Niehues *et al. BMC Genomics* (2022) 23:546
https://doi.org/10.1186/s12864-022-08771-7

BMC Genomics

RESEARCH Open Access

Metabolomic predictors of phenotypic traits can replace and complement measured clinical variables in population-scale expression profiling studies

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

Abstract

Population-scale expression profiling studies can provide valuable insights into biological and disease-underlying mechanisms. The availability of phenotypic traits is essential for studying clinical effects. Therefore, missing, incomplete, or inaccurate phenotypic information can make analyses challenging and prevent RNA-seq or other omics data to be reused. A possible solution are predictors that infer clinical or behavioral phenotypic traits from molecular data. While such predictors have been developed based on different omics data types and are being applied in various studies, metabolomics-based surrogates are less commonly used than predictors based on DNA methylation profiles. In this study, we inferred 17 traits, including diabetes status and exposure to lipid medication, using previously trained metabolomic predictors. We evaluated whether these metabolomic surrogates can be used as an alternative to reported information for studying the respective phenotypes using expression profiling data of four population cohorts. For the majority of the 17 traits, the metabolomic surrogates performed similarly to the reported phenotypes in terms of effect sizes, number of significant associations, replication rates, and significantly enriched pathways. The application of metabolomics-derived surrogate outcomes opens new possibilities for reuse of multi-omics data sets. In studies 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. Additionally, the availability of such surrogates could be used to correct for potential biological confounding. In the future, it would be interesting to further investigate the use of molecular predictors across different omics types and cohorts.

Keywords: Multi-omics, Metabolomics, Transcriptomics, Surrogates, Predictors, Expression profiling, Population cohort study, Meta-analysis, Clinical surrogates, Surrogate outcomes

*Correspondence: Peter-Bram.tHoen@radboudumc.nl
[1]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Geert Grooteplein Zuid 26-28, 6525 GA, Nijmegen, Netherlands
Full list of author information is available at the end of the article

Background

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 population-scale studies of transcriptomics, epigenomics, and other omics data types.

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].

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 ome-wide association studies often have only moderate effect sizes, a common approach to detect relevant features are cross-cohort meta-analyses [8]. 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. Blood is a key specimen in clinical diagnostics reflecting on the health state of an individual. While blood metabolomics methods partially overlap with classical clinical diagnostics methods, they can measure a wide range of metabolites and have the potential to play an important role in personalized medicine approaches [9]. Recently, Bizzarri et al. [10] 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 these surrogates in metabolome-wide association studies to complement missing clinical data and correct for confounding. They further showed that metabolomic surrogates can help explore independent risk factors of all-cause mortality in older individuals [10].

We here propose the use of metabolomics-derived surrogates in analyses of other omics layers, not only as covariates to account for confounding factors, but also as outcome variables. We 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 population-scale gene expression profile association studies. To this end, we applied 17 metabolomic predictors to metabolomics data from four large population cohorts inferring values for phenotypic, exposure, and clinical traits. We performed association studies 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 metabolomic surrogates based on results reported in literature.

Results

In this study, we performed expression profiling studies to evaluate the performance of 17 surrogate outcomes that are based on molecular predictors. Metabolomics data used to infer surrogate outcome values and RNA-seq data used for the evaluation are part of multi-omics data sets of four large Dutch population cohorts: LifeLines (LL) [11], Leiden Longevity Study (LLS) [12], Netherlands Twin Register (NTR) [13], and Rotterdam Study (RS) [14]. An overview of the study workflow is shown in Fig. 1. For the different outcomes (Table 1), we compared expression profile association study results based on metabolomic surrogate outcomes to those based on reported outcomes whenever possible. Additionally, results were compared to literature findings.

Metabolomic surrogate outcomes

We inferred values for clinical variables by applying molecular predictors to metabolomics data. Recently reported metabolomic predictors trained on up to 22 Dutch population cohorts [10] were applied to infer values for outcome variables (Table 1). All 17 metabolomic predictors had been shown to perform accurately with mean

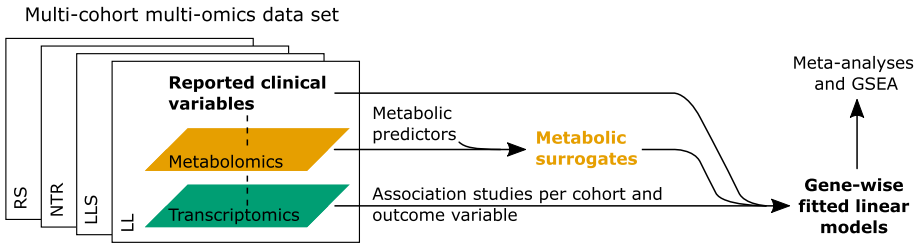

Fig. 1 Overview of study workflow. Gene-wise models are fitted for various outcome variables based on reported information or metabolomic surrogates, respectively

AUC values $>$ 0.7 in a 5-fold cross-validation approach [10]. In order to avoid emphasis on clinical extremes, the metabolomic 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 expression profiling studies to reported or measured outcome values.

Expression profile association studies

In the next step, the 17 metabolomic surrogates were used as outcomes of expression profile association studies. 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 Additional file 1).

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` [15] (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 association study 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 metabolomic surrogate outcomes in two cases, and lower in 10 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 metabolomic 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

Table 1 Overview of phenotypic traits. Availability of variable is indicated by 'x'

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 ($\geq$ 65 y.o.)	x	x

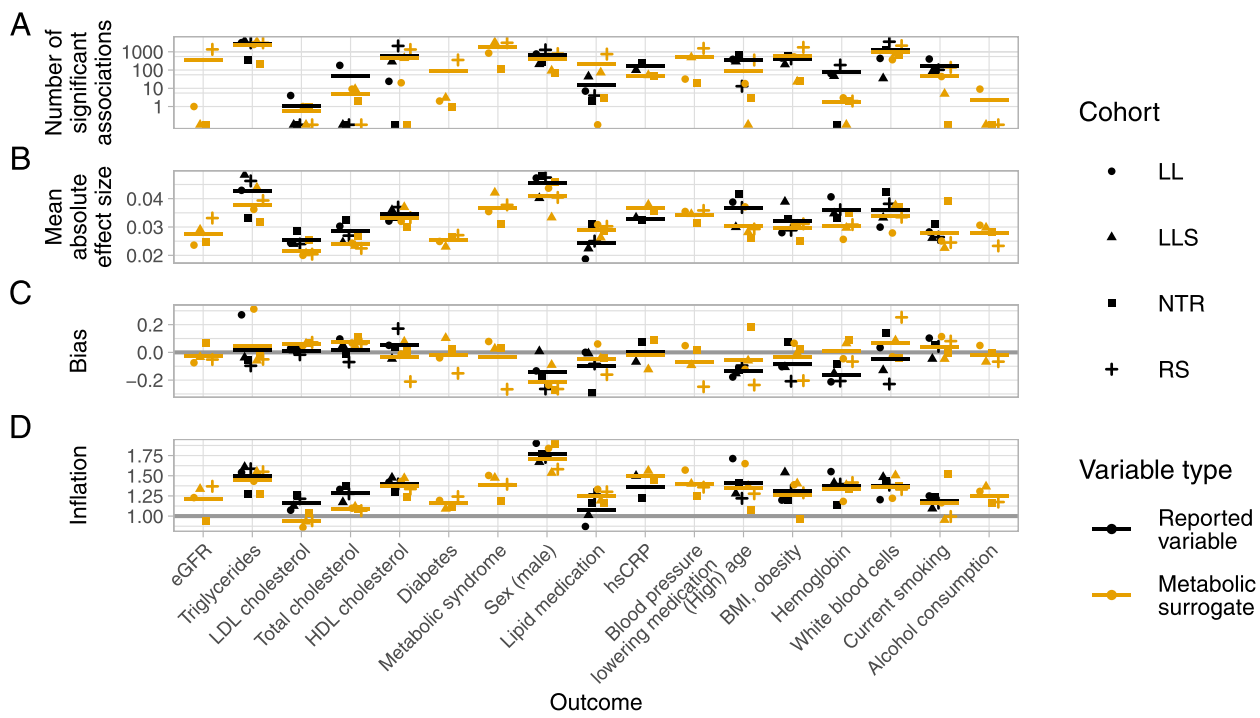

Fig. 2 Comparison of association study result characteristics. Number of significant associations (based on `bacon`-corrected and FDR-corrected *p*-values $p_{b_{adj}} < 0.05$) (A), mean absolute effect sizes (based on `bacon`-corrected effect sizes) across all genes (B), and bias (C) and inflation (D) 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 metabolomic predictors for binary outcome measures. Type of outcome variable is indicated by color: reported or measured variable = black, metabolomic 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

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 metabolomic 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 [16], 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 metabolomic surrogate outcomes instead of reported variables; for the outcomes lipid medication and hsCRP, it slightly increased.

Since the number of significant associations and average effect sizes do not allow drawing conclusions about the similarity of the association study 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. Figure 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 metabolomic predictors. For outcomes based on predictors with reported AUC > 0.9 (triglycerides, LDL cholesterol, total cholesterol, HDL cholesterol, and sex) [10], 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 the highest correlation values, although the predictor's AUC was reported to be lower than those for triglycerides or cholesterol. Lowest similarity of results with average absolute Pearson $r < 0.5$ were observed for the outcomes age, white blood cells, smoking, and hemoglobin, the latter having the lowest correlation values. We often observed that correlations were lower for the NTR cohort. This could be explained by a technical difference in the metabolomic profiles, with NTR missing glutamine [10].

Meta-analyses and replication studies

When comparing two alternative outcome variables, a lower or higher number of found significant associa-

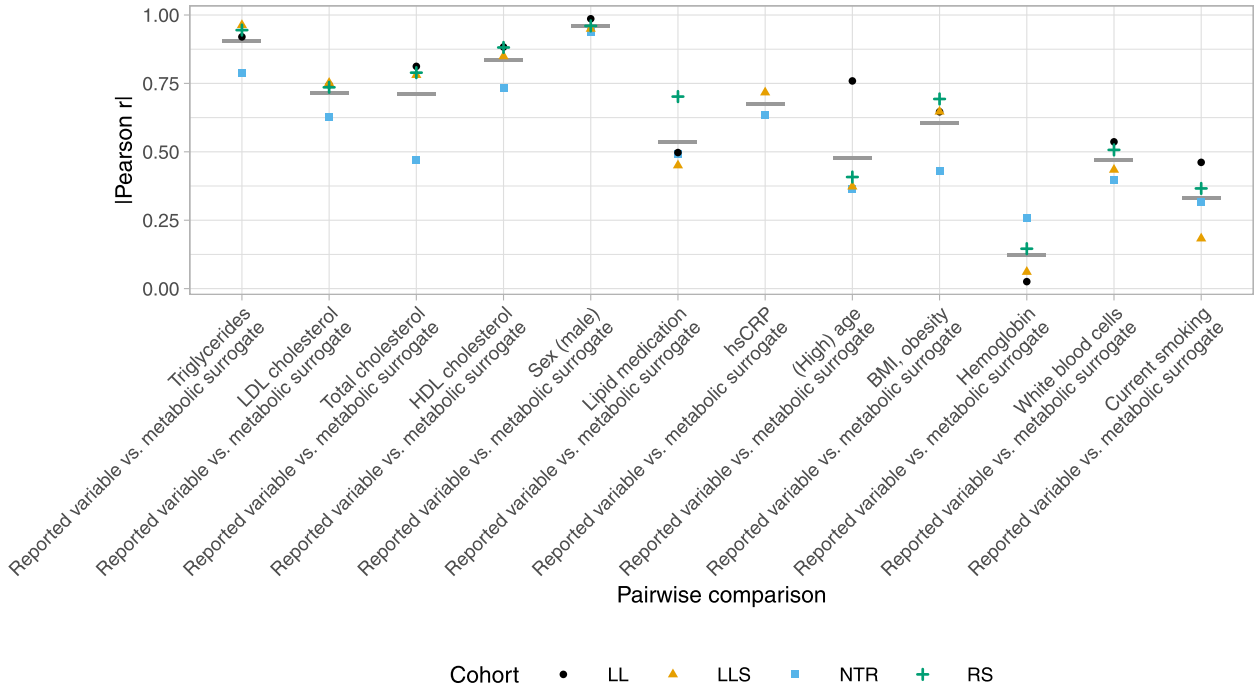

Fig. 3 Pairwise comparisons of association study results. Absolute Pearson correlation coefficients (Pearson *r*) of `bacon`-adjusted regression coefficients of gene-wise linear models (`limma/voom`) for outcome variables in alternative models per comparison and cohort. Comparisons are ordered by performance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

tions 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 expression profile association study 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. [16]. For each comparison, four meta-analyses were performed leaving one cohort out each time, and using the left out cohort for a replication analysis. Figure 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 metabolomic surrogate for lipid medication, the reported variable yielded more meta-analyzed genes for age and smoking. For the latter two outcomes, results based on metabolomic surrogates could not be replicated (Fig. 4B) while on average 30-38% of results based on reported outcomes could be replicated. For these two outcomes we had also observed the 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. Highest average replication rates were observed for triglycerides with 69% for the reported and 67% for the surrogate outcome. 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 (except for reported hemoglobin) in the individual cohorts (Fig. 2).

Gene set enrichment analysis

To arrive at a biological interpretation of the association study results, we performed gene-set enrichment analyses (GSEA) using pathways from the Reactome database. GSEA was applied to both individual cohort

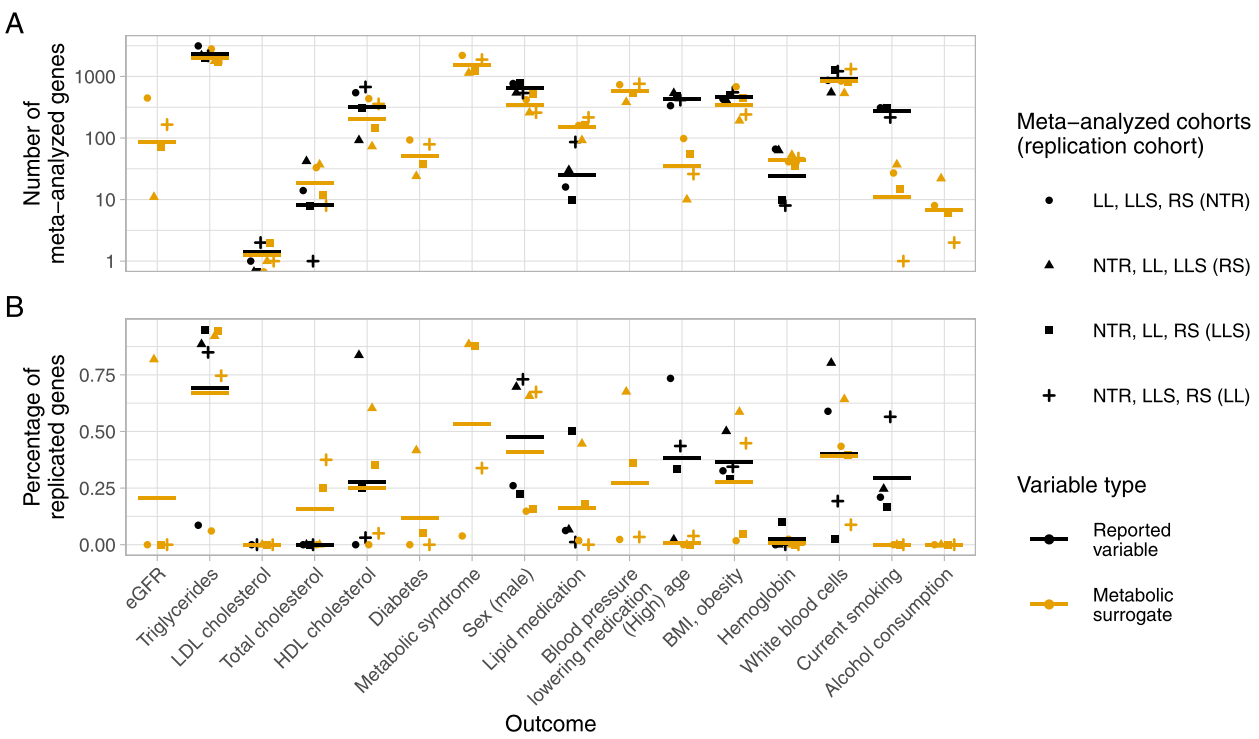

Fig. 4 Meta-analyses and replication studies. Number of meta-analyzed genes (significant associations, `bacon`-adjusted *p*-values FDR-adjusted for multiple testing, $p < 0.05$) in leave-one-cohort meta-analyses (A) and percentage of genes replicated (significant associations, Bonferroni-adjusted for multiple testing, $p < 0.05$) in replication cohort (B). Type of outcome variable is indicated by color: reported or measured variable = black, metabolomic 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

results and results from a meta-analysis of all four cohorts (Fig. 5). For direct comparisons of reported variables and metabolomic 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 Additional file 2). 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 association studies (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.

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 Additional file 2) 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 [17, 18]. 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 [19, 20]. For alcohol consumption, although almost no significant associations were found in individual analyses, several significant gene-outcome associations were found when meta-analyzing multiple cohorts (compare Figs. 2 and 4). Top-ranked positively enriched Reactome pathways from

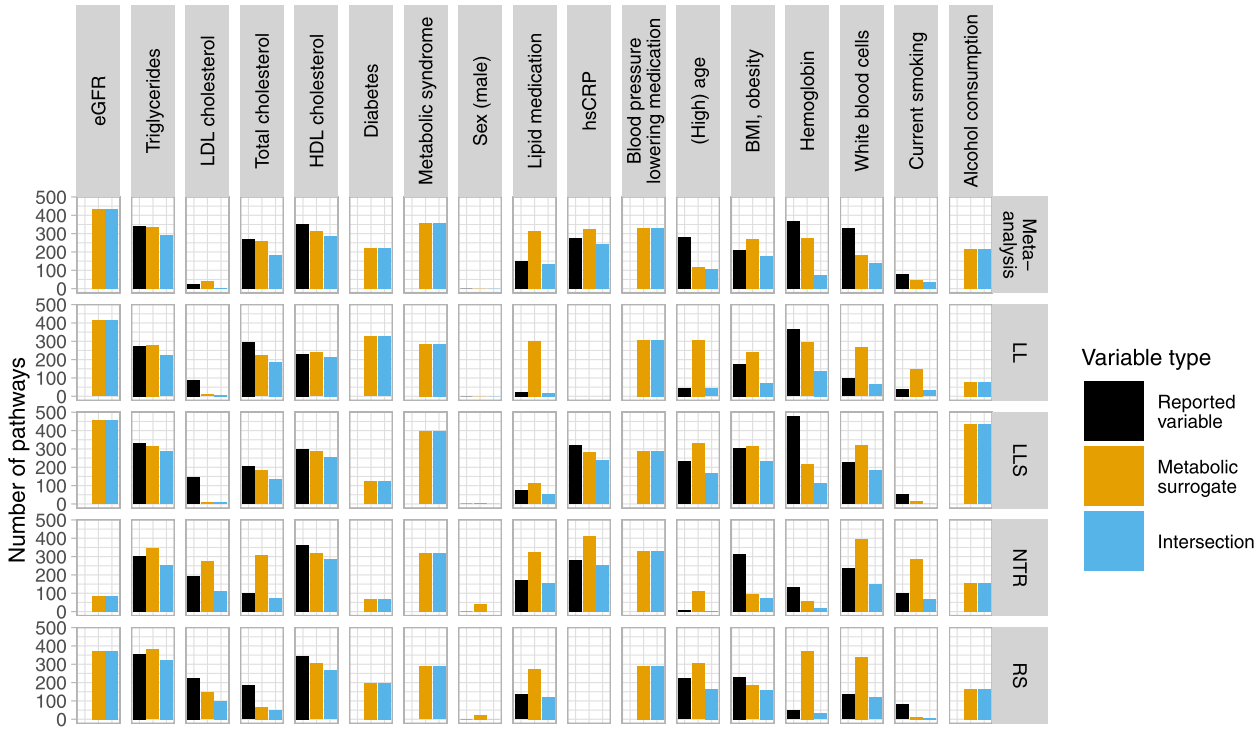

Fig. 5 Gene-set enrichment analyses of association study results. Numbers of significantly enriched (Bonferroni-adjusted $p < 0.05$) pathways (Reactome) for each outcome found in each cohort and in meta-analysis of all four (two for hsCRP) cohorts (top). Values for each type of outcome variable are represented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

gene-set enrichment analysis (Additional file 2), including, e.g., innate immune system, signal transduction, and infectious disease, have been linked previously to chronic alcohol drinking [21].

Discussion

While certain omics predictors especially based on DNA methylation profiles [2–5] are regularly applied in (multi) omics data analyses, metabolomics-based predictors are not commonly used in the analysis of other omics data types. Previously, Bizzarri et al. had shown that metabolomic surrogates can be used to correct for confounding in metabolome-wide association studies [10]. In this study, we investigated the use of these surrogates as outcome variables in the analysis of another omics level. We systematically compared results of population-scale gene expression profile association studies against outcome variables that were either reported or inferred by molecular predictors. The results generally showed good agreement (Fig. 2). Most similar association study 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 [22] and were previously shown to be enriched in a high-cholesterol and high-fat diet induced hypercholesterolemic rat model [23]. The metabolomic predictors for these outcomes are directly related to metabolic markers measured on the Nightingale platform and had shown high performance (AUC $\geq$ 0.95) [10]. 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 metabolomic surrogates for molecular predictors that were known to have lower accuracy (Fig. 3). In the association studies 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 metabolomic 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 [24]. 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 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 Additional file 2). Additionally, the application of metabolomic surrogates for outcomes that were not reported in the data and the comparison of association studies and GSEA results with literature show that these surrogate outcomes allow transferring 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. These results suggest that metabolomic surrogates are a useful tool to complement phenotypic information of multi-omics data sets and enable analyses of clinical outcomes even when they are not reported. This is especially useful when reanalyzing existing data sets. Even though this study comprises four different large population cohorts and 17 metabolomic surrogates, it will be interesting to investigate in the future whether similar results can be observed in other cohorts, for other clinical variables, or for other omics data types.

For both outcome variable types, only few gene-trait associations were significantly associated with the outcome LDL cholesterol. Similarly, low numbers of significant associations were observed for hemoglobin for the surrogate outcome. In contrast to that, the reported outcome yielded more associations which, however, could not be replicated. For high age and current smoking slightly fewer associations were found when using the surrogate outcomes. They 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 metabolomic 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 [25], but the metabolomic predictor for high age trained on binarized data ($\geq$ 65 y.o.) might not be an ideal surrogate to study this. Alternatively, a metabolomics-based biological age predictor based on continuous data [26] might perform better. Differences between GSEA results of different outcome variable types could also arise from different biological information captured by the metabolomic 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 [27]). While many of the top-ranked pathways (Additional file 2) for smoking were found by both outcome variable types, some pathways were solely found by using either reported smoking status ("smoking_current") or metabolomic 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 [28]. Pathways only enriched when using the metabolomic surrogate include "Platelet activation, signaling and aggregation" and "Hemostasis". Increased platelet aggregation has been reported in smokers [29] and platelet-dependent thrombin levels were shown to be increased in smokers and following smoking [30]. This possibly indicates that the reported smoking behavior captures effects of long-term exposure to smoking better, while the metabolomic surrogate captures effects of acute smoking. It would be interesting to further investigate which aspects of the clinical phenotypes are captured by the metabolomic 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.

Conclusions

In our systematic comparison of expression profiling results using either reported variables as outcome or surrogate outcomes inferred from metabolomics profiles, we demonstrated that in many cases metabolomic sur-

rogates yield similar results as reported variables. We showed that 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) [11], Leiden Longevity Study (LLS) [12], Netherlands Twin Register (NTR) [13], and Rotterdam Study (RS) [14, 31]. The data is provided by the Dutch node of the European Biobanking and BioMolecular Resources and Research Infrastructure (BBMRI-NL).

RNA-seq data of all four cohorts 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. [32]. 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 [33].

Metabolomics data was generated by the BBMRI-NL Metabolomics Consortium in 2014 as described by van den Akker et al. [26]. Briefly, metabolite concentrations were measured in EDTA plasma by proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy on the platform of the Nightingale Health Group (Helsinki, Finland) [34].

Data analysis

All analyses were implemented in an R v4.0.3 [35] 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 in available in GitHub [36] and archived in Zenodo [37].

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 association studies, 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 Additional file 1.

The RNA-seq read counts as provided by the `BBMRIomics` R package were then normalized and transformed based on a previous evaluation of analysis strategies [16] as follows. Scaling factors for library sizes were calculated using the trimmed mean of M-values (TMM) method [38] implemented in the R/Bioconductor package `edgeR` v3.32.1 [39]. Using these scaling factors to adjust for sequencing depth, counts were transformed to $\log_2$ counts-per-million (CPM) reads, their mean-variance relationships were estimated using `voom` [40] 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. [26]. Outliers were identified as the samples having more than 1 missing observation (301 removed), more than 1 data point 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 the metabolomic features were z-scaled 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 Additional file 1.

Metabolomic 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. [10]. 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.

Expression profile association studies

To determine association of profiles with clinical outcomes, gene-wise linear regression models were fitted using `limma` [41]. Known potential biological (sex, age) and technical confounders (flow cell number, white blood cell composition) were included in the models. Association studies were performed separately for the respective type of outcome variable, i.e., reported variable or metabolomic surrogate. Parameters for each linear model are summarized in Additional file 1. We adjusted p-values and effect sizes for statistical bias and inflation using the Bayesian method `bacon` [15], 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 false discovery rate (FDR) [42, 43]. 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 [16].

GSEA

Gene-set enrichment analyses (GSEA) were performed using the R/Bioconductor package `fgsea` v1.16.0 [44] and gene sets retrieved from the Reactome Pathway Database [22]. Genes were ranked by $-\log_{10}(p_b) * |\beta_b|$ with p_b = `bacon`-corrected p-value and $beta_b$ = `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}.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08771-7.

Additional file 1: TWAS model parameters. This .csv file contains variable names, sample and feature numbers, and covariates per TWAS model and cohort.

Additional file 2: Significantly enriched pathways from GSEA. This .html file contains plots showing significantly enriched pathways for each outcome variable. GSEAs are based on meta-analyzed TWAS results.

Acknowledgements

We thank the biobanks and participants of LifeLines [53], the Leiden Longevity Study [54], the Netherlands Twin Register [55], and the Rotterdam Study [56]. 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.

BBMRI-NL BIOS consortium

Management Team Bastiaan T. Heijmans (chair)[3], Peter A.C. 't Hoen[7], Joyce van Meurs[8], Rick Jansen[10], Lude Franke[11].

Cohort collection Dorret I. Boomsma[12], RenÉ Pool[12], Jenny van Dongen[12], Jouke J. Hottenga[12] (Netherlands Twin Register); Marleen M.J. van Greevenbroek[13], Coen D.A. Stehouwer[13], Carla J.H. van der Kallen[13], Casper G. Schalkwijk[13] (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga[11], Lude Franke[11], Sasha Zhernakova[11], Ettje F. Tigchelaar[11] (LifeLines Deep); P. Eline Slagboom[3], Marian Beekman[3], Joris Deelen[3], Diana van Heemst[14] (Leiden Longevity Study); Jan H. Veldink[15], Leonard H. van den Berg[15] (Prospective ALS Study Netherlands); Cornelia M. van Duijn[9], Bert A. Hofman[16], Aaron Isaacs[9], AndrÉ G. Uitterlinden[8] (Rotterdam Study).

Data generation Joyce van Meurs (Chair)[8], P. Mila Jhamai[8], Michael Verbiest[8], H. Eka D. Suchiman[3], Marijn Verkerk[8], Ruud van der Breggen[3], Jeroen van Rooij[8], Nico Lakenberg[3].

Data management and computational infrastructure Hailiang Mei (Chair)[17], Maarten van Iterson[3], Michiel van Galen[7], Jan Bot[18], Dasha V. Zhernakova[11], Rick Jansen[10], Peter van 't Hof[17], Patrick Deelen[11], Irene Nooren[18], Peter A.C. 't Hoen[7], Bastiaan T. Heijmans[3], Matthijs Moed[3].

Data analysis group Lude Franke (Co-Chair)[11], Martijn Vermaat[7], Dasha V. Zhernakova[11], RenÉ Luijk[3], Marc Jan Bonder[11], Maarten van Iterson[3], Patrick Deelen[11], Freerk van Dijk[19], Michiel van Galen[7], Wibowo Arindrarto[17], Szymon M. Kielbasa[20], Morris A. Swertz[19], Erik. W van Zwet[20], Rick Jansen[10], Peter-Bram 't Hoen (Co-Chair)[7], Bastiaan T. Heijmans (Co-Chair)[3].

Affiliations of BIOS members [3] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [7] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [8] Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands. [9] Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands. [10] Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands. [11] Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands. [12] Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands. [13] Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands. [14] Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands. [15] Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands. [16] Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands. [17] Sequence Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [18] SURFsara, Amsterdam, the Netherlands. [19] Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [20] Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

BBMRI-NL Metabolomics consortium

Cohort collection J.M. Geleijnse[21], E. Boersma[22], W.E. van Spil[23], M.M.J. van Greevenbroek[24,25], C.D.A. Stehouwer[24,25], C.J.H. van der Kallen[24,25], I.C.W. Arts[6,26,27], F. Rutters[28,29], J.W.J. Beulens[28,29], M. Muilwijk[28,30], P.J.M. Elders[28,30], L.M. 't Hart[28,29,31], M. Ghanbari[16,32], M.A. Ikram[16], M.G. Netea[33], M. Kloppenburg[34,35], Y.F.M. Ramos[3], N. Bomer[36], I. Meulenbelt[3], K. Stronks[37], M.B. Snijder[37], A.H. Zwinderman[38], B.T. Heijmans[3], L.H. Lumey[39], C. Wijmenga[40], J. Fu[40,41], A. Zhernakova[40], J. Deelen[6,3], S.P. Mooijaart[42], M. Beekman[3], P.E. Slagboom[3,6], G.L.J. Onderwater[43], A.M.J.M. van den Maagdenberg[7,43], G.M. Terwindt[43], C.Thesing[44,28], M. Bot[44,28], B.W.J.H. Penninx[44,28], S. Trompet[45,42], J.W. Jukema[45], N. Sattar[46], I.C.C. van der Horst[47], P. van der Harst[48], C. So-Osman[49,50], J.A. van Hilten[51], R.G.H.H. Nelissen[52], I.E. Höfer[53], F.W. Asselbergs[54,55], P. Scheltens[56], C.E. Teunissen[57], W.M. van der Flier[58,56], J. van Dongen[44,28], R. Pool[44], A.H.M. Willemsen[44,28], D.I. Boomsma[44,28].

Sample logistics, database and catalogue H.E.D. Suchiman[3], J.J.H. Barkey Wolf[3], M. Beekman[3], D. Cats[17], H. Mei[17], M. Slofstra[40], M. Swertz[19,40], M.J.T. Reinders[4,5], E.B. van den Akker[4,3].

Steering committee D.I. Boomsma[44,28], M.A. Ikram[16], P.E. Slagboom[3,6].

Affiliations of BBMRI-NL Metabolomics members [3] Department of Biomedical Data Sciences, Section of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [4] Leiden Computational Biology Center, Leiden University Medical Center, Leiden, the Netherlands. [5] The Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands. [6] Max Planck Institute for Biology of Ageing, Cologne, Germany. [7] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [16] Department of Epidemiology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands. [17] Sequence Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands. [18] SURFsara, Amsterdam, the Netherlands. [19] University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [21] Division of Human Nutrition and Health, Wageningen University, Wageningen, The Netherlands. [22] Thorax centre, Erasmus Medical Centre, Rotterdam, the Netherlands. [23] Department of Rheumatology & Clinical Immunology, University Medical Center Utrecht, Utrecht, The Netherlands. [24] Department of Internal Medicine, Maastricht University Medical Center (MUMC+), Maastricht, The Netherlands. [25] School for Cardiovascular Diseases (CARIM), Maastricht University, Maastricht, the Netherlands. [26] Department of Epidemiology, Maastricht University, Maastricht, the Netherlands. [27] Maastricht Center for Systems Biology, Maastricht University, Maastricht, the Netherlands. [28] Amsterdam Public Health Research Institute, Amsterdam, The Netherlands. [29] Department of Epidemiology and Biostatistics, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands. [30] Department of General Practice and Elderly Care Medicine, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands. [31] Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, the Netherlands. [32] Department of Genetics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. [33] Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, The Netherlands. [34] Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands. [35] Department of Rheumatology, Leiden University Medical Center, The Netherlands. [36] Department of Experimental Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [37] Department of Public Health, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [38] Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands. [39] Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032. [40] Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands. [41] Department of Pediatrics, University Medical Center Groningen, Groningen, The Netherlands. [42] Department of Internal Medicine, Division of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands. [43] Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands. [44] Department of Biological Psychology, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, The Netherlands. [45] Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands. [46] Institute of Cardiovascular and Medical Sciences, Cardiovascular Research Centre, University of Glasgow, Glasgow, UK. [47] Department of Critical Care, University Medical Center Groningen, Groningen, The Netherlands. [48] Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands. [49] Sanquin Blood Bank, Leiden and Department of Haematology, Groene Hart Hospital, Gouda, the Netherlands. [50] International Society of Blood Transfusion (ISBT), Amsterdam, The Netherlands. [51] Unit of Transfusion Medicine, Sanquin Blood Bank, Leiden, The Netherlands. [52] Department of Orthopaedics, Leiden University Medical Center, Leiden, The Netherlands. [53] Department of Clinical Chemistry and Hematology, UMC Utrecht, the Netherlands. [54] Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands. [55] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [56] Department of Neurology & Alzheimer Center, VU University Medical Center, Amsterdam, The Netherlands. [57] Neurochemistry Laboratory, Clinical Chemistry Department, Amsterdam University Medical Center, Amsterdam Neuroscience, The Netherlands. [58] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands.

Authors' contributions

All authors (AN, DB, MJTR, PES, AJG, EBA, and PACH) developed the concept for this study (Conceptualization), designed the applied methodology (Methodology), and critically reviewed the initial draft of the manuscript (Writing – Review & Editing). AJG and PACH acquired funding for this study (Funding Acquisition). AN conducted data preprocessing, association studies, meta-analyses, and GSEA (Investigation), implemented corresponding code (Software), created visualizations of study design and results (Visualization), and wrote the initial draft of this manuscript (Writing – Original Draft). DB conducted metabolomics data preprocessing and inferred metabolomics surrogates (Investigation) and implemented corresponding code (Software). PACH coordinated planning and execution of the research activity (Project Administration). All authors read and approved the final manuscript.

Funding

AJG and PACH received funding from EATRIS-Plus [49], which has received funding from the European Union's Horizon 2020 research and innovation programme [50] under grant agreement no. 871096, and The Netherlands X-omics Initiative [51], which is (partly) funded by the Dutch Research Council [52], project no. 184.034.019. The Biobank-based Integrative Omics Study (BIOS) Consortium [47] and the BBMRI Metabolomics Consortium [48] are funded by BBMRI-NL, a Research Infrastructure financed by NWO, project nos. 184.021.007 and 184033111. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The data underlying the results presented in the study are available from BBMRI-NL and access can be requested via the BIOS Consortium [47] and the BBMRI Metabolomics Consortium [48]. The code used to perform analyses for this study is available in GitHub [36] and archived in Zenodo [37].

Declarations

Ethics approval and consent to participate

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 [11]. 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 [12]. The Netherlands Twin Register [13] 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; [45]) 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 [31, 46].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Geert Grooteplein Zuid 26-28, 6525 GA, Nijmegen, Netherlands. [2]Translational Metabolic Laboratory, Department Laboratory Medicine, Radboud university medical center, Geert Grooteplein Zuid 10, 6525 GA, Nijmegen, Netherlands. [3]Molecular Epidemiology, LUMC, Einthovenweg 20, 2333 ZC, Leiden, Netherlands. [4]Leiden Computational Biology Center, LUMC, Einthovenweg 20, 2333 ZC, Leiden, Netherlands. [5]Delft Bioinformatics Lab, TU Delft, Van Mourik Broekmanweg 6, 2628 XE, Delft, Netherlands. [6]Max Planck Institute for the Biology of Ageing, Cologne, Germany.

Received: 17 March 2022 Accepted: 12 July 2022
Published online: 31 July 2022

References

1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):1005–12. https://doi.org/10.1093/nar/gky1120.
2. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):86. https://doi.org/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. Hum Mol Genet. 2017;26(R2):216–24. https://doi.org/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):757–68. https://doi.org/10.2217/epi-2016-0153.
5. Wang Y, Hannon E, Grant OA, Gorrie-Stone TJ, Kumari M, Mill J, Zhai X, McDonald-Maier KD, Schalkwyk LC. DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. BMC Genomics. 2021;22(1):484. https://doi.org/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):1469–86. https://doi.org/10.2217/epi-2019-0206.
7. Schiffman C, McHale CM, Hubbard AE, Zhang L, Thomas R, Vermeulen R, Li G, Shen M, Rappaport SM, Yin S, Lan Q, Smith MT, Rothman N. Identification of gene expression predictors of occupational benzene exposure. PLoS ONE. 2018;13(10):0205427. https://doi.org/10.1371/journal.pone.0205427.
8. Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. Semin Cancer Biol. 2019;55(May 2018):53–60. https://doi.org/10.1016/j.semcancer.2018.04.008.
9. Li S, Todor A, Luo R. Blood transcriptomics and metabolomics for personalized medicine. Comput Struct Biotechnol J. 2016;14:1–7. https://doi.org/10.1016/j.csbj.2015.10.005.
10. Bizzarri D, Reinders MJT, Beekman M, Slagboom PE, BBMRI-NL, van den Akker EB. [1]H-NMR metabolomics-based surrogates to impute common clinical risk factors and endpoints. eBioMedicine. 2022;75:103764. https://doi.org/10.1016/j.ebiom.2021.103764.
11. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, Swertz MA, Muñoz AM, Deelen P, Cénit MC, Franke L, Scholtens S, Stolk RP, Wijmenga C, Feskens EJM. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5(8):006772. https://doi.org/10.1136/bmjopen-2014-006772.
12. Westendorp RGJ, Van Heemst D, Rozing MP, Frölich M, Mooijaart SP, Blauw G-J, Beekman M, Heijmans BT, De Craen AJM, Slagboom PE. Nonagenarian Siblings and Their Offspring Display Lower Risk of Mortality and Morbidity than Sporadic Nonagenarians: The Leiden Longevity Study. J Am Geriatr Soc. 2009;57(9):1634–7. https://doi.org/10.1111/j.1532-5415.2009.02381.x.
13. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JHDA, Draisma HHM, van Dongen J, van 't Ent D, Geels LM, van Lien R, Ligthart L, Kattenberg M, Mbarek H, de Moor MHM, Neijts M, Pool R, Stroo N, Kluft C, Suchiman HED, Slagboom PE, de Geus EJC, Boomsma DI. The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. Twin Res Hum Genet. 2013;16(1):271–81. https://doi.org/10.1017/thg.2012.140.
14. Hofman A, van Duijn CM, Franco OH, Ikram MA, Janssen HLA, Klaver CCW, Kuipers EJ, Nijsten TEC, Stricker BHC, Tiemeier H, Uitterlinden AG, Vernooij MW, Witteman JCM. The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol. 2011;26(8):657–86. https://doi.org/10.1007/s10654-011-9610-5.
15. 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 Biol. 2017;18(1):19. https://doi.org/10.1186/s13059-016-1131-9.
16. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, Franke L, 't Hoen PAC, Heijmans B, van Meurs JBJ. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. Genome Biol. 2019;20(1):235. https://doi.org/10.1186/s13059-019-1878-x.
17. Holditch SJ, Brown CN, Atwood DJ, Pokhrel D, Brown SE, Lombardi AM, Nguyen KN, Hill RC, Lanaspa M, Hopp K, Weiser-Evans MCM, Edelstein CL. The consequences of increased 4E-BP1 in polycystic kidney disease. Hum Mol Genet. 2019;28(24):4132–47. https://doi.org/10.1093/hmg/ddz244.
18. Collins KS, Eadon MT, Cheng Y-H, Barwinska D, Ferreira RM, Mccarthy TW, Janosevic D, Syed F, Maier B, El-Achkar TM, Kelly KJ, Phillips CL, Hato T, Sutton TA, Dagher PC. Alterations in protein translation and carboxylic acid catabolic processes in diabetic kidney disease. bioRxiv, preprint. 2021. https://doi.org/10.1101/2021.04.18.440341.
19. Misselbeck K, Parolo S, Lorenzini F, Savoca V, Leonardelli L, Bora P, Morine MJ, Mione MC, Domenici E, Priami C. A network-based approach to identify deregulated pathways and drug effects in metabolic syndrome. Nat Commun. 2019;10(1):5215. https://doi.org/10.1038/s41467-019-13208-z.
20. Wongdokmai R, Shantavasinkul PC, Chanprasertyothin S, Panpunuan P, Matchariyakul D, Sritara P, Sirivarasai J. The Involvement of Selenium in Type 2 Diabetes Development Related to Obesity and Low Grade Inflammation. Diabetes Metab Syndr Obes Targets Ther. 2021;14:1669–80. https://doi.org/10.2147/DMSO.S303146.
21. Sureshchandra S, Raus A, Jankeel A, Ligh BJK, Walter NAR, Newman N, Grant KA, Messaoudi I. Dose-dependent effects of chronic alcohol drinking on peripheral immune responses. Sci Rep. 2019;9(1):7847. https://doi.org/10.1038/s41598-019-44302-3.
22. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2019;48(D1):498–503. https://doi.org/10.1093/nar/gkz1031.
23. Feng J, Yang J, Chang Y, Qiao L, Dang H, Luo K, Guo H, An Y, Ma C, Shao H, Tian J, Yuan Y, Xie L, Xing W, Cheng J. Caffeine-free hawk tea lowers cholesterol by reducing free cholesterol uptake and the production of very-low-density lipoprotein. Commun Biol. 2019;2(1):173. https://doi.org/10.1038/s42003-019-0396-4.
24. Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, Overvad K, van der Schouw YT, Spencer E, Moons KGM, Tjønneland A, Halkjaer J, Jensen MK, Stegger J, Clavel-Chapelon F, Boutron-Ruault M-C, Chajes V, Linseisen J, Kaaks R, Trichopoulou A, Trichopoulos D, Bamia C, Sieri S, Palli D, Tumino R, Vineis P, Panico S, Peeters PHM, May AM, Bueno-de-Mesquita HB, van Duijnhoven FJB, Hallmans G, Weinehall L, Manjer J, Hedblad B, Lund E, Agudo A, Arriola L, Barricarte A, Navarro C, Martinez C, Quirós JR, Key T, Bingham S, Khaw KT, Boffetta P, Jenab M, Ferrari P, Riboli E. General and Abdominal Adiposity and Risk of Death in Europe. N Engl J Med. 2008;359(20):2105–20. https://doi.org/10.1056/NEJMoa0801891.
25. 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):933–44. https://doi.org/10.1111/acel.12386.
26. van den Akker EB, Trompet S, Barkey Wolf JJH, Beekman M, Suchiman HED, Deelen J, Asselbergs FW, Boersma E, Cats D, Elders PM, Geleijnse JM, Ikram MA, Kloppenburg M, Mei H, Meulenbelt I, Mooijaart SP,

Nelissen RGHH, Netea MG, Penninx BWJH, Slofstra M, Stehouwer CDA, Swertz MA, Teunissen CE, Terwindt GM, 't Hart LM, van den Maagdenberg AMJM, van der Harst P, van der Horst ICC, van der Kallen CJH, van Greevenbroek MMJ, van Spil WE, Wijmenga C, Zhernakova A, Zwinderman AH, Sattar N, Jukema JW, van Duijn CM, Boomsma DI, Reinders MJT, Slagboom PE. Metabolic Age Based on the BBMRI-NL ^{1}H-NMR Metabolomics Repository as Biomarker of Age-related Disease. Circ Genom Precis Med. 2020;13(5):541–7. https://doi.org/10.1161/CIRCGEN.119.002610.
27. Simpson DJ, Chandra T. Epigenetic age prediction. Aging Cell. 2021;20(9):13452. https://doi.org/10.1111/acel.13452.
28. Bhat M, Robichaud N, Hulea L, Sonenberg N, Pelletier J, Topisirovic I. Targeting the translation machinery in cancer. Nat Rev Drug Discov. 2015;14(4):261–78. https://doi.org/10.1038/nrd4505.
29. FitzGerald GA, Oates JA, Nowak J. Cigarette smoking and hemostatic function. Am Heart J. 1988;115(1):267–71. https://doi.org/10.1016/0002-8703(88)90648-5.
30. Hioki H. Acute effects of cigarette smoking on platelet-dependent thrombin generation. Eur Heart J. 2001;22(1):56–61. https://doi.org/10.1053/euhj.1999.1938.
31. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, Klaver CCW, Nijsten TEC, Peeters RP, Stricker BH, Tiemeier H, Uitterlinden AG, Vernooij MW, Hofman A. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol. 2017;32(9):807–50. https://doi.org/10.1007/s10654-017-0321-4.
32. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van 't Hof P, Mei H, van Dijk F, Westra H-J, Bonder MJ, van Rooij J, Verkerk M, Jhamai PM, Moed M, Kielbasa SM, Bot J, Nooren I, Pool R, van Dongen J, Hottenga JJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, Zhernakova A, Li Y, Tigchelaar EF, de Klein N, Beekman M, Deelen J, van Heemst D, van den Berg LH, Hofman A, Uitterlinden AG, van Greevenbroek MMJ, Veldink JH, Boomsma DI, van Duijn CM, Wijmenga C, Slagboom PE, Swertz MA, Isaacs A, van Meurs JBJ, Jansen R, Heijmans BT, 't Hoen PAC, Franke L. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49(1):139–45. https://doi.org/10.1038/ng.3737.
33. van Iterson M, Cats D. BBMRIomics: R utilities for BBMRI omics data analysis. R package version 3.4.2. 2020. https://github.com/bbmri-nl/BBMRIomics. Accessed 17 Jan 2022.
34. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. Circ Cardiovasc Genet. 2015;8(1):192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216.
35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
36. Niehues A. GitHub repository: niehues/bbmri_surrogates_twas. https://github.com/niehues/bbmri_surrogates_twas. Accessed 25 Jan 2022.
37. Niehues A. niehues/bbmri_surrogates_twas:. Zenodo. 2022. https://doi.org/10.5281/zenodo.5903005.
38. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25. https://doi.org/10.1186/gb-2010-11-3-r25.
39. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
40. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):29. https://doi.org/10.1186/gb-2014-15-2-r29.
41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):47. https://doi.org/10.1093/nar/gkv007.
42. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–88. https://doi.org/10.1214/aos/1013699998.
43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc. 1995;57(1):289–300.
44. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov M, Sergushichev A. Fast gene set enrichment analysis. bioRxiv, preprint. 2021. https://doi.org/10.1101/060012.
45. Netherlands Trial Register. https://www.trialregister.nl/. Accessed 25 Jan 2022.
46. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, Kieboom BCT, Klaver CCW, de Knegt RJ, Luik AI, Nijsten TEC, Peeters RP, van Rooij FJA, Stricker BH, Uitterlinden AG, Vernooij MW, Voortman T. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol. 2020;35(5):483–517. https://doi.org/10.1007/s10654-020-00640-5.
47. BIOS Consortium | BBMRI. https://www.bbmri.nl/acquisition-use-analyze/bios. Accessed 25 Jan 2022.
48. BBMRI metabolomics Consortium | BBMRI. https://www.bbmri.nl/Omics-metabolomics. Accessed 25 Jan 2022.
49. EATRIS-Plus - Flagship in Personalised Medicine - EATRIS. https://eatris.eu/projects/eatris-plus/. Accessed 25 Jan 2022.
50. Horizon 2020 | European Commission (europa.eu). https://ec.europa.eu/programmes/horizon2020/en/home. Accessed 25 Jan 2022.
51. The Netherlands X-omics Initiative. https://x-omics.nl/. Accessed 25 Jan 2022.
52. Homepage | NWO. https://www.nwo.nl/en. Accessed 25 Jan 2022.
53. Lifelines Biobank. https://www.lifelines.nl/. Accessed 25 Jan 2022.
54. Leiden Langleven studie. https://leidenlangleven.nl/. Accessed 25 Jan 2022.
55. Nederlands Tweelingen Register | Nederlands Tweelingen Register (vu.nl). https://tweelingenregister.vu.nl/. Accessed 25 Jan 2022.
56. Dept. of Epidemiology (epib.nl). http://www.epib.nl/research/ergo.htm. Accessed 25 Jan 2022.

Publisher's Note