Metabolic predictors of phenotypic traits can replace and complement measured clinical variables in transcriptome-wide association studies

Anna Niehues^{1,2}, Daniele Bizzarri^{3,4}, Marcel J.T. Reinders^{4,5}, P. Eline Slagboom^{3,6}, Alain J. van Gool², Erik B. van den Akker^{3,4,5}, Peter A.C. 't Hoen^{1*}, with the BBMRI-NL BIOS and Metabolomics Consortia[†]

1 Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands

2 Translational Metabolic Laboratory, Department Laboratory Medicine, Radboud university medical center, Nijmegen, The Netherlands

3 Molecular Epidemiology, LUMC, Leiden, The Netherlands

4 Leiden Computational Biology Center, LUMC, Leiden, The Netherlands

5 Delft Bioinformatics Lab, TU Delft, Delft, The Netherlands

6 Max Planck Institute for the Biology of Ageing, Cologne, Germany

[†]Membership lists can be found in the Acknowledgments section.

* peter-bram.thoen@radboudumc.nl

Abstract

Transcriptome-wide association studies (TWAS) can provide valuable insights into biological and disease-underlying mechanisms. For studying clinical effects, availability of (confounding) phenotypic traits is essential. The (re)use of RNA-seq or other omics data can be limited by missing, incomplete, or inaccurate phenotypic information. A possible solution are molecular predictors inferring clinical or behavioral phenotypic traits. Such predictors have been developed based on different omics data types and are being applied in various studies.

In this study, we applied 17 metabolic predictors to infer various traits, including diabetes status or exposure to lipid medication. We evaluated whether these metabolic surrogates can be used as an alternative to reported information for studying the respective phenotypes using TWAS. Our results revealed that in most cases, the use of metabolic surrogates yields similar results compared to using reported information, making them suitable substitutes for such studies.

The application of metabolomics-derived surrogate outcomes opens new possibilities for reuse of multi-omics data sets, especially in situations where availability of clinical metadata is limited. Missing or incomplete information can be complemented by these surrogates, thereby increasing the size of available data sets. Studies would likely also benefit from the use of such surrogates to correct for potential biological confounding. This should be further investigated.

Author summary

Transcriptome-wide association studies (TWAS) can be used to associate gene expression levels with phenotypic traits. These associations can provide insights into biological mechanisms including those that underlie diseases. Such studies require

molecular profiling data from a large number of individuals as well as information on the phenotypic trait of interest. Biobanks that collect samples and corresponding molecular data, also collect information on phenotypic traits or clinical information. However, this information can be heterogeneous and/or incomplete, or a certain piece of information could be missing entirely. In this study, we apply metabolic predictors to infer various traits, including diabetes status or exposure to lipid medication. We evaluate whether these metabolic surrogates can be used as an alternative to reported information for studying the respective phenotypes using TWAS. Our results reveal that in many cases, the use of metabolic surrogates yields similar results compared to using reported information, making them suitable substitutes for such studies. The possibility of using these surrogate outcomes can thus increase the size of data sets for studies where phenotypic information is incomplete or missing.

Introduction

Genome-wide association studies (GWAS) have proven to be valuable in uncovering links between genes and a wide range of phenotypic traits. Such findings have led to the discovery of new disease-related biomarkers and are often the basis for gaining a better understanding of biological processes or disease-mechanisms. Since the introduction of the first GWAS powered by the availability of genome-wide single-nucleotide polymorphism (SNP) profiling, numerous studies have identified thousands of SNP-trait associations [1]. Technological advancements allowing high-throughput profiling of other molecular features, such as transcripts, or DNA methylation sites, also enabled corresponding transcriptome- (TWAS), epigenome- (EWAS), and other omes-wide association studies.

Such studies are susceptible to confounding by biological and technical factors that can influence omics profiles and phenotypic traits of interest. However, measured values to correct for such confounding are often not available. As a solution, differences in cell type composition are commonly accounted for using information contained in the DNA methylation profiles themselves, by either reference-based imputation [2] or reference-free methods such as surrogate variable analysis (methods reviewed in [3] and [4]). Other well-known examples of inferring values for possible confounding factors from DNA methylation profiles include sex [5] and smoking status prediction. Bollepalli et al. [6] trained a smoking status classifier using multinomial LASSO regression. Machine learning approaches have also been applied to other omics data types to predict environmental exposures [7]. Recently, Bizzarri et al. [8] trained predictors on proton NMR-based metabolomics (Nightingale Health) data. The authors applied logistic regression using elastic net regularization to train models for various clinical variables, including physiological measures, environmental exposures, and clinical endpoints. They demonstrated the use of their predicted values as metabolic surrogates to complement missing phenotypic data when performing metabolome-wide association studies (MWAS).

The value of such predictors is not only evident when complementing missing data to account for technical or biological confounding, but also for using them as outcome variables. These molecular surrogates can be used in association studies in order to link molecular features to clinical phenotypes or exposures. Since identified associations in omes-wide association studies often have only moderate effect sizes, a common approach to detect relevant features are cross-cohort meta-analyses [9]. However, the applicability of meta-analyses can be limited by availability of the respective outcomes of interest. Specific clinical, environmental, or phenotypic traits might not be recorded in every cohort, or the data collection might be based on different protocols, making the reported values for these traits not directly comparable.

As more and more multi-omics data sets become available, it becomes possible to make use of molecular predictors to infer phenotypic traits from specific omics layers. We propose their use in subsequent analyses of other omics layers, not only as covariates to account for confounding factors, but also as outcome variables. We here investigated whether values derived from molecular predictors represent a viable alternative to measured or reported clinical or phenotypic traits to serve as outcome variables in TWAS.

To this end, we applied 17 metabolic predictors to metabolomics data from four large population cohorts inferring values for phenotypic, exposure, and clinical traits. We performed TWAS analyses on corresponding RNA-seq data sets employing either reported/measured or inferred values as outcome variables, and systematically compared the respective results of these analyses. For five of the outcomes, where reported values were not available, we evaluated the performance of the metabolic surrogates based on results reported in literature.

Results

In this study, we performed transcriptome-wide association studies (TWAS) to evaluate the performance of 17 surrogate outcomes that are based on molecular predictors. RNA-seq data used for TWAS and metabolomics data used to infer surrogate outcome values are part of multi-omics data sets of four large Dutch population cohorts: LifeLines (LL) [10], Leiden Longevity Study (LLS) [11], Netherlands Twin Register (NTR) [12], and Rotterdam Study (RS) [13]. An overview of the study workflow is shown in Fig 1. For the different outcomes (Table 1), we compared TWAS results based on metabolic surrogate outcomes to those based on reported outcomes whenever possible. Additionally, results were compared to literature findings.

Fig 1. Overview of study workflow. Gene-wise models are fitted for various outcome variables based on reported information or metabolic surrogates, respectively.

Table 1. Overview of phenotypic traits.

Phenotypic trait	Reported outcome	Surrogate outcome
Low estimated Glomerular Filtration Rate (eGFR)		x
High triglycerides	x	x
High LDL-associated cholesterol	x	x
High total cholesterol	x	x
Low HDL-associated cholesterol	x	x
Diabetes		x
Metabolic syndrome		x
Sex	x	x
Lipid medication	x	x
BMI/obesity status	x	x
High high-sensitivity C-reactive protein (hsCRP)	x	x
Blood pressure lowering medication		x
Low hemoglobin	x	x
Low white blood cells	x	x
Current smoking	x	x
Alcohol consumption		x
High age (≥ 65 y.o.)	x	x

Metabolic surrogate outcomes

We inferred values for clinical variables by applying molecular predictors to metabolomics data. Recently reported metabolic predictors trained on up to 22 Dutch population cohorts [8] were applied to infer values for outcome variables (Table 1). All 17 metabolic predictors had been shown to perform accurately with mean AUC values > 0.7 in a 5-fold cross-validation approach [8]. In order to avoid emphasis on clinical extremes, the metabolic predictors trained by Bizzarri et al. are based on binary representations using clinical thresholds for continuous variables. The values returned by the predictors are continuous posterior probability scores for belonging to one of the two groups. In most cases, this is the clinical risk group. We here used these predicted values as surrogate outcomes and compared their use in TWAS to reported or measured outcome values.

Transcriptome-wide association studies

In the next step, the 17 metabolic surrogates were used as outcomes in TWAS. In addition to analyses employing these surrogate outcomes, analyses using measured or reported outcome values were performed. For five outcomes that had limited availability of reported values (eGFR, diabetes, metabolic syndrome, blood pressure lowering medication, and alcohol consumption) only surrogate outcomes were used. In each linear regression model, known biological (age, sex) and technical (flow cell number, white blood cell composition) confounding factors were included (formulas available in S1 Table).

For an initial assessment of the performance of each model, we compared the numbers of significant associations and the effect sizes between outcomes and outcome variable types. Additionally, test statistic (*t*-statistic) bias and inflation were estimated as parameters (mean and standard deviation) of the empirical null distribution using a Bayesian method implemented in the R package *bacon* [26] (Fig 2). Numbers of identified significant associations (Fig 2A) varied strongly across outcomes. Highest numbers were found for the outcomes triglycerides, metabolic syndrome, and white blood cells. For several outcomes, including eGFR, LDL cholesterol, total cholesterol, and alcohol consumption, no or only few significant gene-trait associations were found. For outcomes where TWAS results based on surrogate outcomes could be compared to results for reported variables, the numbers of identified significant associations averaged across all cohorts were higher for the metabolic surrogate outcomes in three cases, and lower in nine cases. However, the variation across the four cohorts was generally higher than the difference between models employing either reported or surrogate outcome variables. Similarly, high variation across cohorts was observed for the other parameters assessed to evaluate the performance of the models. Absolute effect size averaged across all genes (Fig 2B) were generally small, with the highest values observed for the outcomes triglycerides and sex. In 10 cases, the mean absolute effect size averaged across cohorts was lower when using metabolic surrogate outcomes instead of reported variables; in two cases, it was higher. We observed relatively low test statistic bias (Fig 2C) across all outcomes and types of outcome variables. The bias, i.e., the deviation of the empirical null distribution's mean from zero, averaged across cohorts decreased in four cases, increased in three cases, and remained similar in five cases when employing metabolic surrogate outcomes instead of reported variables. Bias in the RS cohort was often higher than in the other cohorts. This may be explained with the differences in population structure. The RS cohort has a higher average age than the other three cohorts [27], indicating higher bias for the studied clinical variables in older populations. Inflation (deviation of the empirical null distribution's standard deviation from one, Fig 2D) was highest for the outcome sex. In most cases, inflation averaged across

cohorts remained stable when using different outcome variable types. For the outcome total cholesterol, it slightly decreased when using metabolic surrogate outcomes instead of reported variables; for the outcomes lipid medication and hsCRP, it slightly increased.

Fig 2. Comparison of TWAS result characteristics. Number of significant associations (based on **bacon**-corrected and Bonferroni-corrected p -values $p_{b_{adj}} < 0.05$) (A), mean absolute effect sizes (based on **bacon**-corrected effect sizes) across all genes (B), and bias (C) and inflation (D) of test statistics (t -statistic) for alternative models per comparisons and cohort. Horizontal lines for test-statistic mean = 0 and standard deviation = 1 of theoretical null distribution were added. Comparisons are ordered by performance of metabolic predictors for binary outcome measures. Type of outcome variable is indicated by color: reported or measured variable = black, metabolic surrogate = orange. Mean values across four cohorts (two cohorts for hsCRP) are plotted as horizontal bars. Note the $\log_{10}$ scale on the y-axis of the upper plot.

Since the number of significant associations and average effect sizes do not allow to draw conclusions about the similarity of the TWAS results employing different types of variables as outcome, we next performed pairwise comparisons of models with different types of outcome variables, i.e., reported vs. surrogate. Fig 3 shows the correlation of regression coefficients from gene-wise fitted linear models between two different types of outcome variables. Correlation coefficients were generally high for surrogate outcomes based on best-performing metabolic predictors. For outcomes based on predictors with reported AUC > 0.9 (triglycerides, LDL cholesterol, total cholesterol, HDL cholesterol, and sex) [8], the correlation coefficients averaged across cohorts ranged between 0.71 and 0.96. We observed a modest trend for decreasing correlations with decreasing performance of predictors. However, there were exceptions, with sex having highest correlation values, although the predictor's AUC was reported to be lower than those for triglycerides or cholesterol. Lowest similarity of TWAS results with average absolute Pearson $r < 0.5$ were observed for the outcomes age, white blood cells, smoking, and hemoglobin, the latter having lowest correlation values. We often observed that correlations were lower for the NTR cohort. This could be explained by a technical difference in the metabolic profiles, with NTR missing glutamine [8].

Fig 3. Pairwise comparisons of TWAS results. Absolute Pearson correlation coefficients (Pearson r) of **bacon**-adjusted regression coefficients of gene-wise linear models (**limma/voom**) for outcome variables in alternative models per comparison and cohort. Comparisons are ordered by performance (AUC) of metabolic predictors for binary outcome measures. Mean values across four cohorts (two cohorts for hsCRP) are plotted as horizontal bars (gray).

Meta-analyses and replication studies

When comparing two alternative outcome variables, a lower or higher number of found significant associations does not necessarily imply that results are better or worse, since the values alone do not indicate if this is due to a reduction or increase of false positive (noise) or true positive findings, respectively. We observed that the TWAS results differ when surrogate values differ from reported values. However, we do not know which set of results is correct, as reported values could contain inaccuracies. Under the assumption that true positive findings, as opposed to false positive results, can be replicated in different cohorts (validating the results), we performed replication studies to determine which outcome variable type is more consistently reflected in the RNA-seq data. We performed leave-one-cohort-out meta-analyses and replication studies for all

comparisons (except for hsCRP where only two cohorts were available) using the approach described by van Rooij et al. [27]. For each comparison, four meta-analyses were performed leaving one cohort out each time, and using the left out cohort for a replication analysis. Fig 4 shows the numbers of significant associations found in each meta-analysis (number of meta-analyzed genes) and the respective percentage of replicated genes. Overall, we did not find substantial differences in the numbers of meta-analyzed genes (Fig 4A) except for the outcomes lipid medication, (high) age, and current smoking. While more genes were meta-analyzed when using the metabolic surrogate for lipid medication, the reported variable yielded more meta-analyzed genes for age and smoking. For the latter two outcomes, results based on metabolic surrogates could not be replicated (Fig 4B) while on average 11-15% of results based on reported outcomes could be replicated. For these two outcomes we had also observed highest differences between number of significations associations (Fig 2A). For other outcome variables, the percentage of replicated genes was quite similar between outcome variable types, but the cohort which was left out for the meta-analysis had a strong impact on the results. Replication rates were generally low with the highest average replaction rate observed for triglycerides with just under 30%. For a number of outcomes, associations could hardly be replicated: LDL-associated cholesterol, hemoglobin, and alcohol consumption. This is in line with the fact that almost no significant associations were found for these outcomes in the individual cohorts (Fig 2).

Fig 4. Meta-analyses and replication studies. Number of meta-analyzed genes (significant associations, *bacon*-adjusted *p*-values Bonferroni-adjusted for multiple testing, $p < 0.05$) in leave-one-cohort meta-analyses (A) and percentage of genes replicated (significant associations, Bonferroni-adjusted for multiple testing, $p < 0.05$) in replication cohort (B). Type of outcome variable is indicated by color: reported or measured variable = black, metabolic surrogate = orange. Mean values across four meta-analyses/replication studies are plotted as horizontal bars. Note the $\log_{10}$ scale on the y-axis of the upper plot.

Gene set enrichment analysis

To arrive at a biological interpretation of the TWAS results, we performed gene-set enrichment analyses (GSEA) using pathways from the Reactome database. GSEA was applied to both individual TWAS results and results from a meta-analysis of all four cohorts (Fig 5). For direct comparisons of reported variables and metabolic surrogates, we observed a highest overlap of significantly enriched pathways for HDL cholesterol and triglycerides in all cohorts, with 70-84% (meta-analysis 76%) and 65-80% (meta-analysis 77%) of significantly enriched pathways found by both reported and surrogate outcome, respectively. The overlap for the outcomes total cholesterol, lipid medication, hsCRP, BMI/obesity, and white blood cells was more variable across cohorts. Meta-analyzed results had an overlap of 38-69% and the order of significantly enriched pathways was highly comparable (see S1 Appendix). The results for high age, current smoking, hemoglobin, and LDL cholesterol demonstrated a lower overlap than other outcomes and showed higher variation across the four cohorts compared to other outcomes. This is partially in line with the comparison of gene-wise linear models from TWAS (Fig 3), which showed that results for the outcome hemoglobin based on reported and inferred values were not correlated, and high age and current smoking were only moderately correlated. Since hardly any significant associations were found for hemoglobin (see Fig 2A), the observed signal for this outcome was generally very low in the studied cohorts, independent of the type of outcome variable. It is surprising to observe that almost no significantly enriched pathways were observed for the outcome

sex, even though many gene-trait associations were found and could be replicated in the meta-analysis and replication approach.

Fig 5. Gene-set enrichment analyses of TWAS results. Numbers of significantly enriched (Bonferroni-adjusted $p < 0.05$) pathways (Reactome) for each outcome found in each cohort and in meta-analysis of all four (two for hsCRP) cohorts (top). Values for each type of outcome variable are represented as colored bars: reported variable = black, metabolic surrogate = orange, intersection, i.e., pathways found by all outcome variables = blue.

In order to evaluate the performance of surrogates for which results could not be compared to results based on reported outcomes, we compared significantly enriched pathways (see S1 Appendix) to the literature. The top-ranked enriched pathways for low eGFR are related to translation. Since low eGFR is an indicator of kidney disease, this is in line with studies reporting increased translational activity to several kidney diseases [30,31]. For diabetes and metabolic syndrome, which is a risk factor for diabetes, 16 out of the top 20 significantly enriched pathways were found for both outcomes. These include pathways related to translation, signaling, infection, and amino acid deficiency and metabolism. This is in agreement with previously reported results [32,33]. For alcohol consumption, although almost no significant associations were found in individual TWAS, several significant gene-outcome associations were found when meta-analyzing multiple cohorts (compare Fig 2 and Fig 4). Top-ranked positively enriched Reactome pathways from gene-set enrichment analysis (S1 Appendix), including, e.g., innate immune system, signal transduction, and infectious disease, have been linked previously to chronic alcohol drinking [34].

Discussion

Our systematic comparison of results of TWAS against outcome variables that were either reported or inferred by molecular predictors generally showed good agreement (Fig 2). Most similar TWAS results across all assessment parameters are those for the outcomes triglycerides and HDL cholesterol. Many significant gene-trait associations were found, of which many could be replicated, and the majority of significantly enriched pathways were found by reported and surrogate outcomes. Regression coefficients of the models including the reported and surrogate outcomes were strongly correlated, and the majority of pathways found by GSEA were obtained by both outcome types. The top-ranked pathways positively enriched for both high triglycerides and low HDL-cholesterol include “GTP hydrolysis and joining of the 60S ribosomal subunit”, “L13a-mediated translational silencing of Ceruloplasmin expression” and “Formation of a pool of free 40S subunits” which participate in the Eukaryotic translation initiation [29] and were previously shown to be enriched in a high-cholesterol and high-fat diet induced hypercholesterolemic rat model [35]. The metabolic predictors for these outcomes are directly related to metabolic markers measured on the Nightingale platform and had shown high performance ($AUC \geq 0.95$) [8]. It is expected that results based on predicted outcomes will depend on the accuracy of the prediction. Accordingly, we observed a slightly lower correlation between reported outcomes and metabolic surrogates for molecular predictors that were known to have lower accuracy (Fig 3). In the TWAS for lipid medication and BMI/obesity, the molecular predictors yielded even more significant gene-trait associations than the reported outcomes (Fig 2A) and a higher number of significantly enriched pathways were obtained when using the metabolic surrogates (Fig 5). For lipid medication, this may be related to inaccurate recording of this trait in the questionnaires used. For BMI, this may be explained by

the more direct capturing of metabolic processes that are associated with obesity as a combined measure of BMI and waist circumference, because BMI alone is not a perfect indicator of metabolic health [36]. Similarity of results based on surrogate and reported value for sex was high across all assessment parameters, but GSEA did not yield significantly enriched pathways in most cases. It is possible that the genes significantly associated with sex belong to too many different pathways and/or that some genes within a pathway have a positive association while others have a negative association resulting in a failure to identify positively or negatively enriched pathways. For the outcomes total cholesterol and hsCRP, regression coefficients were moderately correlated and GSEA results were very similar. For the outcome white blood cells overlap of significantly enriched pathways in GSEA was smaller. However, the order of top-ranked pathways was similar (see S1 Appendix). The application of metabolic surrogates for outcomes that were not reported in the data and the comparison of TWAS and GSEA results with literature show that these surrogate outcomes allow to transfer information from one data set (the training data) to another, thereby facilitating to study phenotypes or exposures in data sets which would otherwise not be possible.

For the outcomes LDL cholesterol, high age, hemoglobin and current smoking only few gene-trait associations were significantly associated with the outcome, and even fewer in the case of surrogate outcomes. For high age and current smoking the surrogate outcomes also performed worse in the meta-analyses and replication studies compared to reported outcome values. Here, differences between results based on different outcome variable types are reflected in lower correlations of regression coefficients of the gene-wise models and in a smaller overlap in enriched pathways from GSEA. Possible reasons for the differences between surrogates and reported values are lower performance of the metabolic predictors for these outcomes, or a lower biological signal for the respective clinical outcomes and thus increased noise in the studied data. It is known that aging is reflected in transcriptomics data [37], but the metabolic predictor for high age trained on binarized data (≥ 65 y.o.) might not be an ideal surrogate to study this. Alternatively, a metabolomics-based biological age predictor based on continuous data [17] might perform better. Differences between GSEA results of different outcome variable types could also arise from different biological information captured by the metabolic surrogates and by reported or measured values. This phenomenon is known from epigenetic clocks whose age predictions can differ from chronological age, and different clocks can reflect different aspects of biological age [38]). While many of the top-ranked pathways (S1 Appendix) for smoking were found by both outcome variable types, some of the pathways were solely found by using either reported smoking status (“smoking_current”) or metabolic surrogate(“s_current_smoking”). Several pathways related to translation initiation (“Formation of the ternary complex, and subsequently, the 43S complex”, “Translation initiation complex formation”, “Ribosomal scanning and start codon recognition”) were only significantly enriched when using the reported variable as outcome. Translation of mRNA is known to be dysregulated in cancers [39]. Pathways only enriched when using the metabolic surrogate include “Platelet activation, signaling and aggregation” and “Hemostasis”. Increased platelet aggregation has been reported in smokers [40] and platelet-dependent thrombin levels were shown to be increased in smokers and following smoking [41]. This possibly indicates that the reported smoking behavior captures effects of long-term exposure to smoking better, while the metabolic surrogate captures effects of acute smoking. It would be interesting to further investigate which aspects of the clinical phenotypes are captured by the metabolic surrogates. This requires additional phenotypic information. To understand which aspect of smoking behavior is reflected in the omics data current smoking status alone might not be sufficient. More information including pack years and years since smoking cessation could help better understand the

information captured by the predictors. It is also possible that different omics types capture different effects better, e.g., short-term and long-term effects. In that case, combining reported outcome variables, and/or molecular surrogates from different omics layers could be very useful, not only to study the effect of a certain exposure, but also to better adjust for confounding factors.

Conclusion

Based on the systematic comparison of TWAS results using either reported variables as outcome or surrogate outcomes inferred from another omics profile, we conclude that metabolic surrogates yield similar results as reported variables. The availability of these surrogate outcomes extends the possibilities of studying various clinical outcomes in population cohorts. It can enable the reuse of existing multi-omics data with limited reported clinical (meta)data. This allows for inclusion of more cohorts in meta-analyses, even when outcomes of interest were not reported for all cohorts. This approach also increases possibilities to study clinical outcomes by allowing to infer important confounding factors. Especially investigations that rely on reuse of existing data, .e.g. in the case of rare disease studies which often also suffer from low sample sizes, will benefit from this approach.

Methods

Data

In this study, we analyzed RNA-seq and metabolomics data from four large Dutch population cohorts: LifeLines (LL) [10], Leiden Longevity Study (LLS) [11], Netherlands Twin Register (NTR) [12], and Rotterdam Study (RS) [13,14]. The data is provided by the Dutch node of the European Biobanking and BioMolecular Resources and Research Infrastructure (BBMRI-NL).

RNA-seq data was generated by the BBMRI-NL Biobank-based Integrative Omics Study (BIOS) Consortium at the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands. RNA sample processing and sequencing is described in detail by Zhernakova et al. [15]. Briefly, total RNA was extracted from whole blood, depleted of globin transcripts, and paired-end sequencing of 2x50-bp reads was conducted using the Illumina HiSeq 2000 platform. Read alignment to reference genome hg19 was performed using STAR (v2.3.0). We used the “Freeze2 unrelated data sets”, which contain maximum sets of unrelated individuals and are available within the BIOS workspace at the SURF Research Cloud via the R package `BBMRIomics` v3.4.2 [16].

Metabolomics data was generated by the BBMRI-NL Metabolomics Consortium in 2014 as described by van den Akker et al. [17]. Briefly, metabolite concentrations were measured in EDTA plasma by proton nuclear magnetic resonance (^1H -NMR) spectroscopy on the platform of the Nightingale Health Group (Helsinki, Finland) [18].

Ethics statement

Written informed consent was obtained previously from all participants of the LL, LLS, NTR and RS biobanks in accordance with the ethical and institutional regulations. The LL study was approved by the ethics committee of the University Medical Centre Groningen, document no. METC UMCG LLDEEP: M12.113965. All participants signed an informed consent form prior to study enrolment [10]. The Leiden Longevity Study was approved by the Medical Ethical Committee of the Leiden University Medical Center (METC LDD) and informed consent was obtained from all subjects. All

procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards [11]. The Netherlands Twin Register [12] study protocol was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board registered with the U.S. Office of Human Research Protections (Institutional Review Board no. IRB00002991, Federal-wide Assurance no. FWA00017598). The study is registered with the Dutch Data Protection Authority (no. m1412317). In accordance with the Declaration of Helsinki, the Netherlands Twin Register obtained informed consent from all participants prior to their entering the study. The Rotterdam Study has been approved by the institutional review board (Medical Ethics Committee) of the Erasmus Medical Center (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). The Rotterdam Study Personal Registration Data collection is filed with the Erasmus MC Data Protection Officer under registration number EMC1712001. The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; <https://www.trialregister.nl/>) under catalogue number Trial NL6645 (NTR6831). All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians [14,21].

Data analysis

All analyses were implemented in an R v4.0.3 [19] workflow employing R packages **renv** v0.14.0 for package management and **drake** v7.13.2 for workflow management. The analyses were run in the BIOS workspace of the SURF Research Cloud which is part of the multi-omics analysis platform of BBMRI-NL. The code to run the analyses is available in GitHub (https://github.com/niehues/bbmri_surrogates_twas) and archived in Zenodo [20].

Data preprocessing

Normalization of values for clinical traits

Numeric values were used for all reported clinical variables. In case of categorical variables, they were binarized as follows. For smoking status, “current smoker” was coded as 1, and “former-smoker” and “non-smoker” were coded as 0; for sex, “male” and “female” were coded as 1 and 0, respectively; for lipid medication, “statins” were coded as 1, and “no” and “yes, but no statins” were coded as 0. In order to be able to compare effect sizes in TWAS, all clinical variables were standardized to zero-mean and unit-variance (z-score normalization).

RNA-seq data preprocessing

Samples with more than 10% missing values in the RNA-seq data were excluded from the analysis. Additionally, for comparisons of models employing either reported or inferred values as outcome, samples missing reported values were excluded. Subsequently, features (transcripts) missing in more than 10% of the samples were removed from the data set. Number of retained samples and features are given in S1 Table.

The RNA-seq read counts as provided by the **BBMRIomics** R package were then normalized and transformed as follows. Scaling factors for library sizes were calculated using the trimmed mean of M-values (TMM) methods [22] implemented in the

R/Bioconductor package `edgeR` v3.32.1 [23]. Using these scaling factors to adjust for sequencing depth, counts were transformed to log₂ counts-per-million (CPM) reads, their mean-variance relationships were estimated using `voom` [24] implemented in the R/Bioconductor package `limma` v3.46.0, and the associated observation-level weights were used in the subsequent linear modeling to adjust for heteroscedasticity.

Metabolomics data preprocessing

The Nightingale Health metabolomics features inquired are the 56 variables selected by van den Akker et al. [17]. Outliers were identified as the samples having more than 1 missing observation (301 removed), more than 1 datapoint under the detection limit (49 removed) and having a value more than 5 standard deviations away from the overall mean observed within BBMRI-NL (0 removed). Remaining with a total of 12926 samples (LLS = 2343, LL = 1475, RS = 5136, NTR = 3972). The remaining 4210 missing values (0.58% of the entire dataset) were imputed as zero and then we z-scaled the metabolic features using the mean and standard deviations observed in BBMRI-NL. The samples were further filtered based on availability of corresponding RNA-seq data. Sample sizes per cohort and model are listed in S1 Table.

Metabolic surrogates

Seventeen logistic regression models trained on up to 22 cohorts, including LL and RS, were applied to the dataset as described in Bizzarri et al. [8]. The surrogates used in this study are the posterior probabilities which represent how likely an individual is at risk for each of the inquired common clinical variables.

Transcriptome-wide association studies (TWAS)

Gene-wise linear regression models were fitted using `limma` [25]. Known potential biological (sex, age) and technical confounders (flow cell number, white blood cell composition) were included in the models. TWAS was performed separately for the respective type of outcome variable, i.e., reported variable or metabolic surrogate. Parameters for each TWAS linear model are summarized in S1 Table. We adjusted *p*-values and effect sizes for statistical bias and inflation using the Bayesian method `bacon` [26], which estimates bias and inflation as parameters from the empirical null distribution of test statistics (*t*-statistic). Additionally, *p*-values were adjusted for multiple testing using the Bonferroni method. TWAS results for two different variable types for the same outcome, were compared by calculating Pearson correlation coefficients of regression coefficients from gene-wise fitted models.

Meta-analysis

Leave-one-cohort-out meta-analyses and replication studies in left out cohort were performed as described in [27].

GSEA

Gene-set enrichment analyses (GSEA) were performed using the R/Bioconductor package `fgsea` v1.16.0 [28] and gene sets retrieved from the Reactome Pathway Database [29]. Genes were ranked by $-\log_{10}(p_b) * |\beta_b|$ with p_b = `bacon`-corrected *p*-value and β_b = `bacon`-corrected effect size. The number of permutations for initial estimation of *p*-values was set to 1×10^4 ; the boundary for calculating *p*-values was set to 1×10^{-50} .

Supporting information

S1 Table. TWAS model parameters. Variable names, sample and feature numbers, and covariates per TWAS model and cohort.

S1 Appendix. Significantly enriched pathways from GSEA. S1 Appendix contains plots showing significantly enriched pathways for each outcome variable. GSEAs are based on meta-analyzed TWAS results.

Acknowledgments

This research was financially supported by EATRIS-Plus (<https://eatris.eu/projects/eatris-plus/>), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 871096, The Netherlands X-omics Initiative (<https://x-omics.nl/>), which is (partly) funded by the Dutch Research Council (NWO), project no. 184.034.019, and by the Biobank-based Integrative Omics Study (BIOS) Consortium (<https://www.bbmri.nl/acquisition-use-analyze/bios>) and the BBMRI Metabolomics Consortium (<https://www.bbmri.nl/Omics-metabolomics>) which are funded by BBMRI-NL, a Research Infrastructure financed by NWO, project nos. 184.021.007 and 184033111. We thank the biobanks and participants of LifeLines (<https://www.lifelines.nl/>), the Leiden Longevity Study (<https://leidenlangleven.nl/>), the Netherlands Twin Register (<https://tweelingenregister.vu.nl/>), and the Rotterdam Study (<http://www.epib.nl/research/ergo.htm>). Special thanks also to Leon Mei, Davy Cats, Martin Brandt and the SURF RSC Team for their support and management of data and computational infrastructure.

Members of the BBMRI-NL BIOS Consortium

Management Team

Bastiaan T. Heijmans (chair)¹, Peter A.C. 't Hoen², Joyce van Meurs³, Rick Jansen⁵, Lude Franke⁶.

Cohort collection

Dorret I. Boomsma⁷, René Pool⁷, Jenny van Dongen⁷, Jouke J. Hottenga⁷ (Netherlands Twin Register); Marleen M.J. van Greevenbroek⁸, Coen D.A. Stehouwer⁸, Carla J.H. van der Kallen⁸, Casper G. Schalkwijk⁸ (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga⁶, Lude Franke⁶, Sasha Zhernakova⁶, Ettje F. Tigchelaar⁶ (LifeLines Deep); P. Eline Slagboom¹, Marian Beekman¹, Joris Deelen¹, Diana van Heemst⁹ (Leiden Longevity Study); Jan H. Veldink¹⁰, Leonard H. van den Berg¹⁰ (Prospective ALS Study Netherlands); Cornelia M. van Duijn⁴, Bert A. Hofman¹¹, Aaron Isaacs⁴, André G. Uitterlinden³ (Rotterdam Study).

Data generation

Joyce van Meurs (Chair)³, P. Mila Jhamai³, Michael Verbiest³, H. Eka D. Suchiman¹, Marijn Verkerk³, Ruud van der Breggen¹, Jeroen van Rooij³, Nico Lakenberg¹.

Data management and computational infrastructure

Hailiang Mei (Chair)¹², Maarten van Iterson¹, Michiel van Galen², Jan Bot¹³, Dasha V. Zhernakova⁶, Rick Jansen⁵, Peter van 't Hof¹², Patrick Deelen⁶, Irene Nooren¹³, Peter A.C. 't Hoen², Bastiaan T. Heijmans¹, Matthijs Moed¹.

Data analysis group

Lude Franke (Co-Chair)⁶, Martijn Vermaat², Dasha V. Zhernakova⁶, René Luijk¹, Marc Jan Bonder⁶, Maarten van Iterson¹, Patrick Deelen⁶, Freerk van Dijk¹⁴, Michiel van Galen², Wibowo Arindrarto¹², Szymon M. Kielbasa¹⁵, Morris A. Swertz¹⁴, Erik. W van Zwet¹⁵, Rick Jansen⁵, Peter-Bram 't Hoen (Co-Chair)², Bastiaan T. Heijmans (Co-Chair)¹.

Affiliations of BIOS members

1 Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

2 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

3 Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands

4 Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands

5 Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

6 Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands

7 Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

8 Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands

9 Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands

10 Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

11 Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands

12 Sequence Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

13 SURFsara, Amsterdam, the Netherlands

14 Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

15 Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Members of the BBMRI-NL Metabolomics Consortium

Cohort collection

J.M. Geleijnse¹, E. Boersma², W.E. van Spil³, M.M.J. van Greevenbroek^{4,5}, C.D.A. Stehouwer^{4,5}, C.J.H. van der Kallen^{4,5}, I.C.W. Arts^{5,6,7}, F. Rutters^{8,9}, J.W.J. Beulens^{8,9}, M. Muilwijk^{8,10}, P.J.M. Elders^{8,10}, L.M. 't Hart^{8,9,11,12}, M. Ghanbari^{13,14}, M.A. Ikram¹³, M.G. Netea¹⁵, M. Kloppenburg^{16,17}, Y.F.M. Ramos¹⁸, N. Bomer¹⁹, I. Meulenbelt¹⁸, K. Stronks²⁰, M.B. Snijder²⁰, A.H. Zwinderman²¹, B.T. Heijmans¹⁸, L.H. Lumey²², C. Wijmenga²³, J. Fu^{23,24}, A. Zhernakova²³, J. Deelen^{25,18}, S.P. Mooijaart²⁶, M. Beekman¹⁸, P.E. Slagboom^{18,25}, G.L.J. Onderwater²⁷, A.M.J.M. van den Maagdenberg^{28,27}, G.M. Terwindt²⁷, C. Thesing^{29,8}, M. Bot^{29,8}, B.W.J.H. Penninx^{29,8}, S. Trompet^{30,26}, J.W. Jukema³⁰, N. Sattar³¹, I.C.C. van der Horst³², P. van der Harst³³, C. So-Osman^{34,35}, J.A. van Hilten³⁶, R.G.H.H. Nelissen³⁷, I.E. Höfer³⁸, F.W. Asselbergs^{39,40}, P. Scheltens⁴¹, C.E. Teunissen⁴², W.M. van der Flier^{43,41}, J. van Dongen^{29,8}, R. Pool²⁹, A.H.M. Willemsen^{29,8}, D.I. Boomsma^{29,8}.

Sample logistics, database and catalogue

H.E.D. Suchiman¹⁸, J.J.H. Barkey Wolf¹⁸, M. Beekman¹⁸, D. Cats⁴⁵, H. Mei⁴⁵, M. Slofstra²³, M. Swertz^{46,23}, M.J.T. Reinders^{47,48}, E.B. van den Akker^{47,18}.

Steering committee

D.I. Boomsma^{29,8}, M.A. Ikram¹³, P.E. Slagboom^{18,25}.

Affiliations of BBMRI-NL Metabolomics members

- 1 Division of Human Nutrition and Health, Wageningen University, Wageningen, The Netherlands
- 2 Thorax centre, Erasmus Medical Centre, Rotterdam, the Netherlands
- 3 Department of Rheumatology & Clinical Immunology, University Medical Center Utrecht, Utrecht, The Netherlands
- 4 Department of Internal Medicine, Maastricht University Medical Center (MUMC+), Maastricht, The Netherlands
- 5 School for Cardiovascular Diseases (CARIM), Maastricht University, Maastricht, the Netherlands
- 6 Department of Epidemiology, Maastricht University, Maastricht, the Netherlands
- 7 Maastricht Center for Systems Biology, Maastricht University, Maastricht, the Netherlands
- 8 Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
- 9 Department of Epidemiology and Biostatistics, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands
- 10 Department of General Practice and Elderly Care Medicine, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands
- 11 Department of Epidemiology and Biostatistics, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands

12 Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, the Netherlands	529 530
13 Department of Epidemiology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands	531 532
14 Department of Genetics, School of Medicine,, Mashhad University of Medical Sciences, Mashhad, Iran	533 534
15 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, The Netherlands	535 536
16 Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands	537 538
17 Department of Rheumatology, Leiden University Medical Center, The Netherlands	539 540
18 Department of Biomedical Data Sciences, Section of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands	541 542
19 Department of Experimental Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands	543 544
20 Department of Public Health, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands	545 546
21 Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands	547 548 549
22 Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032	550 551
23 Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands	552 553
24 Department of Pediatrics, University Medical Center Groningen, Groningen, The Netherlands	554 555
25 Max Planck Institute for Biology of Ageing, Cologne, Germany	556
26 Department of Internal Medicine, Division of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands	557 558
27 Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands	559 560
28 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands	561 562
29 Department of Biological Psychology, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, The Netherlands	563 564
30 Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands	565 566
31 Institute of Cardiovascular and Medical Sciences, Cardiovascular Research Centre, University of Glasgow, Glasgow, UK	567 568

32 Department of Critical Care, University Medical Center Groningen, Groningen, The Netherlands	569 570
33 Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands	571 572
34 Sanquin Blood Bank, Leiden and Department of Haematology, Groene Hart Hospital, Gouda, the Netherlands	573 574
35 International Society of Blood Transfusion (ISBT), Amsterdam, The Netherlands	575 576
36 Unit of Transfusion Medicine, Sanquin Blood Bank, Leiden, The Netherlands	577
37 Department of Orthopaedics, Leiden University Medical Center, Leiden, The Netherlands	578 579
38 Department of Clinical Chemistry and Hematology, UMC Utrecht, the Netherlands	580 581
39 Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands	582 583
40 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands	584 585
41 Department of Neurology & Alzheimer Center, VU University Medical Center, Amsterdam, The Netherlands	586 587
42 Neurochemistry Laboratory, Clinical Chemistry Department, Amsterdam University Medical Center, Amsterdam Neuroscience, The Netherlands	588 589
43 Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands	590 591
44 SURFsara, Amsterdam, the Netherlands	592
45 Sequence Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands	593 594
46 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands	595 596
47 Leiden Computational Biology Center, Leiden University Medical Center, Leiden, the Netherlands	597 598
48 The Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands	599 600

References

1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. 2019;47(D1):D1005–D1012. doi:10.1093/nar/gky1120.
2. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(1):86. doi:10.1186/1471-2105-13-86.

3. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human Molecular Genetics*. 2017;26(R2):R216–R224. doi:10.1093/hmg/ddx275.
4. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*. 2017;9(5):757–768. doi:10.2217/epi-2016-0153.
5. Wang Y, Hannon E, Grant OA, Gorrie-Stone TJ, Kumari M, Mill J, et al. DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC Genomics*. 2021;22(1):484. doi:10.1186/s12864-021-07675-2.
6. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics*. 2019;11(13):1469–1486. doi:10.2217/epi-2019-0206.
7. Schiffman C, McHale CM, Hubbard AE, Zhang L, Thomas R, Vermeulen R, et al. Identification of gene expression predictors of occupational benzene exposure. *PLOS ONE*. 2018;13(10):e0205427. doi:10.1371/journal.pone.0205427.
8. Bizzarri D, Reinders MJT, Beekman M, Slagboom PE, BBMRI-NL, van den Akker EB. ¹H-NMR metabolomics-based surrogates to impute common clinical risk factors and endpoints. *eBioMedicine*. 2022;75:103764. doi:10.1016/j.ebiom.2021.103764.
9. Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*. 2019;55(May 2018):53–60. doi:10.1016/j.semcancer.2018.04.008.
10. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. 2015;5(8):e006772. doi:10.1136/bmjopen-2014-006772.
11. Westendorp RGJ, Van Heemst D, Rozing MP, Frölich M, Mooijaart SP, Blauw GJ, et al. Nonagenarian Siblings and Their Offspring Display Lower Risk of Mortality and Morbidity than Sporadic Nonagenarians: The Leiden Longevity Study. *Journal of the American Geriatrics Society*. 2009;57(9):1634–1637. doi:10.1111/j.1532-5415.2009.02381.x.
12. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JHDA, Draisma HHM, et al. The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. *Twin Research and Human Genetics*. 2013;16(1):271–281. doi:10.1017/thg.2012.140.
13. Hofman A, van Duijn CM, Franco OH, Ikram MA, Janssen HLA, Klaver CCW, et al. The Rotterdam Study: 2012 objectives and design update. *European Journal of Epidemiology*. 2011;26(8):657–686. doi:10.1007/s10654-011-9610-5.
14. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *European Journal of Epidemiology*. 2017;32(9):807–850. doi:10.1007/s10654-017-0321-4.
15. Zhernakova DV, Deelen P, Vermaat M, van Ijcken M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*. 2017;49(1):139–145. doi:10.1038/ng.3737.

16. van Iterson M, Cats D. BBMRIomics: R utilities for BBMRI omics data analysis; R package version 3.4.2. 2020.
17. van den Akker EB, Trompet S, Barkey Wolf JJH, Beekman M, Suchiman HED, Deelen J, et al. Metabolic Age Based on the BBMRI-NL ¹H-NMR Metabolomics Repository as Biomarker of Age-related Disease. *Circulation: Genomic and Precision Medicine*. 2020;13(5):541–547. doi:10.1161/CIRCGEN.119.002610.
18. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circulation: Cardiovascular Genetics*. 2015;8(1):192–206. doi:10.1161/CIRCGENETICS.114.000216.
19. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: <https://www.R-project.org/>.
20. Niehues A. niehues/bbmri_surrogates.twas:v1.0.0 Zenodo. 2022. doi:10.5281/zenodo.5903005.
21. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. *European Journal of Epidemiology*. 2020;35(5):483–517. doi:10.1007/s10654-020-00640-5.
22. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25.
23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–140. doi:10.1093/bioinformatics/btp616.
24. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014;15(2):R29. doi:10.1186/gb-2014-15-2-r29.
25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47–e47. doi:10.1093/nar/gkv007.
26. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*. 2017;18(1):19. doi:10.1186/s13059-016-1131-9.
27. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biology*. 2019;20(1):235. doi:10.1186/s13059-019-1878-x.
28. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov M, Sergushichev A. Fast gene set enrichment analysis. bioRxiv, preprint. 2021;doi:10.1101/060012.
29. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*. 2019;48(D1):D498–D503. doi:10.1093/nar/gkz1031.

30. Holditch SJ, Brown CN, Atwood DJ, Pokhrel D, Brown SE, Lombardi AM, et al. The consequences of increased 4E-BP1 in polycystic kidney disease. *Human Molecular Genetics*. 2019;28(24):4132–4147. doi:10.1093/hmg/ddz244.
31. Collins KS, Eadon MT, Cheng YH, Barwinska D, Ferreira RM, McCarthy TW, et al. Alterations in protein translation and carboxylic acid catabolic processes in diabetic kidney disease. *bioRxiv*, preprint. 2021;doi:10.1101/2021.04.18.440341.
32. Misselbeck K, Parolo S, Lorenzini F, Savoca V, Leonardelli L, Bora P, et al. A network-based approach to identify deregulated pathways and drug effects in metabolic syndrome. *Nature Communications*. 2019;10(1):5215. doi:10.1038/s41467-019-13208-z.
33. Wongdokmai R, Shantavasinkul PC, Chanprasertyothin S, Panpunuan P, Matchariyakul D, Sritara P, et al. The Involvement of Selenium in Type 2 Diabetes Development Related to Obesity and Low Grade Inflammation. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*. 2021;14:1669–1680. doi:10.2147/DMSO.S303146.
34. Sureshchandra S, Raus A, Jankeel A, Ligh BJK, Walter NAR, Newman N, et al. Dose-dependent effects of chronic alcohol drinking on peripheral immune responses. *Scientific Reports*. 2019;9(1):7847. doi:10.1038/s41598-019-44302-3.
35. Feng J, Yang J, Chang Y, Qiao L, Dang H, Luo K, et al. Caffeine-free hawk tea lowers cholesterol by reducing free cholesterol uptake and the production of very-low-density lipoprotein. *Communications Biology*. 2019;2(1):173. doi:10.1038/s42003-019-0396-4.
36. Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, Overvad K, et al. General and Abdominal Adiposity and Risk of Death in Europe. *New England Journal of Medicine*. 2008;359(20):2105–2120. doi:10.1056/NEJMoa0801891.
37. Zierer J, Menni C, Kastenmüller G, Spector TD. Integration of ‘omics’ data in aging research: from biomarkers to systems biology. *Aging Cell*. 2015;14(6):933–944. doi:10.1111/accel.12386.
38. Simpson DJ, Chandra T. Epigenetic age prediction. *Aging Cell*. 2021;20(9):e13452. doi:10.1111/accel.13452.
39. Bhat M, Robichaud N, Hulea L, Sonenberg N, Pelletier J, Topisirovic I. Targeting the translation machinery in cancer. *Nature Reviews Drug Discovery*. 2015;14(4):261–278. doi:10.1038/nrd4505.
40. FitzGerald GA, Oates JA, Nowak J. Cigarette smoking and hemostatic function. *American Heart Journal*. 1988;115(1):267–271. doi:10.1016/0002-8703(88)90648-5.
41. Hioki H. Acute effects of cigarette smoking on platelet-dependent thrombin generation. *European Heart Journal*. 2001;22(1):56–61. doi:10.1053/euhj.1999.1938.

Multi-cohort multi-omics data set

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5