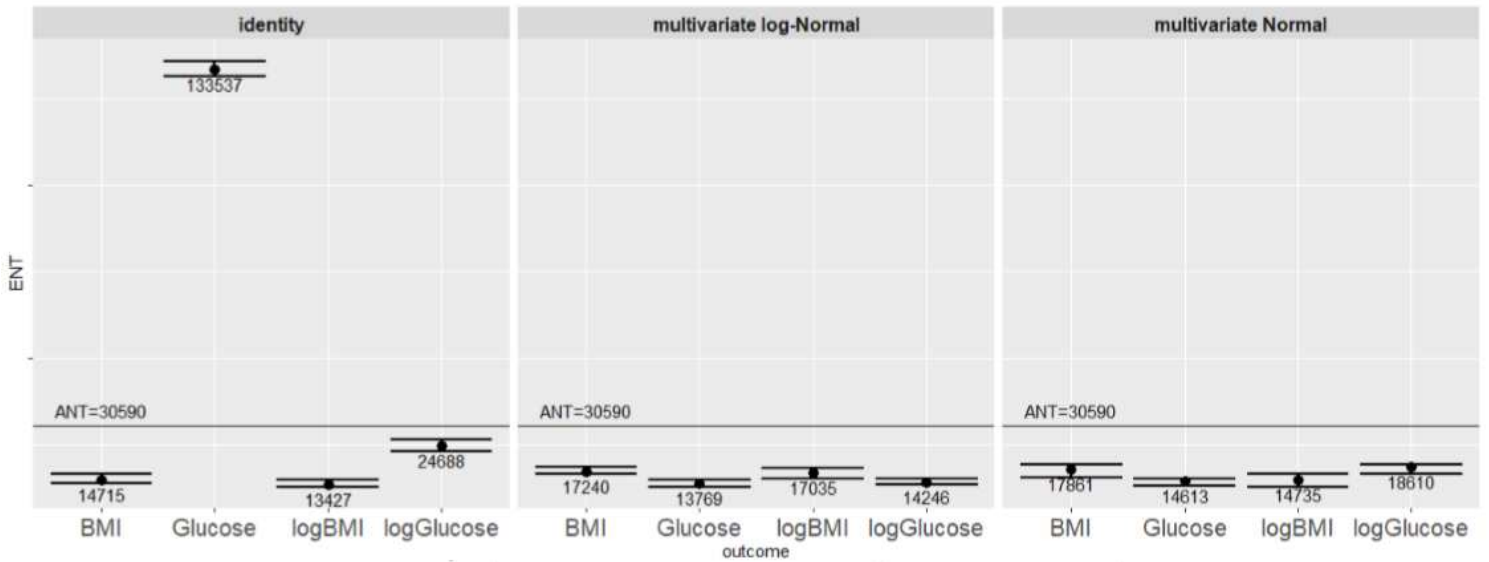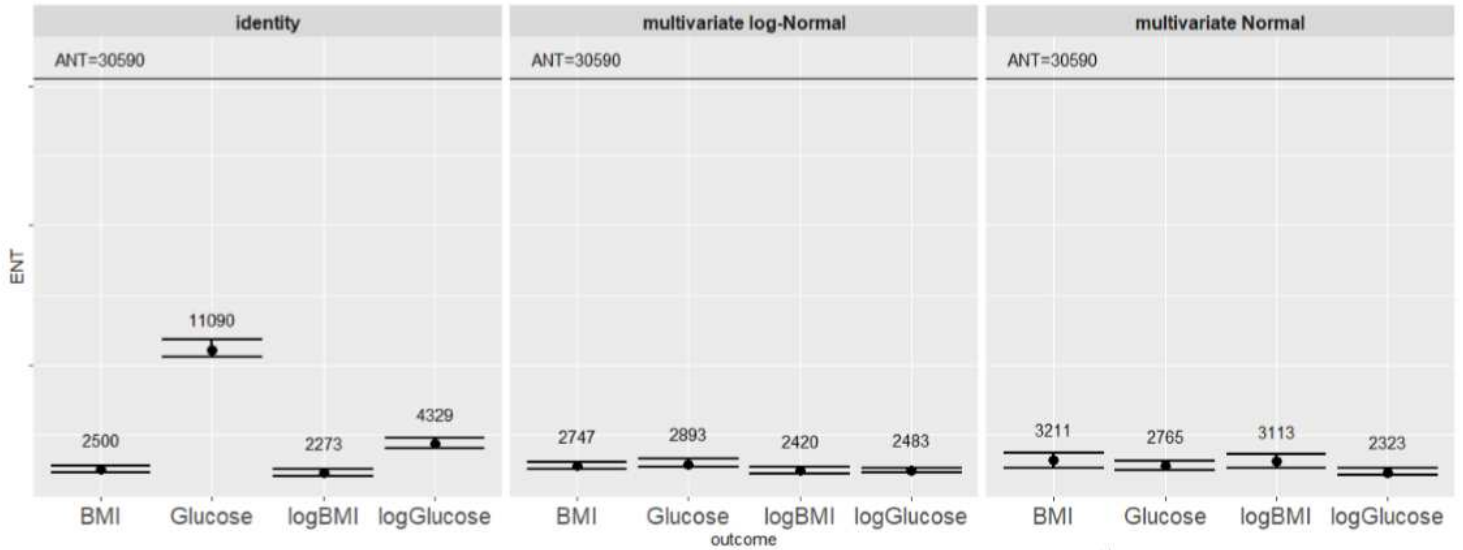

Figure 2

CPMG data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

Figure 3

NOESY data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.
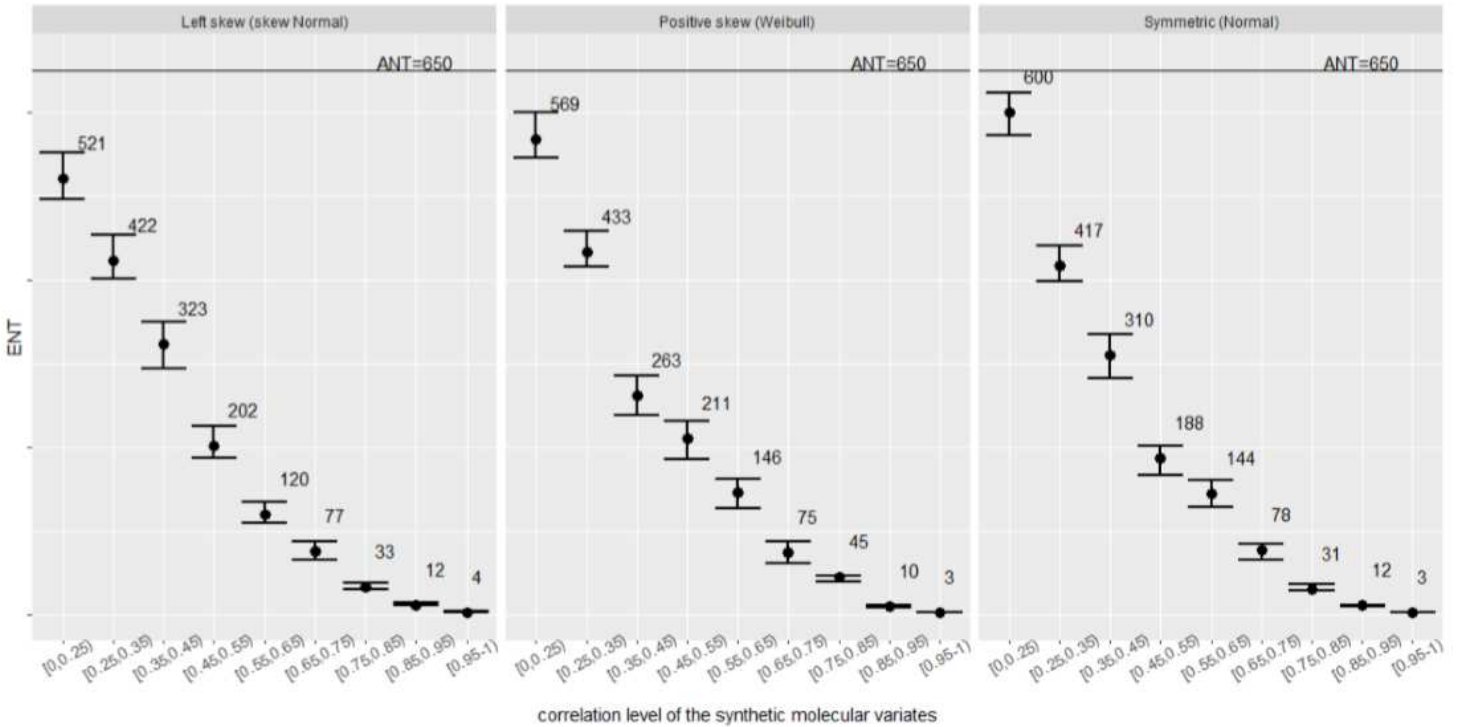

Figure 4

ENT for uncorrelated outcomes across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.
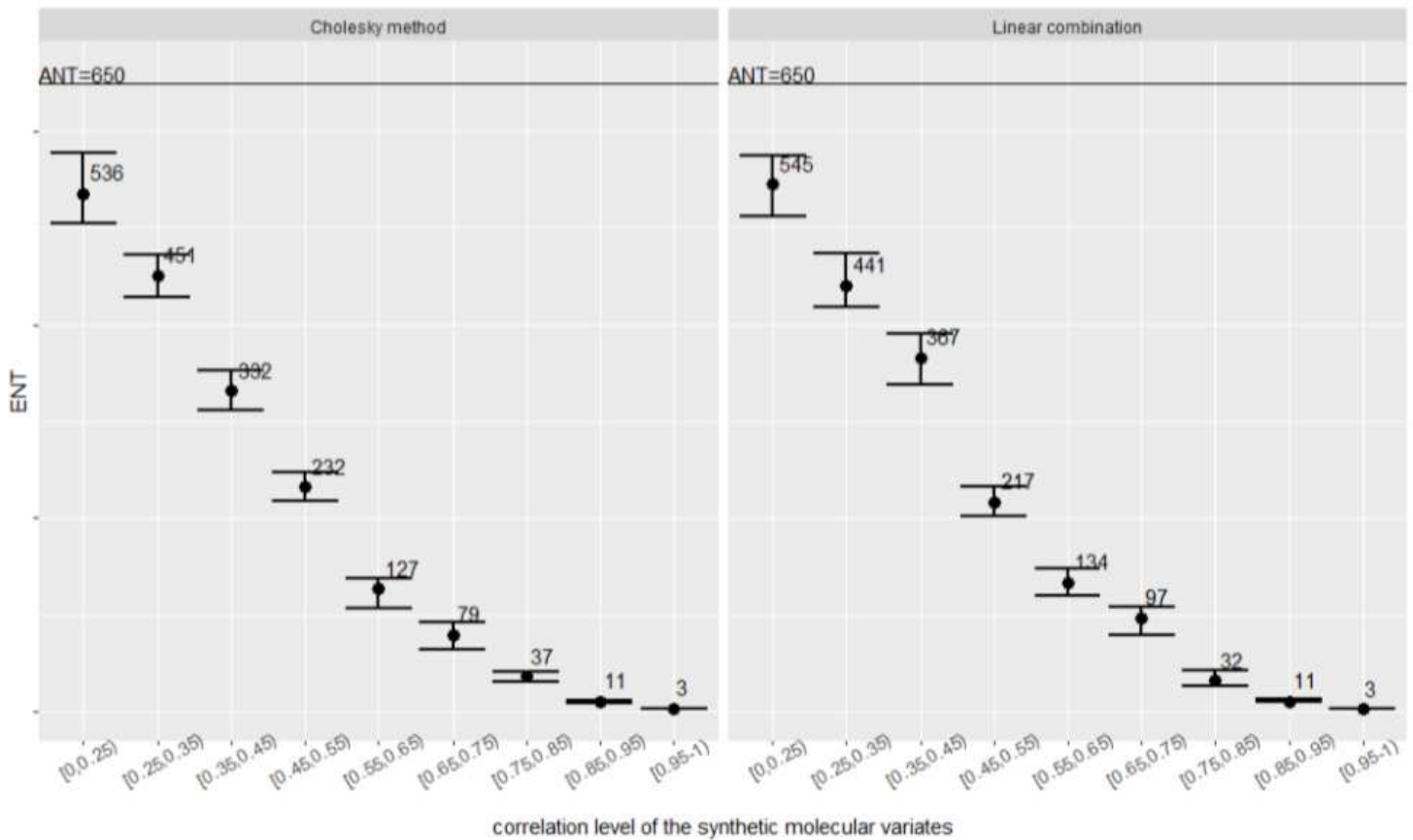
Figure 5

ENT for correlated outcome across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- peluso2020mwslsupp1.pdf
- RtutorialMWSLsupp2.pdf