PLOS GENETICS

RESEARCH ARTICLE

Canonical correlation analysis for multi-omics: Application to cross-cohort analysis

Min-Zhi Jiang[1], François Aguet[2], Kristin Ardlie[3], Jiawen Chen[4], Elaine Cornell[5], Dan Cruz[6], Peter Durda[7], Stacey B. Gabriel[3], Robert E. Gerszten[8], Xiuqing Guo[9], Craig W. Johnson[10], Silva Kasela[11], Leslie A. Lange[12], Tuuli Lappalainen[11], Yongmei Liu[13], Alex P. Reiner[14], Josh Smith[15], Tamar Sofer[16], Kent D. Taylor[9], Russell P. Tracy[7], David J. VanDenBerg[17], James G. Wilson[18], Stephen S. Rich[19], Jerome I. Rotter[20], Michael I. Love[4,21☯*], Laura M. Raffield[21☯*], Yun Li[4,21☯*], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium[¶], TOPMed Analysis Working Group[¶]

1 Department of Applied Physical Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **2** Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, California, United States of America, **3** The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **4** Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **5** Laboratory for Clinical Biochemistry Research, University of Vermont, Burlington, Vermont, United States of America, **6** Department of Medicine, Cardiology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, **7** Department of Pathology & Laboratory Medicine, University of Vermont, Colchester, Vermont, United States of America, **8** Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, **9** Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, University of California at Los Angeles, Torrance, California, United States of America, **10** Department of Biostatistics, University of Washington at Seattle, Seattle, Washington, United States of America, **11** New York Genome Center, New York, New York, United States of America, **12** Department of Epidemiology, Department of Medicine, Division of Biomedical Informatics and Personalized Medicine, Lifecourse Epidemiology of Adiposity & Diabetes Center, Aurora, Colorado, United States of America, **13** Department of Medicine, Cardiology and Neurology, Duke University Medical Center, Durham, North Carolina, United States of America, **14** Department of Epidemiology, University of Washington, Seattle, Washington, United States of America, **15** Northwest Genomic Center, University of Washington, Seattle, Washington, United States of America, **16** Department of Biostatistics, Harvard Medical School, Medicine-Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **17** Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America, **18** Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, **19** Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia, United States of America, **20** Department of Pediatrics, Genomic Outcomes, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, University of California at Los Angeles, Torrance, California, United States of America, **21** Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

☯ These authors contributed equally to this work.
¶ Members are listed in S1 Acknowledgements.
* milove@email.unc.edu (MIL); laura_raffield@unc.edu (LMR); yunli@med.unc.edu (YL)

OPEN ACCESS

Citation: Jiang M-Z, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, et al. (2023) Canonical correlation analysis for multi-omics: Application to cross-cohort analysis. PLoS Genet 19(5): e1010517. https://doi.org/10.1371/journal.pgen.1010517

Editor: Kim-Anh Le Cao, The University of Melbourne, AUSTRALIA

Received: November 9, 2022

Accepted: May 1, 2023

Published: May 22, 2023

Data Availability Statement: JHS data are available at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000964.v5.p1 and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000286.v6.p2. MESA data are available at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001416.v3.p1 and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000209.v13.p3. Our modified SMCCA-GS and SSMCCA functions are available at https://github.com/zjgbz/SMCCA-GS_SSMCCA.

Abstract

Integrative approaches that simultaneously model multi-omics data have gained increasing popularity because they provide holistic system biology views of multiple or all components in a biological system of interest. Canonical correlation analysis (CCA) is a correlation-based integrative method designed to extract latent features shared between multiple assays by finding the linear combinations of features–referred to as canonical variables (CVs)–within each assay that achieve maximal across-assay correlation. Although widely

Funding: APR is funded by National Institutes of Health (NIH) grant R01HL146500 (from National Heart, Lung, and Blood Institute). LMR was supported by NIH grants R01AG075884 (from National Institute on Aging), T32HL129982 (from National Heart, Lung, and Blood Institute) and KL2TR002490 (from National Center for Advancing Translational Sciences). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: LMR is a consultant for the TOPMed Administrative Coordinating Center (through Westat).

acknowledged as a powerful approach for multi-omics data, CCA has not been systematically applied to multi-omics data in large cohort studies, which has only recently become available. Here, we adapted sparse multiple CCA (SMCCA), a widely-used derivative of CCA, to proteomics and methylomics data from the Multi-Ethnic Study of Atherosclerosis (MESA) and Jackson Heart Study (JHS). To tackle challenges encountered when applying SMCCA to MESA and JHS, our adaptations include the incorporation of the Gram-Schmidt (GS) algorithm with SMCCA to improve orthogonality among CVs, and the development of Sparse Supervised Multiple CCA (SSMCCA) to allow supervised integration analysis for more than two assays. Effective application of SMCCA to the two real datasets reveals important findings. Applying our SMCCA-GS to MESA and JHS, we identified strong associations between blood cell counts and protein abundance, suggesting that adjustment of blood cell composition should be considered in protein-based association studies. Importantly, CVs obtained from two independent cohorts also demonstrate transferability across the cohorts. For example, proteomic CVs learned from JHS, when transferred to MESA, explain similar amounts of blood cell count phenotypic variance in MESA, explaining 39.0% ~ 50.0% variation in JHS and 38.9% ~ 49.1% in MESA. Similar transferability was observed for other omics-CV-trait pairs. This suggests that biologically meaningful and cohort-agnostic variation is captured by CVs. We anticipate that applying our SMCCA-GS and SSMCCA on various cohorts would help identify cohort-agnostic biologically meaningful relationships between multi-omics data and phenotypic traits.

Author summary

Comprehensive understanding of human complex traits may benefit from incorporation of molecular features from multiple biological layers such as genome, epigenome, transcriptome, proteome, and metabolome. CCA is a correlation-based method for multi-omics data which reduces the dimension of each omic assay to several orthogonal components–commonly referred to as canonical variables (CVs). The widely-used SMCCA method allows effective dimension reduction and integration of multi-omics data, but suffers from potentially highly correlated CVs when applied to high-dimensional omics data. Here, we improve the statistical independence among the CVs by adopting a variation of the GS algorithm. We applied our SMCCA-GS method to proteomic and methylomic data from two cohort studies, MESA and JHS. Our results reveal a pronounced effect of blood cell counts on protein abundance, suggesting blood cell composition adjustment in protein-based association studies may be necessary. Finally, we present SSMCCA which allows supervised CCA analysis for the association between one phenotype of interest and more than two assays. We anticipate that SMCCA-GS would help reveal meaningful system-level factors from biological processes involving features from multiple assays; and SSMCCA would further empower interrogation of these factors for phenotypic traits related to health and diseases.

Introduction

In recent years, there has been rapid growth in high-dimensional multi-omics datasets (including DNA methylation, RNA-sequencing, metabolomics, proteomics, genomics, microbiome,

etc.). However, careful analyses with integrative methods are needed to fully utilize these rich datasets and provide mechanistic insights into health and disease related outcomes. While many methods have been published [1–3], few studies have evaluated these methods on large-scale datasets from human samples. In addition, despite quite a few successful examples of integrating two omics data-types [4–8], particularly detection of quantitative trait loci using genomic data, there are much fewer such examples of integrative analyses across more than two omics data types.

One promising method for using multi-omics data to explain phenotypic variation in health outcomes is canonical correlation analysis (CCA) [9]. CCA is a statistical technique to identify associations among two assays where each assay contains multiple variables. Specifically, CCA finds a linear combination of variables in each assay that leads to the maximal correlation of the two linear combinations. Principal component analysis (PCA) can be considered as a special case of CCA as the optimization objective is the same in the case that the same data is used for the two assays. CCA is a commonly adopted dimension reduction and information extraction method in genomic studies [1,10–13] as increasingly more modern genomic studies collect data from multiple assays.

An extension of CCA by Witten & Tibshirani [1] called sparse multiple CCA (SMCCA) allows for the input of multiple assays. We hypothesized that this method would be helpful for high-dimensional multi-omics data exploration and for understanding and extracting omics signatures that reflect biologically relevant variations. Specifically, we here leverage our CCA-based method extended from Witten & Tibshirani's SMCCA to extract low-dimensional latent variables from high-dimensional multi-omics data and use them to explain phenotypic traits, focusing on blood cell indices, along with basic demographic and anthropometric characteristics. We perform CCA-based analyses in two studies with rich multi-omics data in hundreds of individuals, the Multi-Ethnic Study of Atherosclerosis (MESA) and the Jackson Heart Study (JHS).

Results

CCA pipeline

A typical CCA-based method generates orthogonal canonical variables (CVs), which are low-dimensional summaries to represent latent variables underlying the multi-assay input data. **Fig 1** is a cartoon illustration where we have three assays (X, Y, and Z) for three samples. Features are assumed to be continuous with no distributional assumptions. For presentation brevity, we only show how we obtain the top 4 CVs. For each assay, CCA infers 4 vectors of weights (e.g., W_{X1}, W_{X2}, W_{X3}, and W_{X4} for assay X), which leads to four CVs. For example, CV_{X1}, the top CV for assay X, is obtained by $X \times W_{X1}$. The weights are inferred by maximizing the correlation of CVs across three assays. Note that in the rightmost CV matrices, each column of a CV matrix is one CV of the corresponding assay. In addition, CVs corresponding to the same column cross assays are expected to have maximal correlation (for instance, CV_{X1}, CV_{Y1}, CV_{Z1},are most correlated), while CVs in different columns are expected to be orthogonal or independent from each other in the same assay.

Modified gram-schmidt algorithm improves orthogonality

SMCCA implemented in the PMA R package does not always provide the expected orthogonal CVs, preventing effective extraction of independent CVs and sometimes causing serious multicollinearity issues in subsequent association analysis. For example, **Fig 2A** and **2B** shows results from PMA's implementation of unsupervised SMCCA when applied to MESA proteomics and methylomics data (detailed in **Methods**) where we observe extensive correlation among the CVs. In the presence of undesired correlated CVs, users will have to perform a

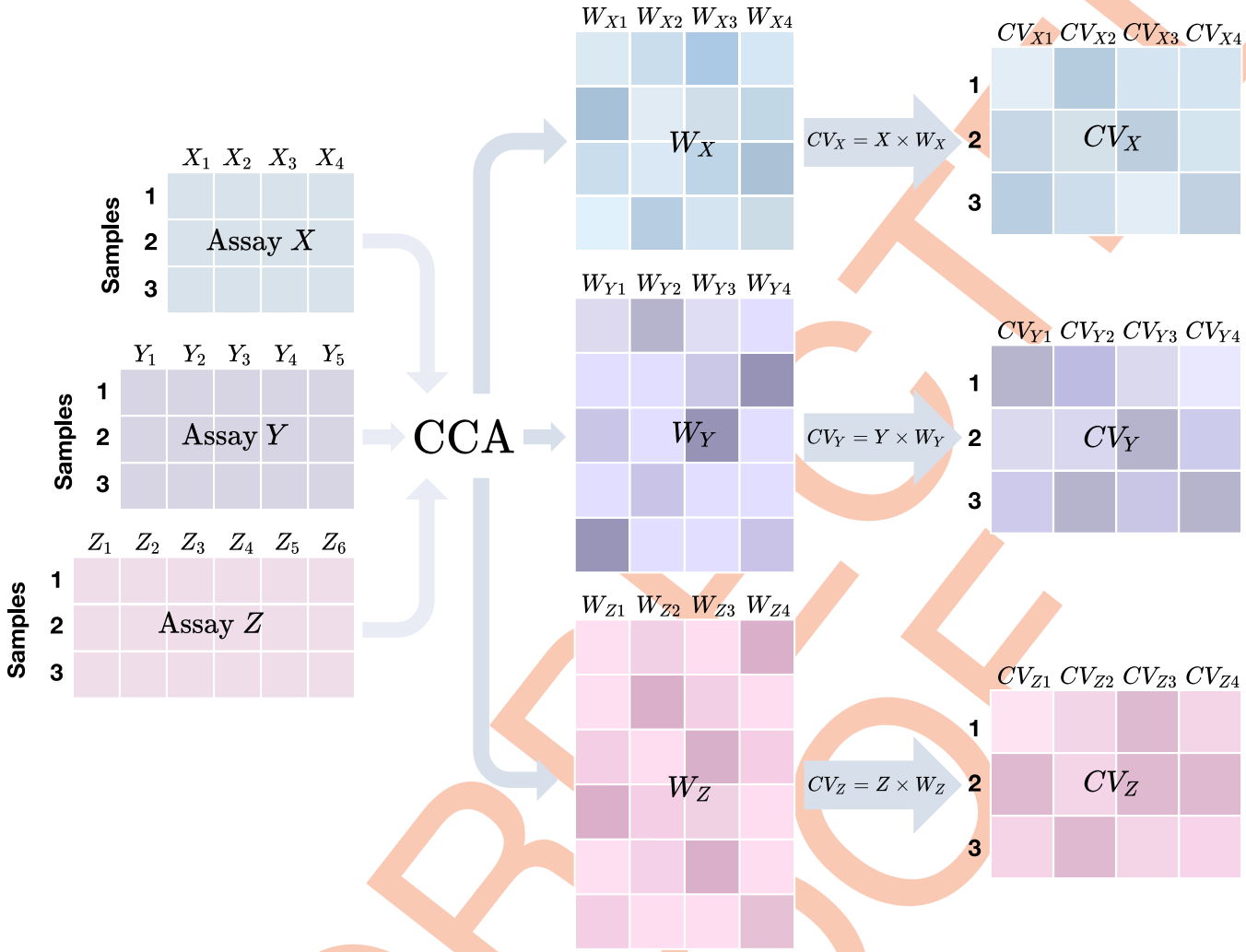

Fig 1. Cartoon illustration of a typical CCA-based method for three assays. X, Y, and Z are three assays with 4, 5, and 6 features respectively. When applying a CCA-based method on them to compute 4 canonical variables (CVs), we would first get their **weight matrices** W_X, W_Y, W_Z, each of which contains 4 weight vectors. By multiplying each assay matrix (left panel) and its corresponding weight matrix (middle panel), we obtain the CV matrix for the assay (right panel) where each column corresponds to one CV.

https://doi.org/10.1371/journal.pgen.1010517.g001

secondary filtering step to generate a list of non-redundant CVs, or else variation in omics data captured by the later CVs may overlap with variance captured by former CVs. Therefore, we sought to improve orthogonality among generated CVs for capturing distinct information from the integrated multi-omics data. Specifically, we follow the Gram–Schmidt (GS) strategy [14] which generates CVs sequentially by progressively subtracting the previous CV from the input matrices (detailed in **Methods**). Fig 2C and 2D shows substantially improved orthogonality among the CVs when applied to the same MESA proteomics and methylomics data. Similar patterns were observed when SMCCA was applied to JHS data (S1 Fig).

Proteomics CVs explain considerable amounts of variation in blood cell counts

We also applied our implementation to proteomics and methylomics data in JHS. As these unsupervised CVs are anticipated to capture shared latent variables underlying the proteomics

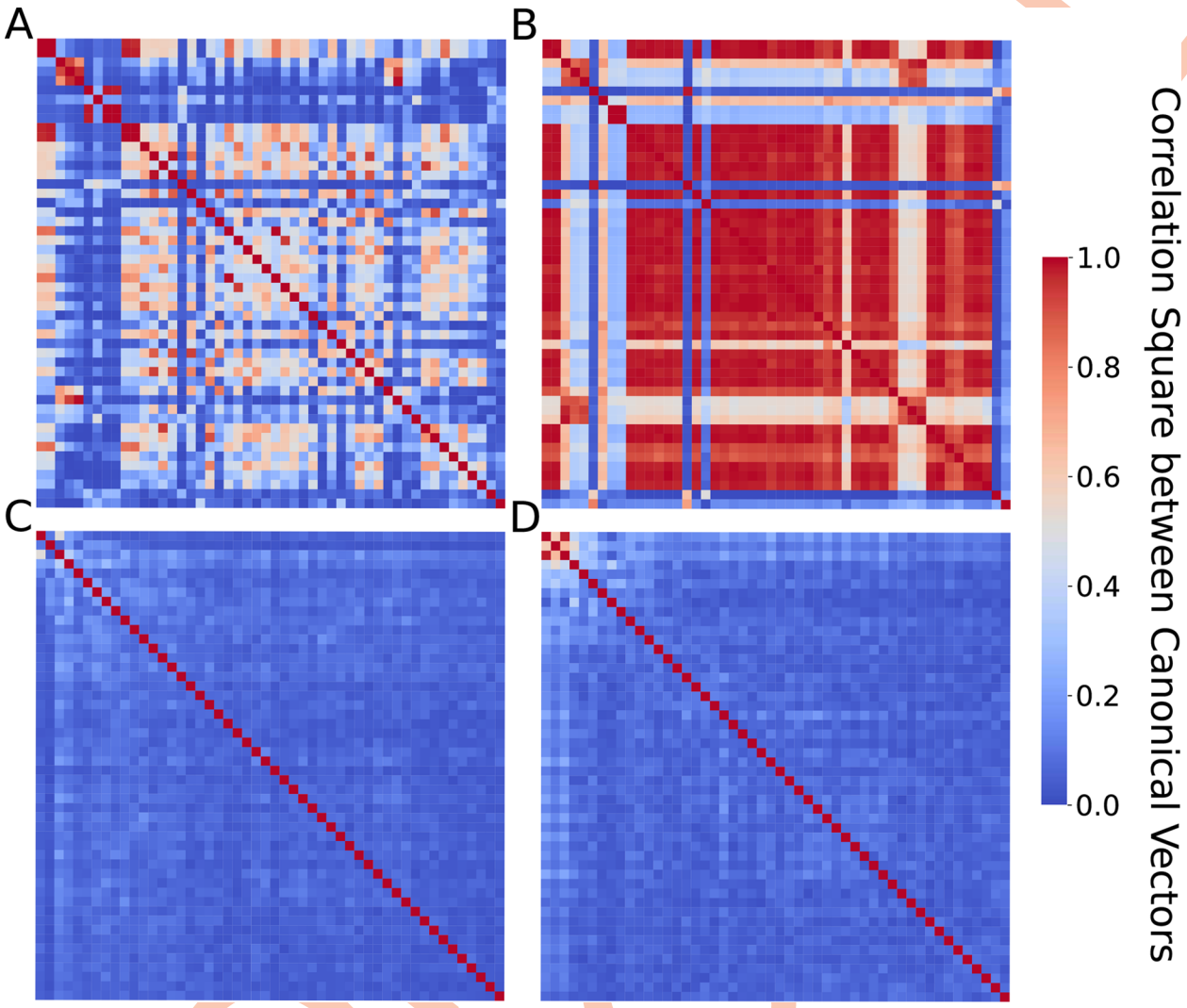

Fig 2. Improved orthogonality among CVs by adopting the Gram–Schmidt (GS) strategy. CVs are inferred from MESA proteomics and methylomics data using unsupervised SMCCA. Each row and column represent one CV, ranging from CV1 to CV50. **(A-B)** Results from the PMA R package, implementation of the original SMCCA methods without the incorporation of GS algorithm. **(C-D)** Results from our SMCCA-GS, with the GS strategy incorporated. Left panel (**A** and **C**) show proteomics CVs, and right panel (**B** and **D**) methylomics CVs.

https://doi.org/10.1371/journal.pgen.1010517.g002

and methylation datasets, we hypothesized that the CVs may explain a non-negligible amount of variation in various phenotypes. Our primary phenotypes of interest in this work are blood cell traits, including white blood cell count (WBC), red blood cell count (RBC) and platelet count (PLT). We also considered age, sex, and body mass index (BMI), as "control" phenotypes which have been widely reported to explain considerable variability in proteomics and methylomics data. For each of the six outcome phenotypes, we fit regression models to estimate the percent of variation explained by the top 50 CVs from each of the two omics data, namely proteomics and methylomics (detailed in **Methods**). For each cohort (MESA or JHS), we had two sets of CVs, one derived from the cohort's own omics data, the other derived from applying the CV weights inferred from the other cohort.

We found that top CVs, from each of the two omics data, explain considerable amounts of variation in almost all of the outcomes evaluated (Fig 3). For example, top 50 methylomics CVs inferred in JHS explained 72%, 100%, 35%, 37%, 34%, 30% of variation in age, sex, BMI,

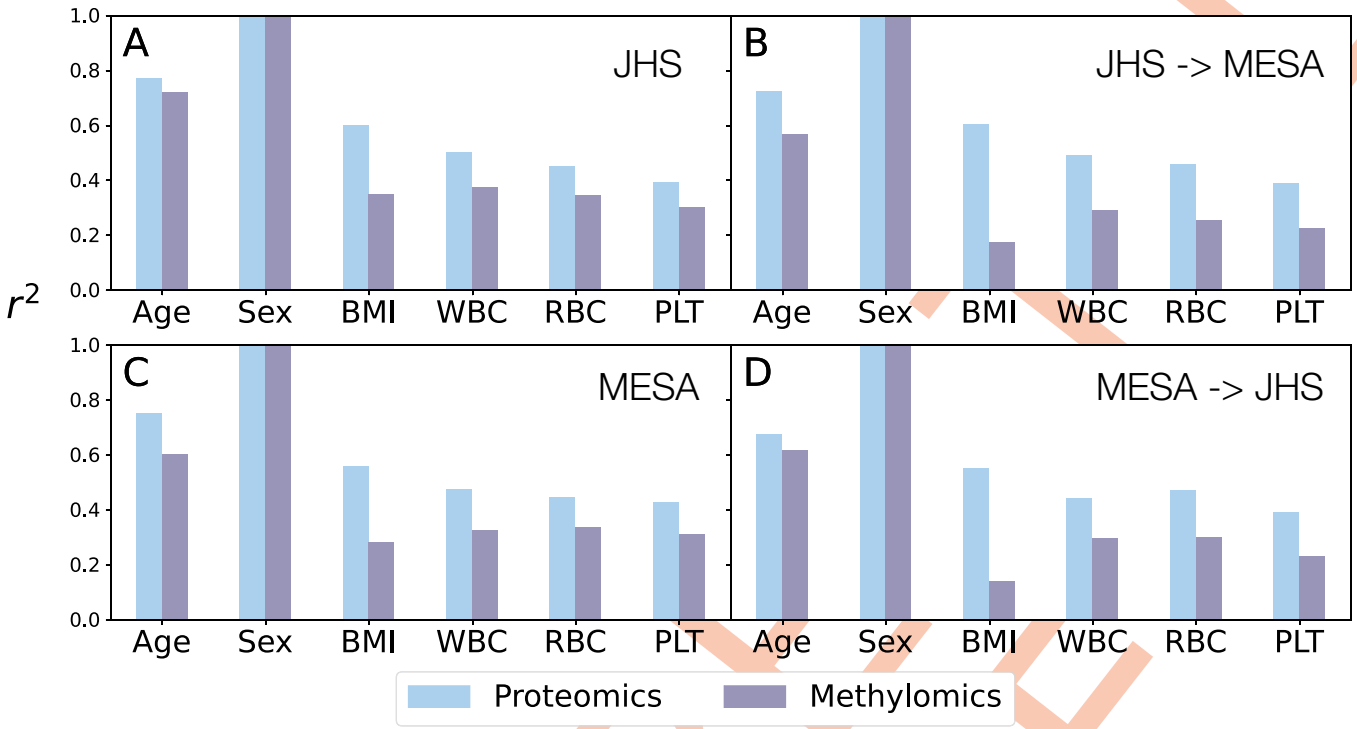

Fig 3. Proportion of variation in outcomes explained by CVs. (A) CVs were inferred using proteomics and methylomics in JHS. The top 50 CVs were used to calculate the r^2 (Y-axis) for each outcome (X-axis). **(B)** We obtained CVs in JHS by applying the weights inferred from MESA, and then calculated r^2 in the same way as in A. **(C)** CVs were inferred using proteomics and methylomics in MESA. **(D)** CVs were obtained in MESA by applying the weights inferred from JHS.

https://doi.org/10.1371/journal.pgen.1010517.g003

WBC, RBC, and PLT respectively, in JHS (Fig 3A). We also observe high transferability between MESA and JHS, by first applying SMCCA-GS separately to each cohort and then transferring the inferred CVs to the other cohort. For example, the top 50 methylomics CVs inferred in MESA explained similar amounts of variation in RBC: 33% in MESA (itself) (Fig 3C) and 30% when applied to JHS (Fig 3D). Such high transferability suggests that latent variables learned by CCA might reflect biological processes shared across cohorts. We also note that these r^2's from methylomics data were most likely under-estimated because the CVs were constructed using the top 10,000 most variable CpG sites (see Methods) instead of the entire ~700,000 sites, for computational reasons. These findings are not surprising: for instance, blood cell composition (notably for white blood cell subtypes) has been long known to influence the methylome. For that reason, in epigenome-wide association studies (EWAS), it has been standard practice to first estimate the leukocyte proportions from methylomics data and adjust for these cell type proportions in subsequent association analysis [15]. Given shared precursors for all hematological cell types, we found it relatively unsurprising that RBC and PLT also had a high percent variation explained by methylomics CVs. Similarly, age [16], sex [17,18] and BMI [19] have been known to explain substantial variability in methylomics data, and are commonly adjusted for as covariates.

More interestingly, the amounts of variation in various outcomes explained by top 50 *proteomics* CVs are even higher, ranging 39% - 100% in JHS and 39% - 100% in MESA. Large r^2 for age, sex, and BMI are expected since all have been reported to rather broadly affect protein profiles [20,21]. However, strikingly, r^2 for blood cell traits are also considerable, and comparable to BMI, 50%, 45%, 39% respectively for WBC, RBC and PLT in JHS using CVs inferred in

JHS. Confirming these results, when applying CV weights inferred from MESA to JHS, we obtained similar r^2's: 44%, 47%, 39% for WBC, RBC and PLT respectively. Similar patterns were also observed in MESA using both MESA and JHS derived weights. These considerable amounts of variations in blood cell counts explained by top proteomics CVs have important implications for association studies involving proteomics data: we should consider adjusting for blood cell proportions in these association studies, under the same rationale in EWAS (variability driven by blood cell subtype abundance is likely not of interest for many disease outcomes of interest whose association with proteomics data is being examined).

CVs vs Principal Components (PCs)

Although CVs are inferred jointly from multi-omics data, we have focused on analyzing CVs from each omics data type separately for their predictive power of outcomes of interest. Thus, we naturally are interested in comparing the CCA-based approach with the standard PCA approach since we can obtain PCs separately for each omics data. Note first that we expect larger and more assay-specific batch effects in JHS than MESA. For example, JHS proteomics data was generated in 3 batches [22], and separately from the methylomics data. In contrast, MESA proteomics and methylomics data were all generated through the MESA TOPMed pilot over a short time period [23,24]. Results shown in **Fig 4** supported our expectations: overall we observe that a lower number of JHS inferred CVs are needed to explain the outcomes with higher r^2 compared to JHS inferred PCs, indicating that top CVs inferred from JHS data tend to capture biological variations while top PCs tend to reflect more assay-specific technical variations. We note that this is supported by the stronger association for CVs vs PCs with technical factors (**S6** and **S7 Figs**), notably for proteomics data which has been subjected to less pre-processing to account for technical effects related to batch/plate (prior to any of the analyses conducted here). The contrasts are most pronounced with age and WBC for proteomics data, and with age for methylomics data. For example, in JHS, proteomics-CV1 explained 33% variation in WBC (blue "+" on the leftmost side of **Fig 4A3**) while proteomics-PC1 only explained 7.7% (purple "+" on the leftmost side of **Fig 4A3**). This noticeable advantage continued until ~20 CVs/PCs. For instance, the top 15 proteomics-CVs in JHS explained 44% variation in WBC (blue "x" in **Fig 4A3**) while top 15 proteomics-PCs only explained 29% (purple "x" in **Fig 4A3**). Similar advantages of CVs over PCs were observed in MESA, but were less pronounced as expected due to the smaller and less assay-specific batch effects in MESA. Reassuringly, applying JHS inferred CV weights to MESA showed advantages similar to those in JHS, more pronounced than using CVs inferred in MESA itself, further demonstrating the power of CVs to capture biologically relevant variations under the presence of assay-specific batch effects.

Supervised sparse multiple CCA

Extending supervised sparse CCA to supervised sparse Multiple CCA. So far, we have generated and evaluated unsupervised CCA where the CVs are inferred from multi-omics data only, without considering any outcomes of interest. Although we assessed the relationship between unsupervised CVs and several outcomes of interest, the CVs themselves were inferred without knowledge of the outcomes. In practice, when we are primarily interested in a particular outcome, supervised approaches can be more effective and powerful. The PMA R package implements a sparse supervised CCA (SSCCA) method. However, this implementation only accepts two omics data at a time, which limits our capabilities in real datasets where there are more than two assays. For instance, in both MESA and JHS, we also have whole genome sequencing (WGS) data [25]. We implemented a sparse supervised multiple CCA (SSMCCA) method to accommodate more than two assays of omics data. Our implementation follows the

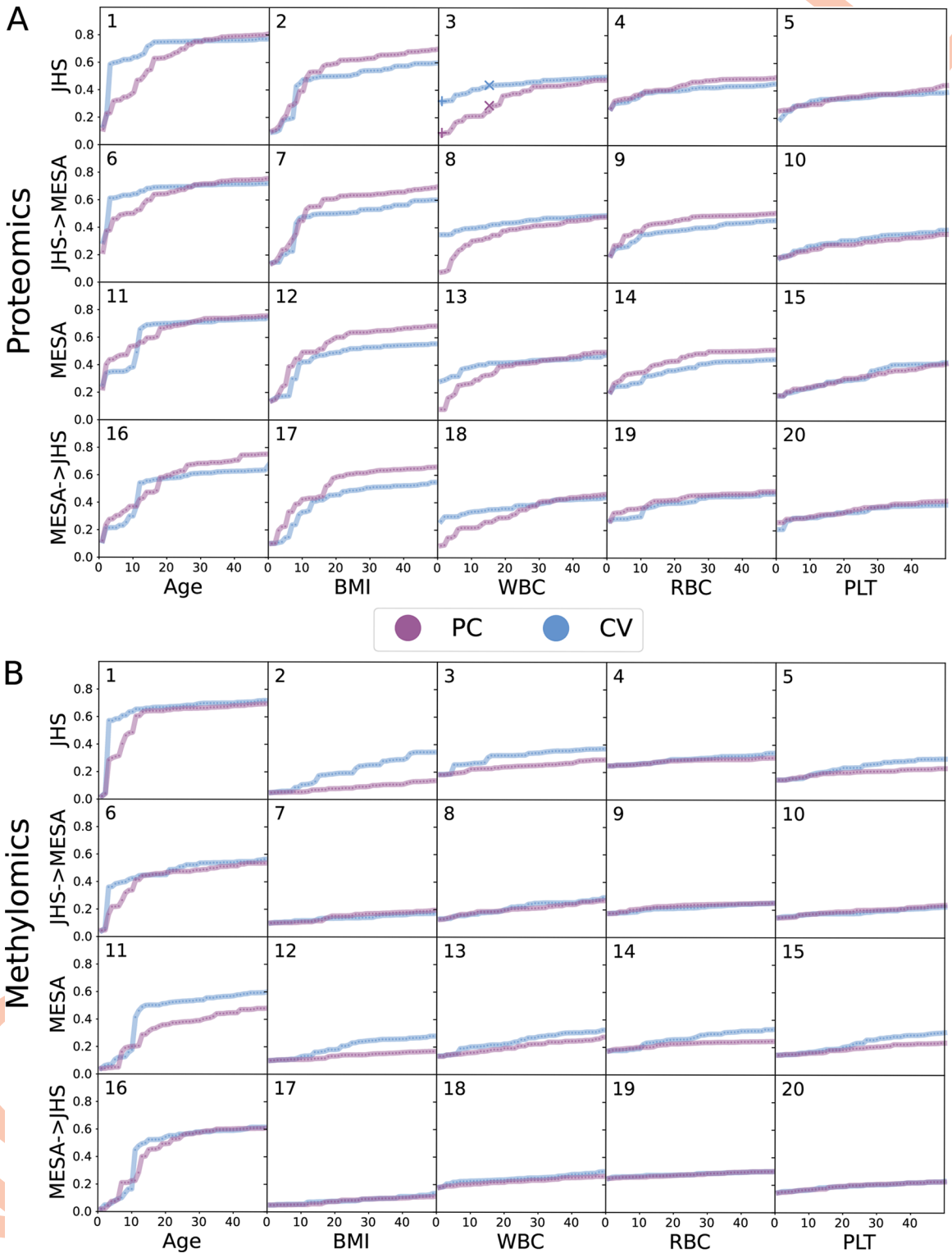

Fig 4. Comparison of r^2, PCs vs CVs. Each column corresponds to one outcome. Within each panel, top row (JHS) shows results in JHS using JHS-inferred CVs. Second row (JHS->MESA) shows results in MESA, also using JHS-inferred weights. Third row (MESA) shows results in MESA, this time using MESA-inferred CVs. Last row (MESA->JHS) shows results in JHS, also using MESA-inferred weights. (**A**) Proteomics. Proteomics CVs explain more variation in white blood cell count (WBC) than PCs. For example, proteomics-CV1 explains 33% of the variation in WBC (blue "+" in **Fig 4A3**), while proteomics-PC1 only explains 7.7% (purple "+" in **Fig 4A3**). This pattern persists until approximately 20 CVs/PCs. The top 15 proteomics-CVs in JHS explain 44% of the variation in WBC (blue "x" in

Fig 4A3), while the top 15 proteomics-PCs explain only 29% (purple "x" in **Fig 4A3**). (**B**) Methylomics. In each sub-figure, X-axis indicates the number of CVs or PCs used and Y-axis the proportion of variation explained in the outcome (i.e., r^2).

https://doi.org/10.1371/journal.pgen.1010517.g004

idea in Witten et al., (2009) [1] where a feature selection step is performed within each assay to retain (by default) top ~80% features most correlated with the outcome of interest. Features selected from each assay form new input matrices to which we then apply our implementation of unsupervised SMCCA with the adapted Gram-Schmidt algorithm.

To ensure our SSMCCA implementation generates sensible supervised CVs, we first compared results from PMA's SSCCA implementation, when there are two assays of data. Specifically, we compared correlations between inferred supervised CV1 and the corresponding outcomes of interest. We compared SSCCA and our SSMCCA by running two methods with 100 different random seeds and for each seed, testing the variation of each outcome explained by supervised proteomics CVs and supervised methylomics CVs (**Fig 5**). We found that in most cases, the amount of variation in outcomes captured by SSMCCA CVs is comparable or significantly higher than SSCCA, indicated by large red circles. For example, **Fig 5A** third row third column (red "x" on **Fig 5A**) shows a large red circle which annotates a case where our SSMCCA outperforms the original SSCCA. In this example, SSMCCA proteomics CV1 explains 4.17% variation in PLT in MESA, while SSCCA 3.48% (p-value = 8E-9 for difference). The larger the difference, the darker the color. In a few cases, the amount of variation captured by SSMCCA CV1 is significantly smaller than SSCCA CV1. For example, **Fig 5B** row 2 column 1 shows a large blue circle (light blue "+" on **Fig 5B**) which indicates a case where the original SSCCA outperforms our SSMCCA. However, although the difference in terms of percent variation explained in RBC by SSCCA vs SSMCCA methylomics CV1 is highly significant (p-value = 3E-28), the absolute difference (4.27E-8 percent variance explained) is tiny, suggesting the difference between the performance of two methods is negligible.

Biologically meaningful features detected by SSMCCA

We applied SSMCCA to three assays–proteomics, methylomics, and genotypes–from MESA to obtain 50 CVs for each assay, and then used standard regression models to assess associations with phenotypes–age, BMI, WBC, RBC, and PLT. CV-phenotype pairs were considered to be significantly associated when p-value $<$ 1E-4 (Bonferroni correction), adjusting for covariates detailed in **S2 Table**. In MESA, we identified 58 significant CV-phenotype pairs, and 5 of them were validated in JHS with the same p-value threshold of 8.62E-4 and same direction of effect (**S3 Table**). For example, WBC and proteomics CV3 were strongly associated in both cohorts (p-value = 2.7E-15 in MESA, 6.8E-16 in JHS, **S3 Table**). Features with high absolute weight coefficients in this CV (**S5 Table**) are biologically relevant for WBC. For example, stem cell factor soluble receptor, which has the highest weight, is known to play a key role in hematopoiesis [26]. Lipocalin 2, with the second highest weight, has been reported to be associated with human neutrophil granules [27].

For each phenotype, we then assembled all features from each assay (i.e., both methylomics and proteomics) with non-zero weight for phenotype-associated CVs in **S3** and **S4 Tables**, and annotated each feature to a gene, on which we performed pathway enrichment analysis (described in **Methods**). For comparison, we also performed the same pathway enrichment analysis using features individually associated with each phenotype, where association is declared when FDR $<$ 5% for each assay-phenotype-cohort combination. Comparing these two sets of pathway enrichment results, we found several pathways only revealed (p. adjust $<$ 0.05) by our SSMCCA, including the growth factor binding gene ontology (GO) [28,29] term and the DisGeNET [30] progressive chronic graft-versus-host disease (GVHD)

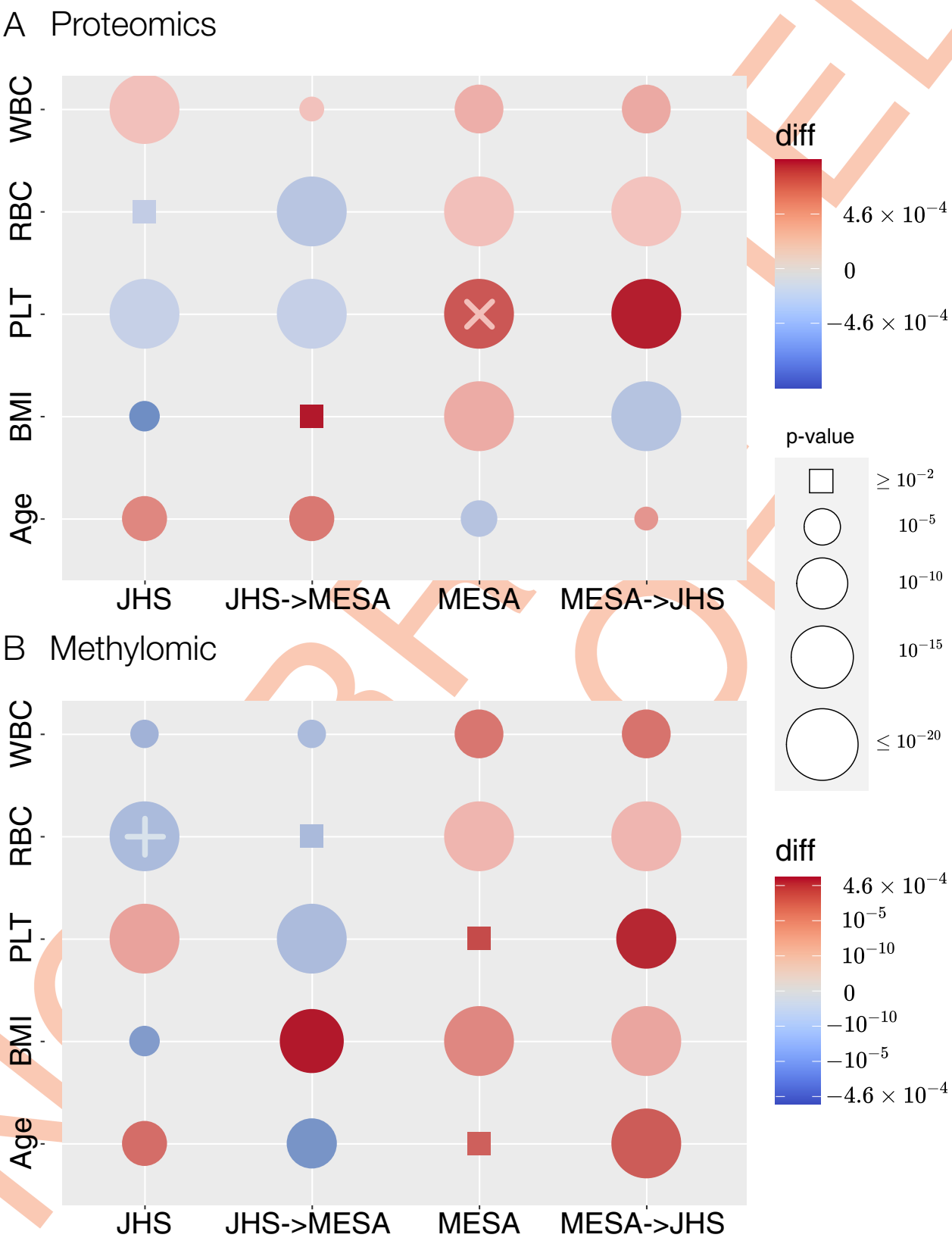

Fig 5. Comparison of SSCCA and SSMCCA. (**A**) proteomics, and (**B**) methylomics. Each row corresponds to a phenotype (from bottom to top, Age, BMI, WBC, RBC, and PLT). Circle size reflects the significance of the difference in variation explained between two methods. Color reflects the size of difference between the variation of phenotype explained by SSCCA and our SSMCCA. Therefore, a larger circle means a more significant difference between the two methods. Note that we use rectangles for insignificant difference with $p > 0.01$. Red means that our SSMCCA explains more phenotypic variation while blue means that SSCCA explains more. The darker the color, the larger the difference (the scale is different for parts A and B, annotated in "diff" column on side of figure).

https://doi.org/10.1371/journal.pgen.1010517.g005

and polypoidal choroidal vasculopathy genesets. All of these pathways have been reported to be related to BMI in previous literature [31–34].

Assigning CpG sites to genes is a challenging task. We adopted the simple nearest gene approach. Other reasonable approaches include promoter-centric assignment [35], leveraging differentially methylated regions [36], or using expression quantitative trait methylation (eQTM) [37] information. We explored the eQTM approach as we have both methylation and gene expression measurements in a subset of samples in JHS and MESA. However, due to limited number of CpGs included in significant CVs, we had only 25–257 genes (with the number of genes implicated by CpGs varying across different outcomes, detailed in **S6 Table**) based on significant eQTMs (**Methods**) for pathway enrichment analysis (results summarized in **S7** and **S8 Tables**). We anticipate to benefit more from this approach when eQTM sample size increases.

Discussion

Large quantities of data across multiple omics (transcriptomics, proteomics, metabolomics, genomics, methylomics, etc.) modalities is currently being generated, for example through efforts funded by NIH's Precision Medicine Initiative [25,35] as well as other large federally funded studies [36]. These high dimensional and complex multi-assay data are unfortunately still too often analyzed only separately (e.g., applying PCA separately to genotype, gene expression, or methylation data) or in a pairwise manner (for example mQTL analysis examining relationships between genome and methylome, or pQTL analysis examining relationships between genome and proteome). Many innovative methods have been proposed (https://github.com/mikelove/awesome-multi-omics [accessed on *2022-07-25*]) for integrative analysis but evaluations in large-scale real omics data are still lacking, with fewer impartial appraisals available to guide method choice in practice.

In the work presented here, we apply CCA-based methods to complex multi-omics datasets to assess their capabilities and limitations. In particular, for the widely used PMA implementation of the SMCCA methods, we identified two limitations: non-orthogonal CVs and inability to accommodate more than two assays for supervised analysis. We provide method extensions, SMCCA-GS and SSMCCA, to address the two limitations. Applying SMCCA-GS to real data in MESA and JHS, we found that CVs are consistent and transferable across cohorts, suggesting that CVs capture constitutive biological relationships shared across cohorts, and are not driven primarily by assay-specific technical variation. This cross-cohort consistency, to our knowledge, has not been well explored in the literature and has important implications for making method choices (e.g., CCA vs PCA) for multi-omics data with or without extensive assay-specific batch effects.

Importantly, our CCA-based analyses reveal that blood cell indices are substantially associated with multiple omics assays including methylomics and proteomics. The former association has been widely appreciated and exerted paradigm-shifting impact on analysis: in methylation association studies, white blood cell composition is adjusted for in methylation analyses in standard practice. The latter association, where CVs from proteomics data showed even more pronounced association with blood cell indices, has been under-appreciated, with

blood cell traits not considered in most current proteomic analyses [22,37–39]. Our findings indicate that blood cell composition should be accounted for (or at least considered) in protein association studies where feasible, similar to what is standard practice for methylation studies.

As demonstrated in **Fig 4**, our SMCCA-GS is in some cases more useful than PCA in explaining variability in phenotypes, using an identical number of PCs/CVs. However, there are also many cases where the methods are nearly equivalent. We hypothesize that our SMCCA-GS demonstrates more consistent advantages in explaining trait variability in JHS versus MESA due to the presence of more substantial JHS batch effects. Due to funding limitations, JHS proteomics and metabolomics data was generated in multiple batches across several years, while the MESA data used here was generated concurrently, funded by NHLBI's TOPMed program. Thus, for proteomics in particular, more batch effects are anticipated in JHS; our SMCCA-GS is particularly advantageous in cases where there is increased assay-specific technical variation.

In multi-omics data, it is commonplace to have drastic differences in the dimension of different omics data. For example, methylomics data, when generated by the widely used Illumina MethylationEPIC BeadChip array, contains almost 10^6 features; transcriptomics data are commonly summarized into ~10^4 expressed genes; and metabolomics and proteomics typically even smaller: only ~10^2–10^3 features depending on the platforms used. Witten et al. (2009) [1], introducing the SMCCA method, analyzed data with 19,672 gene expression measurements and 2,149 comparative genomic hybridization measurements, showing that their method could accommodate such imbalance. Our methods, derived from SMCCA, are also expected to accommodate omics dimension imbalance. In our analyses, results using ~700K CpG sites, while computationally challenging to fit repeatedly, led to similar conclusions as using top 10,000 CpG sites (detailed in **Methods** and **S4** and **S5 Figs**), further suggesting the robustness of sparse CCA methods to imbalance in omics dimension.

We note that CCA-based methods as implemented in our analyses still have several key limitations. Notably, we had to considerably reduce the dimensionality of methylation array and sequencing data in order for our CCA-based method to be computationally feasible (at least for the repeated analyses necessary for methods development and testing). While we were able to fit models for the entire set of CpG sites a single time, with similar overall results in terms of phenotype variance explained (**S4** and **S5 Figs**), our SMCCA-GS approach will require further innovation to be scalable for large-scale datasets. Recently developed methods allow for efficient calculation of generalized CCA solutions across reduced dimensions of each distinct assay, which alleviates some of the computational issues that arise, though sparse identification of individual omics features from the original assay data may still be desired [40].

Methods

Cohorts

Ethics statement. All participants included in this analysis provided written, informed consent for use of genetic and multi-omics data, and all study protocols conform to the 1975 Declaration of Helsinki guidelines. The Jackson Heart Study (JHS) and Multi-Ethnic Study of Atherosclerosis (MESA) studies were approved by the Institutional Review Boards of all participating institutions.

JHS. JHS recruited 5,306 African American participants from the Jackson, Mississippi, metropolitan tri-county area (Hinds, Madison, and Rankin) into a prospective, community-based cohort designed to investigate risk factors for cardiovascular disease among African Americans [41–43]. Demographics of JHS individuals involved in the analysis are displayed in **S1A Table**.

Multi-omics data utilized in JHS analyses including methylomics (n = 1,750, Illumina MethylationEPIC BeadChip array) [44] and proteomics (n = 2,144, SOMAscan 1.3k array) [22], both from the baseline visit, and whole genome sequencing (WGS) data as described below. Methylation levels are quantified by beta values [45]. Traits examined include age, sex, BMI, and hematological traits (WBC, RBC, and PLT). We limited our analyses in JHS to individuals with complete data for proteomics, methylomics, and traits examined (total n = 881, **S2A Fig**).

MESA. The MESA study was initiated in July 2000 to investigate the prevalence, correlates, and progression of subclinical cardiovascular disease (CVD) in a population-based sample of 6,814 men and women aged 45–84 years. The cohort was selected from six US field centers. Based on self-reported race/ethnicity, approximately 38% of the cohort are White, 28% African American, 23% Hispanic, and 11% Chinese American. More demographic information of MESA individuals involved in the analysis is in **S1B Table**.

Longitudinal multi-omics data was generated in MESA through a pilot program from NHLBI's Trans-Omics for Precision Medicine Initiative (TOPMed) at exam 1 (2000–2002) and exam 5 (2010–2011), including ~ 1,000 participants for each exam with methylomics data (Illumina MethylationEPIC BeadChip array) [45] and proteomics (SOMAscan 1.3k array) [22,23]. Methylation levels are quantified by beta values [45]. WGS data are described below. Basic covariates examined include age, sex, BMI, recruitment site, self-reported race/ethnicity, and the same hematological traits as in JHS. We limited our analyses in MESA to individuals with complete data for proteomics, methylomics, and phenotypes examined (total n = 777, **S2B Fig**). Use of the same platforms for multi-omics assessment as in JHS allowed comparison analyses for CVs derived by SMCCA-GS or SSMCCA across cohorts.

Whole Genome Sequencing (WGS) data

Genotypes are derived from TOPMed WGS data (freeze 8). Data harmonization, variant discovery, and genotype calling were previously described [25,46]. In our analysis, to reduce data dimensionality, we first extracted SNPs associated with blood cell traits from Chen et al. (2020) [47] and highly correlated (linkage disequilibrium $r^2 > 0.8$ where r^2 is the in-sample squared Pearson correlation between the corresponding genotype vectors) variants were removed, resulting in 3,789 SNPs for JHS and 3,562 SNPs for MESA in our supervised CCA analysis. Genotypes are coded into numerical values 0, 1, and 2 for our analysis. Population principal components calculated by PC-AiR [48] were adjusted for as covariates. In addition, for WBC, we additionally adjusted for the Duffy null polymorphism (SNP rs2814778 at chromosome 1q23.2) [49].

Transcriptomics

We involve transcriptomics data in eQTM analysis to map our selected 10k CpG sites to genes for pathway enrichment analysis, but we do not include transcriptomics in our multi-omics integration analysis because a considerable number of individuals could not be included in the analysis if we incorporate transcriptomics (**S2 Fig**). For both JHS and MESA, RNA-seq was measured from peripheral blood mononuclear cells and normalized to transcript per million (by Northwest Genomics Center for MESA, as previously described [50], NWGC for JHS using similar pipelines).

Initial quality control (QC) and transformation of multi-omics data

In both cohorts, we applied QC on each assay including sample outlier removal and feature filtering. For each protein in the proteomics data, we first applied log transformation, followed

by inverse normal transformation. After QC, we had 1,317 proteins measured in both cohorts, which made validation across cohorts straightforward.

Methylomics of JHS [44] was normalized using the "noob" normalization method implemented in minfi R package [58,59]. We further removed batch, plate, row, and column effects using the ChAMP R package [51]. For MESA methylation data, which had already been subjected to functional normalization to reduce batch effects [52], we excluded samples with (1) call rate $< 95\%$; (2) sex mismatches; and (3) concordance between SNP probes and genotypes < 0.8. Methylation levels were marked as missing when the detection p-value was > 0.01, and we imputed these missing values using ChAMP R package [51], as our CCA-like methods cannot accommodate missing data. For both JHS and MESA, CpG sites whose probes overlap any SNP with minor allele frequency (MAF) $> 1\%$ were also excluded [53]. After QC, we had 754,767 and 741,727 CpG sites for MESA and JHS respectively. For building validation across cohorts, we only kept the 721,334 CpG sites which passed QC in both cohorts.

Finally, we only kept samples with complete data including proteomics, methylomics, and phenotypes (age, BMI, WBC, RBC, PLT, site, race, sex for MESA; age, BMI, WBC, RBC, PLT, sex for JHS), which led to 881 samples for JHS and 777 samples for MESA. We further identified sample outliers by PCA-IQR plot (**Section 2** in S1 Text and S3 Fig). Four outliers in JHS–one sample with the largest proteomics IQR (wedge pointed on S3A Fig) and three samples with largest methylomics IQR (wedges pointed on S3B Fig)–were removed; and three outliers in MESA were removed–all three with largest methylomics IQR (wedges pointed on S3D Fig).

For each assay, we removed the sex chromosome related proteins and CpG sites. We further removed features that are highly correlated [54], at a squared Pearson correlation 0.8 threshold. We adopted a greedy algorithm (**Algorithm 1** below) to achieve the dual goal of no highly correlated pairs among a maximal number of features retained. For methylomics, we calculated Pearson correlation using the Python package Deep Graph [54] and after removing highly correlated, further kept 10k CpG sites with the highest variance for the computational efficiency. Our CCA-based methods are computationally intensive. For example, even with these 10k CpG sites (~1.3% of all available CpG sites), on a single core of E5-2680 v3 @ 2.50GHz, the wall time of calculating 50 CVs with our SMCCA-GS on proteomics and methylomics is about 8 ~ 14 hours for MESA (774 samples) and about 8 ~ 20 hours for JHS (877 samples); with 20k CpG, the wall time is about 14 ~ 36 hours for MESA and 16 ~ 47 hours for JHS. For validating our variance-based feature selection strategy, we also performed the same analysis as **Figs** 3 and 4 on proteomics and all ~700k CpG sites. The results (S4 and S5 **Figs**) show similar patterns as those from top 10k CpG sites (**Figs** 3 and 4).

Algorithm 1. Remove Highly Correlated Features within Each Assay.
Input: Any assay $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$ ▷ Assay $\mathbf{X}$ with p features
Initiation: $\mathbf{S} = \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{1} \leq i, j \leq p, i \neq j, \mathbf{corr}^2(\mathbf{x}_i, \mathbf{x}_j) \geq \mathbf{0.8}\}$
▷ Each element of $\mathbf{S}$ is a pair of features from $\mathbf{X}$ whose correlation square is no less than 0.8
Definition: Features in each pair $(\mathbf{x}_i, \mathbf{x}_j)$ in $\mathbf{S}$ are viewed as "neighbors".
while $\mathbf{S} \neq \emptyset$ **do**
  dict←{} ▷ Initiate an empty dictionary storing number of neighbors of each feature $\mathbf{x}_i$
  for $\mathbf{x}_i \leftarrow (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ **do** ▷ Loop each feature $\mathbf{x}_i$ of assay $\mathbf{X}$
    Count neighbors of $\mathbf{x}_i$
    dict←{$\mathbf{x}_i$: number of neighbors}
  end for
  Identify $\mathbf{x}_j \in$ dict with minimal number of neighbors.
  Remove any $(\mathbf{x}_j, \cdot)$ from $\mathbf{S}$.

```
    Remove x_j from X.
end while
Output: X  ▷ After removing features above, features remaining in X
are in low correlations while maximizing the feature size.
```

Association analysis between outcomes and CVs/PCs

To quantify the relationship between outcomes and CVs or PCs, we used regression models. Specifically, for continuous outcomes (including age, BMI, WBC, RBC, PLT), we estimated the proportion of variation in outcome that can be explained by CVs or PCs using linear regression models, implemented with the R function "lm", with covariate adjustments outlined in S2 Table. For the binary outcome sex, we employed logistic regression using the "glm" R function and calculated McFadden's pseudo-R-squared using the "PseudoR2" function from the DescTools [55] R package.

Modified gram-schmidt strategy

With PMA implementation, we observed that with our real data where features have complex correlation structure, the weight vectors are sometimes correlated. To mitigate this non-orthogonality issue, we adopt a strategy inspired by Woojoo et al., (2011) [14]. In our implementation, we infer CVs sequentially and remove the effects of the former CVs from the input matrix before calculating weights for the next CVs. In particular, we first follow the PMA approach to generate weights for CV1's of all assays, update input matrices following Eq (1) as the new inputs for calculating weights for CV2's, and sequentially update until we obtain pre-specified numbers of CVs. Eq (1) and Eq (2) show the procedures for inferring the (j+1)'s CVs with input matrices $\{X_{ij}\}_{i=1,\cdots,S}$.

$$\{CV_{ij}\}, \{W_{ij}\} \sim \mathrm{SMCCA}(\{X_{ij}\}_{i=1,\ldots,S}) \tag{1}$$

$$X_{i,j+1} = X_{i1} - CV_{ij} \times (W_{ij})^{\mathrm{T}} \tag{2}$$

Specifically, $\{X_{ij}\}_{i=1,\cdots,S}$ are original input matrices for assay $i = 1,\cdots,S$ where S is the total number of assays, from which W_{i1}'s and CV_{i1}'s, the first weights and CV1's, are inferred by SMCCA implemented in the PMA R package.

eQTM analysis

We used expression quantitative trait methylation (eQTM) results to alternatively map CpG sites to genes (instead of simply mapping to nearest genes). We used transcriptomic and methylomic measurements for 650 and 496 samples from MESA and JHS, respectively, to perform eQTM analysis for the 316 CpG sites contributing to CVs significantly associated with outcomes. We employed the MatrixEQTL R [56] package to assess the association of each CpG site with its nearby genes in the +/- 1Mb neighborhood, while adjusting for age, and sex separately in MESA and JHS. For the multi-ethnic MESA samples, we additionally adjusted for self-reported race/ethnicity and recruitment site. We then conducted meta-analysis using METAL [57], and used a Bonferroni threshold to define significance, identifying 515 significant CpG-gene pairs. Our eQTM analysis successfully mapped 44–112 CpG sites, for each CV significantly associated with outcome, to 25–257 genes (detailed in S6 Table), based on which we further performed pathway enrichment analysis, following the same process detailed in the section below.

Pathway enrichment analysis

For each CCA-prioritized feature of each assay, we first mapped them to genes, and then performed pathway enrichment analysis on these genes utilizing three databases–DisGeNET [30,58] (enrichDGN function in DOSE R package, with default settings), Gene Ontology (GO) [28,29,59,60] (enrichGO function in clusterProfiler R package, with default settings) and Kyoto Encyclopedia of Genes and Genomes (KEGG) [59–63] (enrichKEGG function in clusterProfiler R package, with default settings). For methylomics, we explored two methods for mapping CpG sites to genes: (1) mapping them to the nearest genes using annotations provided by Illumina, and (2) mapping CpG sites to genes with significant signals identified in the eQTM analysis presented above. For proteomics, we mapped proteins to genes using annotations released by SomaScan. For background genes in the enrichment analysis, we included genes annotated from features that are associated with outcome, identified in the feature selection step of our SSMCCA (detailed in Section "Extending Supervised Sparse CCA to Supervised Sparse Multiple CCA" above).

Supporting information

S1 Fig. Improved orthogonality among CVs by adopting the Gram–Schmidt (GS) strategy. CVs are inferred from JHS proteomics and methylomics data using unsupervised SMCCA. Each row and column represent one CV, ranging from CV1 to CV50. **(A-B)** Results from the PMA package, implementation of the original SMCCA methods without the incorporation of GS algorithm. **(C-D)** Results from our SMCCA-GS, with the GS strategy incorporated. Left panel (**A** and **C**) show proteomics CVs, and right panel (**B** and **D**) from methylomics CVs.
(TIF)

S2 Fig. Sample size for each cohort. (A) JHS: 881 participants have complete proteomics, methylation, and phenotype information; 496 participants have complete transcriptomics, methylation, and phenotype information. **(B)** MESA: 777 participants have complete proteomics, methylation, and phenotype information; 650 participants have complete transcriptomics, methylation, and phenotype information.
(TIF)

S3 Fig. PCA-IQR plots. Each dot in the plot represents one individual. X-axis is the interquartile range (IQR) while Y-axis is the top principal component (PC). **(A)** JHS proteomics: one outlier was detected, marked by the wedge pointer; **(B)** JHS methylomics: three outliers were detected; **(C)** MESA proteomics: MESA: no outliers; **(D)** MESA methylomic: three outliers were detected.
(TIF)

S4 Fig. Proportion of variation in outcomes explained by CVs inferred with all CpG sites included. (A) CVs were inferred using proteomics and all ~700k CpG sites in JHS. The top 50 CVs were used to calculate the r^2 (Y-axis) for each outcome (X-axis). **(B)** We obtained CVs in JHS by applying the weights inferred from MESA, and then calculated r^2 in the same way as in A. **(C)** CVs were inferred using proteomics and all ~700k CpG sites in MESA. (D) CVs were obtained in MESA by applying the weights inferred from JHS.
(TIF)

S5 Fig. Comparison of r^2, PCs vs CVs, inferred with all CpG sites included. Each column is for one outcome. Top row (JHS) shows results in JHS using JHS-inferred CVs. Second row (JHS->MESA) shows results in MESA, also using JHS-inferred CV weights. Third row (MESA) shows results in MESA, this time using MESA-inferred CVs. Last row (MESA->JHS)

shows results in JHS, also using MESA-inferred CV weights. **(A)** Proteomics. **(B)** Methylomics. In each sub-figure, X-axis indicates the number of CVs or PCs used and Y-axis the proportion of variation explained in the (i.e., r^2).
(TIF)

S6 Fig. Association with proteomics-specific technical variables, CVs vs PCs. For JHS, the proteomics technical variable is batch-plate combination status. For MESA, the proteomics technical variable is plate.
(TIF)

S7 Fig. Association with methylation-specific technical variables, CVs vs PCs. For JHS, the methylomics technical variable is group-plate combination status. For MESA, the methylomics technical variables are **(A)** "Batch Scan", and **(B)** "Level1 Batch".
(TIF)

S1 Table. Demographics of (A) JHS and (B) MESA.
(XLSX)

S2 Table. Covariate adjustments of each omics data of each cohort.
(XLSX)

S3 Table. Supervised CVs inferred in MESA significantly associated with each phenotype and validated in JHS.
(XLSX)

S4 Table. Supervised CVs inferred in JHS significantly associated with each phenotype and validated in MESA.
(XLSX)

S5 Table. Proteins identified in CV3 with non-zero weights in MESA.
(XLSX)

S6 Table. Mapping CpG Sites to Genes.
(XLSX)

S7 Table. Pathway Enrichment Analysis Results of JHS.
(XLSX)

S8 Table. Pathway Enrichment Analysis Results of MESA.
(XLSX)

S1 Text. Supplementary Information.
(PDF)

S1 Acknowledgement. Members of NHLBI TOPMed Consortium and TOPMed Analysis Working Group.
(PDF)

Acknowledgments

Molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the Northwest Genomics Center (HHSN268201100037C). Genome sequencing for "NHLBI TOPMed: The Multi-Ethnic Study of Atherosclerosis" (phs001416) was performed at Broad Genomics (3U54HG003067-13S1,

HHSN268201600034I). Methylomics for "NHLBI TOPMed: The Multi-Ethnic Study of Atherosclerosis" (phs001416) was performed at the Keck MGC (HHSN268201600034I). RNA-Seq for "NHLBI TOPMed: The Multi-Ethnic Study of Atherosclerosis" (phs001416) was performed at the Northwest Genomics Center (HHSN268201600032I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination, were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I), and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://wwwmesa-nhlbi.org.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Author Contributions

Conceptualization: Michael I. Love, Laura M. Raffield, Yun Li.

Data curation: Min-Zhi Jiang, Laura M. Raffield.

Formal analysis: Min-Zhi Jiang.

Funding acquisition: Michael I. Love, Laura M. Raffield, Yun Li.

Investigation: Min-Zhi Jiang, Michael I. Love, Laura M. Raffield, Yun Li.

Methodology: Min-Zhi Jiang, Michael I. Love, Laura M. Raffield, Yun Li.

Project administration: Michael I. Love, Laura M. Raffield, Yun Li.

Resources: James G. Wilson, Stephen S. Rich, Jerome I. Rotter, Michael I. Love, Laura M. Raffield, Yun Li.

Software: Min-Zhi Jiang.

Supervision: Michael I. Love, Laura M. Raffield, Yun Li.

Visualization: Min-Zhi Jiang, Jiawen Chen.

Writing – original draft: Min-Zhi Jiang, Michael I. Love, Laura M. Raffield, Yun Li.

Writing – review & editing: Min-Zhi Jiang, François Aguet, Kristin Ardlie, Jiawen Chen, Elaine Cornell, Dan Cruz, Peter Durda, Stacey B. Gabriel, Robert E. Gerszten, Xiuqing Guo, Craig W. Johnson, Silva Kasela, Leslie A. Lange, Tuuli Lappalainen, Yongmei Liu, Alex P. Reiner, Josh Smith, Tamar Sofer, Kent D. Taylor, Russell P. Tracy, David J. VanDenBerg, James G. Wilson, Stephen S. Rich, Jerome I. Rotter, Michael I. Love, Laura M. Raffield, Yun Li.

References

1. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009; 8: Article28. https://doi.org/10.2202/1544-6115.1470 PMID: 19572827
2. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann Appl Stat. 2013; 7: 523–542. https://doi.org/10.1214/12-AOAS597 PMID: 23745156
3. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018; 14: e8124. https://doi.org/10.15252/msb.20178124 PMID: 29925568
4. Consortium GTEx. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020; 369: 1318–1330. https://doi.org/10.1126/science.aaz1776 PMID: 32913098
5. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021; 53: 1300–1310. https://doi.org/10.1038/s41588-021-00913-z PMID: 34475573
6. Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat Metab. 2020; 2: 1135–1148. https://doi.org/10.1038/s42255-020-00287-2 PMID: 33067605
7. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018; 558: 73–79. https://doi.org/10.1038/s41586-018-0175-2 PMID: 29875488
8. Zhang J, Dutta D, Köttgen A, Tin A, Schlosser P, Grams ME, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat Genet. 2022; 54: 593–602. https://doi.org/10.1038/s41588-022-01051-w PMID: 35501419
9. Hotelling H. The most predictable criterion. J Educ Psychol. 1935; 26: 139–142.
10. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat Appl Genet Mol Biol. 2009;8: Article 1.
11. Lin D, Zhang J, Li J, Calhoun VD, Deng H-W, Wang Y- P. Group sparse canonical correlation analysis for genomic data integration. BMC Bioinformatics. 2013; 14: 245. https://doi.org/10.1186/1471-2105-14-245 PMID: 23937249
12. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016; 32: 1981–1989. https://doi.org/10.1093/bioinformatics/btw052 PMID: 27153689
13. Tini G, Marchetti L, Priami C, Scott-Boyer M- P. Multi-omics integration-a comparison of unsupervised clustering methodologies. Brief Bioinform. 2019; 20: 1269–1279. https://doi.org/10.1093/bib/bbx167 PMID: 29272335
14. Woojoo L, Donghwan L, Youngjo L, Yudi P. Sparse Canonical Covariance Analysis for High-throughput Data. Stat Appl Genet Mol Biol. 2011; 10: 1–24.
15. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13: 86. https://doi.org/10.1186/1471-2105-13-86 PMID: 22568884

16. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018; 19: 371–384. https://doi.org/10.1038/s41576-018-0004-3 PMID: 29643443

17. Gatev E, Inkster AM, Negri GL, Konwar C, Lussier AA, Skakkebaek A, et al. Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. Nucleic Acids Res. 2021; 49: 9097–9116. https://doi.org/10.1093/nar/gkab682 PMID: 34403484

18. Grant OA, Wang Y, Kumari M, Zabet NR, Schalkwyk L. Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina EPIC array. Clin Epigenetics. 2022; 14: 62. https://doi.org/10.1186/s13148-022-01279-7 PMID: 35568878

19. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017; 541: 81–86. https://doi.org/10.1038/nature20784 PMID: 28002404

20. Zaghlool SB, Sharma S, Molnar M, Matías-García PR, Elhadad MA, Waldenberger M, et al. Revealing the role of the human blood plasma proteome in obesity using genetic drivers. Nat Commun. 2021; 12: 1279. https://doi.org/10.1038/s41467-021-21542-4 PMID: 33627659

21. Lehallier B, Gate D, Schaum N, Nanasi T, Lee SE, Yousef H, et al. Undulating changes in human plasma proteome profiles across the lifespan. Nat Med. 2019; 25: 1843–1850. https://doi.org/10.1038/s41591-019-0673-2 PMID: 31806903

22. Katz DH, Tahir UA, Bick AG, Pampana A, Ngo D, Benson MD, et al. Whole Genome Sequence Analysis of the Plasma Proteome in Black Adults Provides Novel Insights Into Cardiovascular Disease. Circulation. 2022; 145: 357–370. https://doi.org/10.1161/CIRCULATIONAHA.121.055117 PMID: 34814699

23. Schubert R, Geoffroy E, Gregga I, Mulford AJ, Aguet F, Ardlie K, et al. Protein prediction for trait mapping in diverse populations. PLoS One. 2022; 17: e0264341. https://doi.org/10.1371/journal.pone.0264341 PMID: 35202437

24. Raffield LM, Dang H, Pratte KA, Jacobson S, Gillenwater LA, Ampleford E, et al. Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. Proteomics. 2020; 20: e1900278. https://doi.org/10.1002/pmic.201900278 PMID: 32386347

25. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021; 590: 290–299. https://doi.org/10.1038/s41586-021-03205-y PMID: 33568819

26. Broudy VC. Stem Cell Factor and Hematopoiesis. Blood. 1997; 90: 1345–1364. PMID: 9269751

27. Kjeldsen L, Bainton DF, Sengeløv H, Borregaard N. Identification of neutrophil gelatinase-associated lipocalin as a novel matrix protein of specific granules in human neutrophils. Blood. 1994; 83: 799–807. PMID: 8298140

28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25: 25–29.

29. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021; 49: D325–D334. https://doi.org/10.1093/nar/gkaa1113 PMID: 33290552

30. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020; 48: D845–D855. https://doi.org/10.1093/nar/gkz1021 PMID: 31680165

31. Rahman A, Hammad MM, Al Khairi I, Cherian P, Al-Sabah R, Al-Mulla F, et al. Profiling of Insulin-Like Growth Factor Binding Proteins (IGFBPs) in Obesity and Their Association With Ox-LDL and Hs-CRP in Adolescents. Front Endocrinol. 2021; 12: 727004. https://doi.org/10.3389/fendo.2021.727004 PMID: 34394011

32. Saidu NEB, Bonini C, Dickinson A, Grce M, Inngjerdingen M, Koehl U, et al. New Approaches for the Treatment of Chronic Graft-Versus-Host Disease: Current Status and Future Directions. Front Immunol. 2020; 11: 578314. https://doi.org/10.3389/fimmu.2020.578314 PMID: 33162993

33. Woo SJ, Ahn J, Morrison MA, Ahn SY, Lee J, Kim KW, et al. Analysis of Genetic and Environmental Risk Factors and Their Interactions in Korean Patients with Age-Related Macular Degeneration. PLoS One. 2015; 10: e0132771. https://doi.org/10.1371/journal.pone.0132771 PMID: 26171855

34. Kikuchi M, Nakamura M, Ishikawa K, Suzuki T, Nishihara H, Yamakoshi T, et al. Elevated C-reactive protein levels in patients with polypoidal choroidal vasculopathy and patients with neovascular age-related macular degeneration. Ophthalmology. 2007; 114: 1722–1727. https://doi.org/10.1016/j.ophtha.2006.12.021 PMID: 17400294

35. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. N Engl J Med. 2019; 381: 668–676. https://doi.org/10.1056/NEJMsr1809937 PMID: 31412182

36. Sanford JA, Nogiec CD, Lindholm ME, Adkins JN, Amar D, Dasari S, et al. Molecular Transducers of Physical Activity Consortium (MoTrPAC): Mapping the Dynamic Responses to Exercise. Cell. 2020; 181: 1464–1474. https://doi.org/10.1016/j.cell.2020.06.004 PMID: 32589957

37. Png G, Barysenka A, Repetto L, Navarro P, Shen X, Pietzner M, et al. Mapping the serum proteome to neurological diseases using whole genome sequencing. Nat Commun. 2021; 12: 7042. https://doi.org/10.1038/s41467-021-27387-1 PMID: 34857772

38. Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, Wörheide MA, et al. Mapping the proteo-genomic convergence of human diseases. Science. 2021; 374: eabj1541. https://doi.org/10.1126/science.abj1541 PMID: 34648354

39. Williams SA, Kivimaki M, Langenberg C, Hingorani AD, Casas JP, Bouchard C, et al. Plasma protein patterns as comprehensive indicators of health. Nat Med. 2019; 25: 1851–1857. https://doi.org/10.1038/s41591-019-0665-2 PMID: 31792462

40. Brown BC, Wang C, Kasela S, Aguet F, Nachun DC, Taylor KD, et al. Multiset correlation and factor analysis enables exploration of multi-omic data. bioRxiv. 2022. p. 2022.07.18.500246. https://doi.org/10.1101/2022.07.18.500246

41. Taylor HA Jr, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn Dis. 2005; 15: S6-4–17. PMID: 16320381

42. Wilson JG, Rotimi CN, Ekunwe L, Royal CDM, Crump ME, Wyatt SB, et al. Study design for genetic analysis in the Jackson Heart Study. Ethn Dis. 2005; 15: S6-30–37. PMID: 16317983

43. Carpenter MA, Crow R, Steffes M, Rock W, Heilbraun J, Evans G, et al. Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. Am J Med Sci. 2004; 328: 131–144. https://doi.org/10.1097/00000441-200409000-00001 PMID: 15367870

44. Lu AT, Seeboth A, Tsai P- C, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging. 2019; 11: 5895–5923. https://doi.org/10.18632/aging.102173 PMID: 31422385

45. Do WL, Nguyen S, Yao J, Guo X, Whitsel EA, Demerath E, et al. Associations between DNA methylation and BMI vary by metabolic health status: a potential link to disparate cardiovascular outcomes. Clin Epigenetics. 2021; 13: 230. https://doi.org/10.1186/s13148-021-01194-3 PMID: 34937574

46. TOPMed whole genome sequencing methods: Freeze 8. [cited 2 Mar 2022]. Available: https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8

47. Chen M- H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell. 2020; 182: 1198–1213.e14. https://doi.org/10.1016/j.cell.2020.06.045 PMID: 32888493

48. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015; 39: 276–293. https://doi.org/10.1002/gepi.21896 PMID: 25810074

49. Reich D, Nalls MA, Kao WHL, Akylbekova EL, Tandon A, Patterson N, et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. 2009; 5: e1000360. https://doi.org/10.1371/journal.pgen.1000360 PMID: 19180233

50. Kurniansyah N, Wallace DA, Zhang Y, Yu B, Cade B, Wang H, et al. An integrated multi-omics analysis of sleep-disordered breathing traits across multiple blood cell types. medRxiv. 2022. p. 2022.07.09. https://doi.org/10.1101/2022.07.09.22277444.

51. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. 2014; 30: 428–430. https://doi.org/10.1093/bioinformatics/btt684 PMID: 24336642

52. Fortin J- P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014; 15: 503. https://doi.org/10.1186/s13059-014-0503-2 PMID: 25599564

53. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017; 45: e22. https://doi.org/10.1093/nar/gkw967 PMID: 27924034

54. Traxl D, Boers N, Kurths J. Deep Graphs—a general framework to represent and analyze heterogeneous complex systems across scales. arXiv [physics.data-an]. 2016. Available: http://arxiv.org/abs/1604.00971

55. Signorell A, Aho K, Alfons A, Anderegg N, Aragon T, Arachchige C, et al. DescTools: Tools for Descriptive Statistics. 2017. Available: https://cran.r-project.org/package=DescTools

56. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28: 1353–1358. https://doi.org/10.1093/bioinformatics/bts163 PMID: 22492648

57. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26: 2190–2191. https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382

58. Yu G, Wang L- G, Yan G- R, He Q- Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015; 31: 608–609. https://doi.org/10.1093/bioinformatics/btu684 PMID: 25677125

59. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021; 2: 100141. https://doi.org/10.1016/j.xinn.2021.100141 PMID: 34557778

60. Yu G, Wang L- G, Han Y, He Q- Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16: 284–287. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

61. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28: 27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

62. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019; 28: 1947–1951. https://doi.org/10.1002/pro.3715 PMID: 31441146

63. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2022. https://doi.org/10.1093/nar/gkac963 PMID: 36300620