



HAL
open science

MANOCCA: a robust and computationally efficient test of covariance in high-dimension multivariate omics data

Christophe Boetto, Arthur Frouin, Léo Henches, Antoine Auvergne, Yuka Suzuki, Etienne Patin, Marius Bredon, Alec Chiu, Sriram Sankararaman, Noah Zaitlen, et al.

► To cite this version:

Christophe Boetto, Arthur Frouin, Léo Henches, Antoine Auvergne, Yuka Suzuki, et al.. MANOCCA: a robust and computationally efficient test of covariance in high-dimension multivariate omics data. Briefings in Bioinformatics, 2024, 25 (4), pp.bbae272. 10.1093/bib/bbae272 . pasteur-04646136

HAL Id: pasteur-04646136

<https://pasteur.hal.science/pasteur-04646136>

Submitted on 12 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

MANOCCA: a robust and computationally efficient test of covariance in high-dimension multivariate omics data

Christophe Boetto^{1,*}, Arthur Frouin¹, Léo Henches¹, Antoine Auvergne¹, Yuka Suzuki¹, Etienne Patin², Marius Bredon³, Alec Chiu⁴, Milieu Interieur Consortium, Sriram Sankararaman⁴, Noah Zaitlen⁴, Sean P. Kennedy¹, Lluís Quintana-Murci^{2,5}, Darragh Duffy⁶, Harry Sokol^{3,7,8,9}, Hugues Aschard^{1,10}

¹Department of Computational Biology, Institut Pasteur, Université Paris Cité, 25-28 rue du Dr Roux, 75015 Paris, France

²Human Evolutionary Genetics Unit, Institut Pasteur, Université Paris Cité, CNRS UMR2000, 25-28 rue Dr Roux, 75015 Paris, France

³Sorbonne Université, INSERM, Centre de recherche Saint-Antoine, CRSA, Microbiota, Gut and Inflammation Laboratory, Hôpital Saint-Antoine (UMR S938) Sorbonne Université, 27 rue Chaligny, 75012 Paris, France

⁴Department of Human Genetics, University California Los Angeles, 695 Charles E. Young Drive South, Box 708822, Los Angeles, CA 90095-7088, United States

⁵Chair of Human Genomics and Evolution, Collège de France, 11 Pl. Marcelin Berthelot, 75005 Paris, France

⁶Translational Immunology Unit, Institut Pasteur, Université de Paris Cité, 25-28 rue du Dr Roux, 75015 Paris, France

⁷Paris Center for Microbiome Medicine, Fédération Hospitalo-Universitaire, 184 rue du Faubourg Saint-Antoine, 75571 PARIS Cedex 12, France

⁸Gastroenterology Department, AP-HP, Saint Antoine Hospital, 184 rue du faubourg Saint-Antoine, 75012 Paris, France

⁹INRAE Micalis & AgroParisTech, UMR1319, Micalis & AgroParisTech, 4 avenue Jean Jaurès, 78352 Jouy en Josas, France

¹⁰Department of Epidemiology, Harvard TH Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, United States

*Corresponding author. Department of Computational Biology, Institut Pasteur, Université Paris Cité, 25-28 rue du Dr Roux, 75015 Paris, France.

E-mail: boetto.christophe@gmail.com

Abstract

Multivariate analysis is becoming central in studies investigating high-throughput molecular data, yet, some important features of these data are seldom explored. Here, we present MANOCCA (Multivariate Analysis of Conditional Covariance), a powerful method to test for the effect of a predictor on the covariance matrix of a multivariate outcome. The proposed test is by construction orthogonal to tests based on the mean and variance and is able to capture effects that are missed by both approaches. We first compare the performances of MANOCCA with existing correlation-based methods and show that MANOCCA is the only test correctly calibrated in simulation mimicking omics data. We then investigate the impact of reducing the dimensionality of the data using principal component analysis when the sample size is smaller than the number of pairwise covariance terms analysed. We show that, in many realistic scenarios, the maximum power can be achieved with a limited number of components. Finally, we apply MANOCCA to 1000 healthy individuals from the Milieu Interieur cohort, to assess the effect of health, lifestyle and genetic factors on the covariance of two sets of phenotypes, blood biomarkers and flow cytometry-based immune phenotypes. Our analyses identify significant associations between multiple factors and the covariance of both omics data.

Keywords: statistics; covariance; correlation; multivariate analysis; biomarkers

Introduction

Human cohorts commonly collect high-dimensional phenotypic data, including high-throughput omics, extended medical information and biomarkers [1, 2]. A variety of multivariate approaches have been developed to leverage this wealth of data [3–6]. The

joint analysis of multiple outcomes can increase statistical power to detect associations [7, 8], help in deciphering complex biological processes through clustering approaches [9] or improve the prediction accuracy of an outcome of interest [10]. Regarding association testing, existing methods and application have mostly

Christophe Boetto is a PhD student in the field of biostatistics.

Arthur Frouin is a postdoc working on bacterial GWAS.

Léo Henches is a research engineer in the field of statistical genetics.

Antoine Auvergne is a PhD student working on multivariate analysis related to the genetics of psychiatric disorders.

Yuka Suzuki is a postdoc in the field of biostatistics.

Etienne Patin is a research scientist in the field of human evolutionary genetics.

Marius Bredon is a postdoc in the field of metagenomics.

Alec Chiu is a PhD student in the field of biostatistics.

Sriram Sankararaman is a professor in computer science.

Noah Zaitlen is a professor in machine learning.

Sean P. Kennedy is a researcher in the field of clinical metagenomics.

Lluís Quintana-Murci is a professor in human evolutionary genetics.

Darragh Duffy is a research director in the field of translational Immunology.

Harry Sokol is a professor specialized on the human gut microbiome.

Hugues Aschard is a research director in the field of statistical genetics.

Received: November 16, 2023. Revised: April 16, 2024. Accepted: May 28, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

focused on testing the impact of predictors of interest on the mean of a multivariate outcome, typically using a composite null hypothesis such as implemented in a multivariate ANOVA (MANOVA). Conversely, methods to investigate other components of multivariate outcomes remain sparse. One of such components of multivariate outcomes is correlation, which is commonly present in omics data. Although methods to investigate predictors associated with the correlation between multiple outcomes exist [11–14], their performance and robustness have not been assessed, and their efficiency in large-scale agnostic screenings remains unknown. Moreover, they carry substantial inherent limitations, including restriction to binary factors and no adjustment for covariates.

Here, we present a new approach, named MANOCCA (Multivariate ANalysis Of Conditional Covariance), that enables the identification of both categorical and continuous predictors associated with changes in the covariance matrix of a multivariate outcome while allowing for covariates adjustment. We first introduce the key principles and the main characteristics of the approach and demonstrate that, in most realistic scenarios, MANOCCA can outperform existing approaches showing stronger power and robustness. We then describe the challenges faced when analysing high-dimensional data and present a robust solution based on principal components analysis (PCA). We next investigate the power of MANOCCA conditional on alternative parametrizations, providing guidelines for real data application across various settings. Finally, we illustrate the method by studying health, lifestyle and genetic factors associated with variability of blood biomarkers and flow cytometry-based immune phenotypes using data from 1000 healthy subjects from the Milieu Intérieur (MI) cohort.

Methods

The MANOCCA approach

Previous work [15] showed that variability in the correlation between two standardized outcomes Y_1 and Y_2 can be investigated through the element-wise product of those outcomes. The Pearson correlation coefficient between Y_1 and Y_2 is expressed as $\rho_{Y_1 Y_2} = \text{cov}(Y_1, Y_2) / (\sigma_{Y_1} \sigma_{Y_2})$, with $\text{cov}(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1] \mathbb{E}[Y_2]$. For standardized outcomes and a sample size N , it can be re-expressed as the average of the element-wise product across individuals: $\rho_{Y_1 Y_2} = (\sum_{i=1 \dots N} Y_{1i} Y_{2i}) / N$. It follows that the effect of a predictor X on $\text{cor}(Y_1, Y_2)$ can be tested using a standard least-squares regression framework where X is treated as a predictor and the product $Y_1 Y_2$ as the outcome. One can easily demonstrate that, under reasonable assumptions, this test is independent of mean and variance effect. Consider the following models: $Y_1 = \alpha_1 U + \beta_1 X + \epsilon_1$ and $Y_2 = \alpha_2 U + \beta_2 X + \epsilon_2$, where Y_1 and Y_2 are random variables correlated through an unmeasured normally distributed variable U , and depend linearly on a binary predictor X , inducing an effect on the means of Y_1 and Y_2 . The conditional covariance between Y_1 and Y_2 can be expressed as $\text{cov}(Y_1, Y_2 | X = x) = \alpha_1 \alpha_2 \text{var}(U)$ and does not depend on X . Consider the alternative models: $Y_1 = \alpha_1 U + \beta_1 A X + \epsilon_1$ and $Y_2 = \alpha_2 U + \epsilon_2$, where Y_1 and Y_2 are correlated, and the variance of Y_1 depends on the product of a latent continuous variable A multiplied by the binary predictor X . The conditional covariance can again be expressed as $\text{cov}(Y_1, Y_2 | X = x) = \alpha_1 \alpha_2 \text{var}(U)$ and does not depend on X . Finally, consider the models: $Y_1 = \alpha_1 U X + \epsilon_1$ and $Y_2 = \alpha_2 U + \epsilon_2$, where Y_1 and Y_2 are correlated, with the strength of the correlation depending on the predictor X . The covariance can now be expressed as $\text{cov}(Y_1, Y_2 | X = x) =$

$\alpha_1 \alpha_2 x \text{var}(U)$ and does depend on X . Further details on those approximations are provided in the [Supplementary Notes](#).

The approach can easily be extended to more than two outcomes by deriving an $N \times p$ matrix of products between centered outcomes, defined as $\mathbf{P} = P_1, \dots, P_p$, with p , the number of products, equals $k(k-1)/2$ where k is the number of outcomes and N the sample size. The association between a predictor X and \mathbf{P} can then be derived by applying a standard two-way analysis of variance (MANOVA), that is $\mathbf{P} \sim \delta X$. While valid, this approach is limited to situations where the effective sample size N is substantially larger than the number of products p . When this criterion is not met, we use PC analysis to reduce the dimension of the product matrix and use the top m PCs to form an $N \times m$ matrix Ω used as input in our test. Given the independence between the PCs, we first considered using a sum of univariate PC tests to form a joint test; however, this approach was not calibrated (see [Results](#) and [Figures S1](#) and [S2](#)). Instead, we used a MANOVA, that is, $\Omega \sim \beta X$. For fast computation, the joint effect estimates $\hat{\beta} = \hat{\beta}_1 \dots \hat{\beta}_m$ of association between each PC $\Omega_i \in [1 \dots m]$ and the predictor X are first derived using a single matrix operation: $\hat{\beta} = (X^T X)^{-1} X^T \Omega$. The Wilks' lambda statistics, $W = \det(\Omega^T \Omega - \beta X^T X \beta^T) / \det(\Omega^T \Omega)$ is derived in a second step. Under the null hypothesis of no association, W follows a Fisher distribution $F(m, N - m - 1)$. [Figure 1](#) presents an overview of the steps for applying the approach. Note that the Wilk's Lambda statistics can be sensitive to the normality assumption of the outcome data analysed [16]. However, as discussed further in the [Results](#) section, we implemented in the final version of MANOCCA a systematic rank-inverse normal transformation of the input data ensuring that the normality assumption is met.

In a standard MANOVA, potential confounding factors $\mathbf{C} = (C_1 \dots C_c)$ can be incorporated as covariate: $\Omega \sim \mathbf{C} + \beta X$. Again, for fast computation, we used a two-step procedure that consists in adjusting a priori both the outcome and the predictor for the covariates: $\Omega^* = \Omega - \sum_c \hat{\gamma}_c C_i$, where $\hat{\gamma}_i$ is a vector of estimated effect of C_i on Ω , and $X^* = X - \sum_c \hat{\tau}_c C_i$, where $\hat{\tau}_i$ is the estimated effect of C_i on X , and applying the MANOVA on the residual variables: $\Omega^* \sim \beta X^*$.

Type I error rate simulation

The statistical robustness of any covariance test likely depends on the distributional assumptions of the multivariate outcome studied, including the distribution of the outcomes, the number of outcomes analysed jointly, the strength of the correlation across outcomes, the sample size over the number of outcomes ratio and the distribution of the predictors considered (binary or continuous, etc.). We investigated the validity of both MANOCCA and other existing approaches conditional on these factors.

We first assess the calibration under the null of MANOCCA and four existing approaches, the Mantel test [11], the Fisher method [12], the Jennrich test [13] and the BoxM test [14], using fully simulated data and simple scenarios ([Fig. 2](#)). We drew a series of 10 000 replicates with a sample size of 1000, each including a multivariate outcome \mathbf{Y} and a binary predictor $X \sim B(0.4)$ drawn independently of \mathbf{Y} under two different models. Note that we used a binary predictor as the four existing approaches do not allow for the analysis of continuous predictors. In the first model, replicates included five outcomes drawn from a multivariate normal with modest pairwise correlation. In the second model, replicates included 30 highly correlated non-normal outcomes drawn from a multivariate chi-square distribution. The overall calibrations of all tests were derived by testing for association between X

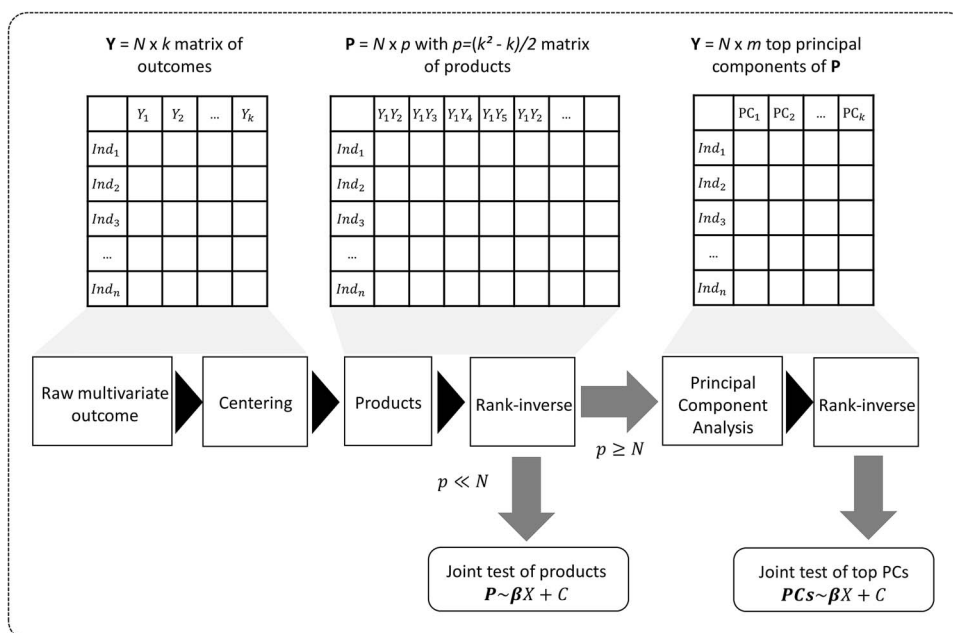


Figure 1. Overview of the MANOCCA approach. Starting with a multivariate outcome matrix of N samples and k variables, the data are first centred. The pairwise product of each of the k outcomes is computed, generating a high-dimensional matrix of size $N \times k(k-1)/2$. If $N \gg k(k-1)/2$, a joint test of all products can be derived; otherwise, the dimension of the product matrix is reduced using a PCA, to form a PC space of size $N \times m$. The final test, including covariates, can be performed on the products or the top m PCs using a Wilk's lambda test.

and the correlation between Y variables and conducting a visual inspection of the P -value distribution.

We next assessed the robustness of MANOCCA under a wider range of scenarios, while modifying some of the modelling parameters. We first compared performance when using a binary or a continuous predictor (Figure S3 available online at <http://bib.oxfordjournals.org/>). We simulated a series of 100 replicates, each including 1000 individuals and a multivariate outcome Y including 400 variables drawn from a multivariate chi-square distribution with a point mass at 0 including 0–50% of the data. For each replicate, we drew 1000 predictors, either from a normal or binary distribution, and applied MANOCCA while varying the number of PCs used and applying no transformation, a rank-inverse normal transformation on the product matrix, the PC matrix or both. The validity of MANOCCA was assessed using a Kolmogorov–Smirnov test for deviation from a uniform [0,1] distribution of the P -values across the 1000 predictors tested. We then conducted simulations using real covariance matrices derived from the MI 169 flow cytometry–based variables and for ranging sample sizes from 1000 to 5000 to draw guidelines on the parametrization of MANOCCA (Figures S6 and S7). For predictors, we considered not only binary predictors with frequencies in [0.01; 0.40] but also categorical ones mimicking genetic variants with minor allele frequency in [0.01; 0.40], both generated independently of the multivariate outcome Y .

Power simulation

To investigate power, we drew a series of 50 replicates with sample size of 1000 including a binary exposure X with frequency of 0.5 and a multivariate outcome Y including 50–169 variables (Fig. 4). For each replicate, we used two covariance matrices, one for the exposed (C_1) and the other for the unexposed (C_2), and tested the association between X and Y using MANOCCA. We generated the outcome from a multivariate normal and real covariance matrix (C) derived from the MI flow cytometry data under three

scenarios. In scenario (i), $C_1 = C$ and $C_2 = |C|^\gamma \times \text{sign}(C)$, inducing a covariance with similar pattern among exposed but variability in the magnitude of covariance. In scenario (ii), $C_1 = C$ and $C_2 = \delta C + (1 - \delta) \Delta$, where Δ is a random covariance generated using the R *randcorr* package [17], thus inducing random noise between exposed and unexposed. In scenario (iii), we first drew $C_1 = C_2 = C$ and then attenuated the covariance in an arbitrary chosen subset ω of C_2 , so that $C_{2(\omega)} = |C_{(\omega)}|^\theta \times \text{sign}(C_{(\omega)})$. We arbitrarily set γ to 1.5, δ to 0.2 and θ to 0.5, as it allowed for a similar average power across scenarios given the other simulation parameters.

MANOCCA association screening in MI

The MI Consortium is a population-based cohort initiated in September 2012 [18]. It comprises 1000 healthy volunteers from western France, with a 1:1 sex ratio. The cohort collected a broad range of variables, including genomic, immunological, environmental and clinical outcomes. We conducted systematic MANOCCA screenings for environmental effects on the covariance of two sets of data: 169 flow cytometry–based immune cell phenotypes [19] and 33 health-related blood biomarkers, including 22 metabolites and 11 cell counts [18] (Tables S1 and S2 available online at <http://bib.oxfordjournals.org/>). We focused on two types of predictors: health and lifestyle factors collected from questionnaires and genome-wide variants. Health and lifestyle factors included demographics, medical and vaccination history, psychological traits, socio-professional information, smoking habits, physiological measurements and nutrition measured as part of the Nutrinet [20] study (Table S3 available online at <http://bib.oxfordjournals.org/>). After the filtering of ancestral outliers individuals [21, 22], the genetic screening was conducted in 894 participants for a total of 5 667 803 variants after filtering and imputation using IMPUTE2 [23]. Except when used as predictor, all analyses were adjusted for age, sex and body mass index (BMI). For blood metabolites, the number of products allowed for a

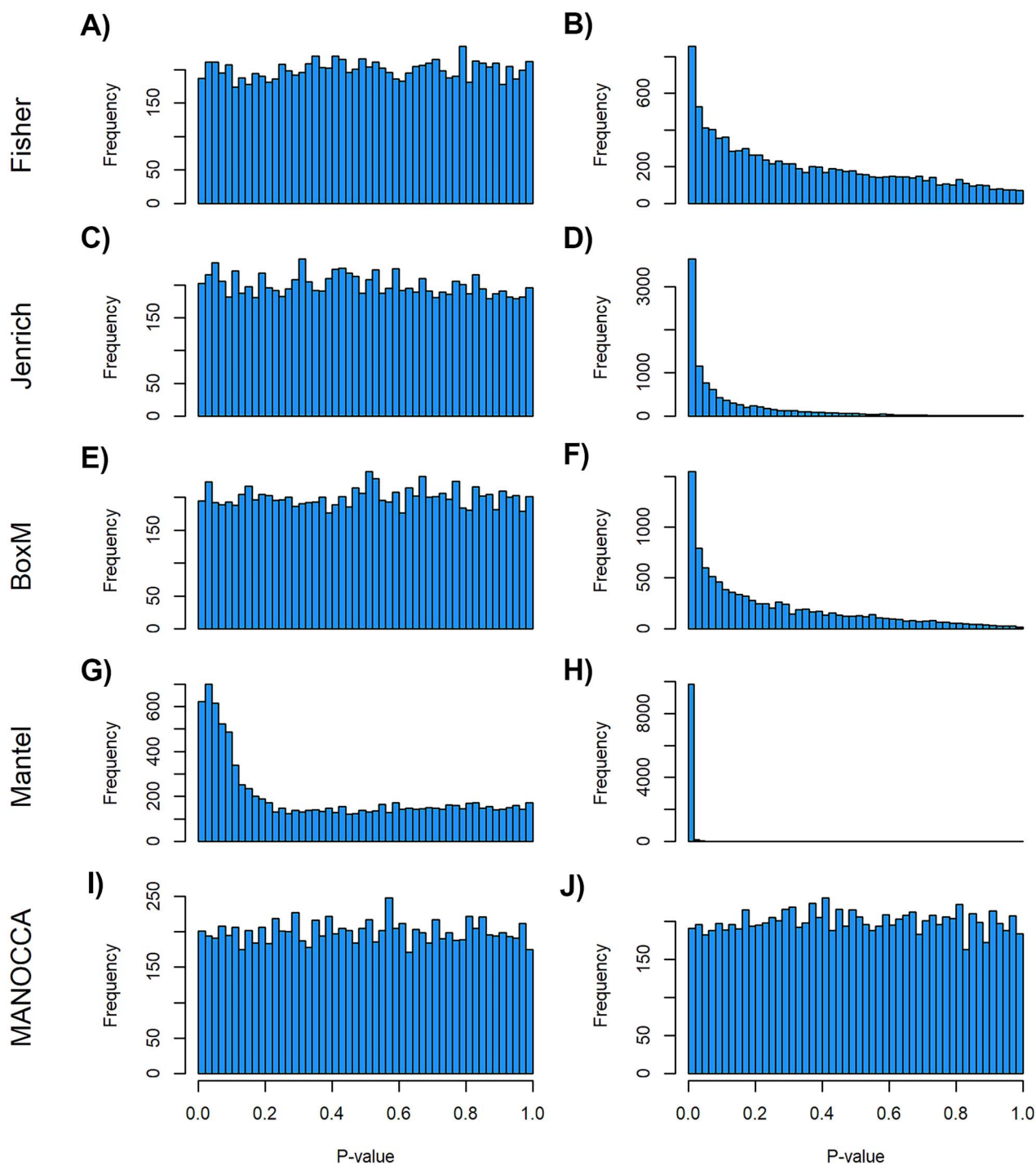


Figure 2. Limitation of existing methods. We assessed the calibration under the null of four approaches representing the state of the art for covariance matrix comparison: the Fisher method (A, B), the Jenrich test (C, D), BoxM (E, F) and Mantel test (G, H), against the proposed MANOCCA approach (I, J). Note that we applied MANOCCA directly on the product matrix thanks to the high sample size compared to the number of products. We simulated a series of 10 000 replicates, with a sample size of 1000 each, under two different null models. In the first model (A, C, E, G, I), replicates included five outcomes Y drawn from a multivariate normal with modest pairwise correlation. In the second model (B, D, F, H, J), closer to the expected distribution of omics data, replicates included 30 non-normal outcomes with high correlation. Calibration was derived by splitting each replicate in two random sets according to a random binary variable $X \sim B(0.4)$ and testing for association between X and the correlation between Y variables. The panels present the distribution of the P-values, expected to be uniformly distributed under this null model, for the five approaches and the two models.

direct analysis of the products without requiring the PCA step, and we considered both the products and the PCs as outcomes. For comparison purposes, we also conducted, for each screening, a standard MANOVA on the mean of the multivariate outcome (see [Supplementary Notes](#)).

Human samples

Samples came from the MI Cohort, which was approved by the Comité de Protection des Personnes—Ouest 6 (Committee for the protection of persons) on 13 June 2012 and by French Agence nationale de sécurité du médicament (ANSM) on 22 June

2012. The study is sponsored by Institut Pasteur (Pasteur ID-RCB Number: 2012-A00238-35) and was conducted as a single centre interventional study without an investigational product. The original protocol was registered under [ClinicalTrials.gov](https://clinicaltrials.gov/study/NCT01699893) (study# NCT01699893). The samples and data used in this study were formally established as the MI biocollection (NCT03905993), with approvals by the Comité de Protection des Personnes—Sud Méditerranée and the Commission nationale de l'informatique et des libertés (CNIL) on 11 April 2018. All donors gave written informed consent. All data used in this study are available at <https://dataset.owey.io/>.

Results

Method comparison and MANOCCA characteristics

We identified four existing approaches allowing to test for the effect of a predictor on the covariance matrix of a multivariate outcome: (i) the Mantel test [11], which consists of deriving a distance metric between two square matrices of the same dimension and comparing this distance to an empirical distribution derived through permutation; (ii) the Fisher method [12], which builds a statistic based on the sum of the squared correlations over all cells from the covariance matrix; (iii) the Jennrich test [13], which, in its simplest form, consists of estimating the statistic based on the Hadamard product of a given correlation matrix and the inverse of a second matrix of the same dimension; and (iv) the BoxM test [14], which extends the Levene's test of homogeneity of variance, an approach often used in human genetics [24]. Further description of each of the four approaches is provided in [Supplementary Notes](#). We conducted a series of simulations to assess their statistical robustness under the null using a binary predictor and no covariates, as these approaches cannot handle continuous predictors and do not allow adjustment for covariates. Except for the Mantel test, all methods performed relatively well for a simple model with a few normally distributed outcomes. Conversely, they all displayed severe type I error rate inflation when confronted with non-normal correlated variables, mimicking omics data (Fig. 2A–H). In comparison, when applied to the same simulated data, MANOCCA was correctly calibrated in all simulations (Figure 2I and J).

The effect of the predictor on the covariance, the mean and the variance of a set of outcomes are expected to be statistically independent (see [Supplementary Notes](#)). We confirmed this orthogonality between mean (derived using a two-way MANOVA), variance (Levene's test) and the proposed covariance tests through simulation. Figure 3 shows that, under realistic modelling assumptions, the MANOCCA test captures only effects on the covariance and can therefore identify effects missed by both mean and variance-based approaches. Figure 3D further illustrates bivariate data where a binary predictor X is associated with covariance but neither the mean nor the variance of the outcomes. Importantly, the independence of the three tests does not imply signal across the three approaches will necessarily be uncorrelated in real data. Indeed, one can easily draw scenarios with, e.g. effect of a predictor on both the mean and covariance of a multivariate outcome. Also, unless specified otherwise, we modelled the effect of a predictor on the covariance through an interaction with a latent variable associated shared across the outcomes tested (see [Methods](#)). Under this modelling, effects on the covariance can in general be transposed to effects on the correlation. However, when the predictor has an effect on the variance of either outcome, this equality is not valid anymore, as

the correlation will depend on X , while the covariance will not ([Supplementary Notes](#)).

Extension to high-dimension data

MANOCCA is readily applicable to the matrix of outcome product in all scenarios where p , the number of products, is substantially smaller than the sample size N . When this criterion is not met, we used PCA to reduce the dimension of P , using the top PCs as the primary outcome. As for all linear models, the maximum number of PCs that can be analysed jointly remained bounded by the sample size. More generally, high-dimension outcomes data bring the question of the latent space dimension, that is, the number of PCs kept in the analysis to achieve maximum power while maintaining a correct type I error rate. Moreover, by construction, both products and PCs tend to display kurtotic distributions, especially for omics-like data, which might also impact performance. We conducted a series of simulations to investigate the robustness of the MANOCCA conditional on these two components. Specifically, we measured the type I error rate while varying the number of top m PCs selected, and applying (i) no transformation, or a rank-inverse normal transformation on (ii) the product, (iii) the PCs or (iv) both products and PCs. As shown in [Figure S3a](#) available online at <http://bib.oxfordjournals.org/>, if the predictor being tested is continuous, MANOCCA shows strong robustness conditional on the outcome's distribution, remaining well calibrated regardless of the transformation applied and allowing for the use of a large number of PCs. Conversely, when analysing binary predictors, the test systematically requires a normalization of the PCs, only allowing a limited number of PCs to be analysed jointly ([Figure S3b](#) available online at <http://bib.oxfordjournals.org/>). [Figures S4](#) and [S5](#) available online at <http://bib.oxfordjournals.org/> further presents the results from extended simulations, providing guidelines to determine the number of PCs that can be analysed jointly conditionally on the predictor frequency and the cohort sample size. From an application perspective, when the predictors to be tested include both continuous and non-continuous variables, we recommend using a systematic rank-inverse normal transformation of the PCs, so that the results from either type of predictors can be interpreted on the same scale.

We next evaluated the power of MANOCCA across different scenarios in which the true covariance depends on a binary predictor X with a frequency of 0.5. We tested up to 400 PCs, and normalized either the products and PCs or the PCs only, the two transformations that display a calibrated null distribution ([Figure S3](#) available online at <http://bib.oxfordjournals.org/>). The optimal number of PCs varied substantially across the simulated scenarios. [Figure 4](#) illustrates three complementary cases. When X acts as a global scaling factor of the covariance, the maximum power is observed when using a limited number of top PCs and decreases after reaching that optimum ([Fig. 4A](#)). When X affects only a subset of outcomes, the maximum power is reached when including a fairly large number of PCs and converge afterwards ([Fig. 4B](#)). As expected, when X induces random noise in the covariance matrix, the power increases continuously with the number of PCs ([Fig. 4C](#)). Among the scenarios we considered, the double normalization (on products and PCs) produced on average larger power, and this transformation was therefore used in all subsequent analyses.

Efficient implementation

The MANOCCA approach requires multiple steps that can be computationally expensive in large-scale data. The main limiting step is the computation of the product matrix followed by the PCA

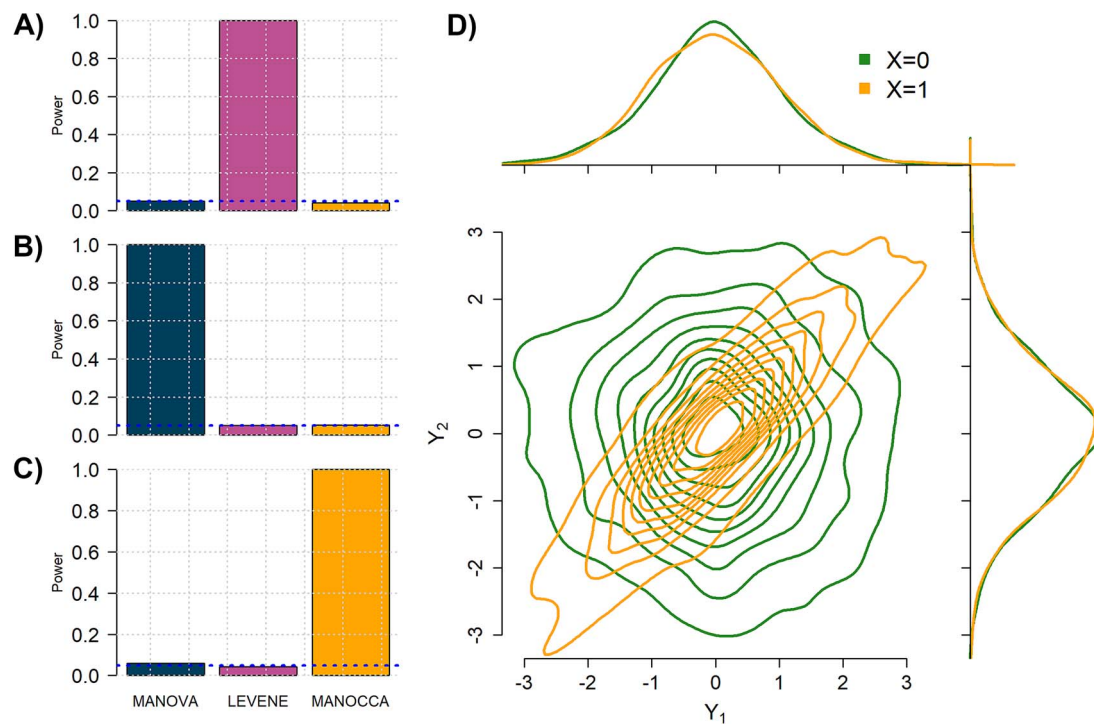


Figure 3. Orthogonality between MANOVCA and other tests. We simulated a series of datasets under three models where a binary predictor X influences orthogonally either the mean, the variance or the covariance of a bivariate outcome Y . In model (A), each outcome Y_i is drawn from a standard additive model: $Y_i = U + \beta_i X + \varepsilon_i$, where U is a normally distributed variable shared across Y and ε_i are independent normal residuals. In model (B), each Y_i is drawn from $Y_i = U + \gamma_i A_i X + \varepsilon_i$, where A_i is normally distributed variables producing heterogeneity in the variance of Y_i conditional on X . In model (C), each Y_i is drawn from the interaction model $Y_i = \delta_i U X + \varepsilon_i$, which produces heterogeneity in the correlation across Y_i conditional on X . For each model, we derived the power at the P -value threshold of 0.05 for a joint mean effect test (MANOVA), a test of variance for a randomly selected Y_i (LEVENE) and the proposed covariance test (MANOVCA). The parameters β_i , γ_i and δ_i were chosen to maximize the power of the at least one of the three tests. The dashed line indicates the P -value threshold of 0.05. (D) shows an example of a bivariate distribution where X is not associated with the mean and variance of the two outcomes but with their covariance.

transformation, but it is a one-time cost regardless of the number of predictors tested. With N the sample size, k the number of outcomes and q the number of predictors to test, the computation time is divided in $O(Nk^2)$ for the computation of the product matrix, $O(\max(N, k^2)^2 * \min(N, k^2))$ for the computation of the PCA and $O(Nq)$ for the test of q predictors (Fig. 5). Most steps were implemented with limited usage to exterior libraries, but ground-proofed against multiple existing tools. The approach is implemented in a Python package with dependencies to *numpy*, *scipy* for the fisher distribution, *scikit-learn* for the PCA and *pandas* for *dataframe* integration, but all computations are performed under *numpy* array to increase performance. Each step was optimized to minimize computational time and, given that most steps are independent, especially the product matrix, the python version allows for a user-friendly parallel computing implementation if multiple cores are available. An R version, though less optimized, is also available and was used to verify our results.

Application to MI omics data

We applied MANOVCA in 1000 healthy individuals from the MI cohort to screen for factors associated with changes in the covariance of 33 blood biomarkers and 169 flow cytometry-based immune phenotypes (Tables S1–S3 available online at <http://bib.oxfordjournals.org/>). Both datasets display high correlation, ranging from -0.71 to 0.99 for the flow cytometry data and from 0.08 to 0.98 for the biomarkers (Figure S6 available online at <http://bib.oxfordjournals.org/>). We first investigated the effect of 49 health-related and lifestyle factors using a subset of 992

participants with complete data. We applied the proposed PCA reduction and investigated power when using 5–200 top PCs with a step of 5, resulting in 40 tests per variable. As shown in Fig. 6A and B, multiple features were associated at a stringent Bonferroni-corrected significance level ($P < 2.5 \times 10^{-5}$ accounting for 1960 tests). Flow cytometry-based phenotypic covariance was associated with age ($P < 6.0 \times 10^{-9}$) and all smoking variables: smoking status ($P < 2.1 \times 10^{-12}$), smoking frequency ($P < 1.2 \times 10^{-11}$), number of years smoked ($P < 3.8 \times 10^{-10}$) and number of years since last smoke ($P < 7.5 \times 10^{-9}$). Likewise, blood biomarkers covariance was strongly associated with age ($P < 3.5 \times 10^{-33}$) and sex ($P < 1.3 \times 10^{-30}$), and to a lesser extent with BMI ($P < 1.6 \times 10^{-7}$) and smoking variables (minimum $P < 5.5 \times 10^{-12}$, for smoking status). Except for the age–flow cytometry association, the maximum association signal was almost reached when including the top 50 PCs and display only modest improvement when including more PCs (Fig. 6).

Thanks to a limited number of outcomes, the blood biomarker dataset could also be analysed by applying MANOVCA directly on the 528 pair-wise products. To investigate the value of using the PCA in such situations, we applied MANOVCA on the product and compared results against the PCA-based approach. Note that the product-based test should be approximately equivalent to the test of all PCs, which was confirmed for these data (Figure S7 available online at <http://bib.oxfordjournals.org/>). Comparing the minimum P -value across the 40 PCA-based test (P_{PCA}) against the P -value from product-based test (P_{prod}), we observed a substantial gain for the PCA-based approach, even when accounting for the multiple testing cost of the PCA approach. The product-based

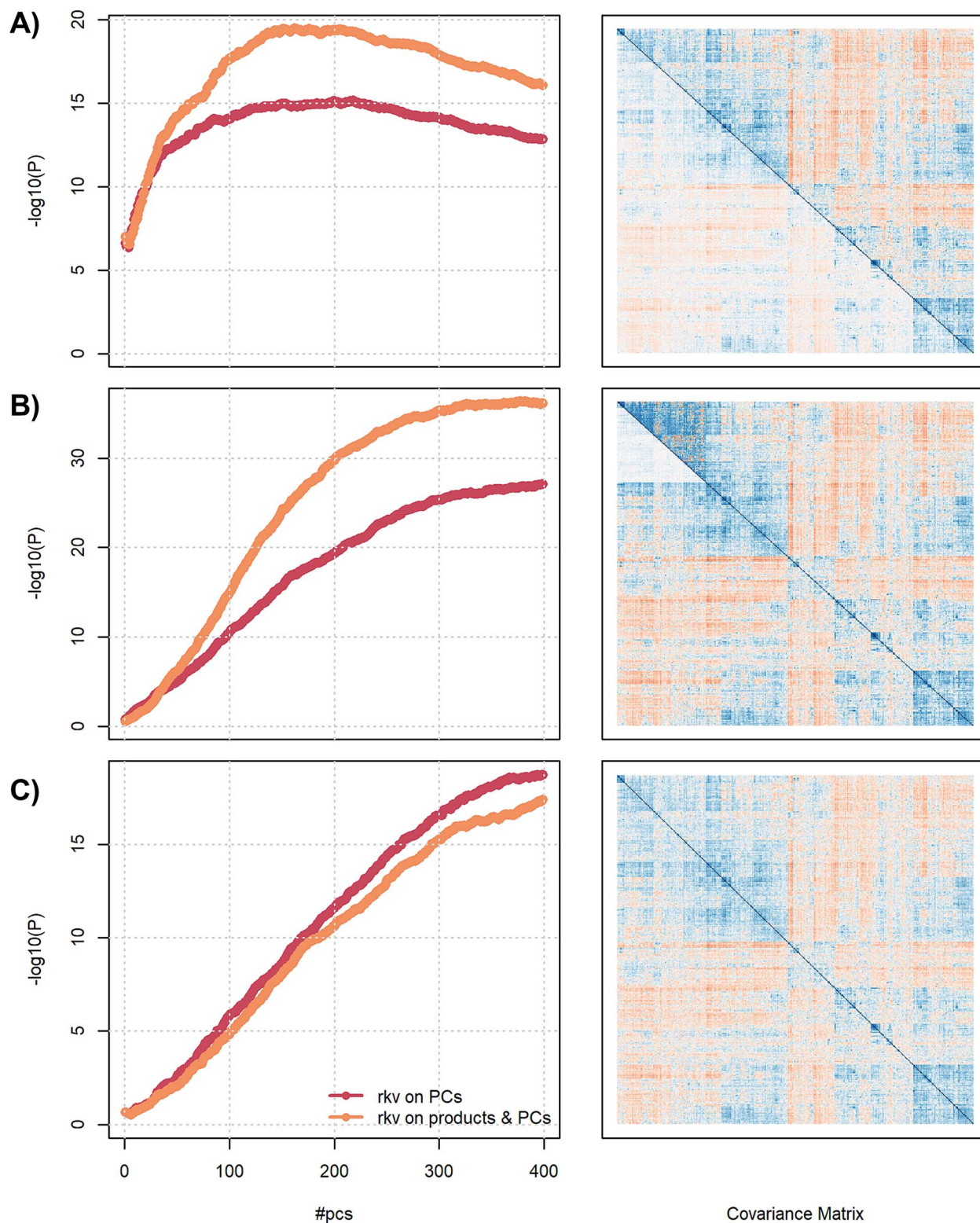


Figure 4. Impact of transformation and number of PCs on statistical power. Power of MANOCCA as a function of the number of PCs retained in the joint test (2–400), while applying two different pre-processings: a rank-inverse normal (rkv) on the PC only or a rank-inverse normal on the PCs and the products. We drew series of 50 replicates with sample size 1000, including a binary exposure with frequency of 0.5 and 30–169 outcomes. For each replicate, we drew two covariance matrices, one for the exposed and one for the unexposed. We generated the outcome under three scenarios using a multivariate normal and covariance derived from real data. In (A), the two matrices are similar but with attenuated covariances among exposed. In (B), random noise is added to the covariance of the exposed group. In (C), the two matrices are equal, except for a subset of outcomes where the covariances have been attenuated. Left panels show the average over the 100 replicates of the $-\log_{10}(P\text{-value})$ derived using MANOCCA for the two pre-processings. Right panels present the matrix produced for each scenario using data from an arbitrarily chosen replicate, with upper and lower triangles showing the true covariance in unexposed and exposed, respectively.

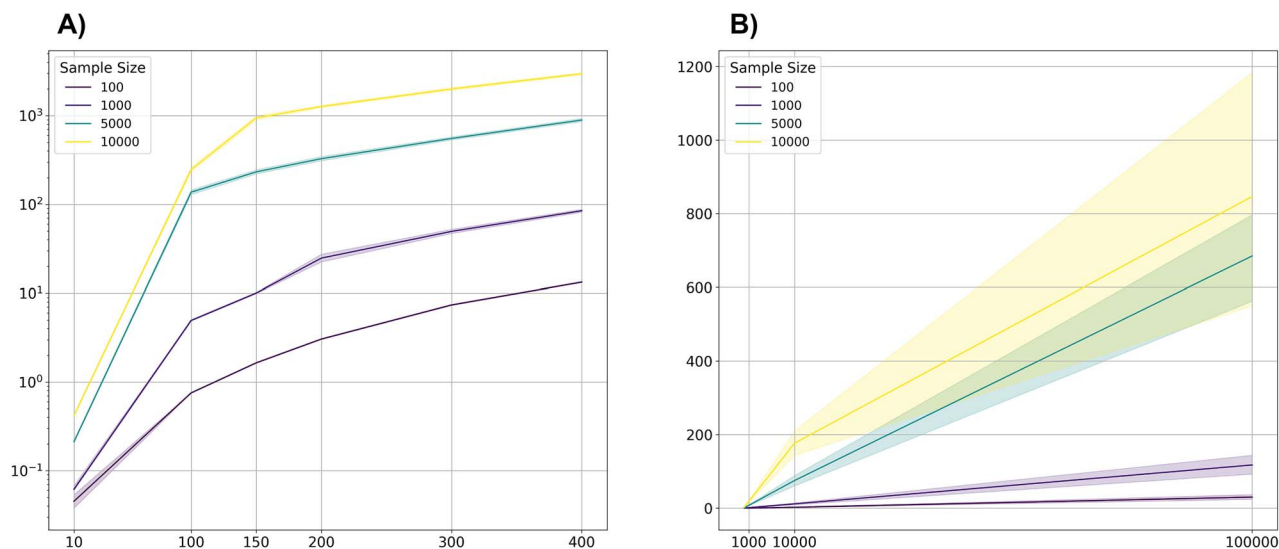


Figure 5. Computational efficiency. Computational time for running MANOCCA on varying sample sizes, numbers of outcomes and numbers of predictors tested. Over 1000 simulations each time, we simulated multivariate normal distributions of various sizes and computed the product matrix $\mathbf{X}^T \mathbf{X}$ rank transform $\mathbf{X} > \text{PCA} > \text{rank transform} > \text{PCA} > \text{rank transform}$ using sequential computation. (A) displays the running time to transform the outcome matrix into the reduced covariance matrix for the test, as a function of the number of outcomes (10, 100, 150, 200, 300, 400) and sample size (100, 1000, 5000 and 10 000). Y-axis is in log₁₀-scale. (B) displays the running time for the testing part for a ranging number of sample sizes (100, 1000, 5000 and 10 000) and ranging number of predictors (1000, 10 000, 100 000).

test identified only five of the seven signals from the PCA-based approach (Fig. 7). For the strongest signals, the association P-value from the best PCA-based test was several orders of magnitude larger than for the product-based test (e.g. for age, $P = 3.5 \times 10^{-33}$ and 2.6×10^{-11} , respectively). These results have three implications. First, it confirms the relevance of using PCA when the dimensionality of the data does not allow the analysis of the product matrix. Using only a subset of top PCs can capture a substantial fraction of the variance of the raw products (e.g. we observed a correlation of 0.85 and higher between the products-based test and the PC-based test, even when using a minimum of 5 PCs, Figure S7). Second, this suggests that, despite the sample size allowing for the full analysis of the product matrix, using PCA remains relevant, with the potential for notable power increase. Third and last, the benefit of testing multiple sets of PCs to search for an optimal threshold can strongly outpace the statistical cost of multiple testing.

When comparing the MANOCCA results to a standard MANOVA applied to both datasets, MANOVA identified significant associations ($P < 1.0 \times 10^{-3}$ to account for the 49 tests conducted) with 8 and 13 predictors associated in the flow cytometry data and the blood biomarkers data, respectively (Table S4 available online at <http://bib.oxfordjournals.org/>). These associations included all associations detected by MANOCCA. More generally, while the two tests are expected to be independent, we observed a strong correlation between association signal as measured by the $-\log_{10}(P\text{-value})$ ($\rho_{\text{biomarkers}} = 0.88$, $\rho_{\text{cytometry}} = 0.65$), suggesting that many of the predictors are associated with effects on both the mean and variance of the outcome studied (Figure S8 available online at <http://bib.oxfordjournals.org/>). However, several predictors display discordant associations, with a significant effect on the mean but no effect on the covariance. This includes, for example, systolic blood pressure ($P_{\text{Manova}} = 2.6 \times 10^{-11}$, $P_{\text{MANOCCA}} = 0.14$) and heart rate ($P_{\text{Manova}} = 4.9 \times 10^{-24}$, $P_{\text{MANOCCA}} = 0.001$) effect on blood biomarkers. On the other hand, one predictor, the number of years of second-hand smoking, displays suggestive significance with MANOCCA on the flow

cytometry data, but did not display the mean effect ($P_{\text{Manova}} = 0.31$, $P_{\text{MANOCCA}} = 3.0 \times 10^{-5}$).

Finally, we conducted genome-wide association studies (GWASs) for both blood biomarkers and flow cytometry datasets, testing 5 699 237 genetic variants with a minor allele frequency (MAF) $> 5\%$ in up to 894 samples where both genetic and phenotypic data were available. All tests were adjusted for age, sex, BMI and the 10 first genetic PCs of the genotyping matrix. Note that for this genetic screening, we only applied the adjustment on the outcomes. Two-sided adjustment has already been used for mean effect tests in GWAS to account for relatedness [21], but would require further investigation to be extended in the MANOCCA test. Following the results from our simulations for genetic variants with an MAF 5% or larger, the type I error will remain robust only for up to 50 PCs (Figure S4c available online at <http://bib.oxfordjournals.org/>), resulting in 10 GWASs per dataset. MANOCCA did not detect any genome-wide significant signals at a stringent Bonferroni-corrected threshold ($P = 5 \times 10^{-9}$, Figure S9 available online at <http://bib.oxfordjournals.org/>). Yet, 46 genetic variants from 11 loci show suggestive significance association ($P = 5 \times 10^{-7}$, Table S5 available online at <http://bib.oxfordjournals.org/>). We conducted a phenome-wide association study on each variant using the ieu database API [25]. Most variants showed strong association with multiple phenotypes from this database (Table S6 available online at <http://bib.oxfordjournals.org/>). In particular, 4 out of the 11 loci harboured genetic variants that were expression quantitative trait loci (eQTL) for one or multiple genes (Table S7 available online at <http://bib.oxfordjournals.org/>), suggesting that our covariance-based approach might capture variants involved in the regulation of gene expression.

Perspective on future applications

Covariance is a feature that is expected to arise commonly in large multidimensional data, including in particular molecular and biomarker data. In general, the relevance of our approach to assess the effect of predictors on such data depends on the a priori

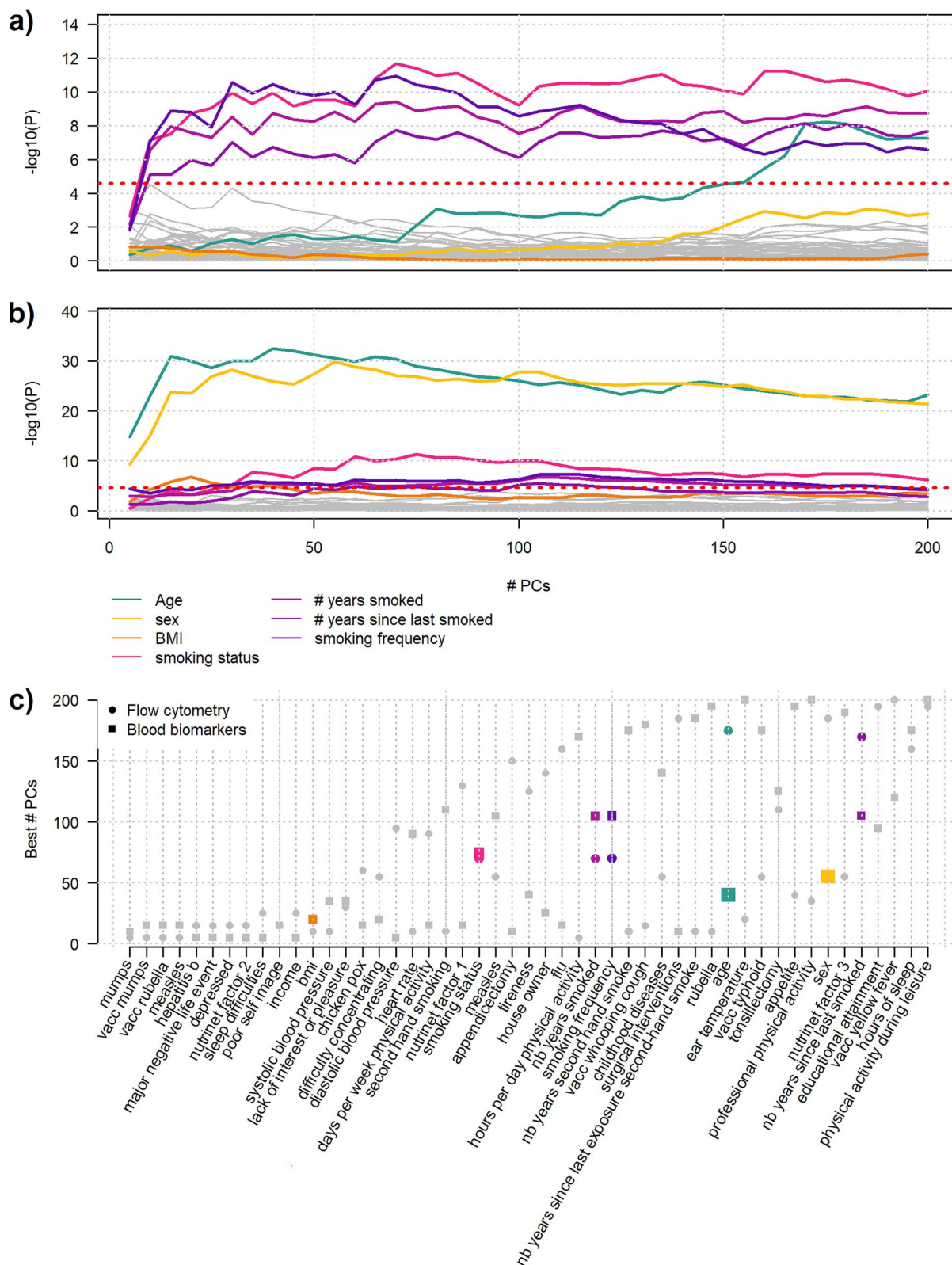


Figure 6. Screenings for host factors screenings for effect of 49 health and lifestyle factors on the covariance of 169 flow cytometry-based immune phenotypes (A) and 33 blood biomarkers (B) using the MANOCCA approach. We ran each screening using 5–200 PCs with a step of 5. Variables with a P-value below the adjusted Bonferroni threshold (2.6×10^{-5} , dashed line) for each screening are displayed in colour: age, sex, body mass index and smoking. (C) presents the list of predictors considered and displays the number of PCs corresponding to the minimum P-value observed.

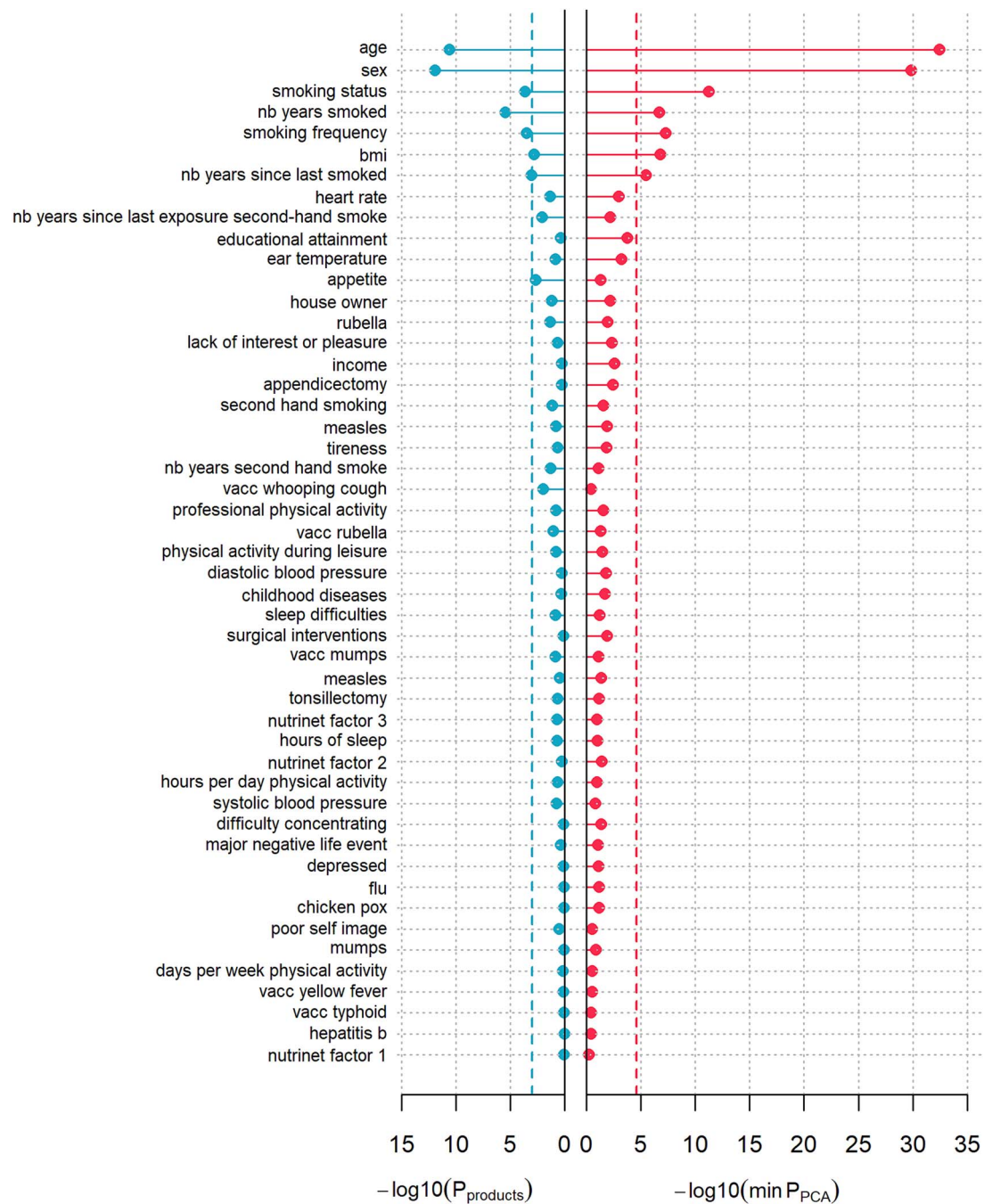


Figure 7. Comparison of PCA-based and product-based MANOCCA screening comparison for covariance signal using MANOCCA on the 33 blood biomarkers as outcomes and using the health and lifestyle electronic case report form (eCRF) questionnaire data as predictors. We ran the MANOCCA screening using the full 528 pair-wise biomarkers products (left) and the PCs derived from the product matrix (right). For the latter, we kept the min(P -value) out of 40 models including 5–200 PCs with a step of 5 PCs. The corresponding Bonferroni correction threshold (dashed line) was derived for each approach based on the number of tests conducted (P -value threshold of 1.0×10^{-3} and 2.6×10^{-5} , respectively).

assumptions of the underlying causal model. To help guiding future applications, we showcase in Figure S10 different models prone to induce or not covariance signals. For clarity, in these simulations, we simply derived the covariance between two outcomes Y_1 and Y_2 across fixed ranged of values of a given predictor X . In general, when the generative model involves the mediation effect, the pleiotropic effect with or without a latent shared factor, the predictor is not expected to produce an effect on the covariance (Figure S10a–c). All models producing a variability of the covariance between Y_1 and Y_2 conditional on X involved some form

of interaction (Figure S10d–f), impacting the association between a latent variable and either ones or both of the outcomes or impacting the association between the two outcomes. The latter generative models can be easily connected to known biological mechanisms. For instance, the model from Figure S10f is fairly appropriate to describe an enzymatic or transporter effect, where X acts as a catalyzer of the transformation between an outcome Y_1 into Y_2 . The model from Figure S10d and e can correspond to scenarios where the relationship between a factor or resource shared by both Y_1 into Y_2 is impacted by X . There is increasing

work in the field of microbiology [26, 27], metabolites [28] and gene expression [29] investigating such relationships and where the proposed MANOCCA approach might be of interest.

Finally, for most statistical models, the relationship between a predictor and an outcome is expected to vary when non-linear transformations are applied to either variable. Here, we devised a pre-processing pipeline that includes some non-linear transformation in order to ensure the statistical robustness of our approach whatever the input outcomes that are analysed. Conversely, the transformation of the raw outcome before deriving the product terms, which might impact both the results and the interpretation of the proposed covariance, remains at the discretion of the user. As for any statistical test, including assessing mean, variance or covariance effect, such transformations will depend on the investigator knowledge about which scale is most relevant to describe the potential underlying biological mechanisms.

Discussion

Covariance is a fundamental feature of omics data. Covariances might be explained by multiple factors, including shared biological mechanisms, shared environmental risk factors or causal effects between the outcomes measured. However, our understanding of the factors involved in covariance has been very limited, partly due to the lack of adapted methodologies and software allowing for systematic screening of large-scale omics datasets. Here, we present MANOCCA, a robust and computationally efficient approach for the identification of predictors associated with the covariance of a multivariate outcome. We show that MANOCCA outperforms existing covariance methods and that, given the appropriate parametrization, it can maintain a calibrated type I error in a range of realistic scenarios when analysing highly multidimensional data. The application of MANOCCA to the MI dataset demonstrates the validity and relevance of our approach, identifying multiple health-related and lifestyle factors significantly associated with the covariance of blood biomarkers and immune phenotypes.

The results from our screenings are in agreement with the existing literature suggesting a role of age, sex, BMI and smoking on the outcomes studied. The marginal association between molecular and cellular biomarkers and ageing has been studied for years and remains topical [30]. There are extensive evidence from clinical and epidemiology studies in healthy individuals that demonstrated associations between age and circulating metabolites [31–33], blood count [31, 33–35] and flow cytometry-based immune phenotypes [36–38]. Association of biomarkers with smoking habits [39] and BMI [40] have been widely described, showing, for example, an impact on platelets glycoprotein [41] and platelet function [42]. Several studies using the same biomarker in MI data have also been conducted, deciphering the short- and long-term effects of smoking on immunity-related biomarkers [43] and describing sexual dimorphism of immunity-related flow cytometry phenotypes [44]. Finally, and more directly related to our work, a recent study suggested association of age and sex with plasma metabolite association networks in healthy subject using correlation approach technics [45].

The MANOCCA approach has four main limitations. First, we used PCA to address situations where the number of covariance terms is larger than the sample size. We defined guidelines that constrain the maximum number of PCs that can be used to ensure the validity of the test when analysing binary or categorical predictors. Regarding power, both simulation and empirical data analyses show that the optimal number of PCs to be

included to maximize power can vary substantially conditional on the true covariance association pattern. Here, we use systematic screenings testing a range of PCs and corrected the association results for multiple testing. Our analyses suggest that the benefit of this strategy largely overcomes its statistical cost. Note that this correction strategy might be further improved as it does not account for correlation between each PC test. Second, our extensive simulation analyses show that when reaching high dimensions, the validity of the test relies on a strong data pre-processing to circumvent the non-normal distribution of products and PCs. Future work is required to identify a refined modelling of these non-normal distributions and avoid this pre-processing. Third, we investigated the performance of our approach on two types of omics data (blood flow cytometry and metabolites) and confirmed its validity and power in these data. Omics data from other sources (e.g. RNAseq) might carry additional complexity that would have to be investigated by simulations before conducting real data applications. Fourth, though applicable to any type of ordinal predictor, MANOCCA is currently restricted to continuous outcomes and unstructured observations. Other fields of application interested in changes in outcome relationships, such as the study of mutation mechanisms in genomic sequences [46], the study of longitudinal data (changes of covariances across different timepoints) or covariances across related individuals, would require further work to be studied using MANOCCA.

Given the increasing number of high-dimensional omics data available in existing human cohorts, our approach provides opportunities to investigate multivariate outcomes from a new perspective. Because MANOCCA is built on a standard linear framework, the approach can be extended in many directions, including the derivation of the individual contribution of outcomes and the development of predictive models. Altogether, we expect the application of our method to produce novel insights on the complex structure linking highly intertwined omics data.

Key Points

- There exist many methods and software to assess the effect of a predictor on the mean and the variance of omics data. Conversely, there is a lack of methodology and tools to assess the effect of predictors on the covariance of multivariate outcomes.
- Under reasonable assumptions on the data distribution, covariance signal is expected to be orthogonal to mean and variance signals and thus can be missed by such approaches.
- MANOCCA provides a robust and scalable test to evaluate the effect of a predictor on the covariance of a multivariate outcome that outperforms the state of the art.
- Our approach identified several significant associations for covariance signal of human metabolites and blood biomarkers with regard to environmental factors.

Supplementary data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

Acknowledgements

The Milieu Intérieur Consortium[†] is composed of the following team leaders: Laurent Abel (Hôpital Necker), Andres Alcover,

Hugues Aschard, Philippe Bousso, Nollaig Bourke (Trinity College Dublin), Petter Brodin (Karolinska Institutet), Pierre Bruhns, Nadine Cerf-Bensussan (INSERM UMR 1163—Institut Imagine), Ana Cumano, Christophe D'Enfert, Ludovic Deriano, Marie-Agnès Dillies, James Di Santo, Gérard Eberl, Jost Enninga, Jacques Fellay (EPFL, Lausanne), Ivo Gomperts-Boneca, Milena Hasan, Gunilla Karlsson Hedestam (Karolinska Institutet), Serge Hercberg (Université Paris 13), Molly A Ingersoll (Institut Cochin and Institut Pasteur), Olivier Lantz (Institut Curie), Rose Anne Kenny (Trinity College Dublin), Mickaël Ménager (INSERM UMR 1163—Institut Imagine), Frédérique Michel, Hugo Mouquet, Cliona O'Farrelly (Trinity College Dublin), Etienne Patin, Sandra Pellegrini, Antonio Rausell (INSERM UMR 1163—Institut Imagine), Frédéric Rieux-Laucat (INSERM UMR 1163—Institut Imagine), Lars Rogge, Magnus Fontes (Institut Roche), Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Frédéric Tangy, Antoine Toubert (Hôpital Saint-Louis), Mathilde Touvier (Université Paris 13), Marie-Noëlle Ungeheuer, Christophe Zimmer, Matthew L. Albert (In Vitro), Darragh Duffy§, Lluís Quintana-Murci§,

¶ unless otherwise indicated, partners are located at Institut Pasteur, Paris

§ co-coordinators of the Milieu Intérieur Consortium

Additional information can be found at: <http://www.milieuinterieur.fr>

Funding

This research was supported by the Agence Nationale pour la Recherche (ANR-20-CE15-0012-01). This work has been conducted as part of the INCEPTION program (Investissement d'Avenir grant ANR-16-CONV-0005). The Milieu Intérieur consortium was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-69-01).

Code availability

All codes are available in Python and R at: <https://gitlab.pasteur.fr/statistical-genetics/manocca>.

Data availability

Data have been deposited in the European Genome-Phenome Archive under accession code EGAC00001001785.

References

- Perez-Riverol Y, Zorin A, Dass G, et al. Quantifying the impact of public omics data. *Nat Commun* 2019;**10**:3512. <https://doi.org/10.1038/s41467-019-11461-w>.
- Wolfe KH, Li WH. Molecular evolution meets the genomics revolution. *Nat Genet* 2003;**33**(Suppl):255–65. <https://doi.org/10.1038/ng1088>.
- Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabolomics* 2013;**1**:92–107.
- Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017;**18**:83. <https://doi.org/10.1186/s13059-017-1215-1>.
- Rahnenfuhrer J, De Bin R, Benner A, et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC Med* 2023;**21**:182. <https://doi.org/10.1186/s12916-023-02858-y>.
- Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* 2013;**4**(5):e201301009. <https://doi.org/10.5936/CSBJ.201301009>.
- Turley P, Walters RK, Maghziyan O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 2018;**50**:229–37. <https://doi.org/10.1038/s41588-017-0009-4>.
- Julienne H, Lechat P, Guillemot V, et al. JASS: command line and web interface for the joint analysis of GWAS results. *NAR Genom Bioinform* 2020;**2**:lqaa003. <https://doi.org/10.1093/nargab/lqaa003>.
- Julienne H, Laville V, McCaw ZR, et al. Multitrait GWAS to connect disease variants and biological mechanisms. *PLoS Genet* 2021;**17**:e1009713. <https://doi.org/10.1371/journal.pgen.1009713>.
- Buergel T, Steinfeldt J, Ruyoga G, et al. Metabolomic profiles predict individual multidisease outcomes. *Nat Med* 2022;**28**:2309–20. <https://doi.org/10.1038/s41591-022-01980-3>.
- Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967;**27**:209–20.
- Steiger JH. Testing pattern hypotheses on correlation matrices: alternative statistics and some empirical results. *Multivar Behav Res* 1980;**15**:335–52. https://doi.org/10.1207/s15327906mbr1503_7.
- Jennrich RI. An asymptotic χ^2 test for the equality of two correlation matrices. *J Am Stat Assoc* 1970;**65**:904–12.
- Box GEP. A general distribution theory for a class of likelihood criteria. *Biometrika* 1949;**36**:317–46. <https://doi.org/10.1093/biomet/36.3-4.317>.
- Lea A, Subramaniam M, Ko A, et al. Genetic and environmental perturbations lead to regulatory decoherence. *elife* 2019;**8**. <https://doi.org/10.7554/eLife.40538>.
- Ates C, Kaymaz O, Kale HE, et al. Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-I error rates. *Comput Math Methods Med* 2019;**2019**:2173638.
- Pourahmadi M, Wang X. Distribution of random correlation matrices: hyperspherical parameterization of the Cholesky factor. *Stat Probabil Lett* 2015;**106**:5–12. <https://doi.org/10.1016/j.spl.2015.06.015>.
- Thomas S, Rouilly V, Patin E, et al. The Milieu Intérieur study - an integrative approach for study of human immunological variance. *Clin Immunol* 2015;**157**:277–93. <https://doi.org/10.1016/j.clim.2014.12.004>.
- Hasan M, Beitz B, Rouilly V, et al. Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping. *Clin Immunol* 2015;**157**:261–76. <https://doi.org/10.1016/j.clim.2014.12.008>.
- Chaltiel D, Adjibade M, Deschamps V, et al. Programme National Nutrition Sante - guidelines score 2 (PNNS-GS2): development and validation of a diet quality score reflecting the 2017 French dietary guidelines. *Br J Nutr* 2019;**122**:331–42. <https://doi.org/10.1017/S0007114519001181>.
- Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9. <https://doi.org/10.1038/ng1847>.
- Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008;**40**:646–9. <https://doi.org/10.1038/ng.139>.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**:e1000529. <https://doi.org/10.1371/journal.pgen.1000529>.

24. Pare G, Cook NR, Ridker PM, et al. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's genome health study. *PLoS Genet* 2010;**6**:e1000981. <https://doi.org/10.1371/journal.pgen.1000981>.
25. Ben E, Lyon MS, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 2020.
26. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol* 2017;**25**(3):217–28. <https://doi.org/10.1016/j.tim.2016.11.008>.
27. Chen L, Collij V, Jaeger M, et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat Commun* 2020;**11**:4018. <https://doi.org/10.1038/s41467-020-17840-y>.
28. Smith CJ, Sinnott-Armstrong N, Cichońska A, et al. Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. *elife* 2022;**11**. <https://doi.org/10.7554/eLife.79348>.
29. Amariuta T, Siewert-Rocks K, Price AL. Modeling tissue co-regulation estimates tissue-specific contributions to disease. *Nat Genet* 2023;**55**:1503–11. <https://doi.org/10.1038/s41588-023-01474-z>.
30. Moqri M, Herzog C, Poganik JR, et al. Validation of biomarkers of aging. *Nat Med* 2024;**30**:360–72. <https://doi.org/10.1038/s41591-023-02784-9>.
31. Sebastiani P, Thyagarajan B, Sun F, et al. Age and sex distributions of age-related biomarker values in healthy older adults from the long life family study. *J Am Geriatr Soc* 2016;**64**:e189–94. <https://doi.org/10.1111/jgs.14522>.
32. Adav SS, Wang Y. Metabolomics signatures of aging: recent advances. *Aging Dis* 2021;**12**:646–61. <https://doi.org/10.14336/AD.2020.0909>.
33. Bortz J, Guariglia A, Klaric L, et al. Biological age estimation using circulating blood biomarkers. *Commun Biol* 2023;**6**:1089. <https://doi.org/10.1038/s42003-023-05456-z>.
34. Crimmins E, Vasunilashorn S, Kim JK, et al. Biomarkers related to aging in human populations. *Adv Clin Chem* 2008;**46**:161–216. [https://doi.org/10.1016/S0065-2423\(08\)00405-8](https://doi.org/10.1016/S0065-2423(08)00405-8).
35. Nakamura E, Miyao K. A method for identifying biomarkers of aging and constructing an index of biological age in humans. *J Gerontol A Biol Sci Med Sci* 2007;**62**:1096–105. <https://doi.org/10.1093/gerona/62.10.1096>.
36. Larbi A, Pawelec G, Wong SC, et al. Impact of age on T cell signaling: a general defect or specific alterations? *Ageing Res Rev* 2011;**10**:370–8. <https://doi.org/10.1016/j.arr.2010.09.008>.
37. Lin Y, Kim J, Metter EJ, et al. Changes in blood lymphocyte numbers with age in vivo and their association with the levels of cytokines/cytokine receptors. *Immun Ageing* 2016;**13**:24. <https://doi.org/10.1186/s12979-016-0079-7>.
38. Valiathan R, Ashman M, Asthana D. Effects of ageing on the immune system: infants to elderly. *Scand J Immunol* 2016;**83**:255–66. <https://doi.org/10.1111/sji.12413>.
39. Yanbaeva DG, Dentener MA, Creutzberg EC, et al. Systemic effects of smoking. *Chest* 2007;**131**:1557–66. <https://doi.org/10.1378/chest.06-2179>.
40. Nimptsch K, Konigorski S, Pischon T. Diagnosis of obesity and use of obesity biomarkers in science and clinical medicine. *Metabolism* 2019;**92**:61–70. <https://doi.org/10.1016/j.metabol.2018.12.006>.
41. Nair S, Kulkarni S, Camoens HM, et al. Changes in platelet glycoprotein receptors after smoking – a flow cytometric study. *Platelets* 2001;**12**:20–6. <https://doi.org/10.1080/09537100120046020>.
42. Inoue T. Cigarette smoking as a risk factor of coronary artery disease and its effects on platelet function. *Tob Induc Dis* 2004;**2**:27–33. <https://doi.org/10.1186/1617-9625-2-1-27>.
43. Saint-Andre V, Charbit B, Biton A, et al. Smoking changes adaptive immunity with persistent effects. *Nature* 2024;**626**:827–35. <https://doi.org/10.1038/s41586-023-06968-8>.
44. Marquez EJ, Chung C-H, Marches R, et al. Sexual-dimorphism in human immune system aging. *Nat Commun* 2020;**11**:751. <https://doi.org/10.1038/s41467-020-14396-9>.
45. Vignoli A, Tenori L, Luchinat C, et al. Age and sex effects on plasma metabolite association networks in healthy subjects. *J Proteome Res* 2018;**17**:97–107. <https://doi.org/10.1021/acs.jproteome.7b00404>.
46. Sun N, Zhao X, Yau SS. An efficient numerical representation of genome sequence: natural vector with covariance component. *PeerJ* 2022;**10**:e13544. <https://doi.org/10.7717/peerj.13544>.