Comprehensive multi-omics profiling of a healthy human cohort

Casper de Visser[1], Anna Niehues[1,2,†], Lukáš Najdekr[3,†], Reka Toth[4,†], Petr V. Nazarov[4,†], Jana Vrbková[3,†], Jarmila Stanková[4], Sara Ekberg[5,6], Elisa Conde Moreno[7], Bishwa Ghimire[8], Bhagwan Yadav[8], Pirkko Matilla[8], Maija Puhka[8], Val F. Lanza[9], Jolein Gloerich[10], Purva Kulkarni[10], Hans Wessels[10], Udo Engelke[10], Emanuela Oldoni[2], Toni Andreu[2], Gary Saunders[2], Arnaud Muller[4], Michaela Bendová[3], Zuzana Rozankova[3], Petr Dzubak[3], Petr Pavlis[3], Bronislav Siska[3], Andreas Scherer[8], M Laura Garcia Bermejo[7], Jessica Nordlund[5,6], Alain van Gool[10,*], Marián Hajdúch[3,*], and Peter A.C. 't Hoen[1,*]

[1]Radboud university medical center, Department of Medical BioSciences, Nijmegen, The Netherlands
[2]EATRIS ERIC, European Infrastructure for Translational Medicine, The Netherlands
[3]Palacký University and University Hospital Olomouc, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Czech Republic
[4]Luxembourg Institute of Health, Bioinformatics and AI Unit, Department of Medical Informatics, Uppsala, Luxembourg, Luxembourg
[5]Uppsala University, Department of Medical Sciences, Uppsala, Sweden
[6]Uppsala University, Science for Life Laboratory, Uppsala, Sweden
[7]Ramon y Cajal Health Research Institute, Biomarkers and Therapeutic Group, Madrid, Spain
[8]University of Helsinki, Institute for Molecular Medicine Finland FIMM, Helsinki, Finland
[9]Ramon y Cajal Health Research Institute, Translational Genomics & Bioinformatics Core Facility, Madrid, Spain
[10]Radboud university medical center, Department of Human Genetics, Nijmegen, The Netherlands
[*]corresponding authors
[†]Shared 2nd author

ABSTRACT

Multi-omics investigations hold promise for uncovering biomarker profiles crucial for disease diagnosis, prognosis, and monitoring interventions. However, these profiles are susceptible to technical and biological variation. Capturing biological variability and removing technical variability are essential for future studies. The EATRIS-Plus consortium analyzed blood samples from 127 healthy adults across twelve omics layers, resulting in one of the most comprehensive omics profiling datasets. These layers were analyzed and integrated with reproducible workflows. Sex significantly impacted all omics profiles, highlighting the need for sex-balanced studies. Age was associated with fewer omics features, but we could accurately predict participants' age with epigenetic predictors($R^2 = 0.90$). And we identified aging biomarkers in other omics layers. Strikingly, omics features including miR-877-5p, CDP-diacylglycerol, *COQ10B*, *SH3GLB1*, and *MFSD14A* mRNAs and immunoglobulin proteins were significantly (False Discovery Rate < 0.05) associated with the difference between biological and chronological age, but not with chronological age itself. The EATRIS-Plus dataset is FAIR and made accessible via the ClinData Portal and Zenodo repository. This reference dataset with normal ranges for abundance and variation of all omics features can guide future biomarker studies. The dataset enables power analyses and sample size estimations for future multi-omics studies and is ideal for benchmarking computational methods that integrate multi-omics data.

Introduction

Multi-omics studies have a large potential for uncovering biomarker profiles that are relevant for disease diagnosis, prognosis, or prediction of the efficacy of medical interventions. Compared to single-omics analysis, multi-omics interrogations of biological systems provide more comprehensive molecular data, deeper insights into biology, and a better understanding of mechanisms underlying health and disease. They also contribute to studying the impact of lifestyle on well-being at the molecular levels, which is useful for identifying individuals or populations at risk and for the stratification of participants in clinical trials. However, these complex biomarker profiles are subject to many different sources of technical and inter- and intra-individual variation. Capturing biological variability in healthy individuals and reducing unwanted technical variation can help to plan and interpret future translational disease and interventional studies in precision medicine[1–3]. Currently, there are many open questions regarding the design and analysis of multi-omics studies: Which omics layers (e.g., genomics, including DNA and DNA methylation profiling, transcriptomics, including mRNA and miRNA sequencing, proteomics, metabolomics, and lipidomics) are the most informative for a particular research question or clinical diagnosis? Which biological and technical confounders should be considered? How are the different omics layers related and functionally connected? Which omics layers capture complementary biological information, and which are largely repetitive and redundant? Which omics layers are the easiest to standardize in routine clinical practice? When is a cross-omics study pertinent and feasible?

To answer these questions, we designed the EATRIS-Plus project (https://www.eatris.eu/projects/eatris-plus), executing one of the most comprehensive multi-omics studies with respect to the number of different omics layers investigated in a human healthy population cohort (see Methods). In this reference study, we investigated twelve different omics layers to uncover sources of technical and biological variation. Mapping the variability of the different omics layers facilitates sample size calculations for future single and multi-omics studies. Assessing the variability of individual features in this reference cohort will help to select robust biomarkers for disease or lifestyle in future diagnostic or prognostic studies. Through testing a range of multi-omics data integration methods, we investigated the complementarity of the different omics layers. This should facilitate the selection of the best combination of omics layers in future multi-omics studies, where budgets or technical implementation in clinical diagnostics may not allow the profiling of multiple omics layers. Finally, we investigated the association of the molecular features with the participants' phenotypes, such as sex, chronological and biological age and body mass index (BMI), and explored the biological nature of the observed variation through pathway analyses.

Results

Study Design

This study profiled blood samples obtained from a cohort of 127 healthy individuals. These individuals were selected from a larger cohort of blood donors who were clinically examined and whose samples were bio-banked at the Institute of Molecular and Translational Medicine of Palacký University and University Hospital Olomouc (see Methods). We selected a group of individuals with equal representation of the sexes and a representative age distribution between 21 and 61 years. All collected samples underwent standardized sampling and comprehensive analysis using 12 different omics technologies, including genomics (whole genome sequencing (WGS) and array comparative genomic hybridization (arrayCGH)), transcriptomics (mRNA and miRNA sequencing), and epigenomics (DNA methylation using enzymatic methylation sequencing (EM-seq))[4,5] on blood cells, and shotgun proteomics, targeted metabolomics (amino acids, very long chain fatty acids (VLCFA) and acylcarnitines), untargeted lipidomics (positive and negative ion mode), and qPCR-based miRNA profiling on heparin or EDTA plasma (Figure 1). Samples were distributed to EATRIS facilities across Europe with relevant expertise for profiling the different omics layers. Quality control and processing of the single omics data were performed locally. A subset of samples in the sequencing-based methods did not meet quality control standards, *e.g.* hemolysis grade for qRT-PCR (see Methods). As a result, all omics layers are available for 60 individuals, with around 90 individuals having at least five different omics layers (Fig. 1). In some omics layers, batch effects originating from the running of samples on different days needed correction (Suppl

Figures S1-9, Tables S1-2). Batch effect correction for lipidomics was performed based on pooled quality control (QC) samples applying the SERRF QC Correction algorithm[6]. For all other omics layers where an internal standard was not available, batch correction was done in a consistent manner using Combat-Seq[7]. Combat-Seq performed better than other batch correction methods on all omics layers, despite being developed for RNA-seq data only. Combat-Seq optimally reduced associations with technical confounders (Suppl Figures S3-9). The unprocessed and batch-corrected pre-processed data were returned for centralized (multi)-omics analyses. This dataset, both before and after correction for differences in cell type proportions, is available via 10.5281/zenodo.10246047 and fdp.radboudumc.nl and provides a rich resource for benchmarking multi-omics data integration methods. Uncorrected raw data and GDPR-sensitive data are available upon request in machine-readable formats in ClinData Portal (https://clindata.imtm.cz).

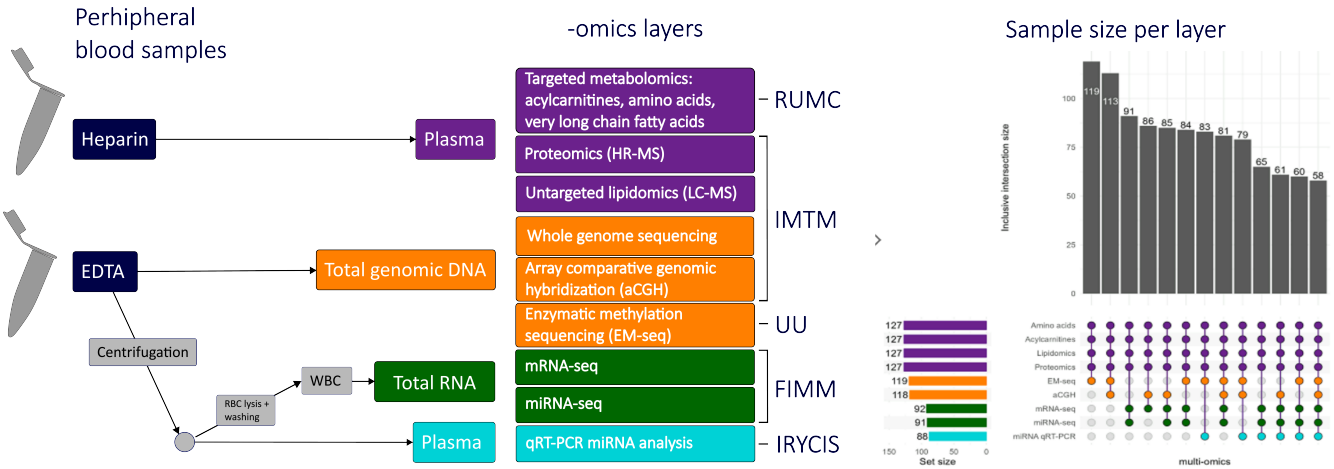

Figure 1. Study design. Sample preparation for the different –omics layers are shown. Colours represent the different blood sample preparations. The EATRIS facilities that analysed the samples are indicated. On the right the sample size per –omics layers are shown, and the sample intersections of different combinations of –omics layers are plotted.

Genomic characterization

Since the genetic make-up of individuals may directly affect other omics layers, we first assessed gross chromosomal aberrations. Through the assessment of the whole genome sequencing, aCGH (Figure 2), and EM-seq data (Suppl Fig S10), we identified one participant with an undiagnosed Klinefelter syndrome (XXY sex chromosomes). This is not surprising given the estimated incidence of 1:500 and because this syndrome does not always lead to evident clinical symptoms. Moreover, aCGH analysis revealed several losses and gains of chromosomal regions in individual participants (Figure 2, Suppl Table 1). We hypothesize that these are the consequences of both genetic heterogeneity in the population and somatic aberrations in clonal expansions of hematopoietic cells[8]. In four genes (*BTNL3*, *BTNL8*, *DUSP22* and *UGT2B17*), we observed significant differences in the mRNA levels among subjects corresponding to gene gain or loss (Suppl Fig S11). Since we regarded these chromosomal aberrations as common genetic variation present in the normal population, we included all subjects in subsequent multi-omics profiling using transcriptomics, methylation sequencing, proteomics, metabolomics and lipidomics.

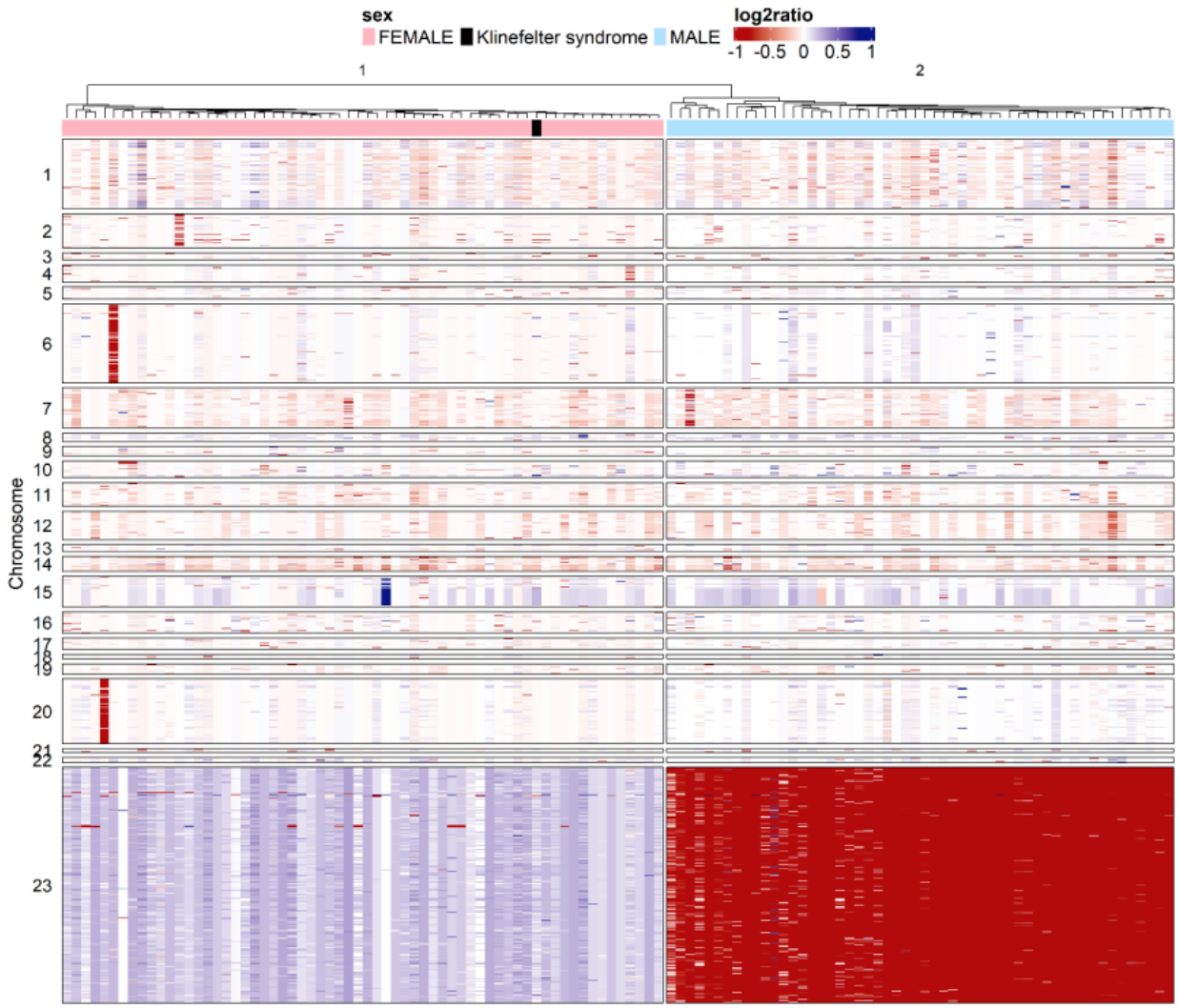

Figure 2. Heatmap of log2 ratio values for genes with LOSS (log2 ratio < -0.5) or GAIN status (log2 ratio > 0.5) grouped by chromosome number. Only 3,707 genes on chromosomes 1-22 and X (labeled as 23) with at least one sample with LOSS/GAIN status presented. The number of selected genes per chromosomes can be found in Supplementary Table 2. The aCGH data was available for 118 samples, hierarchical clustering (complete agglomeration method, Euclidean distance) was performed in this heatmap.

Variability in omics values

To provide researchers with insight into the number, identity and abundance of the features detectable in the different omics layers, we provide the mean, median, standard deviation and coefficient of variation (CV) for each molecular feature in Suppl Table 2. We stratified these values by sex, because we observed large differences between males and females. Figure 3 illustrates the CV for each feature, sex-stratified and categorized by omics layer. CVs demonstrated notable differences among the omics layers. For example, miRNA sequencing exhibited larger variation compared to other data types. A large part of this variation can be explained by variation between individuals in the proportion of the different types of leukocytes expressing these miRNAs, ranging from 10%-50% lymphocytes and 40%-80% neutrophils (Suppl Fig S12). This is the reason why we adjusted mRNA, miRNA and methylation sequencing data for white blood cell proportions in all our subsequent analyses (Suppl Fig S13).

Very long chain fatty acids (VLCFAs) showed more variability in females than in males, whereas most of the other omics

features displayed similar variation in males and females. As expected, EM-seq revealed that most CpGs were constitutively methylated or non-methylated across healthy individuals. However, there was a subset of CpGs with dynamic methylation in the population. Many of the variable sites overlapped with single nucleotide polymorphisms (SNPs), which were excluded from further analysis (Suppl Fig S14).

Since the number of CpGs assessed by EM-seq was very large (> 26 million), and that this large number of features may dominate multi-omics analyses, we decided to continue our analysis with the 100,000 most variable CpG sites that are in closest proximity to 2,498 unique genes.

Phosphatidylcholines were found to be the most variable class of lipids. This was particularly true for the negative ionization mode. Pristanic acid and its metabolite phytanic acid were the most variable metabolites in human plasma. This is not surprising since they are heavily influenced by dietary intake[9]. Among the most variable plasma proteins were ApoL1, Alpha-1-antitrypsin, and Immunoglobulin heavy chains, which is possibly due to variable food intake before blood sampling, allelic[10] and/or antigenic variability of those proteins.

The provided normal ranges for abundance and variation can help to guide power calculations in future biomarker studies and to identify and prioritize biomarkers for disease or lifestyle.

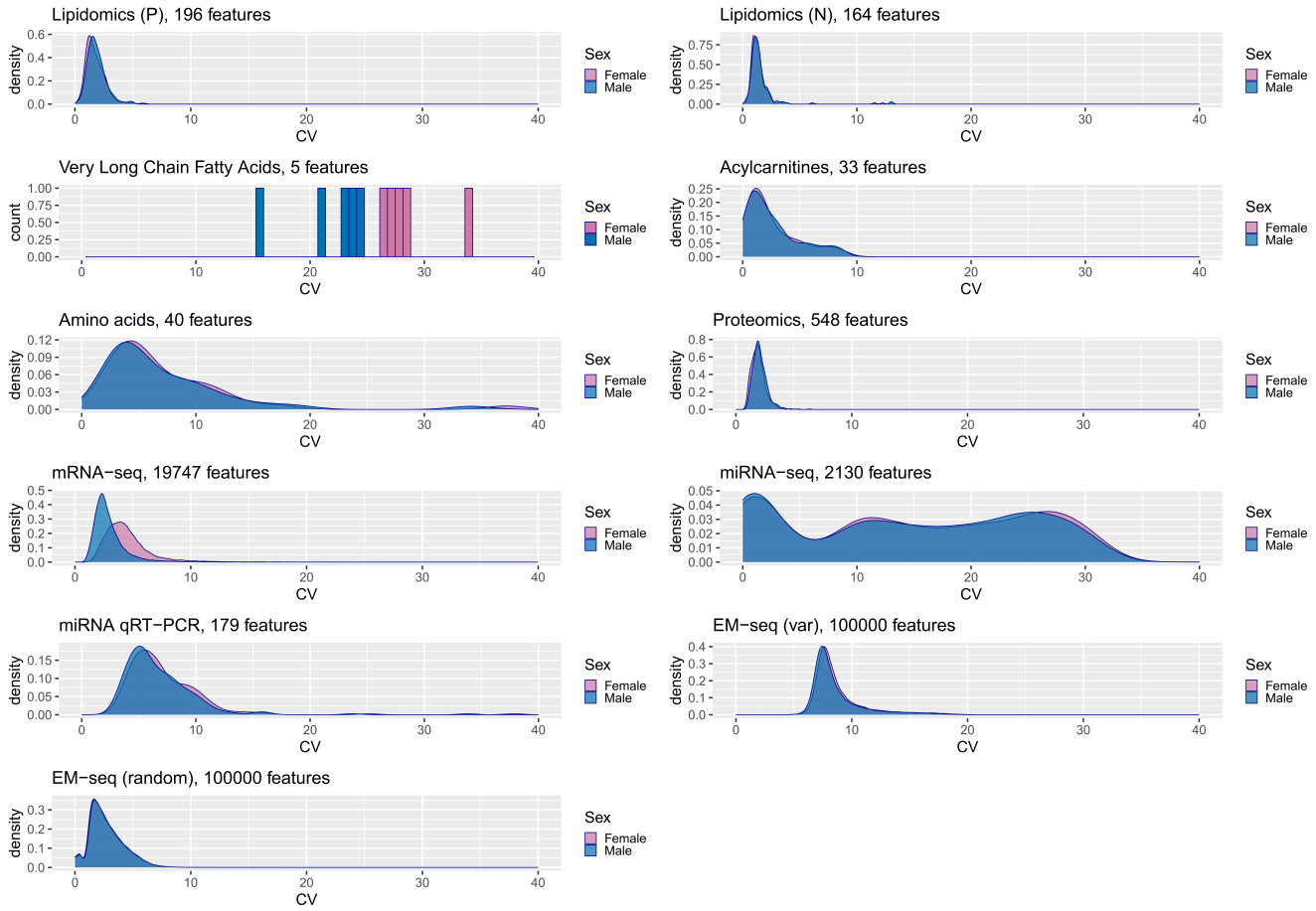

Figure 3. Variability among the –omics data layers. The Coefficient of Variance (CV) is visualized for every feature in each data set using density plots. The CV is calculated by taking the ratio of the standard deviation to the mean, and then multiplying the result by 100 to express it as a percentage. CVs are plotted separately for female/male. Missing metabolomics and proteomics data are imputed. A histogram is used or the VLCFA data, as this data contains only 5 features.

Association of molecular features with phenotypes

Robust linear models (RLM) were applied to discover linear associations of each of the omics features with phenotypic traits and hematological measurements. This analysis included four different models: linear models with and without sex as a covariate, and two separate models for male and female individuals. Note that the number of molecular features varies widely between the different data types (*e.g.* 100,000 methylation levels *vs.* 5 measured very long chain fatty acids (VLCFAs)). Figure 4 summarizes the results of all these different models by plotting the proportion of statistically significant associations (after multiple testing correction) between omics features and phenotypic variables. All significant associations can be found in Suppl Table 3.

A substantial proportion of targeted metabolomics (amino acids, acylcarnitines) and lipidomics analytes were associated strongly with sex, height, weight, BMI and several blood cell types. For example, root mean squared errors (RMSE) of 0.9539 (tiglylcarnitine $\sim$ blood group) and 0.8731 (Phosphoetanolamine $\sim$ Sex) were reached. Their associations with erythrocyte, hemoglobin and hematocrit levels can be largely explained by sex, as they disappeared when sex was added as a covariate and in the sex-stratified models. Notably, some omics features showed associations only within male or female individuals: VLCFAs (pristanic acid and phytanic acid, Suppl Fig S15) were associated with BMI in male but not in female individuals, whereas pristanic acid and a number of other lipids were associated with leukocyte numbers in females only.

Unlike metabolomics and lipidomics, methylation and miRNA profiling mainly showed significant associations with cell type proportions. In both the miRNA-seq and EM-seq data, a large number of features were associated with lymphocyte and neutrophil levels, despite the correction for estimated cell types conducted for these data types. Mean corpuscular hemoglobin concentration (MCHC) levels showed linear relationships with a large number of miRNAs in plasma and measured by qRT-PCR, especially in women. Surprisingly, across all the omics layers, the majority of features did not show significant associations with chronological age. For mRNA-seq, lipidomics, miRNA-qRT-PCR, miRNA-seq and EM-seq data, less than 1% of the features were associated with age, whereas no significant associations were identified in the other omics layers. Specifically, 289 of the 100,000 methylation sites studied were related to age and only 3 mRNA features were associated with age group. We describe associations with biological age in more detail below.

In conclusion, RLMs revealed a large number of significant relations between molecular features and phenotypic traits in all omics layers. Many of the associations can be explained by sex. Hence, balancing sex in future multi-omics- and biomarker studies is essential. The different omics layers exhibited distinct associations with other phenotypic traits, underscoring the potential added value of multi-omics studies.

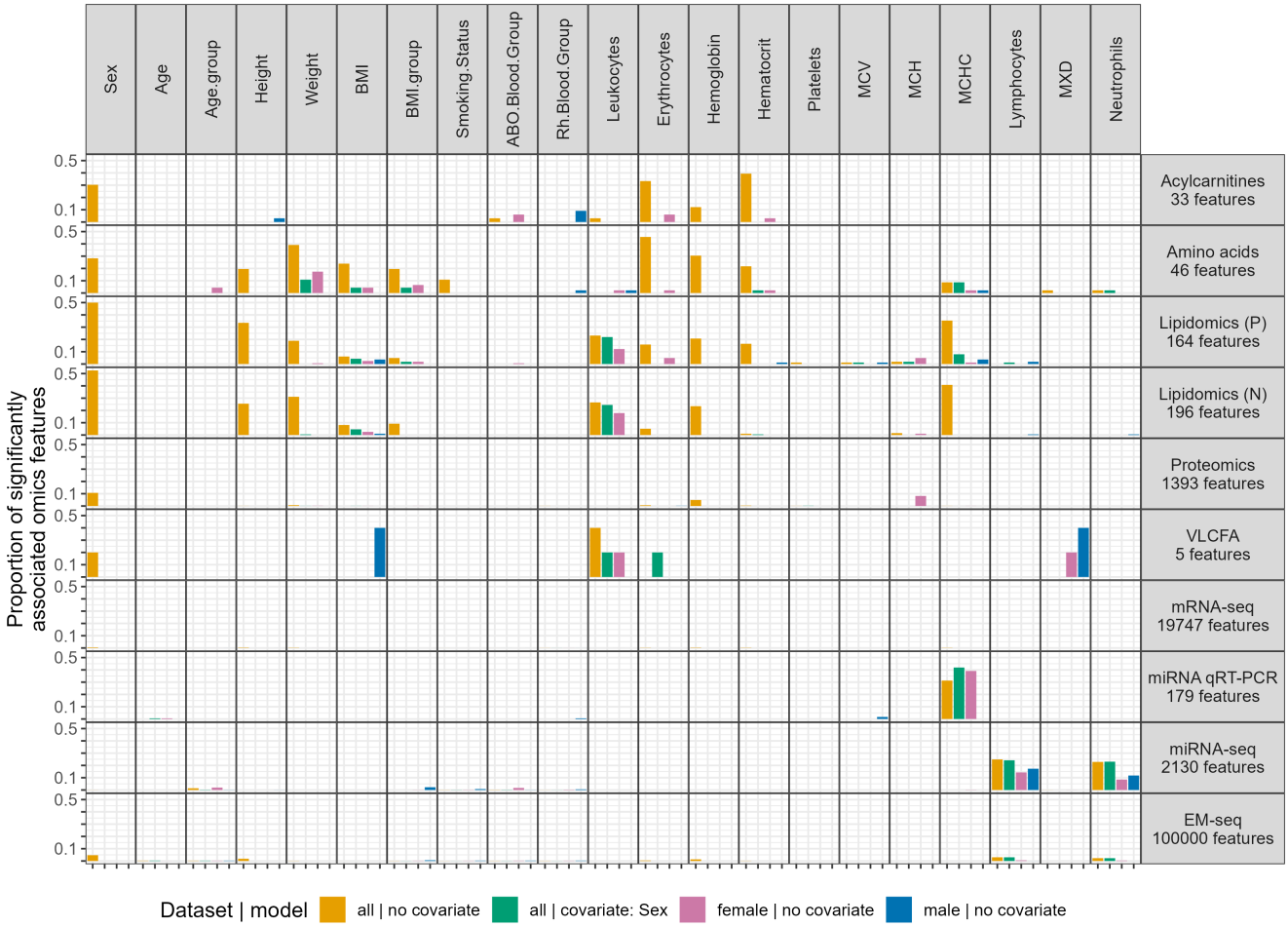

Figure 4. Robust linear models. The proportion of significant associations between -omics features and phenotype variables is depicted, per -omics and per linear model. The maximum value on the y-axis is in all cases 0.5, corresponding to 50% of all features assessed. Numbers of features per omics layer are indicated on the right. Associations with FDR-corrected p-values below 0.05 are considered significant. Missing values in the metabolomics and proteomics data are imputed, and mRNA-seq, miRNA-seq and EM-seq data are corrected for cell-type compositions. For EM-seq, we only considered the 100,000 most variable CpG sites.

Principal component analysis (PCA) on single omics

To reveal multivariate associations with phenotypic traits within single omics data, principal component analysis (PCA) was performed on all individual omics layers. Following this, the correlations between the first ten principal components and the phenotypic and technical variables were evaluated (Suppl Figures S16-15). Figure 5 summarizes these findings by visualizing, for each omics layer, the cumulative variation explained by components that correlated significantly in this analysis.

Patterns observed in the RLM analyses were supported by PCA. For instance, metabolite and lipid levels were mostly associated with sex, body morphology phenotypes and hematological variables. In contrast with the RLM results, the principal components from the EM-seq methylation profiling associated with phenotypic traits, including chronological age. These components jointly captured up to 7% of the variation in the EM-seq data, highlighting the power of PCA and dimension reduction methods in unveiling hidden patterns in high-dimensional data. Moreover, in mRNA-seq and miRNA plasma levels, associations of PCA with age groups were found, which were not detected by the analysis of single omics features.

Technical effects, *i.e.* running the omics analyses in different batches, did not have a major influence. No significant correlations have been detected between one of the first 10 principal components and batch, except for the miRNA-seq data.

This indicates that batch-adjustment methods (see Methods) have mostly been successful in reducing unwanted technical variation in these omics data. However, for miRNA-seq, results should be interpreted with care because of a possible residual influence of the batch in which the samples were run.

In summary, the PCA conducted on the individual omics data not only confirmed findings from the linear models, but also revealed age-dependent associations that could not be identified by testing omics features individually. This indicates the added value of PCA analyses in complex multi-omics data sets.

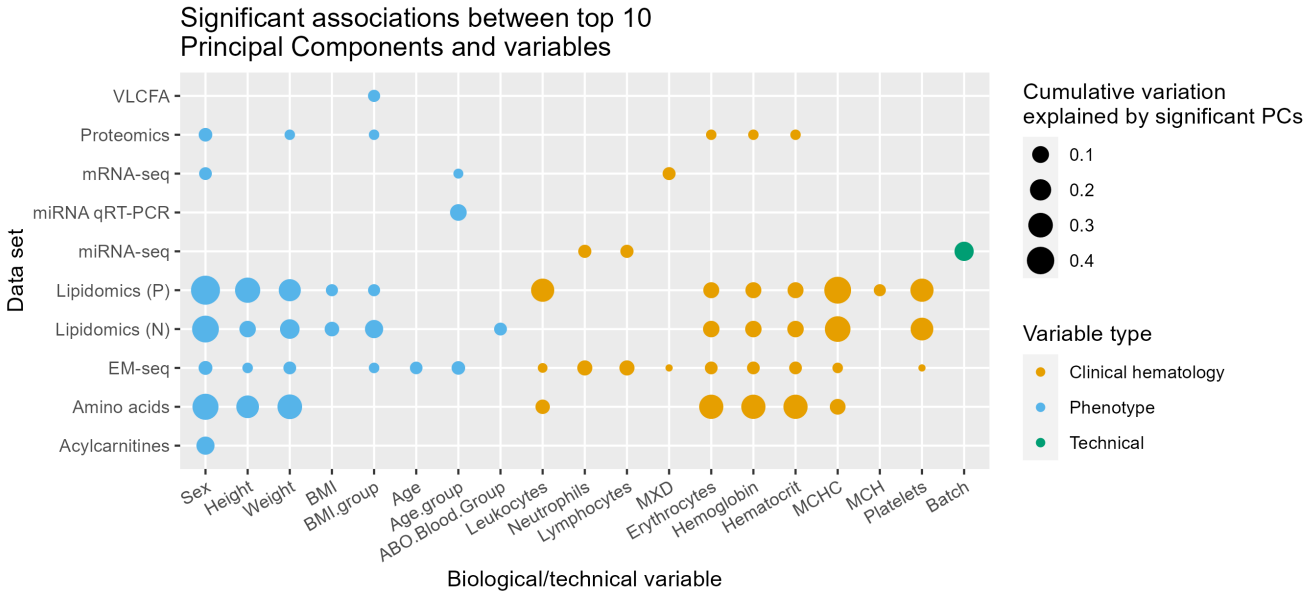

Figure 5. Principal Component Analysis (PCA). Significant associations between one or more PC's and the variable are visualized for each -omics layer. Associations with FDR-corrected p-values below 0.05 are considered significant. The size of each dot represents the amount of variation explained by these components combined. As PCA is not able to handle missing values, imputed datasets were used for this analysis.

Multi-omics factor analysis (MOFA)

To uncover joint sources of variation across different omics layers, we applied the unsupervised multi-omics factor analysis (MOFA) method. MOFA creates latent representations of the multi-omics data and captures them in MOFA factors, much like principal components for single –omics data[11].

Figure 6a depicts the twelve latent dimensions uncovered by MOFA and quantifies the contributions of the individual omics layers to these MOFA factors. This figure clearly illustrates how the MOFA factors comprise information from multiple omics levels. For example, factors 1 and 6 captured variance present in metabolomics, lipidomics, epigenomics and proteomics data. To facilitate the biological interpretation of the MOFA factors, we correlated the factor scores with the continuous phenotype measures (Fig 6b) and performed pathway enrichment analysis (Suppl Figures S17-20) on the weights (Suppl Table 4) of the mRNA, proteins and metabolites on the MOFA factors.

Several significant correlations were identified between MOFA factors and continuous (Figure 6b) and categorical (Figure 6c-e) phenotypic variables. For instance, factors 1 and 6 were associated with sex, height, weight and multiple cell type levels. As the scores for factors 1 and 6 were significantly higher in males than females, the associations with other variables can be driven by sex. Interestingly, the Klinefelter individual was most similar to females in factor 1 and 6, but appeared to be an outlier in factors 3, 5, 10 and 11, suggesting that this individual is different from XX and XY individuals in multiple molecular aspects.

Although MOFA is designed to capture variance across different data sources, factors 2, 3 and 4 represented mainly

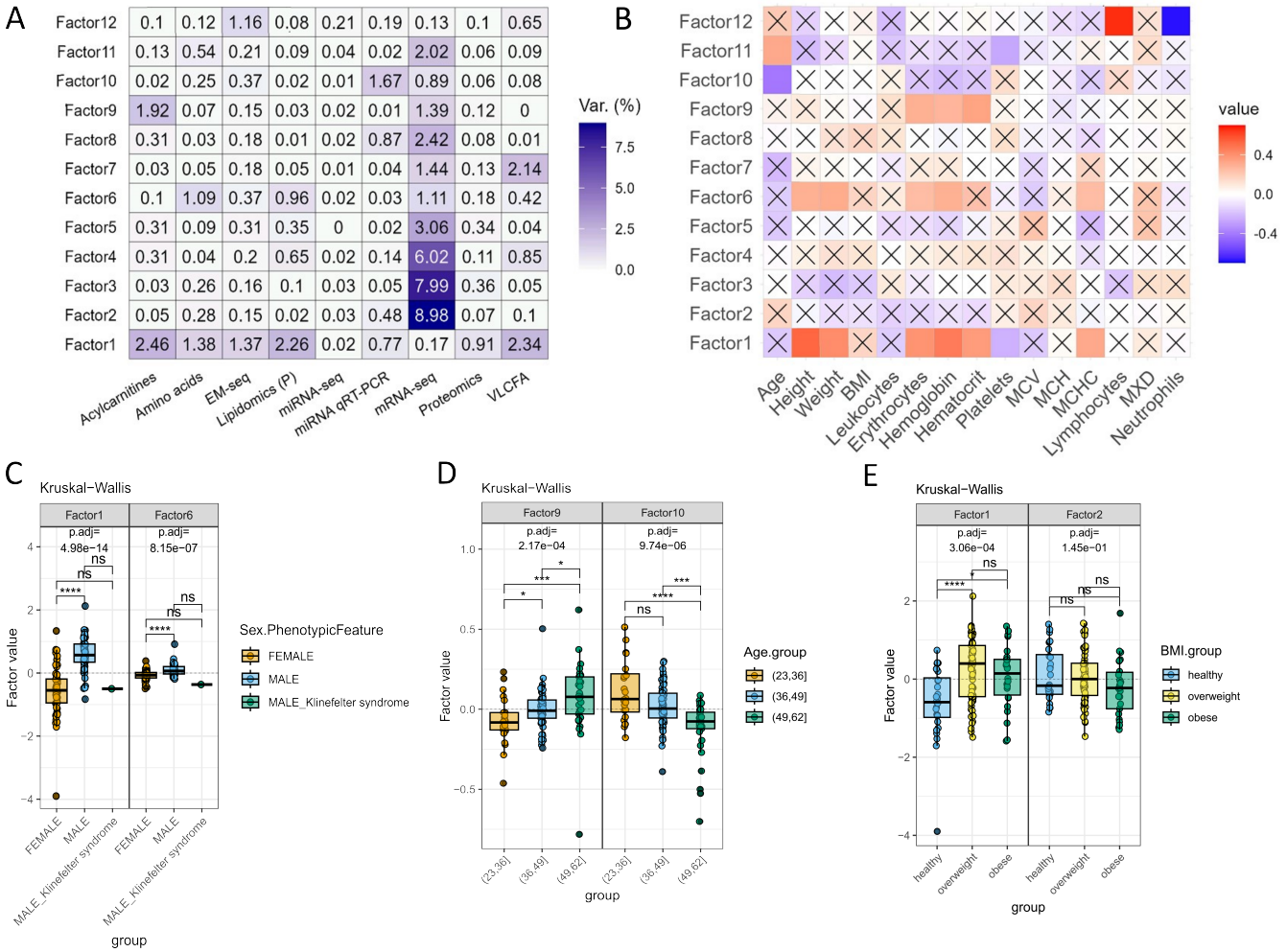

Figure 6. Multi-Omics Factor Analysis (MOFA) results. A) Explained variance across the –omics layers for MOFA factors 1-10. B) Heatmap showing correlations between the MOFA factors and phenotypic variables; significant associations are open squares (not marked with an X). Comparative analysis of the values of the MOFA factors associated with sex (C), age (D) and BMI (E) groups. Significance was tested using Wilcoxon for sex. Kruskal-Wallis was used for age and BMI, which were post-hoc tested using a two-sided Dunn test. All p-values were corrected using Benjamini-Hochberg's false discovery rate (FDR).

contributions from the mRNA-seq data. The MOFA sample scores on these three factors were not correlated with the phenotypic variables, but pathway enrichment analysis revealed several significant associations with immune-related biological pathways (Suppl Table 5). Although factors 2, 3 and 4 were not significantly associated with the overall percentage of lymphocytes and neutrophils, we hypothesized that this variance is due to the cell type proportions that could not be compensated by the cell type correction. This hypothesis was supported by the strong involvement of cell type-specific transcripts in these factors (Supp figures S21). B-cell specific genes demonstrated large negative weights on factor 2, whereas monocyte-specific genes demonstrated large positive weights on factor 4. Furthermore, genes coding for proteins of the large and small subunit of the ribosome had large weights on factor 3. The ribosomal protein gene expression is known to be different between different cell types in the blood[12]. Thus, the cell type correction that we applied may have worked well for the major cell types, based on cell-precursors, but may have been less efficient for corrections of more granular differences in cell types with a common cell precursor, such as the proportion of B- and T-cells.

The mRNA and protein feature weights from factor 5 were associated with immune-related pathways, for example the

complement pathway (Suppl Figures S17 and S19). Since the proteomics and mRNA-seq data were derived from different sample types (plasma vs. white blood cells (WBCs)), this variation is less likely to be affected by the cell type proportions.

Factors 10 and 11 were indicative of the aging process. Factor 10 primarily accounted for variations observed in miRNA qRT-PCR and mRNA-seq data, exhibiting a negative correlation with chronological age. Accordingly, sample scores on factor 10 differed significantly between age groups were observed (Fig 6c). Additionally, the mRNA-seq feature contributions identified enriched pathways associated with chemokine signaling, a hallmark of aging[13], and included mRNAs like *CCR7*, *IRAK2*, *STAT5A* and *CXCR3* (Suppl Fig S18). Remarkably, many of the top 10 qRT-PCR miRNAs (hsa-miR-144-5p, hsa-miR-139-5p, hsa-miR-10b-5p, hsa-miR-32-5p, hsa-miR-99a-5p, hsa-miR-17-5p, hsa-miR-16-5p, hsa-miR-24-3p, hsa-miR-199a-3p, hsa-miR-221-3p) have already been linked to chronological age or biological processes involved in senescence[14]. Some of them have been previously described as potential biomarkers of ageing in different body fluids, such as miR-24-3p, miR-99, and miR-144 in whole blood[15], PBMCs[16], and saliva[17]. Others have been implicated in several age-related diseases like cancer or neurogenerative disorders[18,19]).

Conversely, factor 11 showed a positive correlation with age and a negative correlation with platelet counts, a relationship observed in prior studies[20]. A more detailed analysis of the association between methylation levels and ageing is given below. Pathway enrichment analysis of mRNAs scoring high on factor 11 revealed several upregulated immune-related pathways. Notably, the suppression of the STAT3 signaling pathway was identified, which has been linked to aging[21].

Finally, MOFA factors did not associate with the chromosomal arrangements observed in the aCGH data (Suppl Fig S16), suggesting that they did not have a major impact on the multi-omics profiles except from the XXY configuration in the Klinefelter individual.

In conclusion, MOFA was able to discern shared variation across the omics layers, which could be linked to different phenotypic variables and biological processes. While most sex effects were captured by the first and sixth MOFA factor, other MOFA factors reflected non-sex dependent biological variation that was more difficult to detect in single omics analyses.

Results from other multi-omics data integration methods

In addition to MOFA, we analyzed the multi-omics data with other data integration methods. The results mostly confirmed the results from MOFA and single omics analyses. We applied consensus independent component analysis (consICA), a method and R package we recently developed to disentangle technical biases and important molecular signals, highlighting active biological processes and cell subpopulations[22]. We applied consICA separately to mRNA-seq, miRNA-seq, proteomics, metabolomics, and EM-seq, identified the 20 most important components and integrated these components (Suppl Figures). In line with the PCA and MOFA results, strong associations with sex were identified. Specifically, ICA component g1 from the RNA-seq data (Fig 7a), as well as 2 components each from the EM-seq and metabolomics data showed significant associations with sex (Fig 7d, Fig 8). Additionally, EM-seq component e19 allowed precise recovery of lymphocyte/monocyte abundances through linear regression (Fig 7b), with $R^2 = 0.80$ and $R^2 = 0.77$ for lymphocytes and neutrophils, respectively.

By correlating the weights of single omics consICA results, we built a network of the multi-omics components and identified the biological functions represented by components from different –omics layers (Fig 8, Suppl Table 6). The multi-omics components were linked to general blood characteristics, such as blood coagulation, immune cell function, hematocrit and blood rhesus factor (Blood Rh), and more specific biological functions like nucleosome assembly, mitogen-activated protein kinase (MAPK) signaling, and nitric oxide transport (Fig 8).

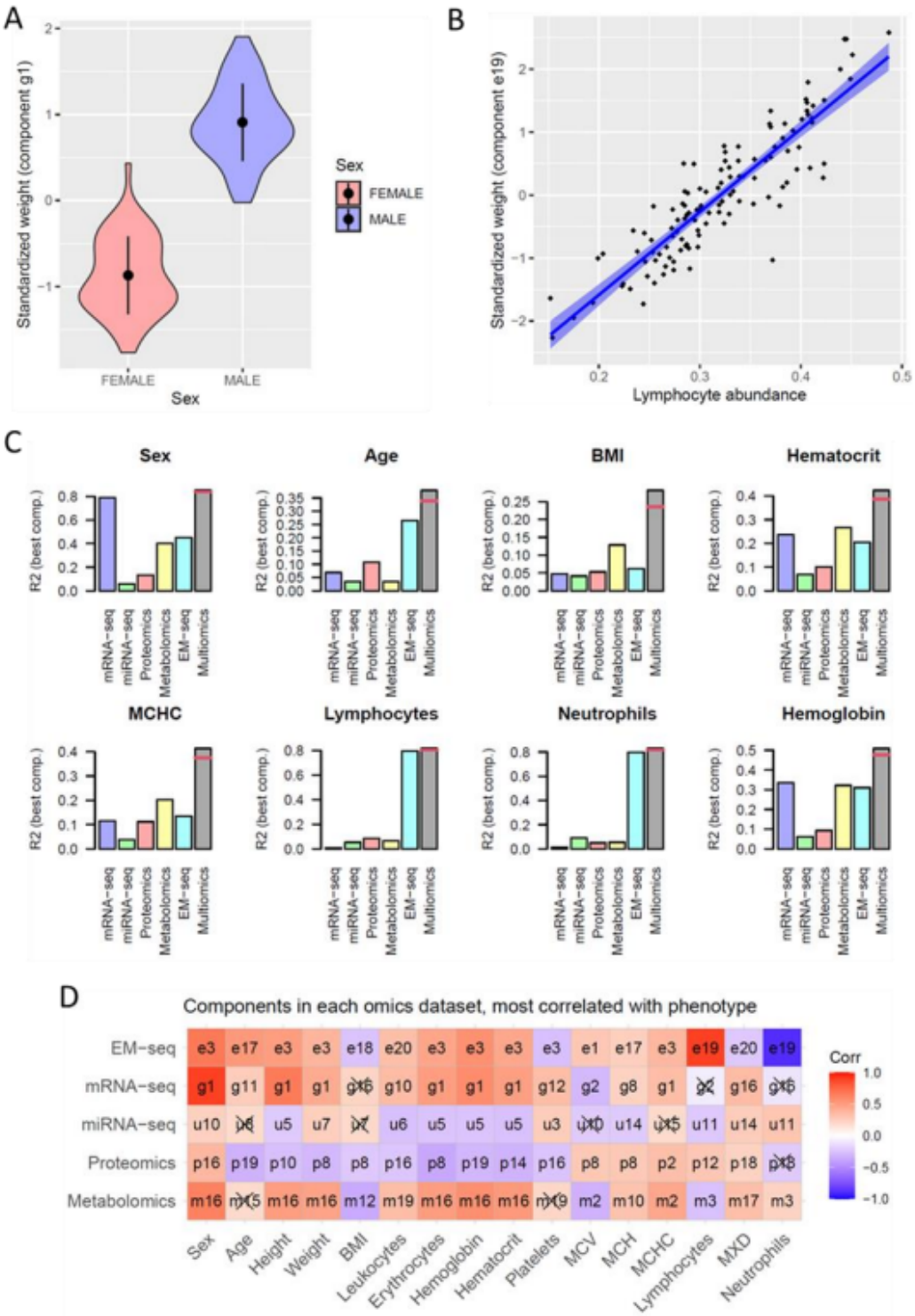

Figure 7. consICA on various -omics layers predicts phenotypes. (A) Violin plot of the weights of mRNA ICA component g1 for males and females (p-value: 3.2e-33; t-test). (B) The weights of DNA methylation component e19 were strongly linked to lymphocyte abundance in the samples. (C) The top R2 observed between experimental variables and independent components of each -omics layer separately, and their combination into a multivariable linear model ("Multi-omics" bar, red lines represent adjusted R2). (D) The top correlated ICA component for each combination of omics and phenotypic variables. Non-significant correlations are marked by X (FDR-corrected p-values of the visualized correlations).

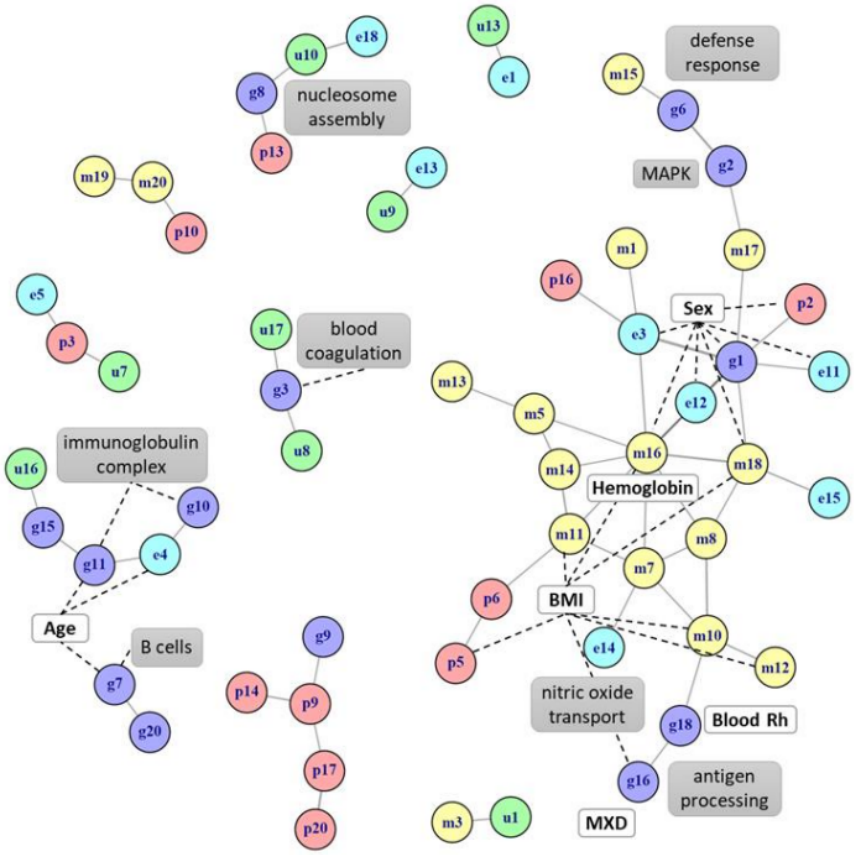

Figure 8. Integration of ICA results as a correlation-based graph. The nodes represent components from epigenomics (e#), transcriptomics (g#), microRNA-omics (u#), proteomics (p#), and metabolomics (m#). The edges show components with correlated weights ($R^2 > 0.1$, p-value < 0.001). Additional annotation based on clinical data (bold, white box) or most important gene sets (grey box) is shown.

To further demonstrate the added value of multi-omics data integration, we built linear regression models to predict phenotypes using the components' weights. For BMI, age hematocrit and hemoglobin levels, the multi-omics components explained more of the variance than the individual –omics components (Fig 7c), demonstrating the complementarity of the information present in the different omics layers. As expected from the results described above, white blood cell proportions could be estimated from the methylation components, with hardly any additional gain from the other omics components (Fig. 7c).

As another way to investigate the complementarity of the different omics layers, we clustered the samples using similarity network fusion (SNF)[23], a method that fuses multiple sample similarity networks constructed from single omics data into one. Here, we showed that clustering based on multiple omics layers did not demonstrate stronger clustering of samples than single omics layers (Suppl Data and Suppl Fig S22). Furthermore, partial least squares 2 (PLS2)[24] analysis demonstrated strongest covariance between the lipidomics in negative and positive ion mode, but also high covariance between other omics layers, which were likely driven by sex (Suppl Data, Suppl Fig S23).

Biomarkers for biological age

Given the large number of omics layers available and the relatively wide age span in our healthy cohort, we aimed to identify biomarkers for biological age. The most robust estimators of biological age are currently derived from methylation data[25]. Notably, these epigenetic clocks were developed on methylation array data, whereas we generated EM-seq data. Therefore, we

first demonstrated the performance of the most popular epigenetic age predictors on our cohort and our EM-seq based data. These clocks will be referred to as: best linear unbiased prediction (BLUP)[26], bayesian neural networks (BNN)[27], elastic net (EN)[26], Hannum[28], Horvath[29], Levine[30] and skin-Horvath[31]. Age predicted with BLUP and EN clocks was most strongly associated with the chronological age of the participants of our cohort (Fig 9a). The high correlations (R^2 of 0.90 and 0.81, respectively) are remarkable given the differences between the methylation profiling platforms and the different geographical origin of our population. The other age predictors performed significantly worse, suggesting they are less generalizable across profiling platforms and/or population. In line with the observed associations between MOFA factors 10 and 11 with chronological age, we observed significant associations between these MOFA factors and predicted age based on the BLUP but not Levine clock (Fig 9b). This confirmed that BLUP was superior in predicting chronological age in our cohort.

To identify candidate omics biomarkers for biological aging, we calculated the age acceleration, *i.e.* the difference between chronological age and age predicted by BLUP and Levine, after adjusting for the calculated cell type composition. Subsequently, we tested for linear associations between age acceleration and all omics features (Fig 9c). Associations of omics features with chronological age were often recapitulated by associations with BLUP-predicted age, which was expected due to the high correlation of chronological age with BLUP-predicted age. However, we also observed a considerable fraction of miRNAs in the plasma with more significant associations with BLUP-predicted age than with chronological age (Suppl Table 3). Moreover, miRNAs such as miR-877-5p, lipids like CDP-diacylglycerol, transcripts such as COQ10B, SH3GLB1, and MFSD14A and immunoglobulin proteins were significantly associated with accelerated age but not with biological age. (Fig 9d, Suppl Figures S25-S28). We observed only weak associations between age acceleration and phenotypic variables such as blood group and smoking status (Suppl Fig S24), making it unlikely that associations with age acceleration are due to confounding of other biological factors. The identified molecules may thus represent candidate biomarkers of biological processes that contribute to the aging process throughout the lifespan or biomarkers of accelerated aging, relevant for pathological contexts like degenerative diseases or cancer.

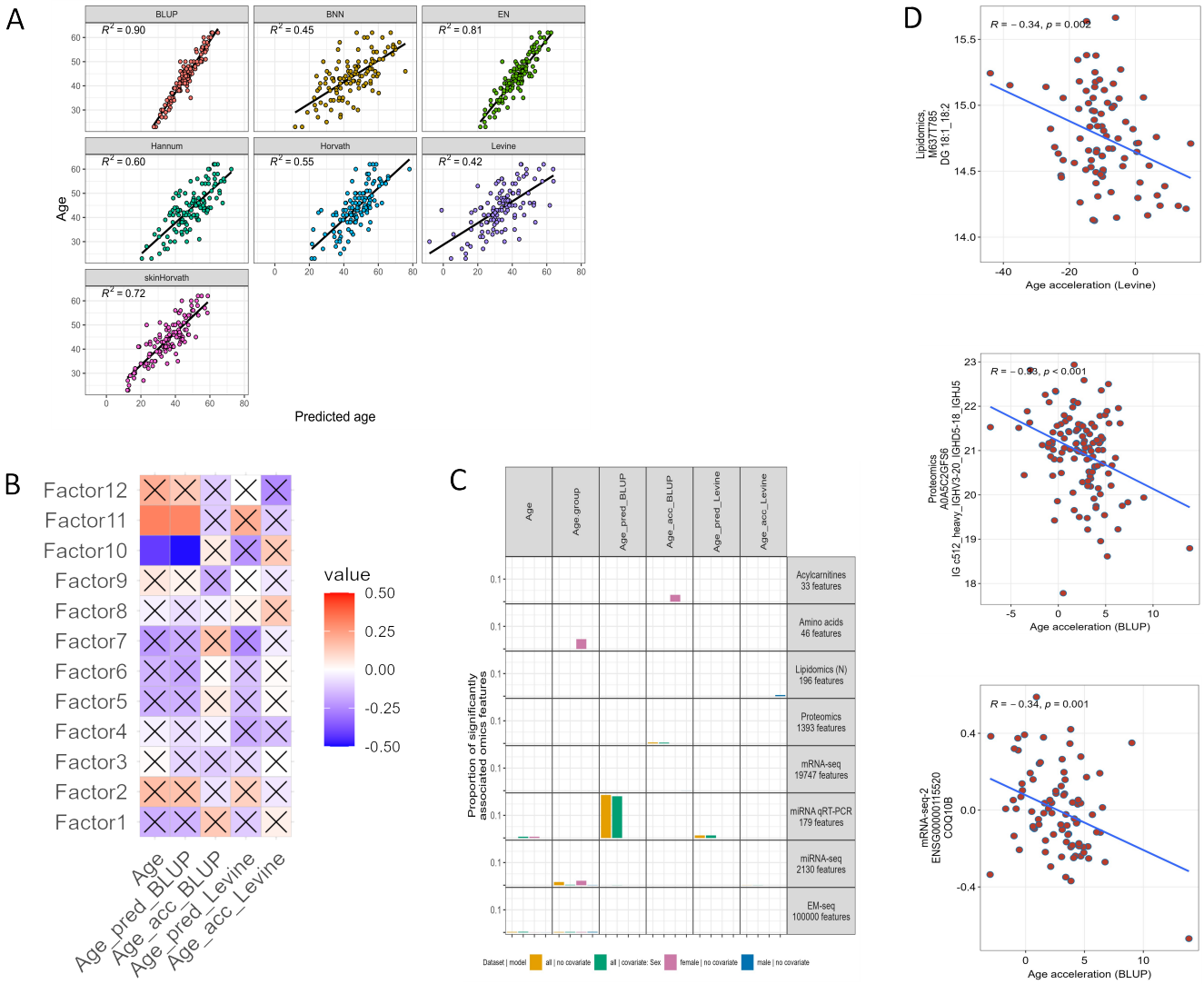

Figure 9. Associations with biological age. A) Performance assessment of the different epigenetics clocks, by comparing them with the chronological age. B) Robust linear model with the age prediction and age acceleration (biological age - chronological age) from BLUP and Levin clocks as outcomes. C) MOFA factor correlations to the same age prediction and age acceleration. D) Individual features correlating with age acceleration.

Discussion

We reported one of the most comprehensive multi-omics studies in a cohort of healthy volunteers performed so far, running twelve different omics techniques on 127 blood samples. Our dataset can be used as a benchmark, providing normal ranges for abundance and variation of tens of thousands of omics features in the healthy population and guiding the design of reproducible omics studies investigating specific (disease) phenotypes or lifestyle interventions. Our study enables power analyses, sample size estimations and sex composition balancing for future multi-omics studies. Moreover, our dataset can be used to benchmark computational methods for multi-omics data integration, as we did in this study. The best-performing multi-omics integration techniques would optimize the number of identified molecular associations with phenotypes that are known to influence molecular omics levels such as sex, age and BMI.

With our benchmark study, we set out to answer the following questions guiding the designs of future multi-omics studies:

Question 1: Which biological and technical confounders should be considered? Answer: Sex-dependent variation

was observed through all omics layers, with strongest effects observed in lipidomics and metabolomics. This suggests that future studies, irrespective of the omics technique applied, should be balanced in the number of females and males included. Additionally, studies may benefit from analyzing the females and males separately, because many diseases progress differently between sexes[32]. Separate analysis can reveal more specific or better prognostic biomarkers for each sex. Cell type composition differences were a major confounder for the omics techniques performed on blood cells but as expected, not so much for plasma-based assays. Our MOFA analyses and linear associations with cellular transcriptomics levels demonstrated that there were residual effects from differences in cell type composition, despite our efforts to correct based on the cell type proportions estimated from the EM-seq data. This can be partially explained by the different sample processing per omics layer (Fig 1). Moreover, we did not consider low abundant subtypes of leukocytes in the estimation of cell type proportions. We suggest using independent analytical technologies such as multiparametric flow cytometry or CyTOF for measuring instead of inferring cell type proportions.

Question 2: How are the different omics layers related and connected, which –omics layers capture complementary biological information, and which are largely redundant? Answer: From the different multi-omics integration techniques performed, we conclude that the omics layers are only weakly connected and contain mostly complementary information. As expected, lipidomics performed in negative and positive ionization modes were strongly correlated, but still showed differences in the number of lipids detected, their variability in the population and their associations with phenotypes. Other omics layers did not strongly correlate with each other nor show significant associations with similar biological pathways. This is partially because they are profiling molecules from different sources: EM-seq, RNA-seq and miRNA-seq were performed on white blood cells, whereas peripheral tissues are the most important source of lipids, proteins and metabolites in the circulation. We therefore recommend selecting omics layers in future omics studies based on specific biological hypotheses, taking into account the relevant cell types and tissues profiled.

Question 3: Which omics layers are easiest to standardize in routine clinical practice? Answer: Our study was run with highly standardized procedures (available through the EATRIS multi-omics toolbox: motbx.eatris.eu) and demonstrated the feasibility of distributed measurements and analyses in different omics laboratories. In establishing reference protocols and materials for multi-omics studies[33–35], we faced several challenges that needed to be addressed when implementing these techniques in multi-center omics analyses. Technical noise, and in particular batch effects, are typically non-biological sources of unwanted variation that obscure the actual biological variation. These also include reagent batches of varying quality, drift of machines, differences between operators and others[5,36–38]. If not addressed either through experimental planning or stringent execution of the study along with careful documentation of experimental steps, such unwanted noise will lead to misinterpretation of results. Experimental design strategies, *e.g.* clinical trial design including selection of appropriate inclusion and exclusion criteria, standardization of biospecimens sampling, preanalytical phase of biospecimens sampling, the inclusion of reference material and the stratification of samples across experimental batches, followed by statistical modelling of the data can alleviate the influence of technical noise[39]. We concluded that batch correction with ComBat-Seq, originally developed for mRNA-seq data, worked equally well for other omics layers. In the batch correction, we added sex as a covariate, to avoid removing sex-dependent effects from the data. However, the partial confounding of batch and sex in some omics layers could have been avoided by running equal numbers of males and females per batch.

Question 4: When is a multi-omics study pertinent and feasible? Answer: From our study, it is evident that running a multi-omics study in a controlled manner in five different laboratories is feasible and provides a comprehensive picture of the underlying biological systems. In most future studies, cost may still be a limiting factor, as each of the omics layers costed between 100 (metabolomics) and ~1,000 (EM-seq) euros per sample at the time of measurement.

We compared our results to those from other large multi-omics cohort studies, Swedish CArdioPulmonary bioImage Study (SCAPIS) in Sweden[40], Biobanking and Biomolecular Resources Research Infrastructure - Netherlands (BBMRI-NL) in the Netherlands[41,42], integrative Personal Omics Profiling (iPOP) in the USA[43,44] and the UK Biobank[45,46]. This revealed consistencies, complementarities and potential discrepancies. Our two miRNA profiling platforms, miRNA sequencing on white

blood cells and qRT-PCR profiling of plasma, have not been applied in SCAPIS, the Biobank-based Integrative Omics Study (BIOS) from BBMRI-NL and iPOP. The two miRNA profiling methods provide complementary data, because the miRNAs in the plasma usually originate from peripheral tissues (bound to lipoproteins[47], and/or extracellular vesicles[48]), with only a minor contribution from blood cells. Still, we observed that miRNA sequencing on white blood cells and qRT-PCR on plasma demonstrated co-variance (*i.e.* a correlation of 0.84 in PLS2). Compared to SCAPIS, we observed different trends in the fraction of omics features associated with sex (SCAPIS: transcripts > proteins > metabolites > lipids; our study: lipids > metabolites > proteins > transcripts). Similar inconsistencies were found for associations with BMI (SCAPIS: proteins > transcripts > metabolites > lipids; our study: metabolites > lipids, with almost no significant associations for proteins and transcripts). This may be attributed to age differences between the two cohorts, with our Czech cohort on average younger than the Swedish SCAPIS cohort, to other differences between the cohorts and the use of targeted antibody based-proteomics techniques *vs.* untargeted mass spectrometry-based proteomics. However, some of the top features associated with sex (L-leucine and *ZFX* and *KDM6A* mRNAs) in the SCAPIS study, were also highlighted as such in our study. The SCAPIS and iPOP studies focused on longitudinal fluctuations of omics features within an individual, whereas we focused on the cross-sectional variation of omics features across individuals. SCAPIS ranked proteins as most stable omics features, followed by lipids, metabolites and transcripts. Also in our cross-sectional analysis, proteins were most stable and transcripts most variable, with metabolites in-between. This suggests that features that vary between individuals are also more variable within an individual. Lipids were among the least variable in our cross-sectional cohort and seem to fluctuate more over time. Furthermore, the iPOP study identified a considerable higher percentage of proteins with seasonal fluctuations than transcripts or metabolites. Analytes with large (environmentally induced) fluctuations and large variation between individuals are not likely to serve as reliable biomarkers and including them in biomarker panels may lead to reduced power. We therefore recommend consulting our tables with coefficients of variations when designing (disease) biomarker assays and to deprioritize analytes with large coefficients of variations. Methylation profiling was only done in BIOS, and not in SCAPIS or iPOP. However, connections between methylation and miRNA levels, metabolomics, proteomics or lipidomics were not established in BIOS. On the contrary, BIOS detected many links between genetic variants and mRNA expression[41] and CpG methylation[42] (expression and methylation quantitative trait loci) that we could not identify due to our smaller sample size. Most notably, we identified candidate miRNA, mRNA, metabolite and lipid biomarkers for biological age and age acceleration predicted from the methylation profiles, underlying many relevant diseases, which have not been observed before. Lastly, our results confirmed some of the strongest associations of plasma protein levels with sex, age and weight found in the UK Biobank resource[46], which used a different platform for detecting protein levels (Olink). For, example, sex hormone-binding globulin (I3L145) strongly associated with both weight and sex in both cohorts. Similarly, Apolipoprotein F (Q13103) associated with sex and weight in both studies. In conclusion, the total overlap of significant associations was limited, compared to the large number of omics features. This highlights the need for more reference materials and standardized guidelines on multi-omics studies for assuring their reproducibility and further implementation in clinical practice.

Extensive methylation profiling using EM-seq is unique to our cohort but is remarkably consistent with the microarray-based methylation profiling performed for the BLUP age predictor[26]. The correlation of the BLUP predicted age with the chronological age in our cohort reaches 0.95, with a slope close to 1. This is even more remarkable since the BLUP predictor is based on methylation array measurements, whereas we used a sequencing-based method. This confirms the superior precision and accuracy of the BLUP predictor over other biological age predictors and the high quality and accuracy of our EM-seq methylation level assessment. Obviously, the much higher resolution and coverage of CpG sites in the EM-seq platform provides more opportunities for fine mapping of causal methylation sites and methylation quantitative trait loci over more commonly used array-based technologies. The age predictions using Levine[30] correlated less with the chronological age. This is not surprising, since this epigenetic clock has been specifically trained on a surrogate healthy age, determined based on clinical measurements. Therefore, age predictions based on Levine are more indicative of a healthy lifespan.

Biological age can also be estimated using other omics data. For instance, proteomic aging clocks have demonstrated

predictive power of age-specific mortality and disease[49], similar to the Levine epigenetic clock. Moreover, different omics age predictors can be integrated into a single multi-omics age predictor, which could be more accurate in predicting healthy lifespan and accelerated aging[50].

In line with results from the iPOP study, we found a number of omics features other than methylation sites associated with age. The metabolite analytes in the two studies do not overlap. Also, the top transcripts and proteins associated with age in the iPOP study (*ACAA2*, *ALDOB*, *CAB39*, *CD14* mRNAs and IGF2R, LV657, IGFALS, SCLT1, APOA4 proteins) were not associated with age in our study. Interestingly, transcripts that we associated with accelerated aging, such as *COQ10B*, *SH3GLB1* and MFSD14A, are closely linked to mitochondrial functions. These findings reinforce the well-established role of mitochondria in the aging process, as previously reported in the literature. Mitochondrial dysfunction has long been implicated in cellular aging and age-related diseases, underlining its critical contribution to maintaining cellular homeostasis and health span[51,52].

Some of the miRNAs associated with aging in our study have also been linked to age-dependent disease progression in other studies. However, it is important to note that miRNA profiles associated with age or disease may be tissue-dependent[18] and may not always be reflected in blood. Firstly, miRNAs miR-10b-5p, miR-17-5p and miR-24-3p, significantly associated with biological age in our single feature analyses, were associated to age of diagnosis in colorectal cancer[53]. Secondly, miR-139 affected senescence in hepatocellular carcinoma[54], and the miR-17-92 cluster is known to act as regulator among aging, cancer, and other diseases[55,56]. Thirdly, miR-144 and miR-221-3p impact brain aging and age-related cataracts, respectively[57]. Lastly, miR-17 and miR-32-5p regulate key senescence processes, including apoptosis and cell cycle[55,58,59]. Furthermore, top miRNAs on age-associated MOFA factor 10 were also associated with age predicted by the BLUP clock: hsa-miR-10b-5p, hsa-miR-17-5p, hsa-miR-24-3p, hsa-miR-199a-3p and hsa-miR-221-3p.

Conversely, none of the 25 miRNAs associated with chronological age identified by Huan *et al.* were associated with age in our study. This is not surprising because Huan *et al.*[15] analyzed miRNAs in whole blood, where platelet levels can alter the miRNA profiles. Nevertheless, we linked a number of these miRNAs to neutrophil and lymphocyte levels. These age associations may be driven by age-dependent variation in blood cell composition[60].

Limitations

In our studies, we mostly used computational methods (robust linear models, PCA, MOFA) that identify linear associations between omics profiles and phenotypes. There are alternative methods for finding latent dimensions across different omics layers. For example, more advanced techniques such as variational autoencoders are promising methods, because they are able to capture non-linear relations across the different omics layers[61]. Potentially, such methods could reveal more complex interactions across the omics layers that have not been found here.

We recognize that we are faced with a huge multiple testing problem in these types of multi-omics studies. While we applied careful multiple testing correction within an omics layer, we did not address multiple testing issues arising from analyzing multiple omics layers and phenotypes. This may have led to larger type I error rates than the 5% that was aimed for. Moreover, we have sought the validation of individual features mostly by comparing to the literature or by comparing with other multi-omics studies on healthy human volunteers. Validation of presented hits will require additional studies. Hence, the value of our current study is mainly in guiding more accurate future studies rather than in drawing conclusions on individual omics features.

Concluding remarks

The presented multi-omics study was undertaken within the context of the EATRIS-Plus project. EATRIS-Plus aimed to provide innovative scientific tools to the research community in the field of Personalized Medicine and increase the quality of the services provided by EATRIS (European Infrastructure for Translational Medicine). Best practices, guidelines and standard operating procedures arising from this and other studies were brought together in EATRIS's multi-omics toolbox (MOTBX at motbx.eatris.eu), a public resource for multi-omics researchers. Moreover, we put significant efforts in making both our data

and computational workflows FAIR (see data availability section), facilitating the use of this data for comparative analyses in future multi-omics studies, and studies developing or benchmarking multi-omics integration algorithms.

Methods

Clinical trial design

The project used a well-characterised Czech cohort of 1,109 healthy volunteers recruited during blood donations who were clinically examined and whose samples were bio-banked within the clinical trial “Assessment of Multi-omics Profiles in Health and Disease (ENIGMA)", NCT04427163. The inclusion criteria for participants in the study were: (a) age between 18-68, (b) absence of severe genetic diseases or a family history of such diseases, (c) absence of premature onset of civilization diseases such as hypertension, diabetes, autoimmune and cancer diseases, ongoing infectious or cancerous diseases, clinically manifest cardiovascular or pulmonary disorders, (d) not on long-term medication at the time of collection of biological material. An exclusion criterion was considered a failure to meet any of the inclusion criteria mentioned above. Participation in the study involved a visit for sample collection, including peripheral blood (BD Vacutainer EDTA and heparin blood collection tubes). All medical examinations are considered safe and minimally invasive.

Sample collection

Samples have been collected at the University Hospital Olomouc and biobanked at the Institute of Molecular and Translational Medicine (IMTM, Palacký University Olomouc). Blood plasma was separated by centrifugation and stored at $-80°C$ until analysed. Blood leukocytes were isolated after osmotic lysis of red blood cells, lysed in TRIzol™ Reagent (ThermoFischer Scientific) and stored at $-80°C$ until the total RNA or DNA extraction and analyses. The average time from blood sampling to processing did not exceed one hour. All samples were distributed from the IMTM biobank to EATRIS-Plus partner sites on dry ice.

A representative subset (age and sex balanced) of 127 samples was selected. The samples were selected according to the parametric or non-parametric distribution (including box-cox transformation strategies, 95% confidence interval). The subset cohort was sex balanced, with an age span ranging from 21 to 61 years. The IMTM distributed these samples to EATRIS-Plus centres for further omics measurements.

Ethics statement

The multi-omics cohort used in EATRIS-Plus project was approved by the Ethics Committee of the University Hospital in Olomouc and the Faculty of Medicine of the Palacký University in Olomouc under the reference number 119/18 on 9 July 2018.

The EATRIS-Plus project is supervised by EATRIS-Plus Ethics Committee, chaired by Michaela Bendova. EATRIS-Plus Ethics Committee ensures that the EATRIS-Plus project’s ethical compliance is guaranteed, and that the project meets the highest ethical standards while maintaining compliance with national regulations. More information can be found in EC Annual Reports.

Genomics - WGS

Sample preparation

The TruSeq DNA PCR-Free LT Kit (Illumina) was used to prepare samples for whole genome sequencing. WGS libraries were prepared according to the manufacturer’s protocol. The input DNA amount for WGS library preparation with genomic DNA for TruSeq-PCR-free libraries was 1 µg. The same fragmentation conditions for all samples on Covaris instrument were used with a targeted size of 350 bp. The concentration of the TruSeq DNA PCR-Free libraries for WGS was measured by fluorometry using Qubit 2.0 fluorometer with HS dsDNA Assay kit (ThermoFisher Scientific, Q32854). The samples were equimolarly pooled based on the measured concentration and final quality check of sample pool was measured by qPCR with the KAPA Library Quantification Complete Kit (Universal) (Roche, KK4824); by fluorometry using Qubit 2.0 fluorometer with

HS dsDNA Assay kit (ThermoFisher Scientific, Q32854); and by capillary electrophoresis using 2100 Bioanalyzer instrument (Agilent) with the High Sensitivity DNA Kit (Agilent, 5067-4626).

For the WGS library preparation, the sequencing was performed using Illumina platform NovaSeq 6000 at 2 x 161 bases read length using the S4 configuration (cat 20028312). Sequencing was performed following the manufacturer's instructions.

Data processing

Raw sequencing data were processed to obtain sequence information. Quality control (QC) was performed using *FastQC (v0.11.9)*. The filtering of low-quality reads and adapter contamination trimming was performed using *fastp (v0.20.1)*. The remaining reads were aligned to the reference human genome hg38/GRCh38 primary assembly with the software package *BWA (v0.7.17-r1188)* using default parameters. This was followed by the of marking duplicate reads and fix mate information using *Picard tools (2.27.5)*[62] .

QC steps and the coverage were checked using the in-house software Genovesa https://www.bioxsys.eu/#/genovesa (developed by Bioxsys, s.r.o). Single-nucleotide variants (SNVs) and insertion/deletion variants (indels) were called using the VarScan v2.4.4[63] (with parameters –min-coverage 20; –min-var-freq 0.1; —p-value 0.5, —min-avg-qual 15).

The variant calling process is based on a Fisher's Exact test (applying algorithms Samtools, Varscan, Freebayes, and their combinations), a statistical test procedure that calculates an exact probability value for the relationship between two dichotomous variables, as found in two by two crosstable. The program calculates the difference between read counts supporting reference and variant alleles with a p-value threshold 0.05. Only SNVs and indels passing the quality filter (a minimal quality of coverage ≥ 30, base quality ≥ 10, mapping quality ≥ 5) and with an alternative allele frequency $\geq 20\%$ per sample, p-value (Fisher exact test) < 0.05 were included for further variant filtering.

Genomics - aCGH

The method is based on complementary binding of tested DNA to 2.6 million oligonucleotide markers immobilized on microarray cartridge. Obtained fluorescent intensities are statistically processed and sample genotype changes are obtained by comparing to the reference model. The array type CytoScan HD (Thermo Fisher Scientific) was used.

Total genomic DNA (130 ng) was fragmented, labelled with fluorescent dyes and hybridized onto the microarray according to the manufacturer instructions. Arrays were scanned using GeneChip Gene Scanner 3000 7G (Thermo Fisher Scientific), the fluorescent intensity for each oligonucleotide markers were measured and .DAT files (image file) were generated. Each marker has its own known position on the array, using annotation library the fluorescent intensity for each marker were calculated and .CEL files were subsequently generated. The scanning process and generating of .DAT and .CEL files were performed using Affymetrix GeneChip Command console version.

The generated .CEL files were annotated and analyzed using Chromosome analysis suite (CHAS) (Applied Biosystems by Thermo Fisher Scientific), version 3.3.0.139 (r10838) and genome version hg38. The generated .CYCHP files were viewed using CHAS. Table of Array Data Quality Control Metrics for all 126 processed CELL files (Filename column) corresponding to the Czech Cohort samples (ID column) was exported from CHAS. 118 samples passed MAPD, SNPQC and waviness.SD and QC thresholds. For all samples passing QC, a w.log2ratio R Data object was created. This object served as an input of segmentation process.

The log2 ratio values were processed using the *rCGH R/Bioconductor*[64] package with default parameter settings, using circular binary segmentation algorithm (CBS)[65], Expectation-Maximization algorithm (EM) for genomic profile centralization[64] and then converting into DNA copy number table over genes. For multi-omics integration process only the subset of 2,549 autosomal genes with at least one sample log2 ratio value outside the interval for normal gene status <-0.5,0.5> was selected (Suppl Table 1). The missing values imputation by Random Forest method implemented in missForest R package[66] was applied on the complete by-gene table.

Epigenomics – Enzymatic methylation sequencing

EM-seq sequencing libraries were generated from 100 ng of DNA using the NEBNext Enzymatic Methyl-seq sample preparation kit (cat E7120S/E7120L, NEB), incorporating unique dual indexes. Library preparation strictly followed the manufacturer's instructions (manualE7120), inclusive of the addition of control DNAs (pUC19 and Lambda). Quality control assessment was conducted using the Fragment Analyzer equipped with the DNF-910 dsDNA kit (Agilent) to ensure library integrity. Libraries were quantified by qPCR, utilizing the Library Quantification Kit for Illumina (KAPA Biosystems) on a CFX384 Touch instrument (Bio-Rad).

Sequencing was performed on an Illumina NovaSeq6000 system, with a paired-end 151 bp read configuration (Illumina, San Diego, CA, USA). All libraries were sequenced to a minimum of 300 M paired-end reads per sample with read length 2x151 bp. Sequencing data was demultiplexed and converted from base calls into FASTQ files using Illumina bcl2fastq2 Conversion Software v2.20.0.422. Quality control after sequencing was performed using the in-house quality control tools CheckQC www.github.com/Molmed/checkQC and Seqreports www.github.com/Molmed/seqreports as well as FastQ Screen from Babraham Bioinformatics (www.bioinformatics.babraham.ac.uk/projects/fastq_screen/). Primary methylation sequencing data analysis was performed with nf-core analysis pipeline methylseq (version 1.6.1), the reference genome used was Ensembl GRCh38, release 105. Pipeline output includes alignment to the reference genome, methylation calls and quality metrics. The controls pUC19 and lambda were aligned using Bismark version 0.23.0 (Babraham Bioinformatics) and the methylation rate was compared against threshold provided by vendor. Quality control after primary analysis consisted of PCA (Fig S1) on the sites (overlapping (3000) a set of array probes selected for sex classification[67].

We used the *methrix (v1.10.0)* R/Bioconductor package for data processing using BSgenome. Hsapiens.NCBI.GRCh38 as reference genome. Samples failed on prior QC measures for sequencing for low conversion rate (n=2) or low alignment rate (n=3) were excluded from further analysis. Information from the different strands on the same CpG site was merged. The resulting object was further filtered by removing all CpG sites covered by < 5 reads in >50% of the samples (n= 1,525,790). CpG-level DNA methylation measurements were indicated as missing values (NA) if the coverage was <5 reads for a given CpG site and sample. The CpG sites not covered by any samples were removed from the dataset (n = 12,775). The sites overlapping with common SNPs as defined by a minor allele frequency > 0.01 in the 1kGenomes based on *MafDb.1Kgenomes.phase3.hs37d5* data package were removed using the *remove_snps* function as implemented in the methrix R package (n = 3,452,226). Cell type composition was assessed using an in-house implementation of the reference-based Houseman algorithm[68] using DNA methylation from the recently published DNA methylation atlas[69] (Fig S2).

To avoid the large number of CpG sites dominating the multi-omics analyses, we selected the 100,000 most variable CpG sites. To adjust the dataset for cell type composition, we first fitted a linear model between the cell type composition and the beta values for each CpG site. If the given site was associated with the cell type composition with a $p < 0.01$, we exchanged the beta values with the residuals of the model. If not, we used the centered beta value.

Transcriptomics – mRNAseq

Library preparation from 800 ng of total RNA was performed according to Illumina stranded mRNA prep Reference Guide (Illumina, San Diego, CA, USA). Library quality check was performed using LabChip GX Touch HT High Sensitivity assay (PerkinElmer, USA). Normalized and pooled libraries were quantified using KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA, USA) and sequenced on the Illumina NovaSeq6000 system to a minimum of 50 M paired-end reads for each sample (Illumina, San Diego, CA, USA). Read length for the paired-end run was 2x151 bp. Sequencing data is demultiplexed and converted from base call files into FASTQ files with Illumina *bcl2fastq2 Conversion Software (v.2.20)*. Sequencing was performed in multiple batches, each with a reference sample Brain RNA for normalization purposes (Human brain total RNA (Thermo Fisher AM7962). Primary RNA-sequencing data analysis is performed with in-house Nextflow-based FIMM-RNAseq data analysis pipeline version 2.0.1. Pipeline outputs include alignments to reference genome and various quality metrices. The gene counts that were produced by the pipeline underwent a variance stabilization transformation from R/Bionconductor's package *DESeq2 (v 1.36.0)*, which also included correction for size and normalization factors and provided

log-transformed counts.

Transcriptomics – MicroRNAseq

A clean up of the samples was performed using the miRNEasy micro kit from QIAGEN (cat. no. 74004). Total RNA was eluted using 15 μL RNase-free water. Library preparation from 600 ng of total RNA was performed in batches of 6-24 samples according to Lexogen Small RNA-Seq library prep Reference Guide (Lexogen GmbH, Vienna, Austria). Library quality check was performed using Agilent Bioanalyzer High Sensitivity DNA assay (Agilent Technologies, Waldbronn, Germany) and libraries were pooled based on the concentrations acquired from the assay. The pooled libraries were subjected to size selection targeting fragments that were 130-180 bp in size using BluePippin (Sage Science, Beverly, Massachusetts, USA). The size selected library pools were quantified using KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA, USA) and sequenced on the Illumina NovaSeq6000 system for 100 cycles using the S1 flow cell with a lane divider to a minimum read depth of 10 M reads for each sample (Illumina, San Diego, CA, USA). Read length for the single-end run was 101 bp.

Sequencing data is demultiplexed and converted from base call files into FASTQ files with Illumina *bcl2fastq2 Conversion Software (v.2.20)*. Primary miRNA-sequencing data analysis is performed with Nextflow-based FIMM-smRNAseq data analysis pipeline. General output includes a sample-specific read count file describing the number of reads mapping to each miRNA and an alignment visualization file.

Transcriptomics – microRNA qRT-PCR

Total RNA enriched in miRNAs was isolated using miRNeasy Serum/plasma advanced kit (217204, QIAGEN) starting from 250 μL of plasma and eluted in a final volume of 25 μL, following the manufactures instructions. Before RNA isolation, 1 μL of synthetic RNA Spike-in was added to each sample (miRCURY RNA Spike-in kit). Spike in amplification serves as technical control of RNA extraction efficiency homogeneity. We used the Universal RT miRNA PCR System (Qiagen) for cDNA synthesis. Briefly, 16 μL of RNA was used as a template for RT in a final volume of 80 μL. An external RNA (cel-miR-39) was added and further amplified to cDNA synthesis efficiency control. Sample hemolysis was checked by miR-23a-3p and miR-451a amplification. Samples in which the subtraction of the expression of these miRNAs was higher than 7.5 were discarded due to elevated levels of hemolysis. 179 miRNA expression was evaluated by qRT-PCR using microRNA Array Profiling (miRCURY LNA miRNA Focus Panels; YAHS-106YG, Qiagen). All reactions were carried out in triplicate using a Light Cycler 480 instrument (Roche, Basel, Switzerland), and Cq values were calculated using a 2nd derivative method (Light Cycler 480 Software 1.5, Roche). Raw Cq values were transformed to Z-scores and normalized among plates, replicates and samples. Replicates with missing values higher than 18 miRNAs (10% of the miRNAs) were removed from the analysis. Internal controls (mean of UniSP2 $>$ 25, mean of cel-miR39 $>$ 30 and Nº of missing values > 40%) and hemolysis were used as quality controls to remove samples from the study.

Proteomics

Plasma peptide mixtures were prepared from plasma samples by the FASP protocol adapted from[70]. The protein concentration was determined by BCA assay, and 100 μg of total proteins were processed. All processed plasma samples were diluted into a 0.1 $\mu g/\mu l$ concentration. Dilution was based on the results from the peptide concentration assay (see MOTBX protocol "Sample preparation for plasma analysis"[71]). During dilution, PROCAL peptides (JPT, cat. n. RTK-1-100pmol) are added to obtain a final concentration of 10 fmol$/\mu l$. Each sample was analyzed in two technical replicates. The 1 µg of peptides and 100 fmol of PROCAL peptides in injection volume 10 µl for 1 technical replicate were used. Peptides were separated by liquid chromatography (UltiMate 3000 RS series, Thermo Scientific™) and detected by mass spectrometry (Orbitrap Fusion™ Tribrid™, Thermo Scientific™). Measurements of all 127 Czech Cohort samples consisted of 4 batches. Each batch started with an SST Test (5 injections of BSA standard); samples were released for analysis based on the results of the SST. The 10 injections of plasma peptide samples (5 samples in duplicate) and 1 BSA standard were running repeatedly until all plasma samples in the batch were measured. All detailed parameters of liquid chromatography and mass spectrometry settings are

described in MOTBX protocol "LC-MS (HCD OT) Plasma Analysis"[72]. Raw data of each technical replicate were searched in Proteome Discoverer (PD) software, (v. 2.5) (Thermo Scientific) in two subsequent steps -the processing and consensus steps. The processing workflow included a search against the complete human UniProtKB database, including Swiss-Prot reviewed and TrEMBL computationally analyzed but unreviewed proteins (https://www.uniprot.org/uniprot/, downloaded in July 2020). In the consensus step, the normalization of abundances was done on specific protein amounts of PROCAL peptides. Detailed settings of both steps have been described in the SOP_Search_PD2.5_Plasma analysis. Results with identified and quantified proteins were exported from PD 2.5 in tab-separated format and processed with software R.

The raw abundances were log2 transformed to make data distribution close to normal. Replicates abundances were aggregated by mean and three filters after exclusion of QC feature were applied: 979 proteins completely missing in at least one batch were excluded, and 29 low FDR confidence (q-value ≥ 0.05) proteins were removed. Only features with no more than 30% of missing values were kept in the dataset. After filtering the dataset consisted of 1393 proteins (features) abundances in rows and 127 samples in columns with 7.9% of missing values in total. A local similarity random forest method implemented in missForest R package[66] was chosen for imputation and a dataset with complete observations was created.

Targeted Metabolomics

Data acquisition

Quantitative targeted metabolomics data were acquired at Radboud university medical center (Radboudumc) and Maastricht UMC+ (MUMC+). Plasma acylcarnitine analysis (ACRN, 44 metabolites) was performed at Radboudumc according to a protocol previously described[73]. Plasma amino acid analysis (AA, 53 metabolites) and very long chain fatty acid analysis (VLCFA, 5 metabolites) were performed at MUMC+ according a previously described[74] and an in-house protocol, respectively. Metabolite levels were reported as µM concentrations.

The 127 plasma samples from the Czech cohort were measured in multiple batches: four batches for ACRN, two batches for AA, and six batches for VLCFA analysis. For ACRN analysis, a mixture of compounds with known concentrations (quality control / QC sample) was measured along with each batch. For AA analysis, two different mixtures of compounds with known concentrations were measured along with each batch. For ACRN analysis, the limit of detection (LOD) is 0.01 µM. For AA analysis, the limits of quantification (LOQ) are below 1 µM, except for ethanolamine (hmdb:HMDB0000149) where LOQ is 2.5 µM, and for formiminoglutamic acid (hmdb:HMDB0000854), 3-methylhistidine (hmdb:HMDB0000479), 1-piperideine-6-carboxylic acid (chebi:49015), and allysine (chebi:17027) for which LOQ were not determined. Metabolites that could not be detected (true value < LOD) were reported with a concentration of 0 µM.

Data pre-processing and preparation

Metabolites missing in $\geq 30\%$ of the samples were removed. Since metabolite concentrations follow a log-normal distribution more closely than a normal distribution, concentrations were log-transformed using the natural logarithm after adding a constant value to each value (log(x+1)). For downstream analyses, the metabolite features were mean-centered to zero and scaled to unit-variance. Missing values were imputed using Low abundance resampling (Suppl Table S3, Suppl Fig S29-31).

Untargeted Lipidomics

Sample preparation

Before the analysis, 50 μL of plasma (thawed on ice) was mixed with 150 μL of ice-cold LCMS isopropanol containing Avanti SPLASH internal standard mixture, FA 20:4(d8), Cer d18:1(d7)/15:0 and vortexed for 20 s. Samples were centrifuged (22,400g, 20 mins at 4°C), and 150 μL of the supernatant was loaded into a low recovery volume HPLC vial with insert. Blank samples were prepared using the same procedure instead of gradient fraction LCMS water. Pooled quality control samples (QC) were prepared by mixing 10 µL of plasma from every sample and thoroughly vortexed (5 min). Each QC (n=18) sample was prepared individually in the same way as the biological samples. QC samples were injected as every 6th sample through the analytical batch. The standard reference material samples (SRM NIST 1950) were prepared the same way as other biological samples

and injected at the beginning and end of the analytical batch. Samples were randomised for the first time before the sample preparation. Second, separate randomization was done for the data acquisition.

LC-MS Analysis:

Separation of non-polar compounds was done using reversed-phase chromatography separation mechanisms on an Accucore C30 column (150 mm ×2.1 mm, 2.6 μm; Thermo Fisher Scientific, MA, USA). The constitution of mobile phase A was 20 mM ammonium formate in 60:40 (v:v) acetonitrile:water and mobile phase B consisted of 20 mM ammonium formate in 85.5:9.5:5 isopropanol:acetonitrile:water (v:v:v).

The gradient was as follows: t= 0.0, 20% B; t= 2.5, 20% B, t= 2.6, 55% B; t= 12, 60% B; t= 12.1, 80% B; t= 19.0, 90% B; t= 21.0, 100% B; t= 23.0, 100% B; t= 23.1, 20% B; t= 30.0, 20% B. All changes were linear (curve = 5), and the flow rate was 0.400 mL/min. The column temperature was 55°C, and the injection volume was 2 μL. Data were acquired separately in positive and negative ionization modes (150–2000 m/z) with a resolution of 60,000. Ion source parameters: sheath gas = 48 arbitrary units, aux. Gas = 15 arbitrary units, sweep gas = 0 arbitrary units, spray voltage = 3.5 kV (positive ion)/2.5 kV (negative ion), capillary temp. = 350 °C, aux. gas heater temp. = 400 °C. The full scan data were acquired separately in files without any MS/MS scans.

The detailed settings were as follows: mass resolution =17,500 (FWHM at m/z 200); isolation width =3.0 m/z; stepped normalised collision energies (stepped NCE) = 20, 40, 100 for positive ionization mode and (stepped NCE) = 40, 60, 130 for negative ionization mode. Spectra were acquired with different precursor m/z ranges in different segments of the chromatogram: t =0–6 min, 150–1100 m/z; t =6–12 min, 550–1250 m/z; t =12–17 min, 550–1500 m/z; t =17–30 min, 550–1100 m/z in case of 30 min elution gradient and for the 50 min elution gradient the segments were as follows: t =0–10 min, 150–1100 m/z; t =10–20 min, 550–1000 m/z; t =20–25 min, 550–1100 m/z; t =25–30 min, 550–1500 m/z; t =30–40 min, 550–1050 m/z; t =40–50 min, 550–1500 m/z. Thermo Tune Plus (2.7.0.1112 SP2) software controlled the instrument. All data were acquired in profile mode[75].

Data were processed and evaluated using Compound Discoverer 3.3 SP1 software (Thermo Fisher Scientific, MA, USA). The workflow was set as follows: Chromatograms were aligned using the self-learning function "ChromAlign" using a QC sample as a reference. Peak detection & Feature grouping was done directly from Thermo Fisher Scientific profile data (*.raw). The missing values were filled using the "Fill Gaps node – Real Peak detection" option. Signal/Batch correction was performed by SERRF QC Correction (Systematical Error Removal using Random Forest, n=200)[6]. The data matrix was filtered using criteria as follows: Compounds with > 50% missing in QC samples and with relative standard deviation (RSD) > 30% were removed; Compounds with corrected areas with RSD > 25% were removed. Blank substruction was done with the ratio of intensity sample/blank = 20/1. Finally, the data matrix was normalised using Median Absolute Deviation (MAD). For the purposes of statistical evaluation, the data matrix was log10 transformed, and univariate scaling was performed. The complexity of metabolite identification is depicted in Suppl Fig S32.

The identification of the compounds was performed on various levels using mzVault and LipidBlast (V68) spectral libraries as well as ChemSpider and LipidMAPS online databases (Suppl Fig S32).

Batch effect adjustment

Batch effect adjustment following the Combat method[7] was performed on proteomics, metabolomics, mRNA-seq and miRNA data using R/Bioconductor's *sva (v3.44.0)*. The parametric method was applied using sex as covariate to retain information related to sex in an unbalanced batch design and without covariate. For the mRNA-seq data, Combat was applied on the gene counts table with the *Combat_seq* function, which is designed specifically for bulk RNA-seq count data. More information on batch-effect adjustment can be found in Suppl Figures S4-9).

Cell type correction

The estimated cell type proportions derived from the EM-seq data (see Methods Epigenomics), were used to correct for the influence of cell types on the transcriptomics and epigenomics data. This correction process involved employing a linear model

where the estimated cell type numbers were used as covariates, after which the residuals of this model were retrieved for downstream analysis. The estimated cell type compositions can be found in Suppl Fig S12.

Reference values

For reporting of reference values, mean, median, standard deviation, minimum, and maximum –omics levels were calculated across all individuals in the cohort. Additionally, the CV was calculated by dividing the standard deviation with the mean for every feature and multiplied by 100 to derive percentages. The reference values were reported for male and female individuals separately and are available in Suppl Table 2.

Single-omics data analyses

Single –omics analyses were performed using R version 4.2.1. The robust linear models were implemented using R package *MASS (version 7.3-60)*. For each feature, linear models were run with no covariates, sex as covariate and for only females and males. PCA was performed using the built-in R function *prcomp*. Correlations of the principal component scores with the continuous phenotypic variables were tested using Pearson pairwise correlation. All p-values were corrected with Benjamini-Hochberg's false discovery rate (FDR).

Multi-omics data analyses

Multi–omics analyses were also performed using R version 4.2.1. Multi-Omics Factor Analysis (MOFA) was implemented in R using a combination of the R/Bioconductor library MOFA2 (version 1.6.0) and the conda library *mofapy2 (v 0.6.7)*, which was called upon by the reticulate package (version 1.3.1). Correlations of the individuals factor scores with the continuous phenotypic variables were tested using Spearman pairwise correlation, where all p-values were corrected using Benjamini-Hochberg's false discovery rate (FDR). The factor loadings of the mRNA-seq, proteomics, metabolomics and EM-seq data were used for gene set enrichment analysis. The R library *multiGSEA (v1.11.2)* was used to find significantly enriched pathways. The multiGSEA function was customized to enable parameter tuning. Standard deviation was used as gene set enrichment analysis (GSEA) score type and the boundary for calculating p-values was set to 1e-100. For the EM-seq data, gene set enrichment was carried out using the *gseGO* function of the R library *clusterProfiler (v4.2.2)*, focusing on the Gene Ontology (GO). CpG sites were mapped to nearest genes using the R/Bioconductor library *biomaRt (v2.50.3)*. As many CpG sites were mapped to the same gene, we selected the CpG site with the highest loading for every gene.

Independent Component Analysis was implemented in R/Bioconductor package *consICA (v.2.0.0)*[22] in order to find 20 independent components per –omics layers. Linear models that were used to predict phenotypic outcomes from the independent component, were created with standard R *lm* tool. Overrepresentation analysis was performed with R library *topGO (v 2.48.0)* within *consICA* package, independently for positively and negatively contributing genes from each component. Network of the correlated components was built using the threshold of $R^2 > 0.1$, calculated on component weights over all common patients (corresponding p-value <0.001).

Similarity Network Fusion was implemented in Python version 3.9.5, using the library *SNFpy (v0.2.2)*[23]. A fused network was constructed for every combination of –omics data, with a minimum of 2 and maximum of 10 –omics layers. The numbers of sample clusters were selected by a built-in function of the library, which finds the optimal number of clusters based on the eigengap method. The goodness of clustering for every combination was quantified with the silhouette score, which was determined with the *SNFpy* library as well. We assessed whether phenotypic differences could be found among the different SNF clusters, using Mann-Whitney U tests (two clusters) or Kruskal-Wallis tests (more clusters).

PLS2 was applied using the R/Bioconductor library *mixOmics (v6.20.0)* to find pairwise correlations among all –omics data. The sparse Partial Least Squares (sPLS) method was implemented using the canonical analysis algorithm, where 2 latent dimensions (PLS components) were calculated for each model. A maximum of 1,000 features per –omics layers was included for each model, which the model selects using LASSO penalization. Pearson correlations were tested across the sample scores on the first PLS component.

Epigenetic clocks

Epigenetic clocks were calculated using the *methylclock (v1.2.1* Bioconductor package. Since the methylation clocks are based on Illumina arrays, we filtered our dataset and identified the CpG sites for each site on the Illumina array. After calculating the biological age, we used the age acceleration, the difference between biological and chronological age, and applied robust linear models as described above.

Data availability

Both clinical and multi-omics data were collected and are stored in ClinData Portal https://clindata.imtm.cz/. ClinData is a software solution for collecting of GDPR compliant clinical and laboratory data on patients included in research projects and clinical studies. ClinData is based on client/server architecture, the application runs on a server, a user is connected through a web browser. All communication between the client and the server is secured through SSL encryption, an industry standard. The Java 8 programming language is used for the server, which allows sustainability and long-term development. The client interface uses open source technologies – HTML, Java Script, jQuery, Angular, Bootstrap.

Moreover, the processed multi-omics data, has been publicly available through Zenodo (10.5281/zenodo.10246047). However, genomics data has been excluded to preserve participant privacy.

Code availability

All scripts that were used to perform both the single- and multi-omics analyses, are made available through a Nextflow[76] pipeline: https://github.com/EATRIS/Multi_Omics_Workflow.git. The pipeline can be re-used for other data sets, as it has been developed following the FAIR4RS principles[77].

References

1. Gallego-Paüls, M. *et al.* Variability of multi-omics profiles in a population-based child cohort. *BMC Medicine* **19**, DOI: 10.1186/s12916-021-02027-z (2021).
2. Brooks, T. G., Lahens, N. F., Mrčela, A. & Grant, G. R. Challenges and best practices in omics benchmarking. *Nat. Rev. Genet.* **25**, 326–339, DOI: 10.1038/s41576-023-00679-6 (2024).
3. Olshansky, G., Giles, C., Salim, A. & Meikle, P. J. Challenges and opportunities for prevention and removal of unwanted variation in lipidomic studies. *Prog. Lipid Res.* **87**, 101177, DOI: 10.1016/J.PLIPRES.2022.101177 (2022).
4. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects dna methylation at single-base resolution from picograms of dna. *Genome Res.* **31**, 1280–1289, DOI: 10.1101/gr.266551.120 (2021).
5. Foox, J. *et al.* The seqc2 epigenomics quality control (epiqc) study. *Genome Biol.* **22**, DOI: 10.1186/s13059-021-02529-2 (2021).
6. Fan, S. *et al.* Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data. *Anal. Chem.* **91**, 3590–3596, DOI: 10.1021/acs.analchem.8b05592 (2019).
7. Zhang, Y., Parmigiani, G. & Johnson, W. E. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics Bioinforma.* **2**, DOI: 10.1093/nargab/lqaa078 (2020).
8. Žilina, O. *et al.* Somatic mosaicism for copy-neutral loss of heterozygosity and dna copy number variations in the human genome. *BMC Genomics* **16**, DOI: 10.1186/s12864-015-1916-3 (2015).
9. Zomer, A. W. M. *et al.* Pristanic acid and phytanic acid: naturally occurring ligands for the nuclear receptor peroxisome proliferator-activated receptor. *J. Lipid Res.* **41**, 1801–1807, DOI: 10.1016/S0022-2275(20)31973-8 (2000).
10. Jager, S., Cramer, D. A. & Heck, A. J. Normal alpha-1-antitrypsin variants display in serum allele-specific protein levels. *J. Proteome Res.* **22**, 1331–1338, DOI: 10.1021/acs.jproteome.2c00833 (2023).

11. Argelaguet, R. *et al.* Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, DOI: 10.1186/s13059-020-02015-1 (2020).

12. Sanz, E., Bean, J. C., Carey, D. P., Quintana, A. & McKnight, G. S. Ribotag: Ribosomal tagging strategy to analyze cell-type-specific mrna expression in vivo. *Curr. Protoc. Neurosci.* **88**, DOI: 10.1002/CPNS.77 (2019).

13. Li, X. *et al.* Inflammation and aging: signaling pathways and intervention therapies. *Signal Transduct. Target. Ther.* **8**, DOI: 10.1038/s41392-023-01502-8 (2023).

14. Potter, M. L., Hill, W. D., Isales, C. M., Hamrick, M. W. & Fulzele, S. Micrornas are critical regulators of senescence and aging in mesenchymal stem cells. *Bone* **142**, DOI: 10.1016/j.bone.2020.115679 (2021).

15. Huan, T. *et al.* Age-associated microrna expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell* **17**, DOI: 10.1111/acel.12687 (2018).

16. Noren Hooten, N. *et al.* microrna expression patterns reveal differential expression of target genes with age. *PLOS ONE* **5**, 1–13, DOI: 10.1371/journal.pone.0010724 (2010).

17. Machida, T. *et al.* Micrornas in salivary exosome as potential biomarkers of aging. *Int. J. Mol. Sci.* **16**, 21294–21309, DOI: 10.3390/ijms160921294 (2015).

18. Kinser, H. E. & Pincus, Z. Micrornas as modulators of longevity and the aging process. *Hum. Genet.* **139**, 291–308, DOI: 10.1007/s00439-019-02046-0 (2020).

19. Čarna, M. *et al.* Pathogenesis of alzheimer's disease: Involvement of the choroid plexus. *Alzheimer's Dementia* **19**, 3537–3554, DOI: 10.1002/alz.12970 (2023).

20. Balduini, C. L. & Noris, P. Platelet count and aging. *Haematologica* **99**, 953–955, DOI: 10.3324/haematol.2014.106260 (2014).

21. Mohamed, E. A. & Sayed, W. M. Implication of jak1/stat3/socs3 pathway in aging of cerebellum of male rat: Histological and molecular study. *Sci. Reports* **10**, DOI: 10.1038/s41598-020-64050-z (2020).

22. Chepeleva, M., Kaoma, T., Zinovyev, A., Toth, R. & Nazarov, P. V. consica: an r package for robust reference-free deconvolution of multi-omics data. *Bioconductor* **4**, DOI: 10.1093/bioadv/vbae102 (2024).

23. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337, DOI: 10.1038/nmeth.2810 (2014).

24. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixomics: An r package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, 1–19, DOI: 10.1371/journal.pcbi.1005752 (2017).

25. Choi, H., Joe, S. & Nam, H. Development of tissue-specific age predictors using dna methylation data. *Genes* **10**, DOI: 10.3390/genes10110888 (2019).

26. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Medicine* **11**, DOI: 10.1186/s13073-019-0667-1 (2019).

27. Alfonso, G. & Gonzalez, J. R. Bayesian neural networks for the optimisation of biological clocks in humans. *bioRxiv* DOI: 10.1101/2020.04.21.052605 (2020).

28. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367, DOI: 10.1016/j.molcel.2012.10.016 (2013).

29. Horvath, H. & Horvath, S. Dna methylation age of human tissues and cell types. *Genome Biol.* **14**, 115, DOI: 10.1186/gb-2013-14-10-r115 (2013).

30. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *AGING* **10**, DOI: 10.18632/aging.101414 (2018).

31. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to hutchinson gilford progeria syndrome and ex vivo studies. *AGING* **10**, DOI: 10.1186/gb-2013-14-10-r115 (2018).

32. Loeffler, I. & Ziller, N. Sex-related aspects in diabetic kidney disease—an update. *J. Clin. Medicine* **12**, DOI: 10.3390/jcm12082834 (2023).

33. Scherer, A. Multi-omics quality assessment in personalized medicine through eatris. *bioRxiv* DOI: 10.1101/2023.10.25.563912 (2023). https://www.biorxiv.org/content/early/2023/11/08/2023.10.25.563912.full.pdf.

34. Scherer, A. Technological Reference Protocols for Transcriptomic, Proteomic and Metabolomic Analysis in EATRIS-Plus Project, DOI: 10.5281/zenodo.5513674 (2021).

35. Najdekr, L. *et al.* Sample Analysis Using Individual Omic Techniques and Data Outputs, DOI: 10.5281/zenodo.10300981 (2023).

36. Piehowski, P. D. *et al.* Sources of technical variability in quantitative lc-ms proteomics: Human brain tissue sample analysis. *J. Proteome Res.* **12**, 2128–2137, DOI: 10.1021/pr301146m (2013).

37. Xiao, W. *et al.* Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* **39**, 1141–1150, DOI: 10.1038/s41587-021-00994-5 (2021).

38. Su, Z. *et al.* A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.* **32**, 903–914, DOI: 10.1038/nbt.2957 (2014).

39. Yu, Y. *et al.* Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol.* **24**, DOI: 10.1186/s13059-023-03047-z (2023).

40. Bergström, G. *et al.* The swedish cardiopulmonary bioimage study: Objectives and design. *J. Intern. Medicine* **278**, 645–659, DOI: 10.1111/joim.12384 (2015).

41. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145, DOI: 10.1038/ng.3737 (2017).

42. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138, DOI: 10.1038/ng.3721 (2017).

43. Li-Pook-Than, J. & Snyder, M. Ipop goes the world: Integrated personalized omics profiling and the road toward improved health care. *Chem. Biol.* **20**, 660–666, DOI: 10.1016/j.chembiol.2013.05.001 (2013).

44. Huang, W. *et al.* Ipop: An integrative plant multi-omics platform for cross-species comparison and evolutionary study. *Mol. Biol. Evol.* **40**, DOI: 10.1093/molbev/msad248 (2023).

45. Sudlow, C. *et al.* Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12**, DOI: 10.1371/journal.pmed.1001779 (2015).

46. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the uk biobank. *Nat.* | **622**, 329, DOI: 10.1038/s41586-023-06592-6 (2023).

47. Michell, D. L. & Vickers, K. C. Lipoprotein carriers of micrornas. *Biochimica et Biophys. Acta - Mol. Cell Biol. Lipids* **1861**, 2069–2074, DOI: 10.1016/j.bbalip.2016.01.011 (2016).

48. Lässer, C. Mapping extracellular rna sheds lights on distinct carriers. *Cell* **177**, 228–230, DOI: 10.1016/j.cell.2019.03.027 (2019).

49. Argentieri, M. A. *et al.* Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat. Medicine* **30**, 2450–2460, DOI: 10.1038/s41591-024-03164-7 (2024).

50. Chen, Q. M. *et al.* Omicmage: An integrative multi-omics approach to quantify biological age with electronic medical records. *bioRxiv* DOI: 10.1101/2023.10.16.562114 (2023).

51. Sun, N., Youle, R. J. & Finkel, T. The mitochondrial basis of aging. *Mol. Cell* **61**, 654–666, DOI: 10.1016/j.molcel.2016.01.028 (2016).

52. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194, DOI: 10.1016/j.cell.2013.05.039 (2013).

53. Moreno, E. C. *et al.* Novel molecular characterization of colorectal primary tumors based on mirnas. *Cancers* **11**, DOI: 10.3390/cancers11030346 (2019).

54. Wang, K. *et al.* Ezh2-h3k27me3-mediated silencing of mir-139-5p inhibits cellular senescence in hepatocellular carcinoma by activating top2a. *J. Exp. Clin. Cancer Res.* **42**, DOI: 10.1186/s13046-023-02855-2 (2023).

55. Grillari, J. & Grillari-Voglauer, R. Novel modulators of senescence, aging, and longevity: Small non-coding rnas enter the stage. *Exp. Gerontol.* **45**, 302–311, DOI: 10.1016/j.exger.2010.01.007 (2010).

56. Bobbili, M. R., Mader, R. M., Grillari, J. & Dellago, H. Oncomir-17-5p: alarm signal in cancer? *Oncotarget* **8**, 71206–71222, DOI: 10.18632/oncotarget.19331 (2017).

57. Persengiev, S., Kondova, I., Otting, N., Koeppen, A. H. & Bontrop, R. E. Genome-wide analysis of mirna expression reveals a potential role for mir-144 in brain aging and spinocerebellar ataxia pathogenesis. *Neurobiol. Aging* **32**, 2316.e17–2316.e27, DOI: 10.1016/J.NEUROBIOLAGING.2010.03.014 (2011).

58. Du, W. W. *et al.* Mir-17 extends mouse lifespan by inhibiting senescence signaling mediated by mkp7. *Cell Death Dis.* **5**, DOI: 10.1038/cddis.2014.305 (2014).

59. Sayed, D. & Abdellatif, M. Micrornas in development and disease. *Physiol. reviews* **91**, 827–887, DOI: 10.1152/physrev.00006.2010 (2011).

60. Pellegrino, R. *et al.* Temporal trends, sex differences, and age-related disease influence in neutrophil, lymphocyte count and neutrophil to lymphocyte-ratio: results from inchianti follow-up study. *Immun. Ageing* **20**, DOI: 10.1186/s12979-023-00370-8 (2023).

61. Hira, M. T. *et al.* Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci. Reports* **11**, DOI: 10.1038/s41598-021-85285-4 (2021).

62. DOI: http://broadinstitute.github.io/picard/ (2018). Picard Toolkit.

63. Koboldt, D. C. *et al.* Varscan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285, DOI: 10.1093/bioinformatics/btp373 (2009).

64. Commo, F. *et al.* Rcgh: A comprehensive array-based genomic profile platform for precision medicine. *Bioinformatics* **32**, 1402–1404, DOI: 10.1093/bioinformatics/btv718 (2016).

65. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* **5**, 557–572, DOI: 10.1093/biostatistics/kxh008 (2004).

66. Stekhoven, D. J. & Bühlmann, P. Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118, DOI: 10.1093/bioinformatics/btr597 (2012).

67. Wang, Y. *et al.* Dna methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC Genomics* **22**, DOI: 10.1186/s12864-021-07675-2 (2021).

68. Houseman, E. A. *et al.* Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinforma.* **13**, DOI: 10.1186/1471-2105-13-86 (2012).

69. Loyfer, N. *et al.* A dna methylation atlas of normal human cell types. *Nature* **613**, 355–364, DOI: 10.1038/s41586-022-05580-6 (2023).

70. Wiśniewski, J. R. Filter-aided sample preparation for proteome analysis. *Methods molecular biology (Clifton, N.J.)* **1841**, 3–10, DOI: 10.1007/978-1-4939-8695-8_1 (2018).

71. Stanková, J. Sample preparation for proteomic plasma analysis. https://motbx.eatris.eu/download/695/?tmstv=1730200709 (2023). Accessed 29-10-2024.

72. Stanková, J. Lc-ms (hcd ot) plasma analysis. https://motbx.eatris.eu/download/691/?tmstv=1730200709 (2023). Accessed 29-10-2024.

73. Vreken, P. *et al.* Quantitative plasma acylcarnitine analysis using electrospray tandem mass spectrometry for the diagnosis of organic acidaemias and fatty acid oxidation defects. *J. Inher. Metab. Dis* **22**, 302–306, DOI: 10.1023/a:1005587617745 (1999).

74. Waterval, W. A., Scheijen, J. L., Ortmans-Ploemen, M. M., van der Poel, C. D. H. & Bierau, J. Quantitative uplc-ms/ms analysis of underivatised amino acids in body fluids is a reliable tool for the diagnosis and follow-up of patients with inborn errors of metabolism. *Clin. Chimica Acta* **407**, 36–42, DOI: 10.1016/J.CCA.2009.06.023 (2009).

75. Jankevics, A., Jenkins, A., Dunn, W. B. & Najdekr, L. An improved strategy for analysis of lipid molecules utilising a reversed phase c30 uhplc column and scheduled ms/ms acquisition. *Talanta* **229**, 122262, DOI: 10.1016/J.TALANTA.2021.122262 (2021).

76. Di Tommaso, P., Chatzou, M. & Floden, E. e. a. Nextflow enables reproducible computational workflows. *Nat Biotechnol* DOI: https://doi.org/10.1038/nbt.3820 (2017).

77. Hong, N. P. C. C. *et al.* Fair principles for research software (fair4rs principles). *Zenodo* DOI: 10.15497/RDA00068 (2022).

Acknowledgements

This work was supported by European Union's Horizon 2020 under grant number 871096 (EATRIS-Plus) and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) under grant number 184.034.019 (X-omics).

Author contributions statement

Conceptualization: A.A., C.V., A.S., J.S., J.N., L.N., P.A.H., A.G., G.S., M.H., E.O., T.A.; **Data curation:** A.A., C.V., J.S., J.N., S.E., L.N., P.A.H., B.G., V.F.L., M.H., M.B., J.V., Z.R., P.D., J.Sm., P.P., B.S.; **Formal analysis:** A.A., C.V., J.S., J.N., S.E., L.N., B.G., B.Y., V.F.L., J.G., H.W., P.K., U.E., P.V.N., J.V., R.T.; **Funding acquisition:** P.A.H., A.G., M.H.; **Investigation:** A.A., C.V., J.S., J.N., S.E., E.C.M., L.N., V.F.L., P.M., M.P., J.G., H.W., P.K., U.E., M.B., Z.R., P.D., J.Sm., P.P., B.S.; **Methodology:** A.A., C.V., P.A.H., M.P., A.G., M.H., M.B., P.V.N., A.M., R.T., Z.R., P.D., J.Sm.; **Project administration:** A.A., A.S., M.L.G.B., J.N., E.C.M., L.N., P.A.H., P.M., M.P., A.G., G.S., M.H., E.O., T.A.; **Resources:** M.P., A.G., J.G., H.W., U.E., M.B., Z.R., P.D., J.Sm., P.P., B.S.; **Software:** A.A., C.V., L.N., V.F.L., P.V.N., J.V., R.T.; **Supervision:** A.A., P.A.H., P.M., M.P., A.G., G.S., M.H., E.O., T.A.; **Validation:** P.M., M.B., Z.R., P.D., J.Sm.; **Visualization:** A.A., C.V., E.C.M., P.A.H., V.F.L., M.H., P.V.N., A.M., R.T.; **Writing – original draft:** A.A., C.V., J.S., M.L.G.B., J.N., L.N., P.A.H., V.F.L., M.P., A.G., M.H., P.V.N., J.V., A.M., R.T.; **Writing – review & editing:** A.A., C.V., J.S., E.C.M., L.N., P.A.H., V.F.L., M.H., P.V.N., J.V., A.M., R.T.