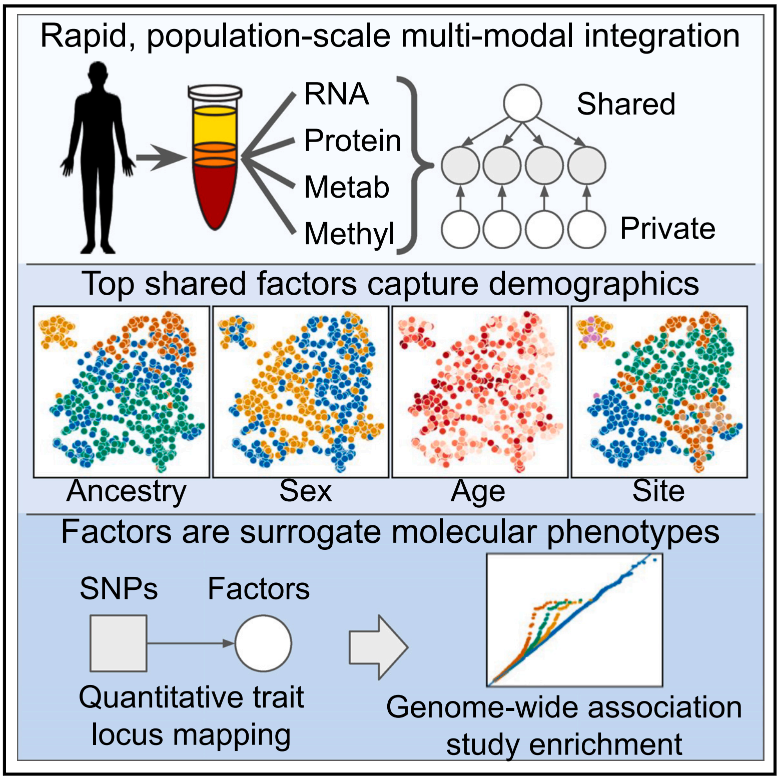
Cell Genomics

Multiset correlation and factor analysis enables exploration of multi-omics data

Graphical abstract


Rapid, population-scale multi-modal integration
RNA
Protein
Metab
Methyl
Shared
Private
Top shared factors capture demographics
Ancestry
Sex
Age
Site
Factors are surrogate molecular phenotypes
SNPs
Factors
Quantitative trait locus mapping
Genome-wide association study enrichment

Highlights

- Rapid, unsupervised multi-modal data integration with self-inferred tuning parameters
- 614 ancestry-diverse participants from MESA/TOPMed with 5 omics types
- Top shared components capture ancestry, even without genetic information
- Further components are enriched for GWAS hits and related to metabolic disease

Authors

Brielin C. Brown, Collin Wang, Silva Kasela, ..., Kristin G. Ardlie, David A. Knowles, Tuuli Lappalainen

Correspondence

bbrown@nygenome.org

In brief

Brown, Wang et al. introduce MCFA, an approach to multi-modal dataset integration that generalizes canonical correlation analysis. MCFA is broadly applicable to data integration challenges but has been designed to handle issues in population-scale multi-omics data. A variety of analyses on the TOPMed/MESA multi-omics pilot demonstrate the power of this method.

Brown et al., 2023, Cell Genomics *3*, 100359
August 9, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.xgen.2023.100359

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Short article

Multiset correlation and factor analysis enables exploration of multi-omics data

Brielin C. Brown,[1,2,21,22,*] Collin Wang,[1,3,21] Silva Kasela,[1,4] François Aguet,[5,6] Daniel C. Nachun,[7] Kent D. Taylor,[8] Russell P. Tracy,[9] Peter Durda,[9] Yongmei Liu,[10] W. Craig Johnson,[11] David Van Den Berg,[12] Namrata Gupta,[6] Stacy Gabriel,[6] Joshua D. Smith,[13] Robert Gerzsten,[14] Clary Clish,[6] Quenna Wong,[11] George Papanicolau,[15] Thomas W. Blackwell,[16] Jerome I. Rotter,[8] Stephen S. Rich,[17] R. Graham Barr,[18] Kristin G. Ardlie,[6] David A. Knowles,[1,2,3,4,20] and Tuuli Lappalainen[1,4,19,20]

[1]New York Genome Center, New York, NY, USA
[2]Data Science Institute, Columbia University, New York, NY, USA
[3]Department of Computer Science, Columbia University, New York, NY, USA
[4]Department of Systems Biology, Columbia University, New York, NY, USA
[5]Illumina Incorporated, San Francisco, CA, USA
[6]The Broad Institute of MIT and Harvard, Boston, MA, USA
[7]Department of Pathology, Stanford University, Stanford, CA, USA
[8]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA
[9]Department of Pathology and Laboratory Medicine, Larner College of Medicine, University of Vermont, Burlington, VT, USA
[10]Department of Medicine, Duke University Medical Center, Durham, NC, USA
[11]Department of Biostatistics, University of Washington, Seattle, WA, USA
[12]Department of Clinical Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[13]Northwest Genomics Center, University of Washington, Seattle, WA, USA
[14]Beth Israel Deaconess Medical Center, Division of Cardiovascular Medicine, Boston, MA, USA
[15]Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, Bethesda, MD, USA
[16]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA
[17]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[18]Mailman School of Public Health, Columbia University, New York, NY, USA
[19]Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden
[20]Senior author
[21]These authors contributed equally
[22]Lead contact
*Correspondence: bbrown@nygenome.org

https://doi.org/10.1016/j.xgen.2023.100359

SUMMARY

Multi-omics datasets are becoming more common, necessitating better integration methods to realize their revolutionary potential. Here, we introduce multi-set correlation and factor analysis (MCFA), an unsupervised integration method tailored to the unique challenges of high-dimensional genomics data that enables fast inference of shared and private factors. We used MCFA to integrate methylation markers, protein expression, RNA expression, and metabolite levels in 614 diverse samples from the Trans-Omics for Precision Medicine/Multi-Ethnic Study of Atherosclerosis multi-omics pilot. Samples cluster strongly by ancestry in the shared space, even in the absence of genetic information, while private spaces frequently capture dataset-specific technical variation. Finally, we integrated genetic data by conducting a genome-wide association study (GWAS) of our inferred factors, observing that several factors are enriched for GWAS hits and *trans*-expression quantitative trait loci. Two of these factors appear to be related to metabolic disease. Our study provides a foundation and framework for further integrative analysis of ever larger multi-modal genomic datasets.

INTRODUCTION

Recent years have seen an explosion in multi-omics data, with studies simultaneously profiling RNA expression, protein levels, chromatin accessibility, and more.[1] By providing complementary views into the underlying biology, these datasets promise to illuminate molecular processes and disease states that cannot be gleaned from any lone modality.[2] However, joint inference methods are lacking in either the number or type of modes that can be used or in flexibility and efficiency.[1] Multi-omics data bring substantial challenges: distributions differ between modes, the sample size is typically small relative to features, efficient algorithms are needed, and each mode has contributions from factors that are shared between modes and unique to itself.[3,4]

Canonical correlation analysis (CCA) is a statistical technique that infers shared factors between two data modes by finding correlated linear combinations of the features in each.[5] CCA has enjoyed substantial attention in genomics[6–9]; however, extending CCA to additional modes is fraught: at least 10 different formulations are equivalent in the two-mode case,[10] and many are challenging to fit.[11] Equivalently, CCA can be conceptualized as a probabilistic model (pCCA), revealing a connection to factor analysis.[12]

We have developed multi-set correlation and factor analysis (MCFA; Figures 1A and S1), an unsupervised integration method that generalizes pCCA and factor analysis, enabling fast inference of shared and private factors in multi-modal data. MCFA is designed to overcome challenges that are common with genomics data such as the large number of features relative to the sample size, the disparate data types, and the unknown contributions of dataset-specific technical factors. MCFA is based on two insights: (1) unlike traditional CCA, pCCA has only one natural extension to multi-modal data, which is both conceptually elegant and efficient to fit, and (2) after fitting pCCA, the residual in a mode represents private structure, which is well modeled by factor analysis. Our method combines these insights to fit factors that are shared across modalities and are private to each simultaneously. For efficiency and regularization, MCFA uses the top principal components (PCs) of each mode.[6,7] It allows the use of random matrix techniques[13] to choose the shared dimensionality and number of PCs, eliminating tuning parameters. Finally, MCFA is a natural approach to integration: as detailed in Methods S1, there is a theoretical connection between our model and multi-set CCA.

We have applied MCFA to 614 ancestry-diverse individuals from the Multi-Ethnic Study of Atherosclerosis (MESA).[14] The Trans-Omics for Precision Medicine (TOPMed)[15] program instituted a multi-omics pilot study to evaluate the utility of long-term stored samples for discovery related to heart, lung, blood, and sleep disorders. MESA provided samples for five omics types: (1) whole-genome sequencing (WGS), (2) RNA sequencing of peripheral blood mononuclear cells (PBMCs), (3) DNA methylation array profiling from whole blood, (4) protein mass spectrometry of blood plasma, and (5) metabolite mass spectrometry of blood plasma. In addition, MESA has collected comprehensive phenotypic metadata. These data include demographic markers such as self-reported ancestry (SRA), sex, age, and education level; morphological features including height, weight, and hip circumference; clinical measures including those related to atherosclerosis, lipid levels, kidney function, and inflammatory biomarkers; and behavioral features regarding smoking, drinking, and exercise frequency.

RESULTS

We integrated RNA sequencing, methylation, protein, and metabolite data using MCFA, which inferred a 14-dimensional shared space. We found that shared structure explained a large proportion of the variance in each mode (Figure 1B, right). Protein levels had the highest sharing with 29.2% of the variance explained (VE) by the shared space, followed by RNA and metabolite levels (16.6% and 17.1%, respectively). Methylation showed the least sharing, with only 8.1% VE by the shared space. Due to the high dimensionality of the data and the limited sample size, about half of the variance in each dataset is unmodeled to reduce overfitting. Using MCFA, it is possible to further infer the variance in each modality explained by the individual factors, thus determining which modalities contribute to each (Figure 1B, left). Our top factor has contributions from all modalities, but their respective contributions to the other factors vary substantially.

We used uniform manifold approximation and projection (UMAP)[16] to construct a 2D embedding of the shared and private spaces (Figure 1C). We noticed a striking clustering of the individuals by SRA and sex in the shared space, even though the top PCs of individual modes do not cluster by these factors (Figure S2), and the shared space was inferred without genetic or sex chromosome features. Shared factor 1 separates Black and White individuals, with Hispanic individuals in between, while factor 3 separates Chinese individuals, and factor 2 differentiates by sex (Figures S2 and S3). We validated this structure via leave-one-out cross-validation, indicating our PC selection strategy mitigated over-fitting (Figure S4).

Next, we evaluated the total phenotypic VE by each of our inferred spaces (Figures 1D and S2; Tables S1, S2, and S3). The shared space captured 95.3% of the variation in sex, 83.3% in site, 80.0% in SRA, and 60.2% in age. The shared space also captured anthropomorphic differences such as BMI (51.0% VE) and clinical measures including those related to kidney function (creatine, 64.8% VE) and inflammation (tumor necrosis factor (TNF)-alpha receptor-1 69.1% VE). We used CIBERSORT[17] and the Houseman method[18] to estimate the cell-type composition of our RNA (PBMC) and methylation (whole blood) samples, respectively. Both shared and privates spaces contributed to the relative proportions of PBMC-abundant cell types (e.g., T cells and natural killer (NK) cells) estimated from both data modalities, while the proportion of PBMC-depleted types (e.g., neutrophils) estimated from the methylation data was only captured by the methylation private space. Modality-private spaces frequently captured technical factors: 100% of the variance in sequencing center and 71.6% of the variance in 3′ bias are captured by the RNA private space, while 76.8% of the methylation array batch is captured by its private space. Many phenotypes that are themselves measurements of metabolites were captured by the metabolite private space; however, the strongest association was with the month of sample collection (85.8% VE). We noticed no large associations between the protein private space and any of our metadata, despite several of our phenotypes being clinical protein markers; however, several of these factors are partially captured by the shared space.

We compared the results obtained on MESA using MCFA with other multi-modal analysis approaches. We focused on two alternative methods: (1) MOFA2[4] and (2) a multi-modal autoencoder (MMAE, see STAR Methods and Figure S5). In the MOFA2 analysis, the methylation batch and cell-type proportions dominated the inferred shared space, likely owing to the very large number of features in that modality compared with the other modalities (Figure 2). The MMAE mitigated this overfocus on methylation somewhat and additionally captured RNA

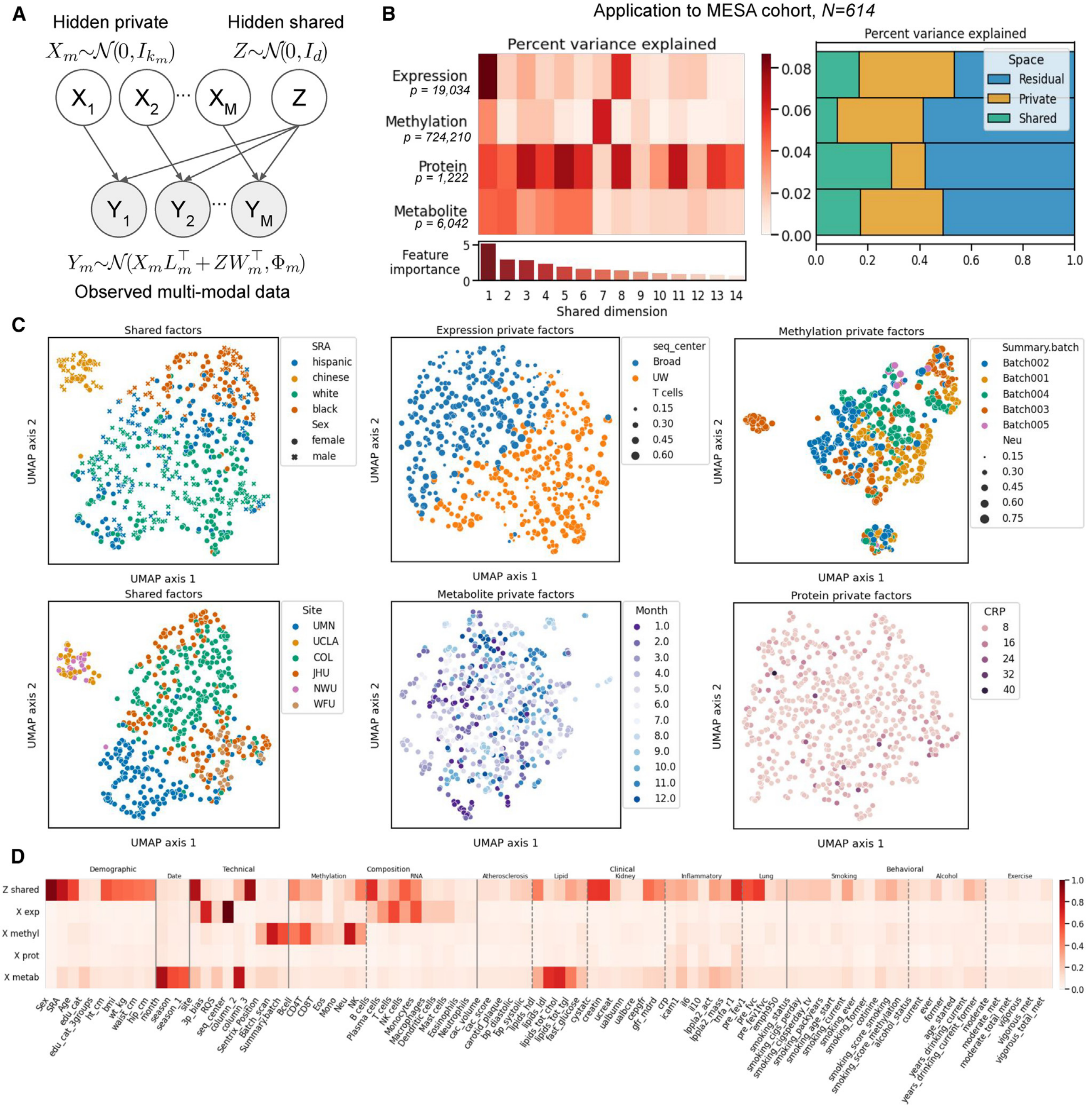


Figure 1. Overview of MCFA integration results

(A) The MCFA model. Each observed data mode (Ym) has contributions from two latent factors, one private to it (Xm) and one shared with other modes (Z).

(B) Breakdown of the variance in four omics types captured by the inferred space, as well as the per-mode contribution to each shared factor.

(C) UMAP embedding of the shared and private spaces, annotated with the most relevant feature set. Broadly, the top shared factors capture demographics, while the top private factors capture technical variation.

(D) Variance in sample metadata explained by each learned space. This shows that the shared space also captures inferred cell-type composition estimates as well as clinical biomarkers.

sequencing center and RNA cell-type proportions (Figure 2). Thus, neither MOFA2 nor the MMAE were able to infer shared variation while discarding dataset-specific technical artifacts. Moreover, using up to 8 cores of an Intel Xeon E5-2697v3 CPU on our cluster, MOFA2 took approximately 56 min to run when set to "medium" tolerance, while our MMAE took approximately 109 min to converge. In contrast, MCFA is able to process the same dataset in around 2 min.

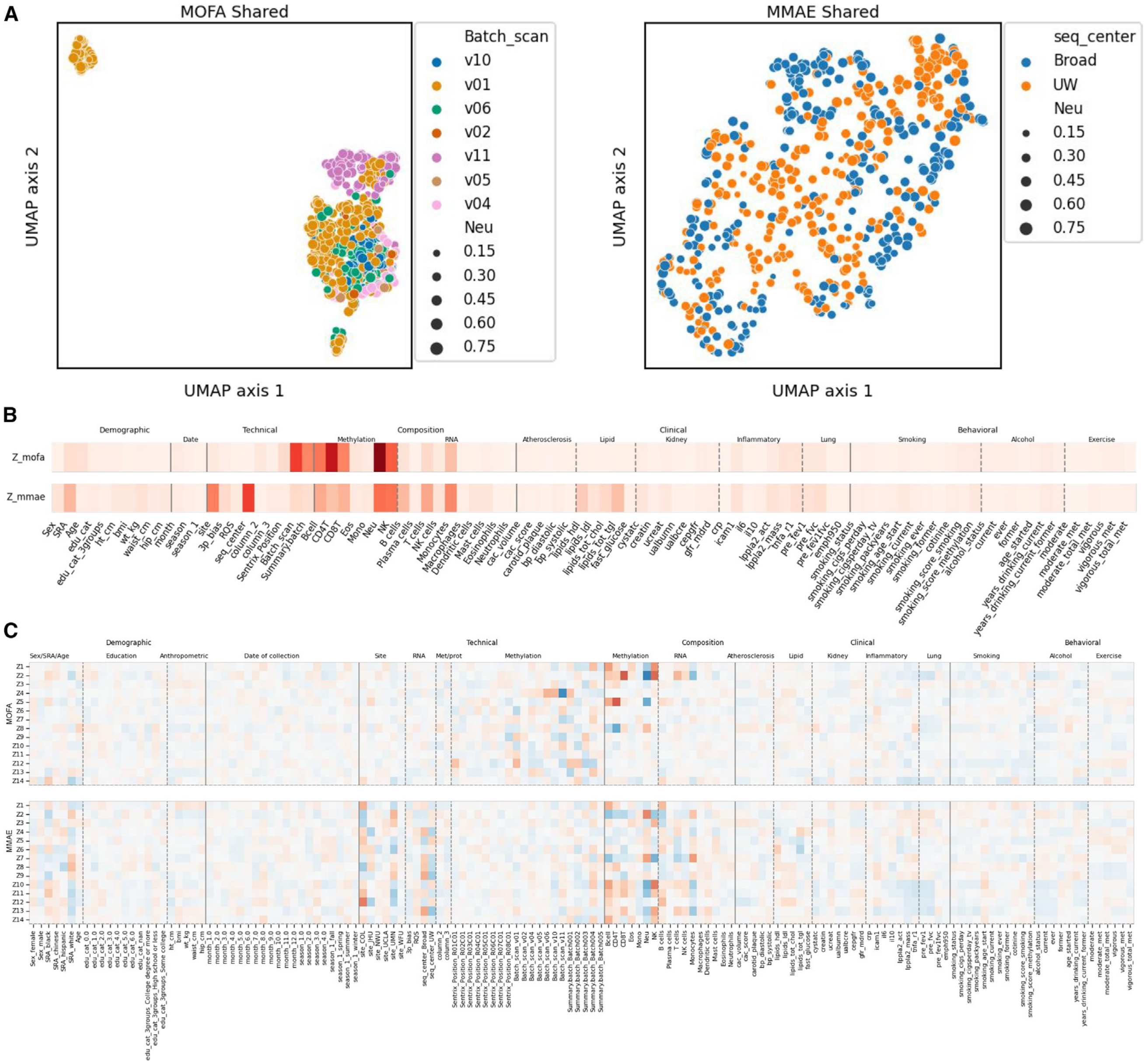


Figure 2. Comparison of MCFA with other methods

(A) UMAP embeddings of MOFA (left) and MMAE (right) shared space show that these methods fail to separate meaningful information from technical variation.

(B) Variance in sample metadata explained by the MOFA2 (top) and MMAE (bottom) shared spaces. MOFA2 primarily learns factors related to the methylation dataset, while the MMAE additionally incorporates some factors related to RNA sequencing.

(C) Correlation of each inferred factor with each metadata sample for MOFA (top) and the MMAE (bottom).

Finally, we integrated WGS data by conducting a genome-wide association study (GWAS) of the inferred factors while controlling for site, age, sex, and 11 genotype PCs. We hypothesized that genetic associations with our inferred factors, which represent major axes of molecular variation, may be enriched for known GWAS hits or *trans*-expression quantitative trait loci (eQTLs). We obtained a list of 10,174 such associations from the eQTLgen consortium,[19] of which 3,854 are *trans*-eQTLs, and further defined a more limited set of 1,107 “influential” *trans*-eQTLs that affect at least 10 genes. We tested the GWAS of each factor for enrichment of these three categories and found 9 significant enrichments (mean $\chi^2_{cat} > 1$, false discovery rate [FDR] 5%; Figures 3A and S6).

Factor 7 showed the strongest enrichment for reported GWAS hits and *trans*-eQTLs. The top SNPs associated with factor 7 are from blood lipid studies and are located primarily around the FADS1 and FADS2 genes, which are known to regulate lipid metabolism.[20] These include rs174541 ($p = 4.3\times10^{-5}$ for factor 7 association), which is also reported in GWASs of type 2 diabetes[21]; rs174549 ($p = 5.6\times10^{-5}$), which is also reported in

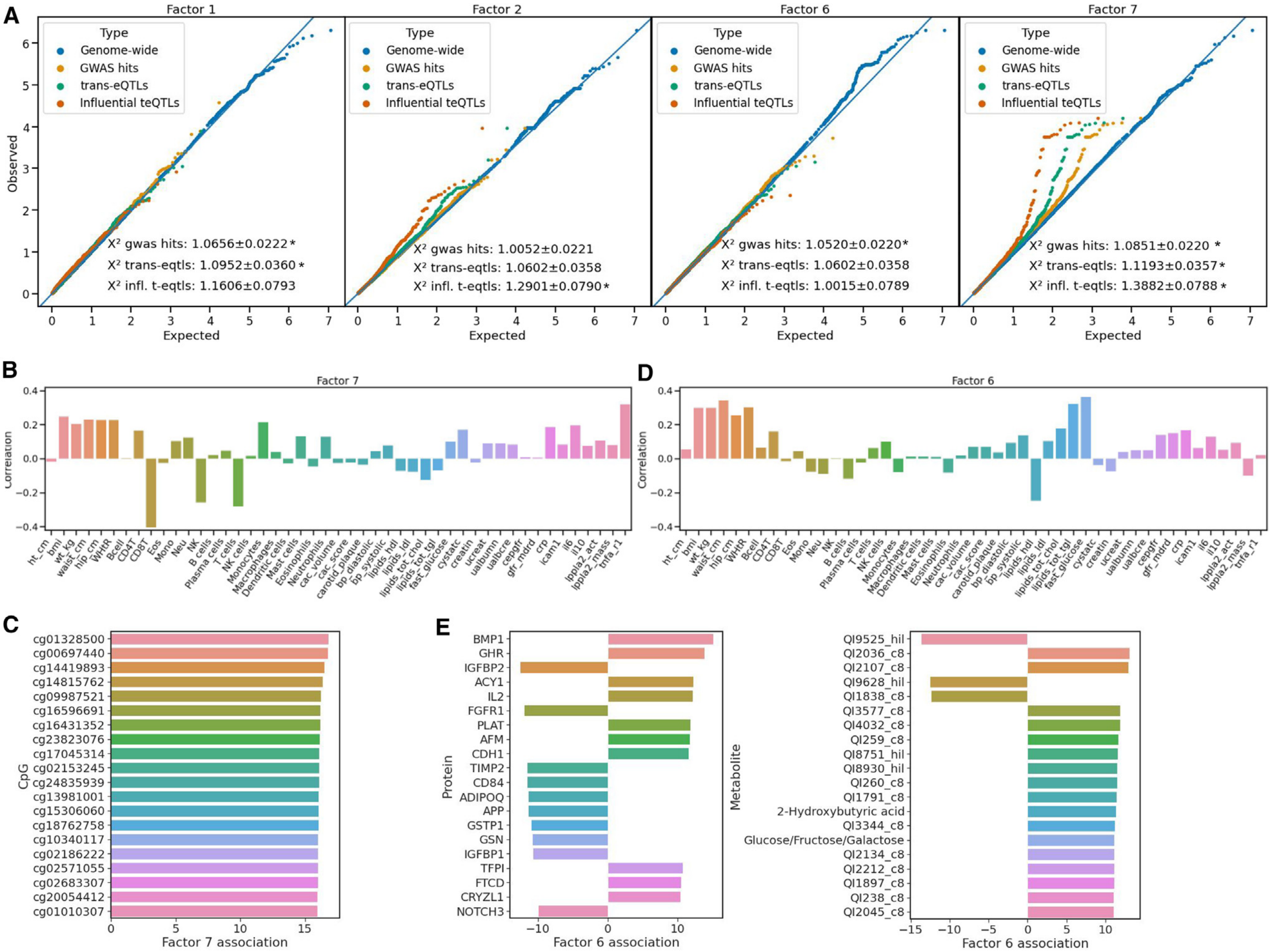


Figure 3. Factor interpretation and integration with GWAS data

(A) QQ-plot of a GWAS for factors 1, 2, 6, and 7. Genetic associations with these factors are enriched for known GWAS loci (1, 6, and 7), *trans*-eQTLs (1 and 7), or highly influential *trans*-eQTLs (2 and 7).

(B and C) Correlation of factors 6 (B) and 7 (C) with morphological, immune-composition, and clinical metadata reveals that factor 6 is related to body composition and lipid profile, while factor 7 is related to body composition, inferred blood cell-type composition, and inflammatory biomarkers.

(D) Z-transformed correlation of individual protein and metabolite data with factor 6 reveals genes and metabolites related to insulin resistance and metabolic syndrome.

(E) Z-transformed correlation of individual methylation values with factor 7. Many genes colocated to these CpGs are involved in lipid metabolism.

GWASs of white blood cell count[22]; and rs1535 ($p = 8.3 \times 10^{-5}$), which is also reported in a GWAS of inflammatory bowel disease[23] (Table S4). Factor 7 explains 6.7% of the modeled variation in methylation, the largest of any factor, and is anti-correlated with sample proportion of CD8$^+$ T cells and NK cells estimated from methylation data ($\rho = -0.41$ and -0.25), and correlated with BMI ($\rho = 0.25$) and measures of inflammation including TNF-R1 ($\rho = 0.33$) and interleukin-6 ($\rho = 0.20$) (Figure 3B).

To assess the contribution of individual CpGs, we calculated the Z-transformed correlation of individual CpG values with factor 7 (Figure 3C). As epigenome-wide association studies remain small, generally little is known about the effects of individual CpGs and their associations with traits. Instead, we linked each gene to the CpGs falling in a window from 1.5 kb upstream of the transcription start site to the transcription termination site. Many of the genes colocated to CpGs with high weights for factor 7 have been implicated in lipid metabolism GWASs including IQCG and TMEM178A (cg01328500 and cg02571055; phosphatidylcholine levels[24]), DSCAML1 (cg02571055; triglyceride levels[25]), PTK2 (cg02153245; ApoB and low-density lipoprotein [LDL] levels[26]); TULP4 (cg02571055; lipoprotein A levels[27]), and C7orf50 (cg20054412; LDL, high-density lipoprotein [HDL], and total cholesterol levels[28]). Interestingly, our second strongest hit, cg00697440, is colocated with CD86. Recent work has suggested that B7 molecules including CD86 play an important role in regulating CD8$^+$ T cell population dynamics.[29] While further research is needed to establish causal relationships of these genetic effects and methylation patterns in *cis* and *trans* on gene regulation and diverse traits, DNA methylation patterns have been previously associated with lipid metabolism and metabolic disease.[30,31] Further research is required to determine whether the immune-cell component of this factor is related to the lipid metabolism component or whether these are simply independent biological functions captured by the same factor.

We used the same strategy to interpret factor 6. Factor 6 is correlated with fasting glucose, waist circumference, and triglycerides ($\rho = 0.36$,0.34, and 0.32, respectively) and anti-correlated with HDL cholesterol ($\rho = -0.25$; Figure 3D). Factor 6 explains 6% of the variance in protein levels and 4.1% of the variance in metabolite levels. Many of the top-weighted metabolites are uncharacterized products from untargeted metabolomics, but the two top characterized targets are 2-hydroxybutyric acid, a known marker of insulin resistance and glucose intolerance,[32,33] and glucose itself (Figure 3E). Several of the top-weighted proteins in this factor have known roles in growth and development including BMP1, GHR, IGFBP2, and FGFR1. GWASs have implicated BMP1 in coronary artery disease,[34,35] IGFBP2 in type 2 diabetes and BMI,[36] and FGFR1 in triglyceride levels[28] and waist-hip ratio.[37] Other notable highly weighted proteins include TFPI, which is involved in blood coagulation and is associated with BMI-adjusted waist-hip ratio,[38] and ADIPOQ, which is involved in regulating glucose levels[39] (Figure 3E). Many of the top GWAS hits associated with this factor corroborate these observations, including rs4805885, which is associated with adiponectin (ADIPOQ) levels[40]; rs9787485, which is associated with insulin-carbohydrate interaction[41]; and rs7679, which is associated with HDL, LDL, and triglyceride levels[42] (Table S5).

Interestingly, the strongest genetic association with this factor comes from GWASs of schizophrenia (rs112973353; $p = 1.6\times 10^{-4}$ for factor 6 association), and we find 5 independent schizophrenia risk loci with factor 6 association p values below 0.01 (Table S5). Insulin resistance and schizophrenia have been consistently associated for nearly 100 years,[43] and while the association signal of each locus with factor 6 is relatively weak, the probability of finding 5 independent loci with these p values under the null is approximately 4×10^{-13}. While further research is needed, our results suggest that these particular loci may confer schizophrenia risk via insulin resistance. Another notable signal in our GWAS associations is related to erythrocyte and platelet traits. These hits include rs12451471 ($p = 8\times 10^{-4}$; mean corpuscular hemoglobin concentration[44]; platelet count[45]) and rs13224082 ($p = 9\times 10^{-4}$; platelet distribution width, platelet count, plateletcrit[44]), among others (Table S5). Again, further research is required to establish causality and direction of effect between genetics, metabolite and protein levels, and traits, but we note that there is an established link between insulin resistance and platelet dysfunction.[46]

DISCUSSION

MCFA has several advantages compared with other multi-omics integration approaches. Compared with group factor analysis methods,[4] MCFA separates modality-specific from dataset-shared factors. Compared with non-negative matrix factorization-based methods that share a feature weight set across modalities,[3] MCFA is able to use all data types. As we have shown, MCFA is also substantially faster and is able to handle datasets with unbalanced numbers of features across the modes.

While our top factors captured ancestry and sex, these factors are usually observed and considered confounding in clinical applications. In that context, one could fit the model conditional on known confounding factors. Since we see exploratory data analysis as a primary application of MCFA, our goal instead was to map the primary axes of biological variation contained within these population-scale multi-omics data. It is important that these factors are a primary driver of variation within such data, as it implies that sampling across race and sex is critical for equitable discovery in medical genomics. Still, because these factors are captured by the top components, and the components themselves are orthogonal, further components can still capture disease-relevant information.

Integration with GWAS is biased toward well-powered studies that will typically have more hits, some of which may be acting indirectly through another phenotype.[47] Interpretability of factors is also biased toward the metadata collected in the study. In MESA, the goal was evaluation of risk factors for heart disease, and thus MESA focused metadata collection on lipid phenotypes, inflammatory biomarkers, and body morphology. It is therefore unsurprising that we are most easily able to interpret factors related to metabolic syndrome, lipid metabolism, and immune function in this study. Still, the ability of MCFA to produce results that are correlated with these factors demonstrates the utility of broad-scale sample metadata when interpreting results from multi-omics studies.

Careful consideration is required when analyzing multi-omics datasets that include WGS or genotype data. There are two primary ways that one can think about integrating these data: (1) include genetic information as a mode in the fit model, interpretable as inferring a latent state that affects genotype as well as molecular factors, or (2) look for genetic associations with inferred molecular factors, interpretable as mapping QTLs for inferred molecular phenotypes. In this study, we chose the latter due to the improved causal interpretation and to demonstrate the utility of surrogate molecular phenotypes. In other cases, for example the analysis of genetic copy-number variation data in tumor samples, the former analysis approach may be preferred. Future work with larger sample sizes may allow for network inference and Mendelian randomization methods to generate directed hypotheses.[47,48] Genetic associations are particularly valuable in this, with the inferred axes of molecular variation providing promising future traits for GWAS and phenome-wide association studies. TOPMed is among the most ambitious current efforts to collect multi-omics population-level data; thus, given the results of this pilot analysis, we expect future integration studies in this cohort to be fruitful.

Limitations of the study

Due to the use of observational data and unsupervised methods, all analyses should be considered exploratory; they can find structure in the data while generating hypotheses but cannot be used to make causal claims and may reflect technical properties of the underlying data. For example, in MESA, the sample collection site is strongly correlated with SRA. We repeated our analysis of the VE by the learned space while additionally controlling for site (Table S3) and noticed a small decrease in the proportion of VE in SRA (from 80.0% to 71.6%).

We observed that estimated cell-type composition had a strong association with both shared and private spaces. Since cell-type composition was inferred from the data, there may be circularity in composition estimation itself. In addition, complex interactions exist between cell-type composition in tissue samples and clinical, environmental factors as well as technical factors related to biospecimen collection. Thus, caution is necessary for biological interpretation in this aspect of the analysis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Multiset correlation and factor analysis
  - ○ Model initialization
  - ○ High dimensionality and selection of hyperparameters
  - ○ Calculating the variance explained
  - ○ Calculating relative feature importance
  - ○ SNP set enrichment analysis
  - ○ The MESA multi-omics pilot
  - ○ Cross-validation
  - ○ Comparison to MOFA2 and MMAE
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100359.

ACKNOWLEDGMENTS

B.C.B. would like to thank Lior Pachter and Nicholas Bray for numerous insightful conversations about CCA over the years. B.C.B. would also like to thank Andrew Stirn for reviewing the auto-encoder code. Funding for D.A.K. and B.C.B. is provided by NIA U01AG068880. Funding for B.C.B. is provided by NHGRI K99HG012373 and the Columbia Data Science Institute. Funding for T.L. and S.K. is provided by National Heart, Lung, and Blood Institute (NHLBI) R01HL142028. Funding for T.L. is provided by NIH R01AG057422 and NIMH R01MH106842. Funding for MESA lung measures is provided by NHLBI R01HL077612 and R01HL093081. WGS for the TOPMed program was supported by the NHLBI. WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity quality control (QC), and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1) and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). The MESA projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators (accession number phs000209.v13.p3). Support for the MESA projects are conducted and supported by the 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, UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. 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://www.mesa-nhlbi.org.

AUTHOR CONTRIBUTIONS

Conceptualization, B.C.B., D.A.K., and T.L.; methodology, B.C.B. and C.W.; software, B.C.B. and C.W.; validation, B.C.B.; formal analysis, B.C.B., C.W., and D.A.K.; investigation, B.C.B. and C.W.; resources, S.K., F.A., D.C.N., K.D.T., R.P.T., P.D., Y.L., W.C.J., D.V.D.B., N.G., S.G., J.D.S., R.G., C.C., Q.W., G.P., T.W.B., J.I.R., S.S.R., R.G.B., K.G.A., D.A.K., and T.L.; data curation, B.C.B., S.K., F.A., D.C.N., K.D.T., R.P.T., P.D., Y.L., W.C.J., D.V.D.B., N.G., S.G., J.D.S., R.G., C.C., Q.W., G.P., T.W.B., J.I.R., S.S.R., R.G.B., K.G.A., D.A.K., and T.L.; writing – original draft, B.C.B.; writing – review & editing, B.C.B., C.W., S.K., J.I.R., S.S.R., D.A.K., and T.L.; supervision, J.I.R., S.S.R., D.A.K., and T.L.; funding acquisition, B.C.B., R.G.B., D.A.K., and T.L.

DECLARATION OF INTERESTS

T.L. is a paid adviser or consultant of GSK, Pfizer, and Goldfinch Bio and has equity in Variant Bio. F.A. is an employee and shareholder of Illumina, Inc.

Received: September 28, 2022
Revised: April 26, 2023
Accepted: June 14, 2023
Published: July 10, 2023

REFERENCES

1. Krassowski, M., Das, V., Sahu, S.K., and Misra, B.B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front. Genet. *11*, 610798. https://doi.org/10.3389/FGENE.2020.610798.
2. Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol. *18*, 83. https://doi.org/10.1186/S13059-017-1215-1.
3. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873–1887.e17. https://doi.org/10.1016/j.cell.2019.05.006.
4. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124. https://doi.org/10.15252/msb.20178124.
5. Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika *28*, 321–377. https://doi.org/10.2307/2333955.
6. Brown, B.C., Bray, N.L., and Pachter, L. (2018). Expression reflects population structure. PLoS Genet. *14*, e1007841. https://doi.org/10.1371/journal.pgen.1007841.
7. Soneson, C., Lilljebjörn, H., Fioretos, T., and Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinformatics *11*, 1–20. https://doi.org/10.1186/1471-2105-11-191.
8. Naylor, M.G., Lin, X., Weiss, S.T., Raby, B.A., and Lange, C. (2010). Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants. PLoS One *5*, e10395. https://doi.org/10.1371/JOURNAL.PONE.0010395.
9. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions,

technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

10. Kettenring, J.R. (1971). Canonical analysis of several sets of variables. Biometrika *58*, 433–451.

11. Asendorf, N.A. (2015). Informative Data Fusion: Beyond Canonical Correlation Analysis. https://deepblue.lib.umich.edu/handle/2027.42/113419.

12. Bach, F.R., and Jordan, M.I. (2005). A Probabilistic Interpretation of Canonical Correlation Analysis. https://www.di.ens.fr/∼fbach/probacca.pdf.

13. Marčenko, V.A., and Pastur, L.A. (1967). Distribution of Eigenvalues for Some Sets of Random Matrices. Math. USSR. Sb. *1*, 457–483. https://doi.org/10.1070/sm1967v001n04abeh001994.

14. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881. https://doi.org/10.1093/AJE/KWF113.

15. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.

16. Mcinnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

17. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

18. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86. https://doi.org/10.1186/1471-2105-13-86.

19. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

20. Schaeffer, L., Gohlke, H., Müller, M., Heid, I.M., Palmer, L.J., Kompauer, I., Demmelmair, H., Illig, T., Koletzko, B., and Heinrich, J. (2006). Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. Hum. Mol. Genet. *15*, 1745–1756. https://doi.org/10.1093/HMG/DDL117.

21. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116. https://doi.org/10.1038/NG.520.

22. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415–1429.e19. https://doi.org/10.1016/J.CELL.2016.10.042.

23. Liu, J.Z., Van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986. https://doi.org/10.1038/NG.3359.

24. Rhee, E.P., Ho, J.E., Chen, M.H., Shen, D., Cheng, S., Larson, M.G., Ghorbani, A., Shi, X., Helenius, I.T., O'Donnell, C.J., et al. (2013). A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab. *18*, 130–143. https://doi.org/10.1016/J.CMET.2013.06.013.

25. Pollin, T.I., Damcott, C.M., Shen, H., Ott, S.H., Shelton, J., Horenstein, R.B., Post, W., McLenithan, J.C., Bielak, L.F., Peyser, P.A., et al. (2008). A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science *322*, 1702–1705. https://doi.org/10.1126/SCIENCE.1161524.

26. Richardson, T.G., Sanderson, E., Palmer, T.M., Ala-Korpela, M., Ference, B.A., Davey Smith, G., and Holmes, M.V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. *17*, e1003062. https://doi.org/10.1371/JOURNAL.PMED.1003062.

27. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194. https://doi.org/10.1038/s41588-020-00757-z.

28. Graham, S.E., Clarke, S.L., Wu, K.H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. https://doi.org/10.1038/S41586-021-04064-3.

29. Zenke, S., Palm, M.M., Braun, J., Gavrilov, A., Meiser, P., Böttcher, J.P., Beyersdorf, N., Ehl, S., Gerard, A., Lämmermann, T., et al. (2020). Quorum Regulation via Nested Antagonistic Feedback Circuits Mediated by the Receptors CD28 and CTLA-4 Confers Robustness to T Cell Population Dynamics. Immunity *52*, 313–327.e7. https://doi.org/10.1016/J.IMMUNI.2020.01.018.

30. Mittelstraß, K., and Waldenberger, M. (2018). DNA methylation in human lipid metabolism and related diseases. Curr. Opin. Lipidol. *29*, 116–124. https://doi.org/10.1097/MOL.0000000000000491.

31. Gomez-Alonso, M.D.C., Kretschmer, A., Wilson, R., Pfeiffer, L., Karhunen, V., Seppälä, I., Zhang, W., Mittelstraß, K., Wahl, S., Matias-Garcia, P.R., et al. (2021). DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures. Clin. Epigenetics *13*, 7. https://doi.org/10.1186/s13148-020-00957-8.

32. Gall, W.E., Beebe, K., Lawton, K.A., Adam, K.P., Mitchell, M.W., Nakhle, P.J., Ryals, J.A., Milburn, M.V., Nannipieri, M., Camastra, S., et al. (2010). α-Hydroxybutyrate Is an Early Biomarker of Insulin Resistance and Glucose Intolerance in a Nondiabetic Population. PLoS One *5*, e10883. https://doi.org/10.1371/JOURNAL.PONE.0010883.

33. Ferrannini, E., Natali, A., Camastra, S., Nannipieri, M., Mari, A., Adam, K.P., Milburn, M.V., Kastenmüller, G., Adamski, J., Tuomi, T., et al. (2013). Early Metabolic Markers of the Development of Dysglycemia and Type 2 Diabetes and Their Physiological Significance. Diabetes *62*, 1730–1737. https://doi.org/10.2337/DB12-0707.

34. Van Der Harst, P., and Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ. Res. *122*, 433–443. https://doi.org/10.1161/CIRCRESAHA.117.312086.

35. Aragam, K.G., Jiang, T., Goel, A., Kanoni, S., Wolford, B.N., Atri, D.S., Weeks, E.M., Wang, M., Hindy, G., Zhou, W., et al. (2022). Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nat. Genet. *54*, 1803–1815. https://doi.org/10.1038/S41588-022-01233-6.

36. Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.J., Butterworth, A.S., Howson, J.M.M., Assimes, T.L., Chowdhury, R., Orho-Melander, M., Damrauer, S., et al. (2017). Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat. Genet. *49*, 1450–1457. https://doi.org/10.1038/NG.3943.

37. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum. Mol. Genet. *28*, 166–174. https://doi.org/10.1093/HMG/DDY327.

38. Justice, A.E., Karaderi, T., Highland, H.M., Young, K.L., Graff, M., Lu, Y., Turcot, V., Auer, P.L., Fine, R.S., Guo, X., et al. (2019). Protein-coding variants implicate novel genes related to lipid homeostasis contributing to body-fat distribution. Nat. Genet. *51*, 452–469. https://doi.org/10.1038/s41588-018-0334-2.

39. Martinez-Huenchullan, S.F., Tam, C.S., Ban, L.A., Ehrenfeld-Slater, P., Mclennan, S.V., and Twigg, S.M. (2020). Skeletal muscle adiponectin induction in obesity and exercise. Metabolism *102*, 154008. https://doi.org/10.1016/j.metabol.2019.154008.

40. Dastani, Z., Hivert, M.F., Timpson, N., Perry, J.R.B., Yuan, X., Scott, R.A., Henneman, P., Heid, I.M., Kizer, J.R., Lyytikäinen, L.P., et al. (2012). Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. *8*, e1002607. https://doi.org/10.1371/JOURNAL.PGEN.1002607.

41. Zheng, J.S., Arnett, D.K., Lee, Y.C., Shen, J., Parnell, L.D., Smith, C.E., Richardson, K., Li, D., Borecki, I.B., Ordovás, J.M., et al. (2013). Genome-wide contribution of genotype by environment interaction to variation of diabetes-related traits. PLoS One *8*, e77442. https://doi.org/10.1371/JOURNAL.PONE.0077442.

42. Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet. *41*, 56–65. https://doi.org/10.1038/NG.291.

43. Henkel, N.D., Wu, X., O'Donovan, S.M., Devine, E.A., Jiron, J.M., Rowland, L.M., Sarnyai, Z., Ramsey, A.J., Wen, Z., Hahn, M.K., et al. (2022). Schizophrenia: a disorder of broken brain bioenergetics. Mol. Psychiatry *27*, 2393–2404. https://doi.org/10.1038/s41380-022-01494-x.

44. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, 1214–1231.e11. https://doi.org/10.1016/j.cell.2020.08.008.

45. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14. https://doi.org/10.1016/j.cell.2020.06.045.

46. Vinik, A.I., Erbas, T., Park, T.S., Nolan, R., and Pittenger, G.L. (2001). Platelet Dysfunction in Type 2 Diabetes. Diabetes Care *24*, 1476–1485. https://doi.org/10.2337/DIACARE.24.8.1476.

47. Brown, B.C., and Knowles, D.A. (2020). Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. Preprint at bioRxiv. https://doi.org/10.1101/2020.06.18.160176.

48. Brown, B.C., and Knowles, D.A. (2021). Welch-weighted Egger regression reduces false positives due to correlated pleiotropy in Mendelian randomization. Am. J. Hum. Genet. *108*, 2319–2335. https://doi.org/10.1016/J.AJHG.2021.10.006.

49. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. *39*, 1–22.

50. Parra, L.C. (2018). Multiset Canonical Correlation Analysis simply explained. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03759.

51. Witten, D.M., and Tibshirani, R.J. (2009). Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. Stat. Appl. Genet. Mol. Biol. *8*, Article28. https://doi.org/10.2202/1544-6115.1470.

52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Preprint at arXiv. https://doi.org/10.48550/arXiv.1201.0490.

53. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics, pp. 105–142.

54. Wu, D., and Smyth, G.K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. *40*, e133. https://doi.org/10.1093/nar/gks461.

55. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience *4*, 7–16. https://doi.org/10.1186/s13742-015-0047-8.

56. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

57. Kasela S., Aguet F., Kim-Hellmuth S., Brown B.C., Nachun D.C., Tracy R.P., Durda P., Liu Y., Taylor K.D., Johnson W.C., et al. Interaction molecular QTL mapping discovers cellular and environmental modifiers of genetic regulatory effects. bioRxiv 2022. doi:10.1101/2023.06.26.546528. https://www.biorxiv.org/content/10.1101/2023.06.26.546528v1

58. Stilp, A.M., Emery, L.S., Broome, J.G., Buth, E.J., Khan, A.T., Laurie, C.A., Wang, F.F., Wong, Q., Chen, D., D'Augustine, C.M., et al. (2021). A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. Am. J. Epidemiol. *190*, 1977–1992. https://doi.org/10.1093/aje/kwab115.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
MESA TOPMed multi-omics pilot data	dbGaP	dbGaP: phs001416.v3.p1
Software and algorithms		
Multiset Correlation and Factor Analysis	Zenodo	https://doi.org/10.5281/zenodo.7951370
MOFA2	github	https://github.com/bioFAM/MOFA2
eQTLgen *trans*-eQTL summary statistics	eQTLgen	https://www.eqtlgen.org/trans-eqtls.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Brielin Brown (bbrown@nygenome.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The MESA TOPMed multi-omics pilot data have been deposited on dbGap and are publicly available as of the date of publication. The accession number is listed in the key resources table. All original code has been deposited on zenodo and is publicly available as of the date of publication. The DOI is listed in the key resources table. The code is also available on github at https://github.com/collinwa/MCFA. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Multiset correlation and factor analysis

Let $Y = \{Y_m\}_{m=1}^{M}$ be a set of $N \times p_m$ observed data matrices: N individuals measured in M data modalities consisting of p_m features each. We model each observed mode as having contributions from two low-dimensional hidden factors (Figures 1A and S8)

$$z_n \sim N(0, I_d)$$

$$x_n^m \sim N(0, I_{k_m})$$

$$y_n^m \sim N\left(W_m z_n + L_m x_n^m, \Psi_m\right)$$

where d is the shared hidden dimensionality, k_m are the dataset-private hidden dimensionalities, W_m are $p_m \times d$ shared space loading matrices, L_m are $p_m \times k_m$ private space loading matrices and $\Psi_m = diag(\psi_m^1, \ldots, \psi_m^{p_m})$ are the diagonal residual covariance matrices. Given Y, d and k_m, our goal is to infer the hidden factors Z and X_m and loading matrices W_m and L_m. This can be accomplished using a straightforward application of expectation maximization (EM).[49] For a derivation of the EM update equations, as well as a more detailed exposition including the relationship to pCCA, factor analysis and other multiset CCA (MCCA) methods, see Methods S1. In practice, we center and scale all data variables. This is not strictly required, however it enables simple estimation of the number of PCs to include and simplifies explained variance calculations, see below.

Model initialization

An important aspect of EM optimization is choosing a good initialization. We benchmarked three approaches to initializing W: random initialization and two versions of MCCA that correspond to maximizing the sum of pairwise correlations with the average variance and average norm constraints. These MCCA formulations can be solved via simple eigendecompositions. We found that the sum of pairwise correlations with average variance constraint produced the best initial estimates (Figure S7). This can be solved with a simple two step procedure: 1) whiten each data matrix using the singular value decomposition (SVD), 2) perform a second SVD on the concatenated whitened data matrices[50]:

```
Input: $Y_1, \ldots, Y_M, d$
Result: $\widehat{\boldsymbol{W}} = [\boldsymbol{W}_{\boldsymbol{1}}^\top : \ldots : \boldsymbol{W}_{\boldsymbol{M}}^\top]^\top$
$U_{all} \leftarrow$ concatenate(SVD($Y_1$). $U, \ldots,$ SVD($Y_M$). $U$);
$\widehat{W} \leftarrow$ SVD($U_{all}$). $V[:, 0 : d]$;
$\widehat{\rho} \leftarrow$ SVD($U_{all}$). $\lambda[0 : d]$;
return $\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{\rho}}$
```

We initialize L and Ψ using probabilistic PCA on the residual data matrices after fitting MCCA. Specifically:

```
Input: $\boldsymbol{Y_i}, \boldsymbol{W_i}, \boldsymbol{N}, \boldsymbol{k_i}$
Result: $\widehat{\boldsymbol{L}}_{\boldsymbol{i}}, \widehat{\boldsymbol{\Psi}}_{\boldsymbol{i}}$
$\Sigma_{\boldsymbol{i}}^\perp \leftarrow \boldsymbol{Y_i}^\top \boldsymbol{Y_i}/\boldsymbol{N} - \boldsymbol{W_i}\boldsymbol{W_i}^\top$;
$\widehat{L}_i \leftarrow$ eigh($\Sigma_i^\perp$). $V[:, 0 : k_i]$;
$\sigma^2 \leftarrow$ mean(eigh($\Sigma_i^\perp$). $\lambda[k_i :]$);
$\widehat{\Psi}_i \leftarrow \sigma^2 1_{k_i}$;
return $\widehat{L}_i, \widehat{\Psi}_i$
```

High dimensionality and selection of hyperparameters

There are two primary approaches to control for over-fitting in applications of CCA-type methods to high-dimensional ($N \ll p$) problems. The first is to use penalized optimization techniques, where the objective function additionally contains an l_1 constraint on the weight matrices.[51] The second is to project each dataset onto its informative principal components.[6,7,11] In this application, we choose the latter approach in order to find components with broad effects on the structure of the data, rather than specific effects on small numbers of molecular features.[11] We choose the number of principal components of each dataset using the Marchenko-Pasteur law,[13] which states that for mean 0, variance 1 data, principal components with corresponding eigenvalues above $\lambda_m = 1 + \sqrt{p_m/N}$ should be considered non-noise. We are not aware of a corresponding law for the cross-covariance matrices used in CCA, however, the empirical spectral distribution of the cross-covariance of matrices of random noise can be easily estimated in practice:

```
Input: $N, k = \{k_m\}_{m=1}^M, n_{it}$
Result: $\rho$
for $it \leftarrow 0$ to $n_{it}$ do
    for $k_m \in k$ do
        $[Y_m]_{i=1, j=1}^{N, k_m} \sim N(0, 1)$;
    end
    $\rho[\boldsymbol{it}] \leftarrow$ max(InitializeMCFA($Y_1, \ldots, Y_M$). $\rho$
end
return mean($\rho$)
```

Then we keep all components where $\rho_{init} > \rho$.

Calculating the variance explained

The linear-Gaussian nature of the model simplifies estimation of the variance explained. That is, if the features of each mode $Y_m^{(:,j)}$ are normalized to variance 1, the model $Y_m^{(:,j)} = \sum_d W_m^{(j,d)} Z^{(:,d)} + \sum_{k_m} L_m^{(j,k_m)} X_m^{(:,k_m)} + \epsilon$ implies that the variance in feature j of mode m explained by shared factor d is $W_m^{(j,d)2}$. Likewise, the variance explained by the k_m-th private factor of mode m is $L_m^{(j,k_m)2}$. The total variance in mode m explained by a given shared factor d (respectively, private factor k_m) is thus given by $\sum_j W_m^{(j,d)2}$ (respectively, $\sum_j L_m^{(j,k_m)2}$), and the total variance in the mode explained by the factors are $\sum_{j,d} W_m^{(j,d)2}$ and $\sum_{j,k_m} L_m^{(j,k_m)2}$, respectively. Note that when working in PC-space, the raw W and L features correspond to variance in PCs explained, rather than modality features. Thus, we calculate the variance explained after projecting back into the original feature space $W_m \leftarrow V_m W_m, L_m \leftarrow V_m L_m$ where V_m are the right singular vectors of mode m.

To calculate the variance in a metadata feature explained by a particular space, we regressed the trait value T on the shared or private space, $T \sim Z$ or $T \sim X_m$. For continuous-valued traits we used linear regression as implemented in SciKitLearn v1.0 `linear_model.LinearRegression` and report the coefficient of determination.[52] For discrete-valued traits, we used multinomial logistic regression as implemented in SciKitLearn v1.0 `linear_model.LogisticRegression`.[52] We fit two models: a null model including only intercept or intercept and site, and one including the factor variables. We report the variance explained as the McFadden pseudo-R^2, $1 - \frac{ll_{alt}}{ll_{null}}$, with ll_{null} and ll_{alt} being the model negative log likelihood for the null and alternative model respectively.[53]

Calculating relative feature importance

Feature importance in traditional CCA is defined by the correlation of the variables in the reduced space $\rho = cor(Y_1 f_1, Y_2 f_2)$. Unfortunately this notion breaks down in higher dimensions. As we discuss further in Methods S1, the degree of sharing in MCCA is defined by functions of the cross-correlation matrix in the reduced space,

$$S = cor(Y_1 f_1, \ldots, Y_m f_m) \in R^{m \times m}.$$

We seek to define an analogous quantity for our graphical model. In MCFA, the data in the reduced (shared) space is given by the posterior mean of Z, $\widehat{Z} = E[Z|W, \Psi, L, Y] = Y(WW^\top + LL^\top \Psi)^{-1} W$. We can also calculate the posterior mean of Z conditional on observing a single mode, $\widehat{Z}_m = E[Z|W_m, \Psi_m, L_m, Y_m] = Y_m(W_m W_m^\top + L_m L_m^\top \Psi_m)^{-1} W_m$. This latter quantity is analogous to the reduced variables $Y_m f_m$ in MCCA. Thus we can summarize the importance of each dimension of the shared space by calculating functions of the cross-correlation of columns of $\widehat{Z}_m$,

$$S_d = cor\left(\widehat{Z}_1^{(:,d)}, \ldots, \widehat{Z}_m^{(:,d)}\right).$$

The relevant function in our model is the generalized variance $|S|$, see Methods S1. The determinant of a correlation matrix is bounded between 0 and 1, with lower values indicating more correlation, and higher values less. Thus to aid interpretability, we report $\rho_d = -log|S_d|$ and reorder columns of Z and W with decreasing ρ_d.

SNP set enrichment analysis

For SNP set enrichment analysis, we broadly follow the approach of CAMERA.[54] In brief, enrichment statistics can be inflated due to correlations in the sample - in this case, linkage disequilibrium between two GWAS SNPs. This results in an under-estimate of the standard error of the enrichment test statistic and an increase in false positives. We calculate the variance inflation factor by using plink v 1.9[55] to estimate linkage disequilibrium between annotation SNPs in 337, 781 unrelated individuals from the UK Biobank.[56] The variance inflation factor is $\nu = 1 + (p_A - 1)\rho_{\not{A}}$, with $\rho_{\not{A}}$ the average person correlation between features in set A. We test the known GWAS mean χ^2 statistic $h_0 : \chi_A^2 = 1$ against the alternative $h_1 : \chi_A^2 > 1$. The standard error of the test statistic is $\sigma_t = \sigma\sqrt{\frac{\nu}{p_A} - \frac{1}{p_m}}$ with σ the pooled empirical standard deviation of the test statistics.

The MESA multi-omics pilot

The Multi-Ethnic Study of Atherosclerosis (MESA) is a prospective cohort study with the goal to identify progression of subclinical atherosclerosis.[14] MESA recruited 6,814 participants, ages 45–84 years and free of clinical cardiovascular disease, during 2000–2002. The participants are 53% female, 38% non-Hispanic white, 28% Black, 22% Hispanic and 12% Asian-American. The Multi-Omics pilot dataset includes 30x whole genome sequencing (WGS) through the Trans-Omics for Precision Medicine (TOPMed) Project.[15]

Blood samples for multi-omic analysis of participants were collected at two time points (exam 1 and exam 5). RNA expression was profiled using poly-A RNA sequencing of PBMCs, and methylation was quantified by the Illumina 750K EPIC array in whole blood. The levels of 1,305 proteins were measured from plasma samples using the standard SOMAscan DNA aptamer–based platform, and metabolite levels were determined from targeted and untargeted mass spectrometry of blood plasma. The MESA Multi-Omics pilot biospecimen collection, molecular phenotype data production and quality control (QC) are described in detail in Kasela et al.[57]

Cross-validation

We used leave-one-out cross-validation (CV) to evaluate our model. The primary reason we chose leave-one-out CV over k-fold CV is that our hyperparameter selection method depends on the sample size. With $n - 1$ individuals, the same parameters used for the full inference procedure are likely to be valid. For small k, fitting with $\frac{k-1}{k}n$ individuals while using the same number of PCs may result in over-fitting in the training set, and using a smaller number of PCs may not capture the same variation as the full model.

To perform cross-validation we hold out a set of individuals, fit the MCFA model, then project the held out individuals into the learned space. If W_{tr}, L_{tr} and Φ_{tr} are the model parameters learned from the training set, the projections of the test data into the learned spaces are given by

$$\widehat{Z}_{te} = Y_{te}\left(W_{tr} W_{tr}^\top + L_{tr} L_{tr}^\top \Psi_{tr}\right)^{-1} W_{tr} \; \widehat{X}_{te} = Y_{te}\left(W_{tr} W_{tr}^\top + L_{tr} L_{tr}^\top \Psi_{tr}\right)^{-1} L_{tr}$$

The full data reconstruction is

$$\widehat{Y}_{te} = \widehat{Z}_{te} W_{tr}^\top + \widehat{X}_{te} L_{tr}^\top$$

We evaluate model fit by calculating the normalized root mean squared error (NRMSE). In order to provide a fair evaluation across modes with a highly variable number of features, we calculate NRMSE on a per mode basis

$$NRMSE = \sqrt{\frac{1}{p_m}\sum_{i=1}^{p_m}\frac{\left(Y_m^{(:,i)} - \widehat{Y}_m^{(:,i)}\right)^2}{varY_m^{(:,i)}}}$$

and potential over-fitting can be assessed by comparing the median training set NRMSE against the median test set NRMSE over many cross-validation iterations.

Comparison to MOFA2 and MMAE

We installed MOFA2 version 0.6.7 using `pip install mofapy2`. We used the options `scale_groups = False`, `scale_views = False`, `ard_weights = True` and `spikeslab_weights = True`. We set the convergence tolerance to convergence_mode = 'medium'. For comparison purposes we set the number of factors equal to the hidden dimensionality inferred by `MCFA (factors = 14)`.

Our multi-modal auto-encoder architecture is visualized in Figure S4. We used two hidden layes per dataset, with the first layer having dimensionality equal to 8 times that modalities MCFA-inferred number of PCs, and the second layer having dimensionality equal to that modalities MCFA-inferred number of PCs. These layers are then concatenated, and sent through an additional hidden layer with 8 times the MCFA-inferred number of shared dimensions to the final 14-dimensional encoded representation. All layers except the final encoder layer consist of a linear transform followed by ReLU activation, while the final encoder layer omits the ReLU activation. The decoder had identical architecture to the encoder only reversed. The network was implemented in `pytorch` v1.11.0 and optimized with `Adagrad` using 10 batches per epoch until the NRMSE change relative to the total loss was less than 10^{-6}.

QUANTIFICATION AND STATISTICAL ANALYSIS

We analyzed individuals from Exam 1 where all five data types were collected and passed QC. All data modalities were inverse rank normalized prior to sample filtering based on the availability of other data types. There were 614 individuals with observations of WGS, RNA-seq, methylation, metabolomics and proteomics that all pass QC. We further removed all features (CpGs, genes, proteins) located on sex-chromosomes, 0-variance features, CpGs with missing data, and CpGs where the probe was within 5 bases of an SNP, leaving us with $6,042$ metabolites, $1,222$ proteins, $19,034$ genes, and $724,210$ CpGs. We analyzed 28 PCs of RNA expression, 39 PCs of methylation, 27 PCs of protein expression and 63 PCs of metabolite, as determined using the aforementioned method. For sample metadata, we leveraged the rich phenotype data available in MESA that were harmonized by the TOPMed Data Coordinating Center.[58] For details on the estimation of sample cell-type proportions from methylation and RNA-seq data, see Kasela et al.[57] Genetic association analyses were conducted using `plink` v 1.9[55] while controlling for site, age, sex and 11 genotype PCs; reported *p*-values are uncorrected and tested against a null of 0 effect. SNP set enrichment significance was defined as having an FDR *q*-value below 0.05 when corrected for 3 tested sets across 14 factors tested against the null hypothesis that the mean χ^2 test statistic is 1.

Cell Genomics, Volume *3*

Supplemental information

Multiset correlation and factor analysis

enables exploration of multi-omics data

Brielin C. Brown, Collin Wang, Silva Kasela, François Aguet, Daniel C. Nachun, Kent D. Taylor, Russell P. Tracy, Peter Durda, Yongmei Liu, W. Craig Johnson, David Van Den Berg, Namrata Gupta, Stacy Gabriel, Joshua D. Smith, Robert Gerzsten, Clary Clish, Quenna Wong, George Papanicolau, Thomas W. Blackwell, Jerome I. Rotter, Stephen S. Rich, R. Graham Barr, Kristin G. Ardlie, David A. Knowles, and Tuuli Lappalainen

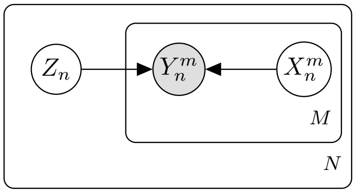


Figure S1: Multset correlation and factor analysis as a plate diagram, related to Figure 1. For each individual n (the outer plate), we sample z_n from the d-dimensional unit Gaussian, $z_n \sim \mathcal{N}(0, I_d)$. For each mode m (the inner plate) and individual n we sample x_n^m from a k_m-dimensional unit Gaussian. The observed features of that mode are then sampled from a p_m-dimensional unit Gaussian, $y_n^m \sim \mathcal{N}(W_m z_n + L_m x_n^m, \Psi_m)$. Here $W_m \in \mathbb{R}^{p_m \times d}$ is the transformation weight matrix, L_m is the private-space loadings matrix, and $\Psi_m = \mathrm{diag}(\Psi_m^1, \ldots, \Psi_m^{p_m})$ is the diagonal residual covariance matrix.

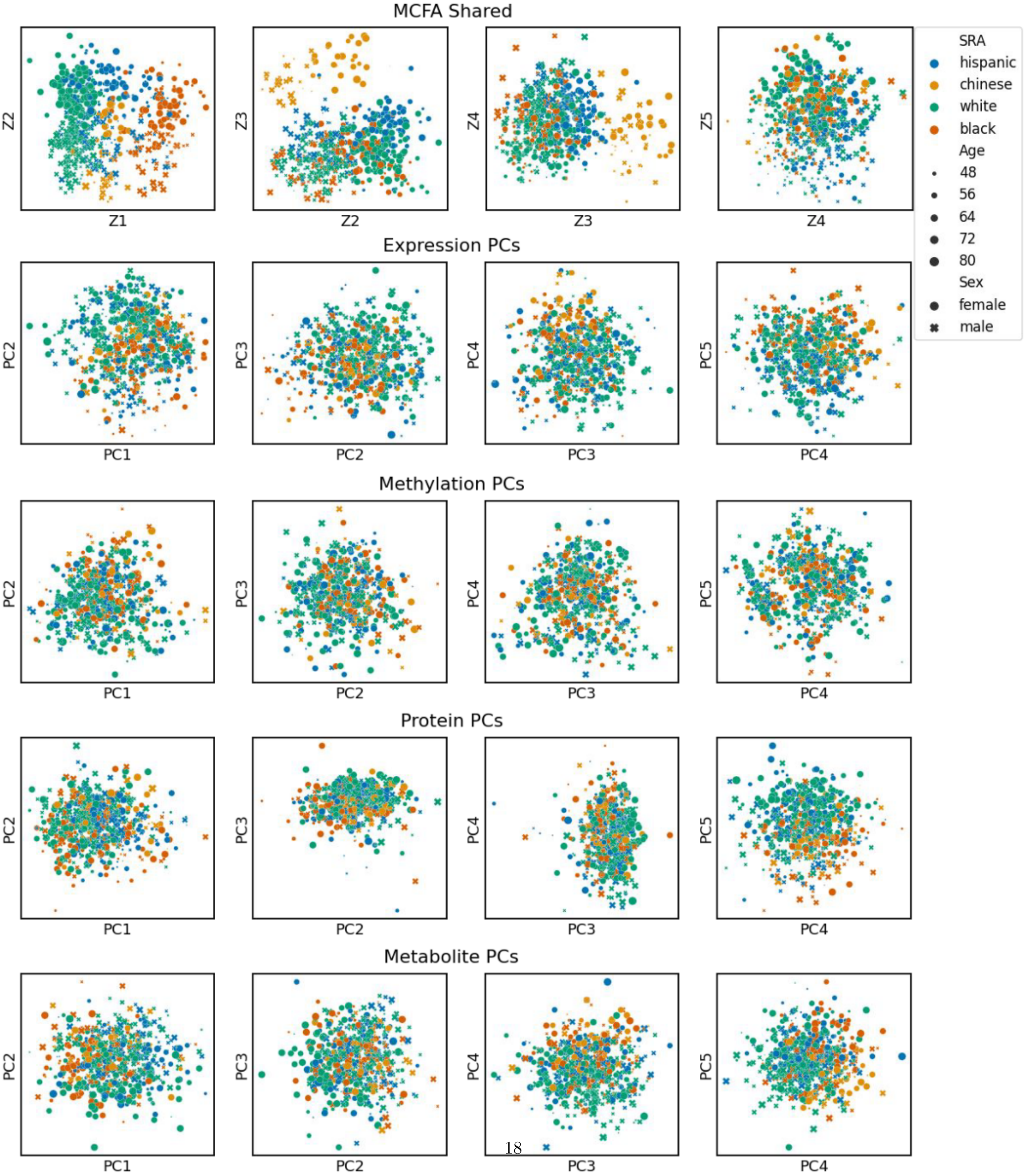


Figure S2: The top shared components learned from the MCFA model and top PCs of individual datasets, related to Figure 1. The MCFA factors clearly reflect self-reported ancestry (SRA), age and sex, while none of the top PCs of any of the datasets show this structure.

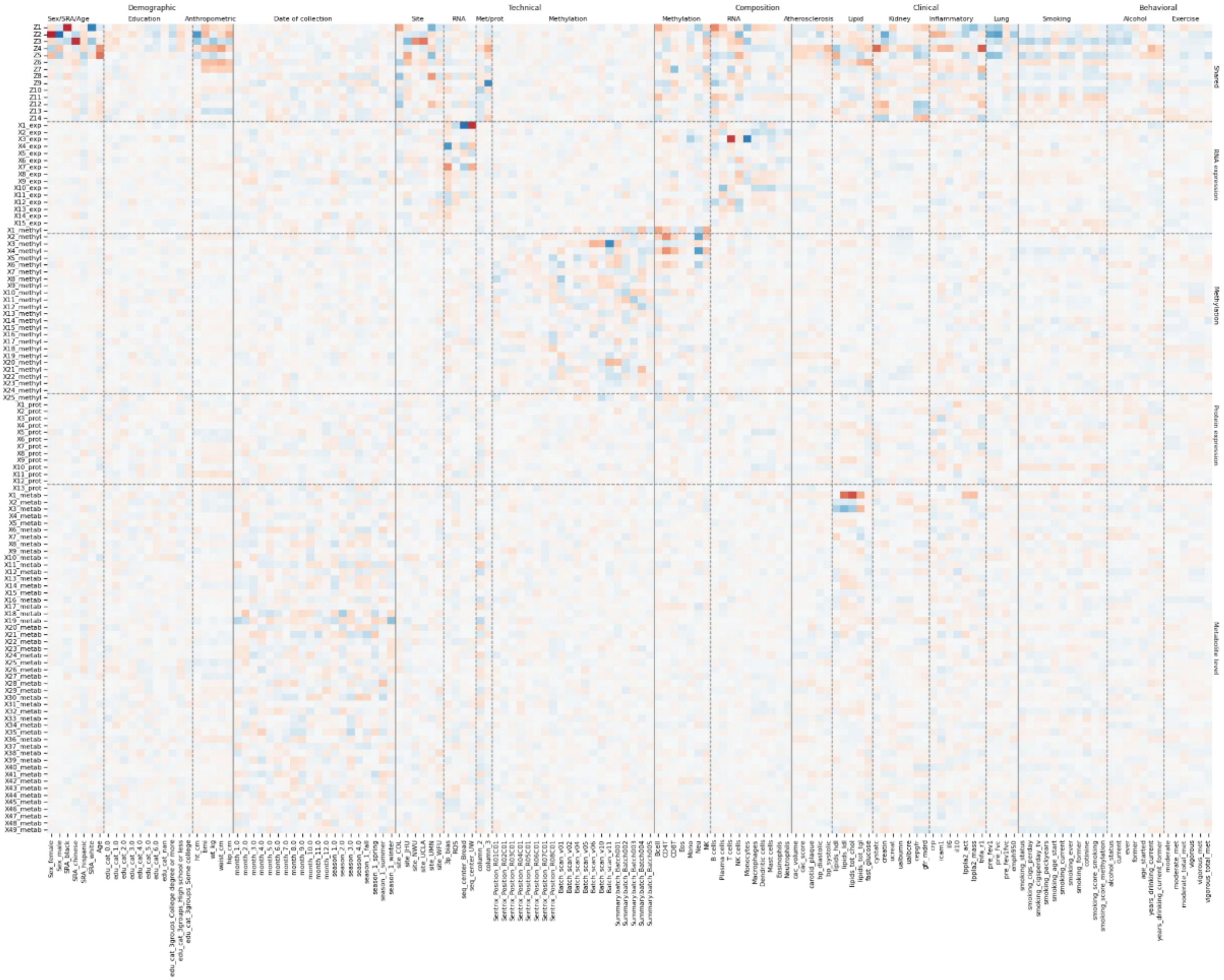


Figure S3: Correlation of each dimension of each learned space (rows) with each metadata factor (columns), related to Figure 1. Red values indicate positive correlation and blue values negative correlation.

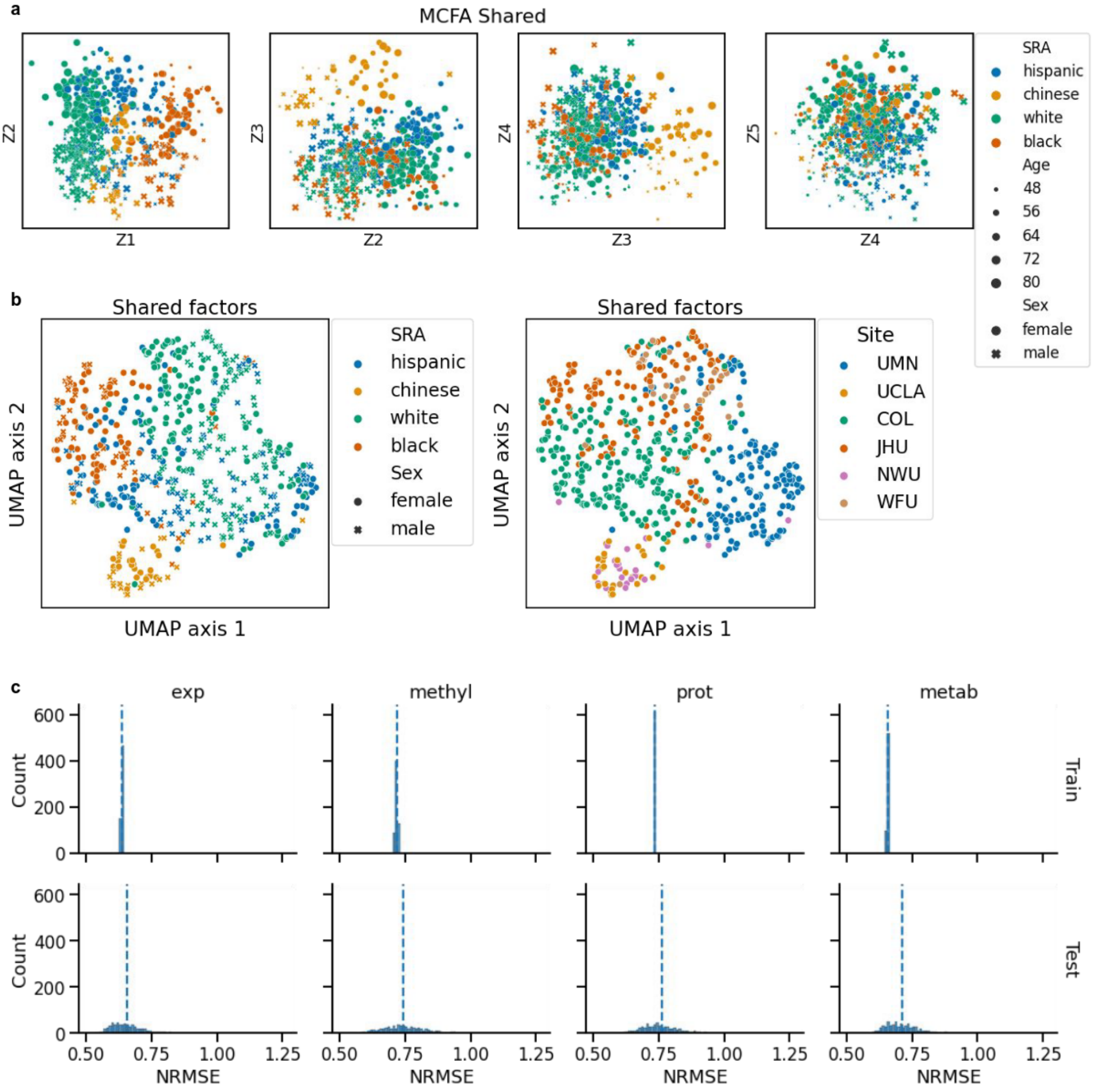


Figure S4: Cross-validation analysis of the MCFA feature space, related to Figure 1. a) The top shared cross-validated MCFA components, annotated by self-reported ancestry (SRA), age and sex. Each point is creating by holding that individual out, fitting MCFA on the remaining individuals, then projecting the held-out individual into the shared space (see Online Methods). b) UMAP embeddings of the cross-validated MCFA components. c) Normalized root mean square error of the 613 training individuals for each dataset (top), versus the held-out individual (bottom), split by data type. Blue dashed line indicates the median.

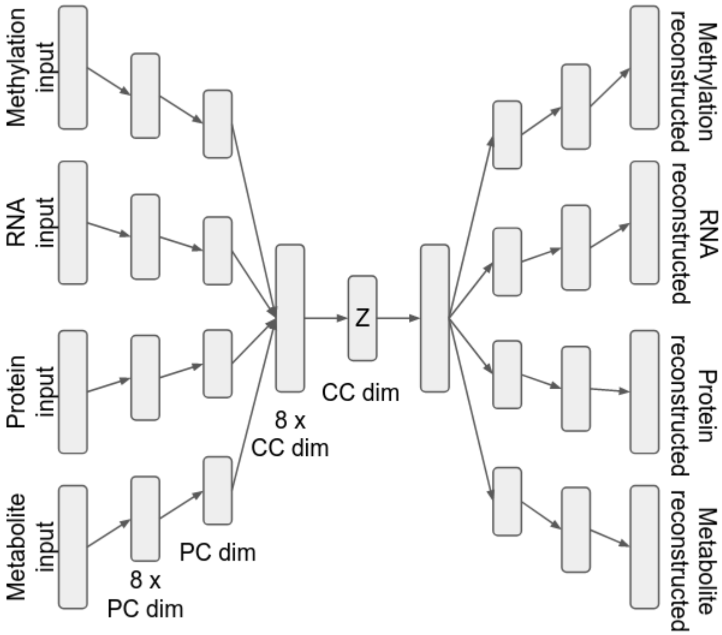


Figure S5: The multi-modal auto-encoder architecture, related to Figure 2.

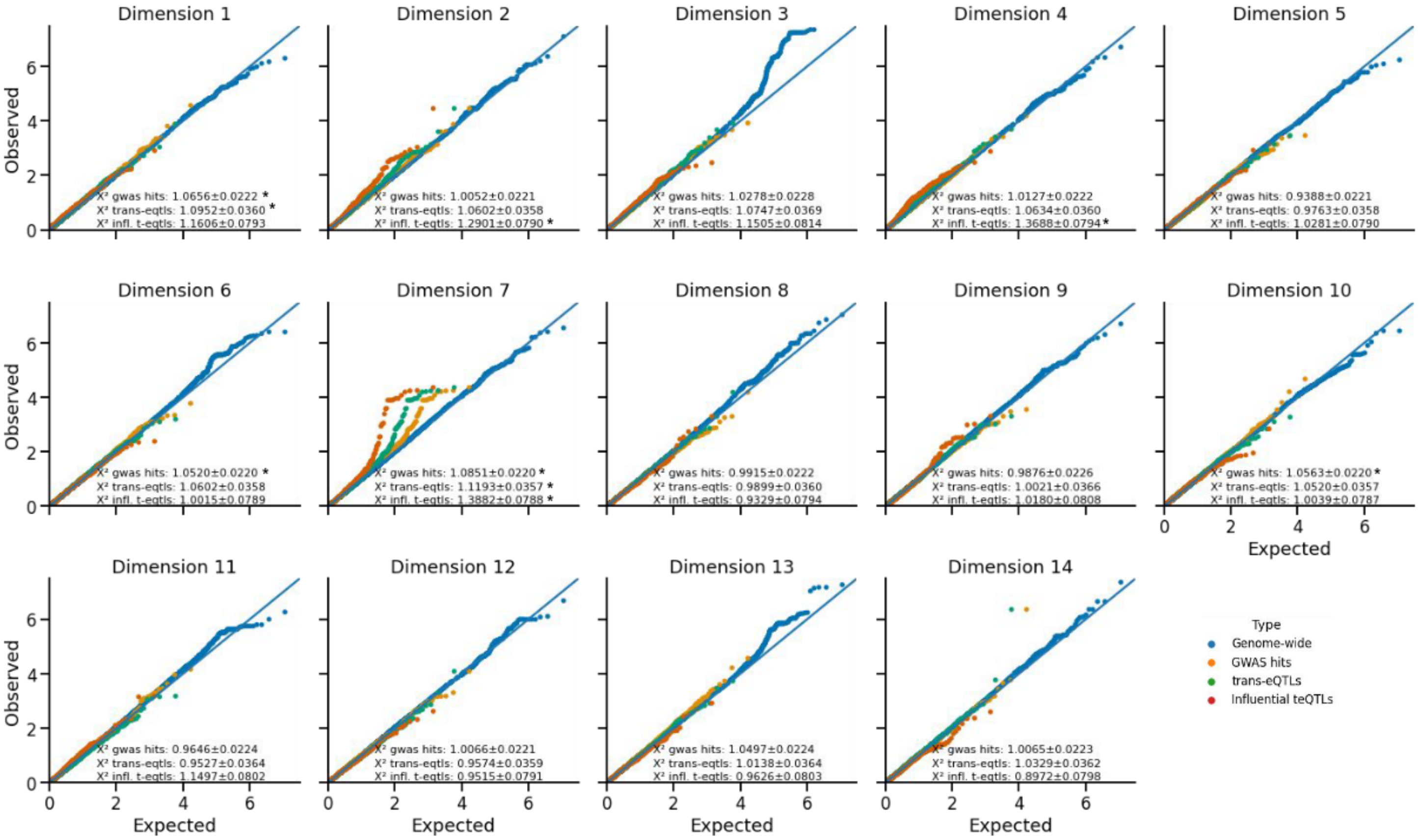


Figure S6: Q-Q plot of GWAS results for each shared factor, related to Figure 3.

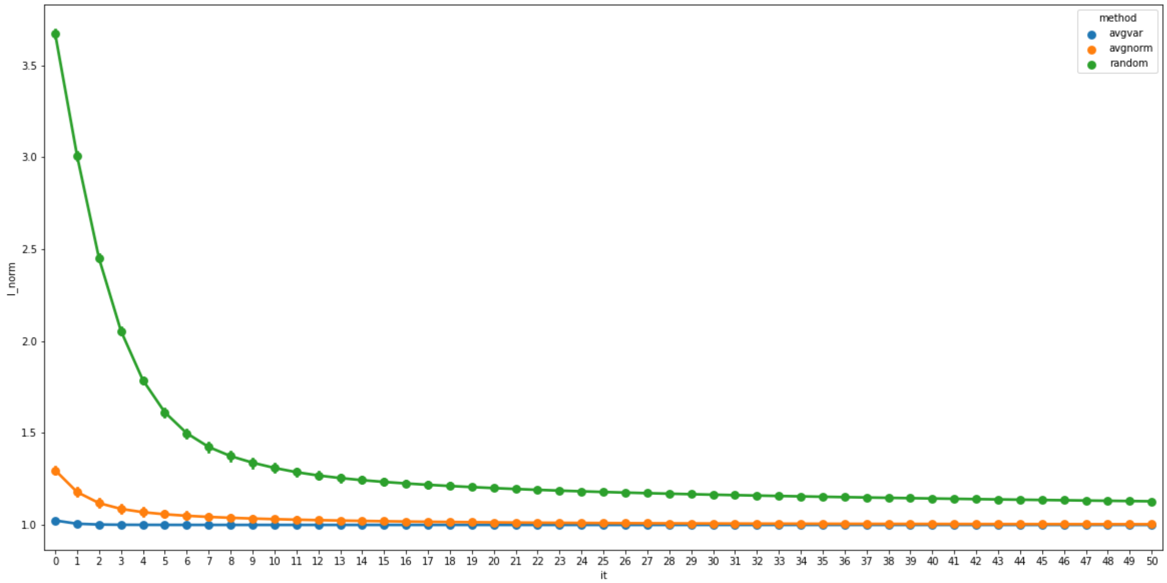


Figure S7: Average normalized model likelihood (y-axis, negative log-likelihood divided by minimum negative log-likelihood) as a function of EM step iteration, related to the STAR Methods. We simulated three datasets with $p = 30, 40, 50$ observed features generated by $k_m = 8, 11, 15$ private and $d = 10$ shared factors and $N = 1000$ individuals in 100 simulations. We compared random, SUMCORR-AVGVAR, and SUMCORR-AVGNORM model initializers. SUMCORR-AVGVAR produced good initial estimates resulting in fast convergence, while other approaches took longer to converge.

Trait(s)	PMID(s)	rs ID(s)	Chr:Pos(s)	F7 p val(s)
Trans fatty acid levels, Red blood cell fatty acid levels, Metabolite levels	25646338, 25500335, 23378610	rs174541, rs174549, rs174556, rs174555, rs1535, rs174536, rs174537, rs174547, rs174546, rs174545, rs174550, rs174548, rs174583, rs2727270, rs2727271, rs174535, rs174601, rs174534, rs108499, rs174576, rs174538, rs174528, rs174574, rs102275, rs509360, rs2072114, rs2845573	11:61565908, 11:61571382, 11:61580635, 11:61579760, 11:61597972, 11:61551927, 11:61552680, 11:61570783, 11:61569830, 11:61569306, 11:61571478, 11:61571348, 11:61609750, 11:61603237, 11:61603358, 11:61551356, 11:61623140, 11:61549458, 11:61547237, 11:61603510, 11:61560081, 11:61543499, 11:61600342, 11:61557803, 11:61548559, 11:61605215, 11:61601908	4e-05, 6e-05, 6e-05, 6e-05, 8e-05, 0.00011, 0.00011, 0.00013, 0.00013, 0.00013, 0.00013, 0.00025, 0.00041, 0.00077, 0.00077, 0.00088, 0.00094, 0.00101, 0.00118, 0.00276, 0.00285, 0.00476, 0.00567, 0.00872, 0.00945, 0.0098, 0.00981
Glycerophospholipid levels, dihomo-gamma-linolenic acid levels	26068415, 24823311	rs174555, rs1535, rs174537, rs174536, rs174547, rs174546, rs174550, rs174576, rs102275	11:61579760, 11:61597972, 11:61552680, 11:61551927, 11:61570783, 11:61569830, 11:61571478, 11:61603510, 11:61557803	6e-05, 8e-05, 0.00011, 0.00011, 0.00013, 0.00013, 0.00013, 0.00276, 0.00872
Blood metabolite levels	24816252	rs174556, rs174550, rs174548, rs2727271, rs174535, rs174601, rs651007, rs174538, rs7642243	11:61580635, 11:61571478, 11:61571348, 11:61603358, 11:61551356, 11:61623140, 9:136153875, 11:61560081, 3:195941216	6e-05, 0.00013, 0.00025, 0.00077, 0.00088, 0.00094, 0.00148, 0.00285, 0.00552
Height	25429064	rs7184046, rs174547, rs7155279, rs8007661, rs3738814, rs7153027, rs7154721, rs7158300	15:75866150, 11:61570783, 14:92485881, 14:92459958, 1:17331676, 14:92427222, 14:92427348, 14:92482948	6e-05, 0.00013, 0.00141, 0.0037, 0.00371, 0.00447, 0.0045, 0.00883
Heart rate, Laryngeal squamous cell carcinoma	23583979, 25194280	rs174549	11:61571382	6e-05
Phospholipid levels (plasma), Inflammatory bowel disease	21829377, 26192919	rs1535, rs174536, rs174547, rs174550, rs174535, rs174538, rs174574, rs102275	11:61597972, 11:61551927, 11:61570783, 11:61571478, 11:61551356, 11:61560081, 11:61600342, 11:61557803	8e-05, 0.00011, 0.00013, 0.00013, 0.00088, 0.00285, 0.00567, 0.00872
Colorectal cancer, gamma-linolenic acid levels, Crohns disease	24836286, 26584805, 26192919	rs174537	11:61552680	0.00011
Cholesterol, total	25961943	rs174554, rs174546, rs174570, rs579459, rs7525649, rs635634, rs1532085	11:61579463, 11:61569830, 11:61597212, 9:136154168, 1:55499156, 9:136155000, 15:58683366	0.00012, 0.00013, 0.00125, 0.00148, 0.00222, 0.0053, 0.00634
linoleic acid levels, Resting heart rate, Metabolic traits, arachidonic acid levels, Sphingolipid levels, Lipid metabolism phenotypes, HDL cholesterol, Triglycerides	26584805, 20639392, 21886157, 24823311, 26068415, 22286219, 19060906, 19060906	rs174547, rs174550, rs2727270, rs11827215	11:61570783, 11:61571478, 11:61603237, 11:61458595	0.00013, 0.00013, 0.00077, 0.00817
adrenic acid, Fasting glucose-related traits (interaction with BMI), Fasting glucose-related traits	24823311, 22581228, 20081858	rs174550	11:61571478	0.00013
Delta-6 desaturase activity	26584805	rs174545, rs174548	11:61569306, 11:61571348	0.00013, 0.00025
LDL cholesterol	24097068	rs174546, rs174570, rs579459, rs7525649, rs11206510, rs41279633, rs635634	11:61569830, 11:61597212, 9:136154168, 1:55499156, 1:55496039, 7:44580876, 9:136155000	0.00013, 0.00125, 0.00148, 0.00222, 0.00325, 0.00468, 0.0053
Hematology traits, Blood metabolite ratios	23303382, 24816252	rs174548	11:61571348	0.00025
Systemic lupus erythematosus	23273568	rs4852324, rs10911628, rs2286672	2:74202578, 1:184649503, 17:4712617	0.00034, 0.00469, 0.00877
QT interval	24952745	rs174583, rs4657178, rs4784934, rs2968863	11:61609750, 1:162210610, 16:58459926, 7:150623137	0.00041, 0.00629, 0.00655, 0.0098

Continued on the next page

Trait(s)	PMID(s)	rs ID(s)	Chr:Pos(s)	F7 p val(s)
Mature red cell;HGB	27863252	rs1256061	14:64703593	0.00046
IgG glycosylation	23382691	rs2186369	22:24170996	0.0005
Liver enzyme levels (alkaline phosphatase)	22001757	rs174601, rs579459	11:61623140, 9:136154168	0.00094, 0.00148
Mean platelet volume	27863252	rs7743045, rs1716505, rs471756	6:119102271, 12:65005079, 9:239313	0.00114, 0.0046, 0.00909
Polycystic ovary syndrome	22885925	rs2059807	19:7166109	0.00129
Proinsulin levels	21873549	rs11603334	11:72432985	0.00142
Schizophrenia	19571811	rs13194053, rs12908161, rs6932590, rs12823424, rs6878284	6:27143883, 15:85207825, 6:27248931, 12:2514112, 5:101769726	0.00144, 0.00309, 0.00492, 0.00661, 0.00822
Coronary artery disease or ischemic stroke, Soluble levels of adhesion molecules, Coronary heart disease, Coronary artery disease or large artery stroke, Urinary metabolites (H-NMR features), Red blood cell traits, Soluble E-selectin levels	24262325, 20167578, 21378990, 24262325, 24586186, 23222517, 19729612	rs579459	9:136154168	0.00148
Serum alkaline phosphatase levels, Iron status biomarkers (ferritin levels), End-stage coagulation, E-selectin levels	24094242, 25352340, 23381943, 20147318	rs651007	9:136153875	0.00148
PR segment	24850809	rs10850409	12:115381740	0.00158
Type 2 diabetes	20581827	rs7578326, rs1552224, rs1387153	2:227020653, 11:72433098, 11:92673828	0.00164, 0.0074, 0.00881
Bone mineral density	24249740	rs227425	14:70456699	0.00201
Red blood cell count, Hematological and biochemical traits, Venous thromboembolism, Angiotensin-converting enzyme activity	20139978, 20139978, 22672568, 20066004	rs495828	9:136154867	0.00232
Psoriasis	25903422	rs7769061, rs28512356	6:111926909, 3:189615475	0.00249, 0.00812
Rheumatoid arthritis	Curated from Immunobase	rs4452313	3:17047032	0.00265
Uterine fibroids	21460842	rs12484776	22:40652873	0.00304
Coronary artery disease, Myocardial infarction (early onset)	26343387, 19198609	rs11206510	1:55496039	0.00325
Epithelial ovarian cancer, Ovarian cancer	25581431, 23535730	rs7651446	3:156406997	0.00327
Myeloid white cell;MYELOID#	27863252	rs4844622, rs7575217	1:208034329, 2:101776932	0.00338, 0.00995
Digestive system disease (Barretts esophagus and esophageal adenocarcinoma combined)	24121790	rs2687201	3:70928930	0.00389
Central corneal thickness	22814818	rs3132306	9:137440212	0.00392
Cholangitis, Primary Sclerosing, Celiac disease	Curated from Immunobase, Curated from Immunobase	rs7426056	2:204612058	0.00406
Autism	24189344	rs6537825	1:114948281	0.00429
Dupuytrens disease	21732829	rs8124695	20:39028436	0.00447
Corneal structure	23291589	rs1536482	9:137440528	0.00464
PR interval	20062060	rs1896312	12:115346424	0.00476
Immature red cell;IRF	27863252	rs59918340	8:142232256	0.00492
Obesity-related traits	23251661	rs12104221	19:3797100	0.00503
Allergic sensitization, IgE and Allergic Sensitization	23817571, Curated from Immunobase	rs9865818	3:188072513	0.0055
Select biomarker traits	17903293	rs2494250	1:159278251	0.00594
Multiple sclerosis, Liver Cirrhosis, Biliary	Curated from Immunobase, Curated from Immunobase	rs9736016	11:118724894	0.00605
Dengue shock syndrome	22001756	rs3132468	6:31475486	0.00627
Metabolic syndrome	22399527	rs1532085	15:58683366	0.00634

Continued on the next page

Trait(s)	PMID(s)	rs ID(s)	Chr:Pos(s)	F7 p val(s)
Joint damage progression in ACPA-negative rheumatoid arthritis	26077402	rs2833522	21:33179371	0.00641
Prostate cancer	21743467	rs2242652	5:1280028	0.00641
Ulcerative colitis	26192919	rs11641184	16:11704651	0.00657
Post bronchodilator FEV1/FVC ratio	26634245	rs62346060	4:145469373	0.0066
Amyotrophic lateral sclerosis	23624525	rs6703183	1:209712889	0.00673
Menarche (age at onset)	25231870	rs7104764	11:229977	0.00678
Bladder cancer	24163127	rs710521	3:189645933	0.00686
Orthostatic hypotension	24124408	rs6736587	2:81855725	0.00687
Exfoliation syndrome	25706626	rs4926244	19:13374913	0.00694
Asthma (childhood onset)	22560479	rs9815663	3:3614887	0.0072
Monocyte count	25096241	rs1991866	8:130624105	0.0079
Body mass index	25673413	rs16951275	15:68077168	0.00801
Age-related hearing impairment (interaction)	24939585	rs727809	5:152610222	0.00818
IgA nephropathy	26028593	rs2074038, rs2738058	11:44087989, 8:6821617	0.00824, 0.00946
Mature red cell;HCT	27863252	rs17476364	10:71094504	0.00842
Myeloid white cell;BASO#	27863252	rs16823866	2:145324977	0.0085
Allergic rhinitis	25085501	rs6583203	3:197079586	0.0087
Oleic acid (18:1n-9) plasma levels, Stearic acid (18:0) plasma levels, Palmitoleic acid (16:1n-7) plasma levels	23362303, 23362303, 23362303	rs102275	11:61557803	0.00872
Optic cup area, Vertical cup-disc ratio	25631615, 25241763	rs5756813	22:38175477	0.00877
Fasting plasma glucose, Glycated hemoglobin levels, Metabolic syndrome (bivariate traits)	19060909, 20858683, 21386085	rs1387153	11:92673828	0.00881
Cognitive function	25644384	rs17522122	14:33302882	0.00936
Cortical thickness	21810643	rs4906844	15:26277545	0.00938
alphalinolenic acid	26584805	rs509360	11:61548559	0.00945
Mature red cell;MCH	27863252	rs13231886	7:44814172	0.0097

Table S4: GWAS hits with nominal *p*-value for association with factor 7 below 0.01, related to Figure 3

Trait(s)	PMID(s)	rs ID(s)	Chr:Pos(s)	F6 p val(s)
Schizophrenia	26198764	rs112973353, rs11874716, rs4801131, rs12966547, rs10425465, rs11682175, rs4129585	14:104537680, 18:52750688, 18:52752700, 18:52752017, 19:33897934, 2:57987593, 8:143312933	0.00017, 0.00604, 0.00604, 0.00604, 0.00634, 0.00749, 0.0085
Age-related macular degeneration	15761122	rs380390, rs1061147, rs1329428, rs10801555, rs1329424, rs1410996	1:196701051, 1:196654324, 1:196702810, 1:196660261, 1:196646176, 1:196696933	0.00045, 0.00073, 0.0012, 0.00276, 0.00446, 0.00691
Vitiligo	Immunobase	rs3814231	10:115481018	0.00047
Febrile seizures, Febrile seizures (MMR vaccine-related)	25344690, 25344690	rs273259	1:79093818	0.00059
Adiponectin levels	22479202	rs4805885	19:33906123	0.00063
Menopause (age at onset)	22267201	rs2517388	8:37977732	0.00078
Mature red cell;MCHC	27863252	rs12451471, rs4238686	17:78102517, 16:88788934	0.0008, 0.00171
Refractive error	23396134	rs1656404	2:233379941	0.00091
Platelet;PDW	27863252	rs13224082, rs74142329	7:116515781, 10:80949988	0.00092, 0.00621
Alzheimers disease (late onset)	26339675	rs75002042	5:15669967	0.00094
Common traits (Other)	20585627	rs17646946	1:152062767	0.001
Protein C levels, Coagulation factor levels, Factor VII levels, Anticoagulant levels	25376901, 20231535, 20231535, 22216198	rs867186	20:33764554	0.00105
Mean platelet volume	27863252	rs11121529, rs6446731, rs11043280, rs17396340, rs6762	1:10271688, 4:3284751, 12:122426643, 1:10286176, 11:838722	0.00117, 0.00406, 0.00449, 0.00785, 0.00809
Menarche (age at onset)	25231870	rs2137289, rs852069	18:44752125, 20:17122593	0.0012, 0.0094
Homeostasis model assessment of insulin resistance (interaction), Fasting insulin (interaction)	24204828, 24204828	rs9787485	10:83566686	0.00132
Neuroticism	25993607	rs35855737	3:65542856	0.00132
Lymphoid white cell;LYMPH#	27863252	rs7090504	10:6091017	0.00139
Hair morphology	19896111	rs11803731	1:152083325	0.00149
Bone mineral density (spine), Dupuytrens disease, Ulcerative colitis, Bone mineral density (hip)	19801982, 21732829, Immunobase, 19079262	rs7524102	1:22698447	0.00177
Prostate cancer	23535732	rs7241993	18:76773973	0.00187
Sudden cardiac arrest	21658281	rs10765792	11:95866700	0.00196
Bone mineral density	22504420	rs6426749, rs34920465	1:22711473, 1:22700351	0.00203, 0.00886
Height	20881960	rs4605213, rs537930, rs17807185, rs11107116, rs2806561, rs6919534	17:49244747, 5:134348703, 7:77308295, 12:93978504, 1:23504795, 6:35246903	0.00205, 0.00304, 0.00488, 0.00716, 0.00745, 0.0075
Rheumatoid arthritis	22446963	rs3781913	11:72373496	0.00218
Age-related hearing impairment (interaction)	24939585	rs2882667	5:138314106	0.00243
Advanced age-related macular degeneration	26691988	rs570618, rs10922109	1:196657064, 1:196704632	0.00244, 0.00554
Inflammatory bowel disease, Immature red cell;IRF	Immunobase, 27863252	rs12568930, rs1363907, rs34920465	1:22702231, 5:96252803, 1:22700351	0.00253, 0.00611, 0.00886
Immature red cell;HLSR%	27863252	rs1193, rs3794738	2:87002229, 17:76119293	0.00279, 0.00313
Obesity-related traits	23251661	rs494558	13:110929162	0.00289
Platelet count	27863252	rs2979489	8:30280833	0.00391
Multiple sclerosis	Immunobase	rs180515, rs1886700, rs3118470, rs9891119	17:58024275, 16:68685905, 10:6101713, 17:40507980	0.00401, 0.00442, 0.00803, 0.00836
Testicular germ cell tumor	23666240	rs210138	6:33542538	0.00403
Lipid metabolism phenotypes, Myocardial infarction	22286219, 26343387	rs55791371, rs6065906	19:11188153, 20:44554015	0.00404, 0.00964
Triglycerides, HDL cholesterol	19060906, 19060906	rs7679, rs439401, rs6065906	20:44576502, 19:45414451, 20:44554015	0.00411, 0.00416, 0.00964
HDL Cholesterol - Triglycerides (HDLC-TG)	21386085	rs439401	19:45414451	0.00416
Circulating myeloperoxidase levels (serum)	23620142	rs800292	1:196642233	0.00421
Coronary artery disease	26343387	rs56289821	19:11188247	0.00452

Continued on the next page

Trait(s)	PMID(s)	rs ID(s)	Chr:Pos(s)	F6 p val(s)
Epilepsy (generalized)	22949513	rs13026414	2:57934055	0.00478
Body mass index	25673413	rs1528435	2:181550962	0.005
Systemic lupus erythematosus	Immunobase	rs9782955	1:236039877	0.00534
Response to tocilizumab in rheumatoid arthritis	22491018	rs11121380	1:9408959	0.00545
Venous thromboembolism (gene x gene interaction)	23509962	rs318497	6:2912277	0.00548
Neuroblastoma	22941191	rs11037575	11:43728330	0.00559
Type 1 diabetes	Immunobase	rs11571316, rs402072, rs3118470, rs425105	2:204731089, 19:47219122, 10:6101713, 19:47208481	0.00565, 0.00725, 0.00803, 0.00828
Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (combined)	23453885	rs12966547	18:52752017	0.00604
Crohns disease	Immunobase	rs2549794, rs1363907, rs9891119	5:96244549, 5:96252803, 17:40507980	0.00611, 0.00611, 0.00836
Self-reported allergy	23817569	rs962993	10:9053132	0.00635
Dilated cardiomyopathy	21459883	rs2234962	10:121429633	0.00657
Resting heart rate	20639392	rs12666989, rs314370	7:100486754, 7:100453208	0.00668, 0.0067
QT interval	24952745	rs2298632	1:23710475	0.0067
Compound white cell;MONO%	27863252	rs2630709	2:8445160	0.00678
End-stage coagulation	23381943	rs1410996	1:196696933	0.00691
Blood pressure, Hypertension, Systolic blood pressure, Diastolic blood pressure	21909110, 21909115, 21909115, 21909115	rs633185	11:100593538	0.00701
Eosinophil counts	19198610	rs1420101	2:102957716	0.00733
Vertical cup-disc ratio	25241763	rs868153	6:122389955	0.00758
Adolescent idiopathic scoliosis	26394188	rs13398147	2:222760279	0.00784
Immature red cell;RET#	27863252	rs3896594	2:8762414	0.0079
Alopecia areata	Immunobase	rs3118470	10:6101713	0.00803
Myeloid white cell;MONO#	27863252	rs2485713	10:8465203	0.00817
Interstitial lung disease	23583980	rs7934606	11:1093945	0.0084
Mature red cell;HGB	27863252	rs5762813	22:29203314	0.00913
Type 2 diabetes	21874001	rs1802295	10:70931474	0.00968
Serum protein levels (sST2)	23999434	rs1420103	2:102948632	0.00999

Table S5: GWAS hits with nominal p-value for association with factor 6 below 0.01, related to Figure 3

Methods S1: Relationship between MCFA, CCA and PCA, related to STAR methods.

PCA, pPCA and FA

Principal components analysis[1] is a classic technique for dimensionality reduction. Assume we have N samples measured at p features each. Let y_n be a p-vector denoting the observations for sample n and let $Y = [y_1 : \ldots : y_N]^\top$ be the corresponding $N \times p$ data matrix. For ease of notation, we assume throughout that each feature has mean 0 but note that this is not a requirement.

There are many ways to derive PCA, but perhaps the most common is to consider the problem of finding a unit projection vector v_i that maximizes the variance in the reduced space $T_i = Yv_i$. The first k principal axes $v_1, \ldots, v_k$ are a sequence of orthonormal vectors which successively maximize the variance in the reduced space $T_i = Yv_i$. Let $\hat{\Sigma} = Y^\top Y/N$ be the empirical covariance matrix of Y. PCA solves the following problem:

$$\begin{aligned} &\texttt{max}_v v^\top \hat{\Sigma} v \\ &\text{s.t. } v^\top v = 1 \end{aligned}$$

The top-k principal axes are thus given by the eigenvectors of $\hat{\Sigma}$ that have the k highest eigenvalues. The data in PC space is thus given by the linear projection of the data into this space. Specifically, let $V_k = [v_1, \ldots, v_k]$ be the projection matrix, and let $Y = U\Lambda V^\top$ be the singular value decomposition of Y. Notice that the eigenvectors of the covariance matrix $\hat{\Sigma}$ and the right singular values of the data matrix Y are the same. The points in PC-space are thus given by $T_k = YV_k = U_k\Lambda_k$ where U_k are the the left singular vectors with the k-highest singular values.

Tipping and Bishop[2] introduced a graphical model called probabilistic Principal Components Analysis that provides a generative framework for understanding PCA. The model is as follows:

$$\begin{aligned} x_n &\sim \mathcal{N}(0, I_k) \\ y_n &\sim \mathcal{N}(Lx_n, \sigma^2 I) \end{aligned}$$

where L is a $p \times k$ weight matrix and $\sigma^2 \geq 0$ is the residual noise. They show that the maximum likelihood estimate of the parameters W and σ^2 are given by

$$\begin{aligned} L_{ML} &= V_k(\Lambda_k^2 - \sigma^2 I_k)^{1/2} R \\ \sigma^2_{ML} &= \frac{1}{p-d} \sum_{j=k+1}^{d} \Lambda_j^2 \end{aligned}$$

where R is an arbitrary $k \times k$ orthogonal rotation matrix. Thus, as $\sigma^2 \to 0$, W represents an orthogonal projection into standard PC space. This defines an equivalence relationship between pPCA and PCA.

Factor analysis is a very similar model, with the only difference being the form of the noise term. Rather than force an isotropic noise model $\sigma^2 I$, factor analysis allows for an arbitrary diagonal positive semi-definite matrix $\Psi = \text{diag}(\psi_1, \ldots, \psi_p) \succeq 0$. This allows each observed feature to have it's own error variance.

CCA and pCCA

Now assume each sample is measured on two different sets of conceptually distinct features y_n^1 and y_n^2 with corresponding $N \times p_1$ and $N \times p_2$ data matrices Y_1 and Y_2. As before let $\hat{\Sigma}_{11} = Y_1^\top Y_1/N$ and $\hat{\Sigma}_{22} = Y_2^\top Y_2/N$ be the empirical covariance matrices for modalities 1 and 2, and let $\hat{\Sigma}_{12} = Y_1^\top Y_2/N$ be empirical cross-covariance matrix between the features in each mode. The first set of canonical vectors f_1, f_2 are those that maximize the correlation

$$\text{cor}(Y_1 f_1, Y_2 f_2) = \frac{f_1^\top \hat{\Sigma}_{12} f_2}{\sqrt{f_1^\top \hat{\Sigma}_{11} f_1}\sqrt{f_2^\top \hat{\Sigma}_{22} f_2}}$$

Note that, similarly to PCA, this definition reveals that CCA is a constrained optimization problem:

$$\begin{aligned}\texttt{max}_{f_1,f_2} f_1^\top \hat{\Sigma}_{12} f_2 \\ \text{s. t. } f_1^\top \hat{\Sigma}_{11} f_1 = f_2^\top \hat{\Sigma}_{22} f_2 = 1\end{aligned}$$

but note that rather than the unit norm constraint $f^\top f = 1$ used in PCA, we have a unit variance constraint $f^\top \hat{\Sigma} f = 1$. This unit variance constraint allows there to be correlation within the features of a dataset that is not explained by correlation between the features across datasets.

Successive components can be found by projecting out the first canonical component and again maximizing the correlation of the residuals[3]. Equivalently, all components can be found by solving an eigenvalue problem. To see this consider the change of variables $g_1 = V_1 \Lambda_1^{-1} f_1$ and $g_2 = V_2 \Lambda_2^{-1} f_2$. The correlation is now given by:

$$\text{cor}(Y_1 f_1, Y_2 f_2) = \frac{g_1^\top U_1^\top U_2 g_2}{\sqrt{g_1^\top g_1}\sqrt{g_2^\top g_2}}$$

which indicates that g_1, g_2 are the top pair of left-right singular vectors of the matrix $U_1^\top U_2$. Further components are further singular vectors of $U_1^\top U_2$, and it's singular values are the correlations. This also reveals that CCA is equivalent to using PCA to whiten the variables of each data matrix, concatenating them, and then performing PCA again on the whitened, concatenated data matrix.

Likewise to probabilistic PCA, probabilistic CCA is a graphical model that provides a generative framework for thinking about CCA[4]. The model is as follows:

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_d) \\ y_n^1 &\sim \mathcal{N}(W_1 x_n, \Psi_1) \\ y_n^2 &\sim \mathcal{N}(W_2 x_n, \Psi_2) \end{aligned}$$

similarly to FA and pPCA, we sample a d-dimensional random normal hidden vector, pass it through a weight matrix, and add random noise. We have two weight matrices W_1 and W_2 of shape $p_1 \times d$ and $p_2 \times d$, and two noise matrices Ψ_1 and Ψ_2, however in this case these noise matrices are arbitrary positive semi-definite matrices ($\Psi_\bullet \succeq 0$). Bach and Jordan show that the maximum likelihood estimate of the parameters of pCCA can be determined from the CCA solution:

$$\begin{aligned} W_{1,ML} &= \hat{\Sigma}_{11} F_{1d} M_1 \\ W_{2,ML} &= \hat{\Sigma}_{22} F_{2d} M_2 \\ \Psi_{1,ML} &= \hat{\Sigma}_{11} - W_{1,ML} W_{1,ML}^\top \\ \Psi_{2,ML} &= \hat{\Sigma}_{22} - W_{2,ML} W_{2,ML}^\top \end{aligned}$$

where $F_{\bullet d} = [f_{\bullet 1}; \ldots; f_{\bullet d}]$ are the first d canonical directions and M_1, M_2 are arbitrary matrices with spectral norm less than 1 such that $M_1 M_2 = \rho_d$.

Multi-set canonical correlation analysis

Now rather than having two sets of conceptually distinct features for each sample, assume we have M different conceptually distinct sets of features $\{y_n^m\}$ with corresponding $N \times p_m$ data matrices $\{Y_m\}$. In MCCA, we are still interested in finding projection vectors $\{f_m\}$ which map our high dimensional data into a one-dimensional space, however there are many formulations that are equivalent to classical CCA with two datasets. Let $\hat{\Sigma}_{kl} = Y_k^\top Y_l / N$ be the empirical cross-covariance matrix between the features in dataset k and dataset l. The covariance of the data in the reduced space is given by

$$S = \begin{bmatrix} f_1^\top \hat{\Sigma}_{11} f_1 & \ldots & f_1^\top \hat{\Sigma}_{1M} f_M \\ \vdots & \ddots & \vdots \\ f_M^\top \hat{\Sigma}_{M1} f_1 & \ldots & f_M^\top \hat{\Sigma}_{MM} f_M \end{bmatrix}$$

The various formulations of MCCA correspond to optimizing different objective functions $J(S)$ subject to the a constraint function $h(f, \hat{\Sigma})$[5,6]. In brief, possible objective functions include:

- SUMCOR: Maximize the sum of pairwise correlations: $J = \sum_{i,j} f_i^\top \hat{\Sigma}_{i,j} f_j = 1^\top S 1$
- SUMSQCOR: Maximize the sum of squares of pairwise correlations: $J = \sum_{i,j} (f_i^\top \hat{\Sigma}_{i,j} f_j)^2 = ||S||_F^2$
- MAVAR: Maximize the largest eigenvalue of S: $J = \lambda_1(S)$
- MINVAR: Minimize the smallest eigenvalue of S: $J = \lambda_d(S)$
- GENVAR: Minimize the determinant of S, also known as the generalized variance: $J = |S| = \prod_i \lambda_i(S)$

while possible constraints include:

- VAR: The canonical directions each have unit variance: $h : \forall_i f_i^\top \hat{\Sigma}_{ii} f_i = 1$
- AVGVAR: The canonical directions have unit variance on average: $h : \sum_i f_i^\top \hat{\Sigma}_{ii} f_i = d$
- NORM: The canonical directions each have unit norm: $h : \forall_i f_i^\top f_i = 1$
- AVGNORM: The canonical directions have unit norm on average: $h : \sum_i f_i^\top f_i = d$

It is straightforward to see that any of the 5 listed objective functions could be combined with either of the first two constraints to create 10 optimization problems that are equivalent to CCA in the two-dataset case. The final two constraints correspond to relaxations of the unit variance constraint which can reveal a simpler optimization problem in some cases, and which does not suffer from being trivially satisfiable when $p > N$.

Some of these can be fit by solving eigenvalue problems, while others require more complicated iterative methods. Of particular note are the SUMCOR and GENVAR objectives. The GENVAR objective was the first considered MCCA approach[7], where a simple solution for the $M = 3$ and $p_1 = p_2 = p_3 = 2$ case is given. GENVAR is a particularly natural way of thinking about MCCA - it is a single value that represents the multidimesnioal scatter of points in space[8]. Smaller values of the generalized variance indicate less scatter, and thus higher "correlation" of the points in the reduced space. Despite this, is has received relatively little attention as a method for MCCA, perhaps because it is challenging to fit[6]. On the other hand, most attention has been focused on the SUMCOR objective[9], which can be solved easily with the AVGVAR and AVGNORM constraints. SUMCOR with AVGVAR constraint can be solved via a simple two-stage procedure: first whiten each data matrix, concatenate the whitened features, and then perform PCA on the whitened, concatenated features. SUMCOR with AVGNORM constraint is even simpler to solve: simply perform PCA on the concatenated feature set. In the two dataset case, this latter method is sometimes called "diagonal CCA" and forms the basis of the original integration approach used in Seurat[10] as well as many sparse CCA approaches[11]. This is also closely related to group factor analysis approaches for multi-modal data, see for example[12] and references therein. The equivalence to PCA on the concatenated feature set makes it straightforward to see that the NORM-based constraints involve an implicit assumption that shared factors are responsible for both covariation across features in different modes and between features within a mode.

Probabilistic graphical model for multi-set CCA

Here, we describe a probabilistic graphical model for multi-set CCA (pMCCA). Note that while this model is an option in our software package, we derive and discuss it primarily to draw connection to traditional multiset CCA. The full model which includes simultaneous factor analysis of the private spaces is described in the next section. Unlike traditional CCA, pCCA has one single obvious generalization to multiple datasets:

$$z_n \sim \mathcal{N}(0, I_d) \tag{1}$$

$$y_n^m \sim \mathcal{N}(W_m z_n, \Psi_m) \tag{2}$$

where again we have weight matrices W_m of shape $p_m \times d$ and arbitrary positive semi-definite noise matrices $\Psi_m \succeq 0$.

We will now see that pMCCA is related to MCCA in the following ways:

Observation 1. *The maximum likelihood solution to the pMCCA model corresponds to the GENVAR objective with VAR constraint in the $M = 3$ dataset case.*

Observation 2. *The maximum likelihood solution to the pMCCA model does not correspond to any of the listed MCCA formulations in the $M \geq 4$ case.*

Let $W = [W_1^\top : \ldots : W_M^\top]^\top$ be the stacked weight matrices and $\Psi = \text{diag}(\Psi_1, \ldots, \Psi_M)$ be the block diagonal covariance matrix. The model covariance is given by:

$$\Sigma = \begin{bmatrix} W_1 W_1^\top + \Psi_1 & W_1 W_2^\top & \ldots & W_1 W_M^\top \\ W_2 W_1^\top & W_2 W_2^\top + \Psi_2 & \ldots & W_2 W_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ W_M W_1^\top & W_M W_2^\top & \ldots & W_M W_M^\top + \Psi_M \end{bmatrix} = WW^\top + \Psi \tag{3}$$

Let $Y = [Y_1 : \ldots : Y_M]$ so that the empirical covariance matrix can be written $\hat{\Sigma} = Y^\top Y/N$. The model negative log-likelihood is given by:

$$l(\Sigma|\hat{\Sigma}) = \frac{Np}{2} \log 2\pi + \frac{N}{2} \log |\Sigma| + \frac{N}{2} \operatorname{Tr}(\Sigma^{-1}\hat{\Sigma}) \tag{4}$$

where $p = \sum_m p_m$ is the total number of features from all datasets. It is straightforward to see that many arguments from Bach and Jordan[4] carry over to the multi-set case, thus we refer readers there for proofs. In particular we have

Lemma 1. *At a stationary point of the likelihood $\Sigma_{mm} = W_m W_m^\top + \Psi_m = \hat{\Sigma}_{mm}$.*

Thus, at a stationary point

$$\operatorname{Tr}(\Sigma^{-1}\hat{\Sigma}) = \operatorname{Tr}\left(\begin{bmatrix} \hat{\Sigma}_{11} & \ldots & W_1 W_M^\top \\ \vdots & \ddots & \vdots \\ W_M W_1^\top & \ldots & \hat{\Sigma}_{MM} \end{bmatrix}^{-1} \hat{\Sigma} \right) = p \tag{5}$$

so that the models minimum negative log-likelihood is proportional to the log generalized variance of the model:

$$l(\Sigma|\hat{\Sigma}) \propto \log |\Sigma| \tag{6}$$

Moreover, Lemma 1 allows us to further factorize Σ

$$\Sigma = \begin{bmatrix} \hat{\Sigma}_{11}^{1/2} & 0 & \ldots & 0 \\ 0 & \hat{\Sigma}_{22}^{1/2} & \ldots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{\Sigma}_{MM}^{1/2} \end{bmatrix} \begin{bmatrix} I & \tilde{W}_1 \tilde{W}_2^\top & \ldots & \tilde{W}_1 \tilde{W}_M^\top \\ \tilde{W}_2 \tilde{W}_1^\top & I & \ldots & \tilde{W}_2 \tilde{W}_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{W}_M \tilde{W}_1^\top & \tilde{W}_M \tilde{W}_2^\top & \ldots & I \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_{11}^{1/2} & 0 & \ldots & 0 \\ 0 & \hat{\Sigma}_{22}^{1/2} & \ldots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{\Sigma}_{MM}^{1/2} \end{bmatrix} \tag{7}$$

and so

$$l(\Sigma|\hat{\Sigma}) \propto \begin{vmatrix} I & \tilde{W}_1 \tilde{W}_2^\top & \ldots & \tilde{W}_1 \tilde{W}_M^\top \\ \tilde{W}_2 \tilde{W}_1^\top & I & \ldots & \tilde{W}_2 \tilde{W}_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{W}_M \tilde{W}_1^\top & \tilde{W}_M \tilde{W}_2^\top & \ldots & I \end{vmatrix} \tag{8}$$

where $\tilde{W}_m = \hat{\Sigma}_{mm}^{-1/2} W_m$. Notice that the off-diagonal blocks are the cross-covariance matrices in the model of individually-whitened datasets. If there exists a set of projection vectors f_m such that $f_m^\top \hat{\Sigma}_{mm} f_m = 1$ and the projection of the data into the space spanned by f_m has covariance equal to the above, then minimizing the above determinant solves the GENVAR MCCA objective with VAR constraint. Let $\tilde{Y}_m = Y_m \hat{\Sigma}_{mm}^{-1/2}$ be the whitened datasets and let $g_m = \hat{\Sigma}_{mm}^{1/2} f_m$ be the change of variables that gives $g_m^\top g_m = 1$. The projection in MCCA space is given by $\tilde{Y}_{||}^m = \tilde{Y}^m g_m g_m^\top$. We seek g_m such that

$$\begin{aligned}
\tilde{Y}_{||}^{k\top} \tilde{Y}_{||}^l &= g_k g_k^\top \tilde{Y}^{k\top} \tilde{Y}^l g_l g_l^\top && (9)\\
&= g_k g_k^\top \hat{\Sigma}_{kk}^{-1/2} \hat{\Sigma}_{kl} \hat{\Sigma}_{ll}^{-1/2} g_l g_l^\top && (10)\\
&= g_k f_k^\top \hat{\Sigma}_{kl} f_l g_l^\top && (11)\\
&= \tilde{W}_k \tilde{W}_l^\top && (12)
\end{aligned}$$

Thus we must satisfy

$$g_k^\top \tilde{W}_k \tilde{W}_l^\top g_l = f_k^\top \hat{\Sigma}_{kl} f_l \tag{13}$$

Notice that for $d = 1$, the left and right side are scalars. This means we can express our necessary criterion as

$$q_k q_l = c_{kl} \tag{14}$$

For $M = 3$, this results in a system of 3 equations in 3 unknowns, which has solutions of the form $q_1 = \sqrt{\frac{c_{12} c_{13}}{c_{23}}}$. Note that these can be found by setting $g_k = \tilde{W}_k / ||\tilde{W}_k||_2$ which yields $q_k = ||\tilde{W}_k||_2$. For $M > 3$, there are more equations than unknowns and they cannot be mutually satisfied in general. Note also that a similar argument can be used in the $d > 1$ case to show that this system is not satisfiable even for $M = 3$. Thus, unlike CCA and pCCA, fitting $d > 1$ components is not equivalent to iteratively fitting single components and projecting them out.

Multiset Correlation and Factor Analysis

The residual covariance matrices Ψ_d deserve additional attention. Put simply, these matrices represent the residual structure in each modality after accounting for shared structure across modes. Instead of allowing this matrix to be arbitrary, we can instead think of this matrix as having some additional structure. For example, Ψ_d might be the sum of a low rank and an isotropic covariance matrix. This suggests that we add an additional latent variable to each dataset, corresponding to a factor model for the "private" structure (e.g. not shared with other datasets). Specifically, we modify pMCCA such that for each dataset, we additionally sample a latent variable from a k_m-dimensional unit Gaussian. The observed data are then sampled from a multi-variate Gaussian where the mean is a linear combination of both private variables, but the residual covariance matrix is now diagonal.

$$\begin{aligned}
z_n &\sim \mathcal{N}(0, I_d) && (15)\\
x_n^m &\sim \mathcal{N}(0, I_{k_m}) && (16)\\
y_n^m &\sim \mathcal{N}(W_m z_n + L_m x_n^m, \Psi_m) && (17)
\end{aligned}$$

where L_m are the $k_m \times p_m$ private space loading matrices and $\Psi_m = \text{diag}(\psi_m^1, \ldots, \psi_m^{p_m})$ are the diagonal residual covariance matrices. Note that in general we allow the entries on the diagonal of Ψ_m to take different values, similarly to factor analysis. One could additionally constrain the entries of the diagonal to be the same, $\Psi = \sigma^2 I_p$, similar to pPCA.

This model can be fit via a straightforward application of the expectation-maximization (EM) algorithm[13]. We first derive conditional expectation of the log-likelihood, $\mathcal{L}$ for the model under the generative process specified in Figure S1. For convenience, let

$$\begin{aligned}
y_n &= [y_n^{1\top} : \ldots : y_n^{m\top}]^\top \in \mathbb{R}^p && (18)\\
x_n &= [x_n^{1\top} : \ldots : x_n^{m\top}]^\top \in \mathbb{R}^k && (19)\\
W &= [W_1^\top : \ldots : W_M^\top]^\top \in \mathbb{R}^{p \times d} && (20)\\
L &= \text{diag}(L_1, \ldots, L_m) \in \mathbb{R}^{p \times k} && (21)\\
\Psi &= \text{diag}(\Psi_1, \ldots, \Psi_M) \in \mathbb{R}^{p \times p} && (22)
\end{aligned}$$

where $k = \sum_m k_m$. At a given time step t during the computation of the EM algorithm, let the conditional expectation for given latent variables z_i and x_i be $\mathbb{E}[\cdot|W_t, \Psi_t, L_t, y_i] = \langle\cdot\rangle$.

The conditional expectation of the log-likelihood (E-step) is:

$$\begin{aligned}
\langle \mathcal{L} \rangle = -\sum_{i=1}^{N} \tilde{C} + \frac{1}{2} \ln |\Psi| + \frac{1}{2} \operatorname{Tr}\left(\Psi^{-1} y_i y_i^\top\right) + \frac{1}{2} \operatorname{Tr}\left(L^\top \Psi^{-1} L \left\langle x_i x_i^\top \right\rangle\right) && (23)\\
+ \frac{1}{2} \operatorname{Tr}\left(W^\top \Psi^{-1} W \left\langle z_i z_i^\top \right\rangle\right) + \operatorname{Tr}\left(L^\top \Psi^{-1} W \left\langle z_i x_i^\top \right\rangle\right) - y_i^\top \Psi^{-1} W \langle z_i \rangle && (24)\\
- y_i^\top \Psi^{-1} L x_i + \frac{1}{2} \operatorname{Tr} \left\langle x_i x_i^\top \right\rangle + \frac{1}{2} \operatorname{Tr} \left\langle z_i z_i^\top \right\rangle && (25)
\end{aligned}$$

At a given timestep t, we compute the update of parameters $t+1$ by differentiating $\mathcal{L}$ with respect to W_t, L_t, and Ψ_t, and setting the derivative of the corresponding expected log-likelihood to 0. The following update steps are derived using standard matrix differentiation results[14].

$$\begin{aligned}
W_{t+1} &= \left(\sum_{i=1}^{N} y_i \left\langle z_i^\top \right\rangle - L_t \left\langle x_i z_i^\top \right\rangle\right)\left(\sum_{i=1}^{N} \left\langle z_i z_i^\top \right\rangle\right)^{-1} && (26)\\
L_{t+1} &= \left(\sum_{i=1}^{N} y_i \left\langle x_i^\top \right\rangle - W_t \left\langle z_i x_i^\top \right\rangle\right)\left(\sum_{i=1}^{N} \left\langle x_i x_i^\top \right\rangle\right)^{-1} \Psi_{t+1} && (27)\\
&= \frac{1}{N} \sum_{i=1}^{N} y_i y_i^\top + L \left\langle x_i x_i^\top \right\rangle L^\top + W \left\langle z_i z_i^\top \right\rangle W^\top + 2L \left\langle x_i z_i^\top \right\rangle W^\top - 2y_i \left\langle z_i^\top \right\rangle W^\top - 2y_i \left\langle x_i^\top \right\rangle L^\top && (28)
\end{aligned}$$

References

1. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science **2,** 559–572.
2. Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B (Methodological) **61,** 611–622. doi:10.5351/CSAM.2017.24.2.143.
3. Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika **28,** 321–377. doi:10.2307/2333955.
4. Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis.
5. Kettenring, J. R. (1971). Canonical analysis of several sets of variables. Biometrika **58,** 433–451.
6. Asendorf, N. A. (2015). Informative Data Fusion: Beyond Canonical Correlation Analysis.
7. Steel, R. G. D. (1951). Minimum Generalized Variance for a set of Linear Functions. Ann. Math. Statist. **22,** 456–460. doi:10.1214/AOMS/1177729594.
8. Kocherlakota, S. and Kocherlakota, K. (2004). Generalized Variance. Encyclopedia of Statistical Sciences. doi:10.1002/0471667196.ESS0869.
9. Parra, L. C. (2018). *Multiset Canonical Correlation Analysis simply explained* tech. rep. (2018). arXiv: 1802.03759v1.
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology 2018 36:5 **36,** 411–420. doi:10.1038/nbt.4096.
11. Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10,** 515–534. doi:10.1093/biostatistics/kxp008.

12. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W. and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology **14.** doi:10.15252/msb.20178124.
13. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39,** 1–22.
14. Petersen, K. B., Pedersen, M. S., *et al.* (2008). The matrix cookbook. Technical University of Denmark **7,** 510.