Multi-set correlation and factor analysis of multi-omic data

Brielin C Brown PhD

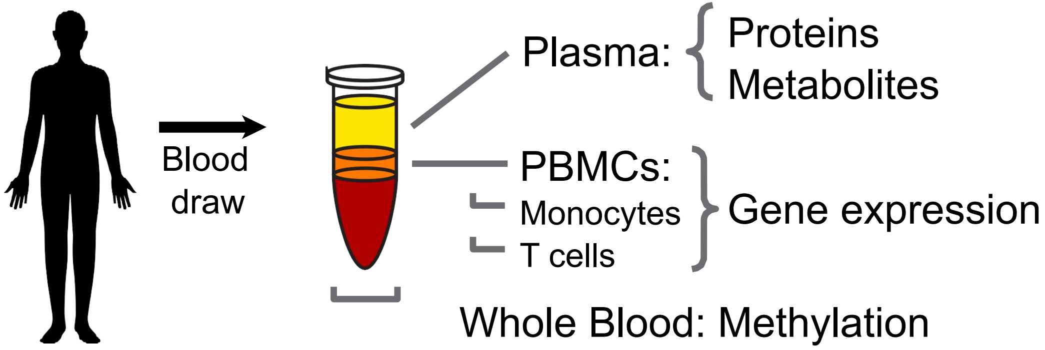
Postdoctoral researcher
Knowles and Lappalainen labs
New York Genome Center

DSI Fellow
Data Science Institute
Columbia University

Multi-omics data promise to revolutionize biomedical research

MESA
Multi-Ethnic Study of Atherosclerosis

NHLBI TOPMed



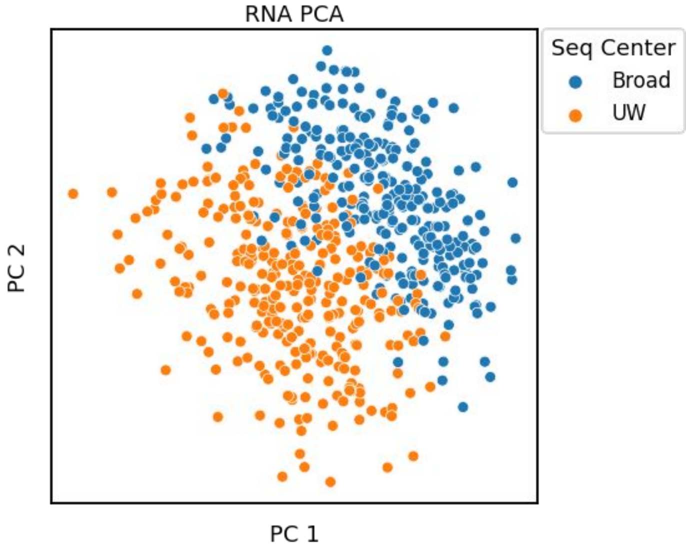
Multi-*omic* but also

- Multi-*phenotype*
- Multi-*tissue*
- Multi-*ethnic*
- Multi-*center*

NHLBI TOPMed

- Generating multi-omic data for tens of thousands of patients
- MESA multi-omic pilot

How can we think about exploring genomic data?



How can we think about exploring genomic data?

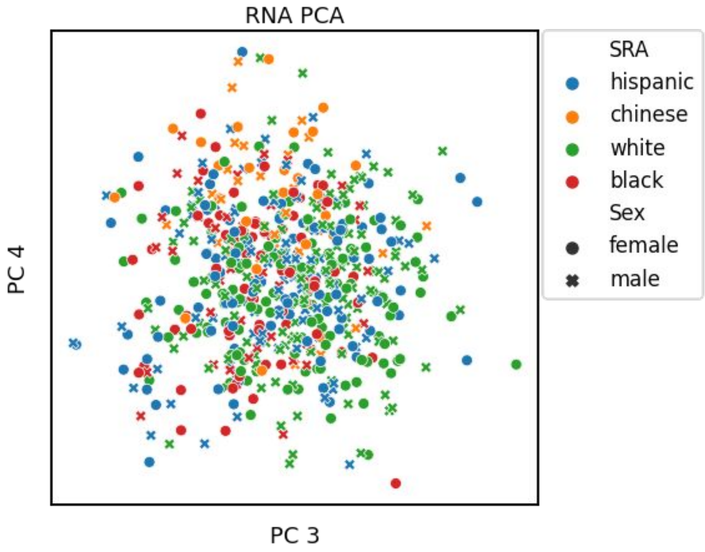
RNA levels — RNA loadings

$$Y = XL^\top + \epsilon$$

Low-dim features — Residual

Includes

- Principal Components Analysis (PCA)
- Factor Analysis (FA)



How can we simultaneously explore two datasets?

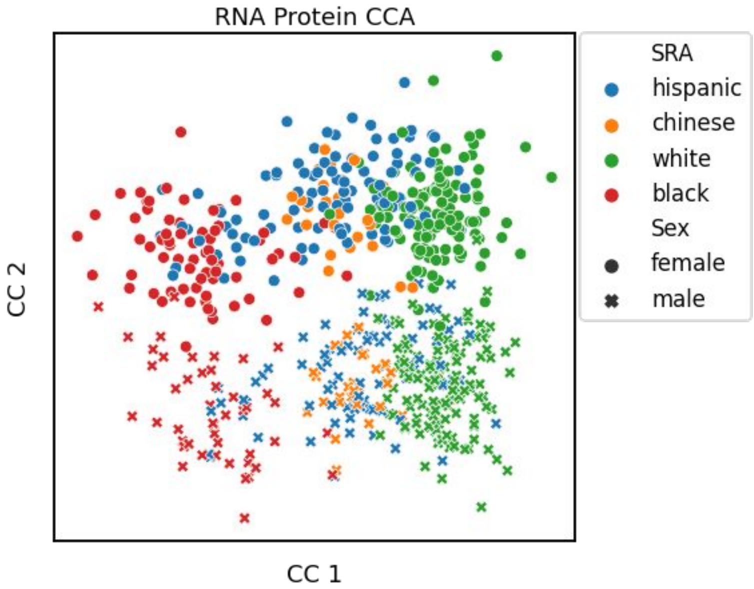
Feature loadings / Residual

RNA levels $$Y_1 = ZW_1^\top + \epsilon_1$$

Protein levels $$Y_2 = ZW_2^\top + \epsilon_2$$

Low-dim shared factors

Includes
- Group Factor Analysis (GFA)
- Canonical Correlation Analysis (CCA)
- *Probabilistic* CCA



Multiset Correlation and Factor Analysis combines these

$$Y_1 \approx Z W_1^\top + X_1 L_1^\top + \epsilon_1$$

$$\vdots$$

Multi-omic data

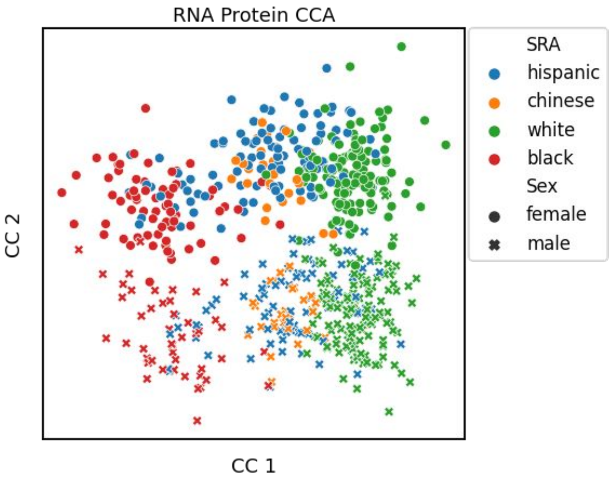
$$Y_m \approx Z W_m^\top + X_m L_m^\top + \epsilon_m$$

Shared factors and loadings

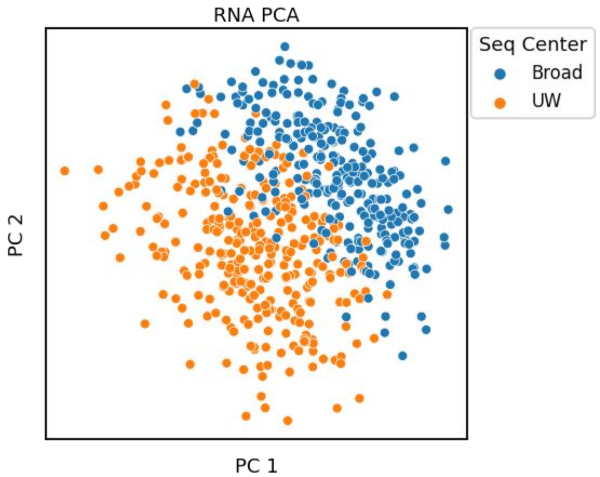
Private factors and loadings

Residual

$$Y_{RNA} \approx$$



$$+$$



Multiset Correlation and Factor Analysis combines these

Multi-omic data

$$Y_1 \approx Z W_1^\top + X_1 L_1^\top + \epsilon_1$$

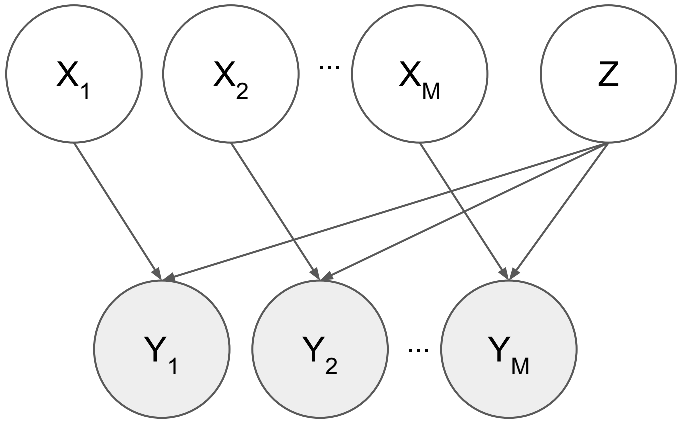
$$\vdots$$

$$Y_m \approx Z W_m^\top + X_m L_m^\top + \epsilon_m$$

Shared factors and loadings

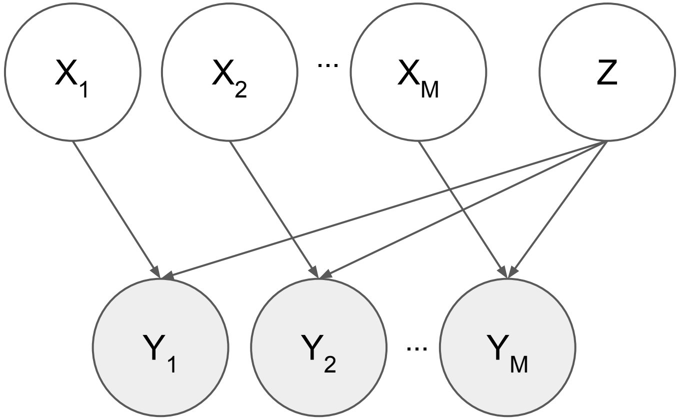
Private factors and loadings

Residual

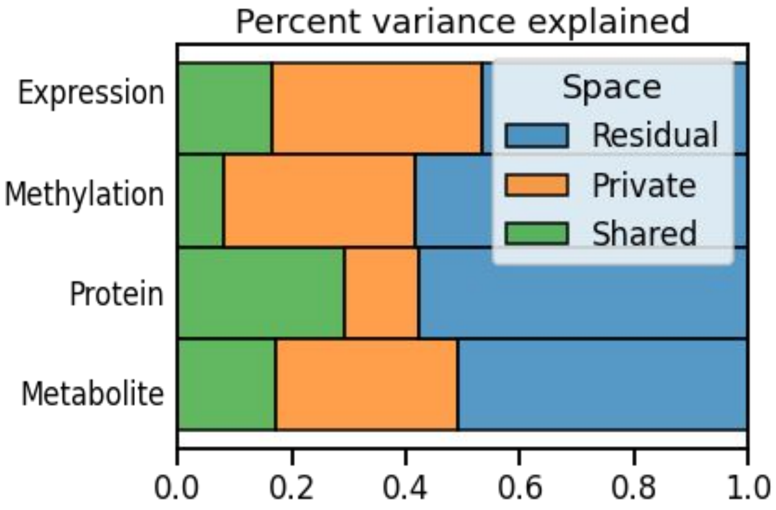


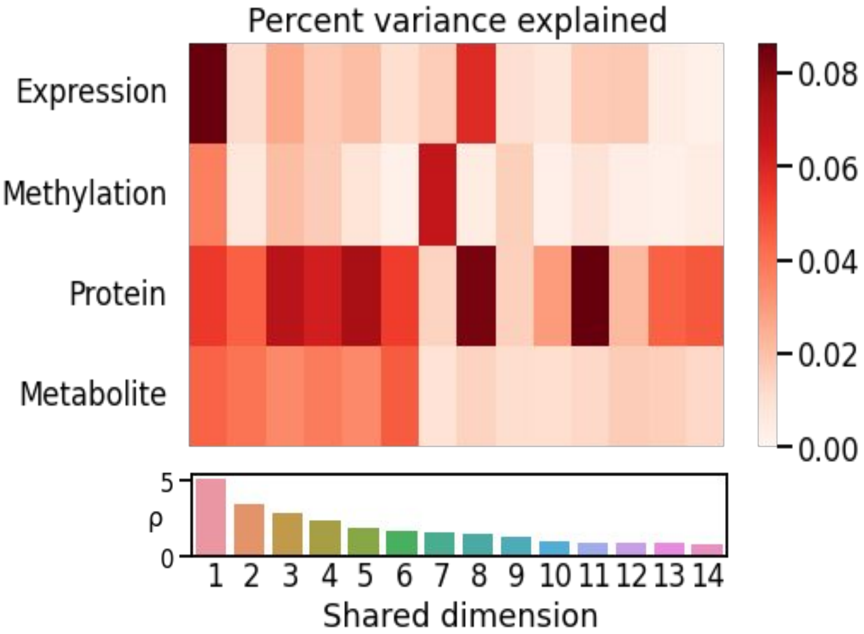
MCFA has numerous advantages

- Elegant approach to multiset CCA
  - Strong theoretical connection
- Calculations done in PC-space
  - Extremely efficient, controls over-fitting
- No hyperparameters to tune
  - Self-selected with random matrix techniques
- Works with any data type
  - No “gene scores”
- Factor loadings are interpretable
  - Similar to regression coefficients
- Fully unsupervised!
  - Supports exploratory analysis

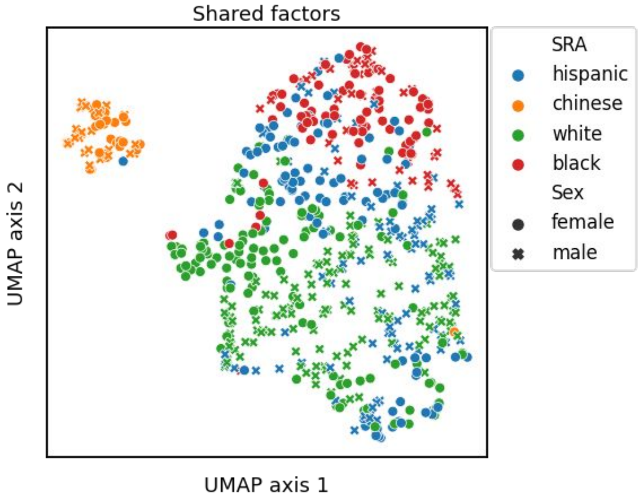


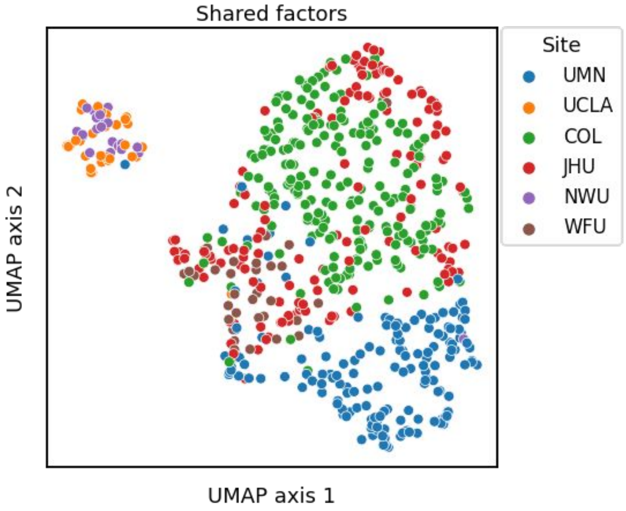
MCFA finds structure in 614 multi-omic samples





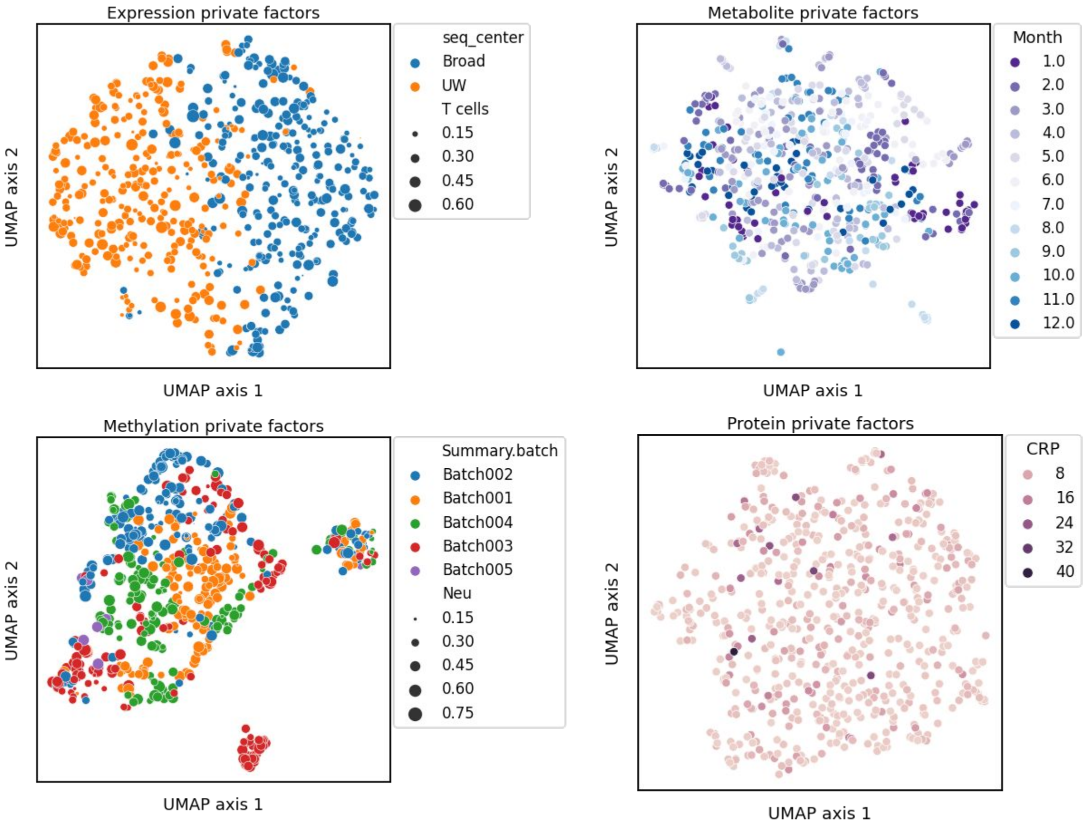
Shared structure reflects demographics and site



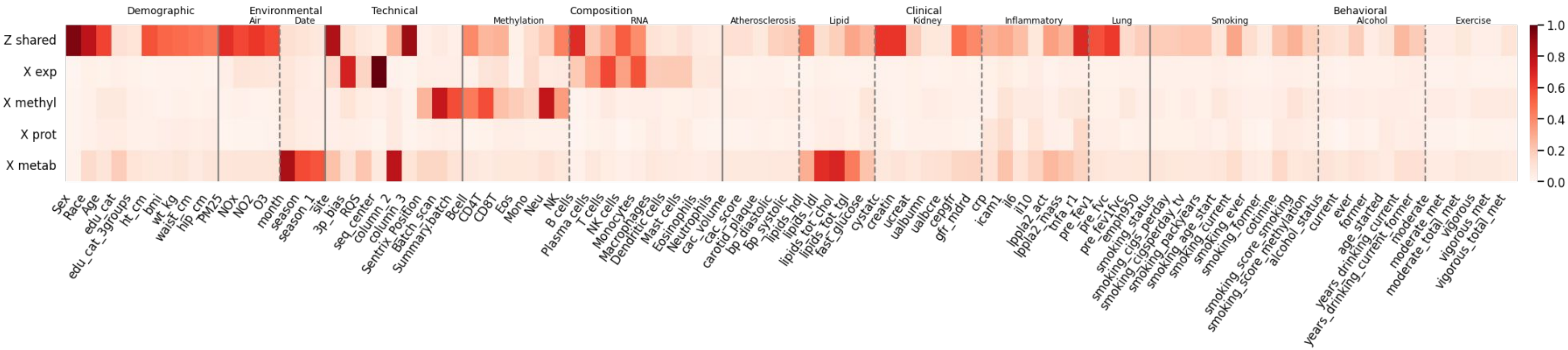


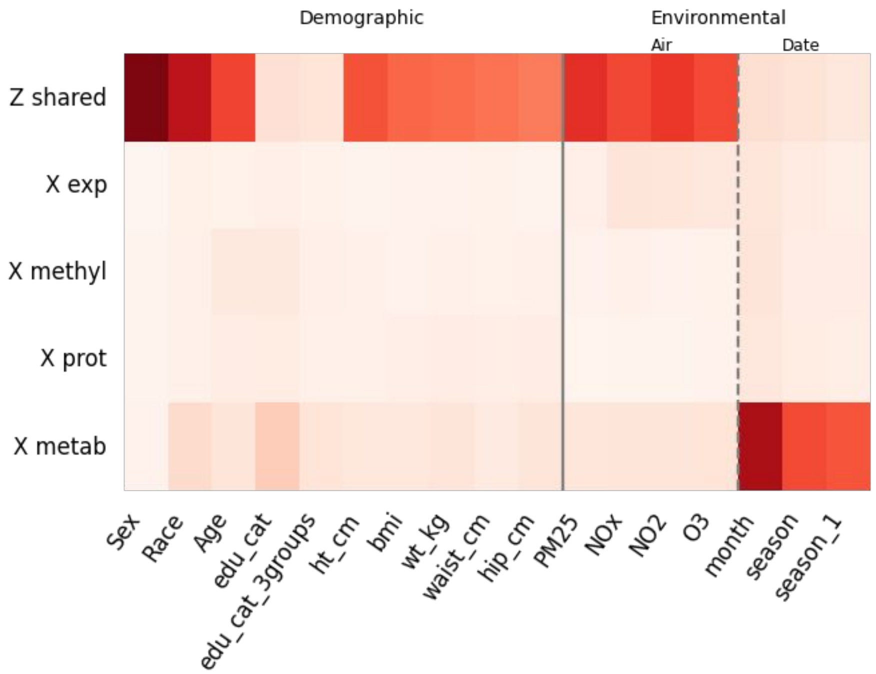
This is without genetic information

Private structure captures omic-specific effects

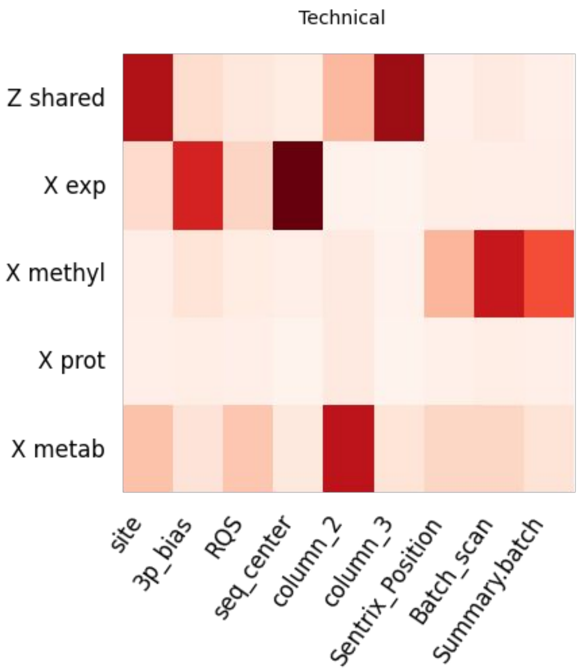


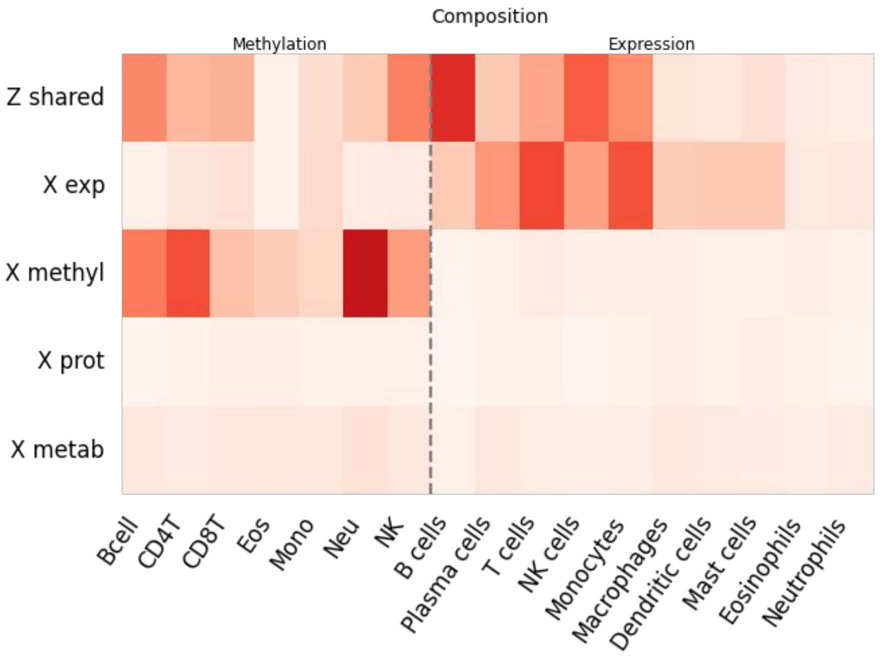
MCFA further captures environmental differences and clinical biomarkers

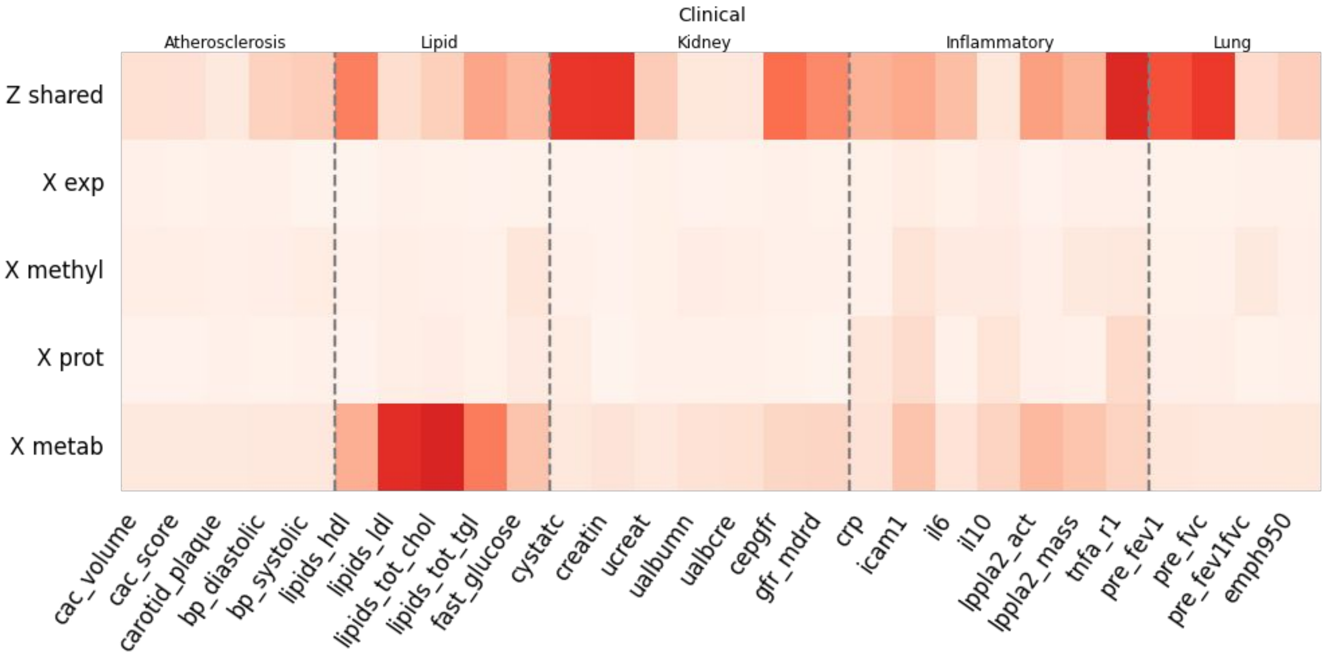




Demographic
Environmental
Air
Date
Z shared
X exp
X methyl
X prot
X metab
Sex
Race
Age
edu_cat
edu_cat_3groups
ht_cm
bmi
wt_kg
waist_cm
hip_cm
PM25
NOx
NO2
O3
month
season
season_1

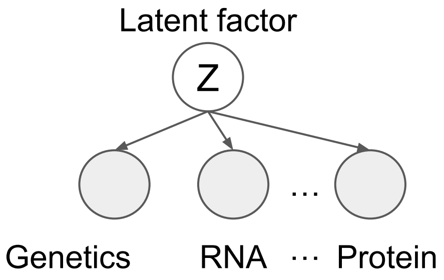

Technical
Z shared
X exp
X methyl
X prot
X metab
site
3p_bias
RQS
seq_center
column_2
column_3
Sentrix_Position
Batch_scan
Summary.batch


Composition
Methylation
Expression
Z shared
X exp
X methyl
X prot
X metab
Bcell
CD4T
CD8T
Eos
Mono
Neu
NK
B cells
Plasma cells
T cells
NK cells
Monocytes
Macrophages
Dendritic cells
Mast cells
Eosinophils
Neutrophils

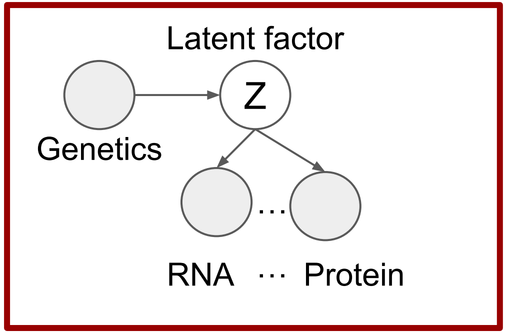

Clinical
Atherosclerosis
Lipid
Kidney
Inflammatory
Lung
Z shared
X exp
X methyl
X prot
X metab
cac_volume
cac_score
carotid_plaque
bp_diastolic
bp_systolic
lipids_hdl
lipids_ldl
lipids_tot_chol
lipids_tot_tgl
fast_glucose
cystatc
creatin
ucreat
ualbumn
ualbcre
cepgfr
gfr_mdrd
crp
icam1
il6
il10
lppla2_act
lppla2_mass
tnfa_r1
pre_fev1
pre_fvc
pre_fev1fvc
emph950

There are multiple ways to integrate genetic data

1. Include genetics in the MCFA model
    a. Somatic variation
    b. Copy number variation
    c. Known functional SNPs (eQTLs, etc)

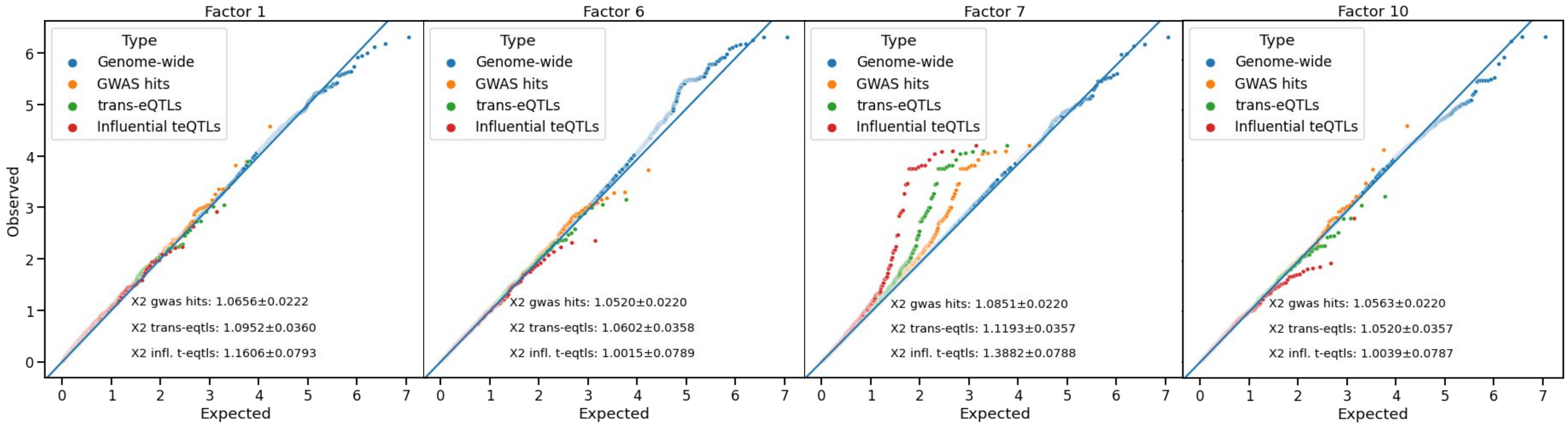


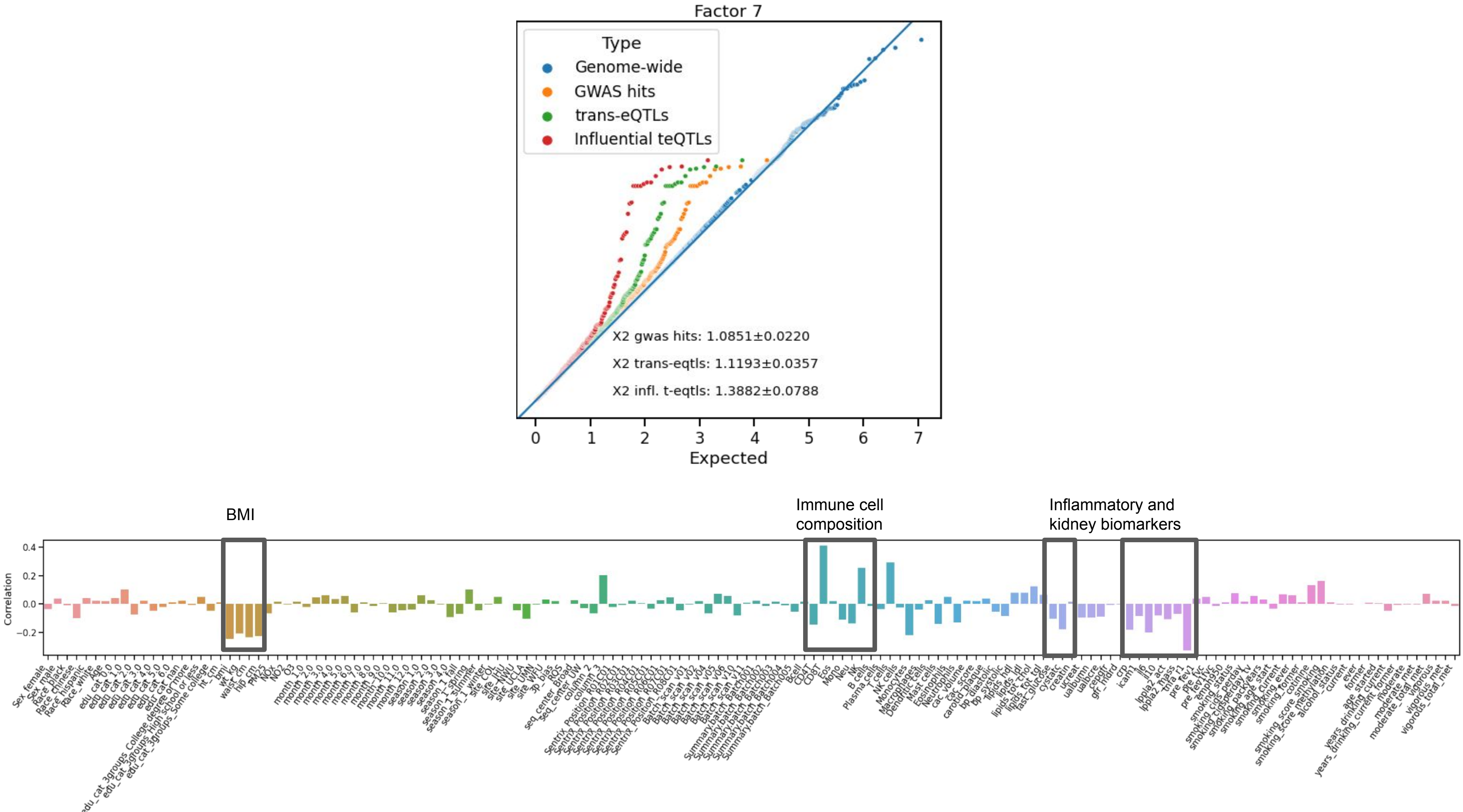
2. Think of factors as "latent phenotypes"
    a. Factors learned on molecular data may represent coordinated biological activity
    b. These may be affected by genetic variation
    c. Look for genetic associations with latent phenotypes (GWAS)



GWAS can be used to integrate genetic data

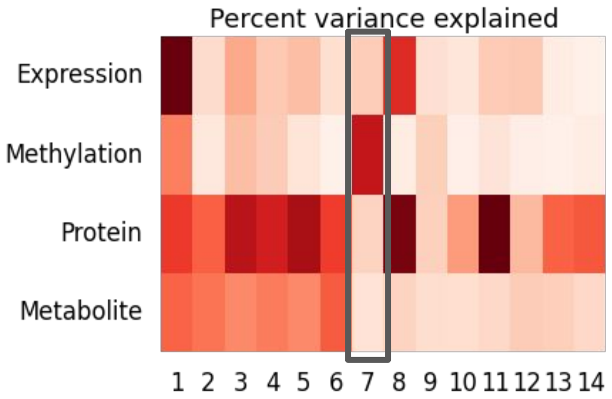
Are genetic associations with shared factors enriched for known GWAS hits or *trans*-eQTLs?




Factor 7
Type
Genome-wide
GWAS hits
trans-eQTLs
Influential teQTLs
X2 gwas hits: 1.0851±0.0220
X2 trans-eqtls: 1.1193±0.0357
X2 infl. t-eqtls: 1.3882±0.0788
Expected
BMI
Immune cell composition
Inflammatory and kidney biomarkers
Correlation

Top SNPs implicate lipid metabolism near FADS1, FADS2 genes

Trait	rs ID	Chr	Position	Gene	p-val_F7
Red blood cell fatty acid levels	rs174541	11	61565908	FADS2	0.000043
Metabolite levels	rs174549	11	61571382	FADS1, FADS2	0.000056
Blood metabolite levels	rs174556	11	61580635	FADS1, FADS2	0.000057
Height	rs7184046	15	75866150	PTPN9	0.000061
Plasma omega-6 polyunsaturated fatty...	rs174555	11	61579760	FADS1, FADS2	0.000064
Phospholipid levels (plasma)	rs1535	11	61597972	FADS2	0.000083
Plasma omega-6 polyunsaturated fatty...	rs174537	11	61552680	MYRF, TMEM258	0.000109
Phospholipid levels (plasma)	rs174536	11	61551927	MYRF,TMEM258	0.000109
Cholesterol, total	rs174554	11	61579463	FADS1, FADS2	0.000117
Plasma omega-6 polyunsaturated fatty...	rs174547	11	61570783	FADS1, FADS2	0.000127
Plasma omega-6 polyunsaturated fatty...	rs174550	11	61571478	FADS1, FADS2	0.000127
Plasma omega-6 polyunsaturated fatty...	rs174546	11	61569830	FADS1, FADS2	0.000127
Red blood cell fatty acid levels	rs174545	11	61569306	FADS1, FADS2	0.000127
Blood metabolite ratios	rs174548	11	61571348	FADS1, FADS2	0.000253
Triglycerides	rs174529	11	61543961	MYRF, TMEM258	0.000264
Systemic lupus erythematosus	rs4852324	2	74202578	DGUOK-AS1	0.000343
Trans fatty acid levels	rs174583	11	61609750	FADS2	0.000405
Mature red cell;HGB	rs1256061	14	64703593	ESR2	0.000464
IgG glycosylation	rs2186369	22	24170996	SMARCB1	0.000505
Plasma omega-6 polyunsaturated fatty...	rs2727270	11	61603237	FADS2	0.000774



GWAS SNPs associated with Factor 7 are also associated with methylation level

Trait	rsid	Chr	Position	Gene	p-val_trait	p-val_F7	CpG	p-val_mqtl	n_snps
Red blood cell fatty acid levels	rs174541	11	61565908	FADS2	3.000000e-19	0.000043	cg19610905	3.964333e-115	29
				PTPN9	2.000000e-10	0.000061	cg01268058	1.333136e-48	1
				[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1

Losol *et al. Genes & Nutrition* (2019) 14:20
https://doi.org/10.1186/s12263-019-0644-8

Genes & Nutrition

RESEARCH Open

Effect of gestational oily fish intake on the risk of allergy in children may be influenced by *FADS1/2*, *ELOVL5* expression and DNA methylation

Purevsuren Losol[1,2], Faisal I. Rezwan[1], Veeresh K. Patil[3,4], Carina Venter[3], Susan Ewart[5], Hongmei Zhang[6], S. Hasan Arshad[3,4], Wilfried Karmaus[6] and John W. Holloway[1,4*]

OPEN ACCESS Freely available online

PLOS ONE

Supplementation with N-3 Long-Chain Polyunsaturated Fatty Acids or Olive Oil in Men and Women with Renal Disease Induces Differential Changes in the DNA Methylation of FADS2 and ELOVL5 in Peripheral Blood Mononuclear Cells

Samuel P. Hoile[1], Rebecca Clarke-Harris[1], Rae-Chi Huang[2], Philip C. Calder[1,3], Trevor A. Mori[4], Lawrence J. Beilin[4], Karen A. Lillycrop[5], Graham C. Burdge[1*]

Gomez-Alonso *et al. Clin Epigenet* (2021) 13:7
https://doi.org/10.1186/s13148-020-00957-8

Clinical Epigenetics

RESEARCH Open Access

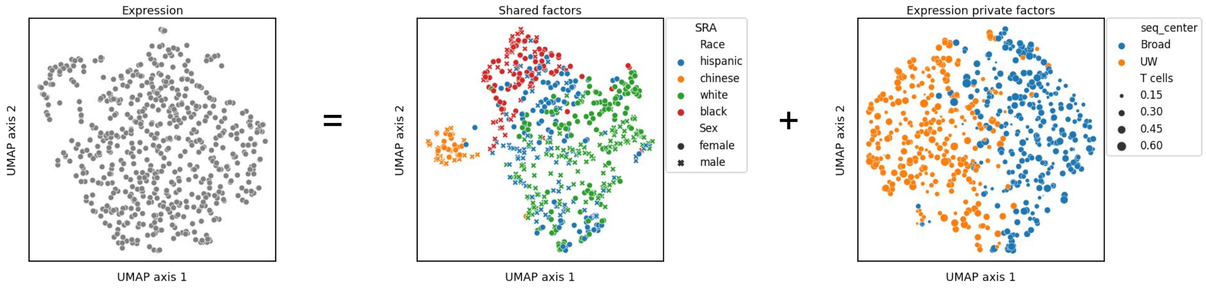
Check for updates

DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures

Monica del C. Gomez-Alonso[1,2†], Anja Kretschmer[1,2†], Rory Wilson[1,2], Liliane Pfeiffer[1,2], Ville Karhunen[3,10], Ilkka Seppälä[4], Weihua Zhang[3,5], Kirstin Mittelstraß[1,2], Simone Wahl[1,2], Pamela R. Matias-Garcia[1,2], Holger Prokisch[6,7], Sacha Horn[1,2], Thomas Meitinger[6,7,8], Luis R. Serrano-Garcia[1,2,9], Sylvain Sebert[10], Olli Raitakari[11,12,13], Marie Loh[3,14], Wolfgang Rathmann[15,16], Martina Müller-Nurasyid[17,18,30], Christian Herder[16,19,20], Michael Roden[16,19,20], Mikko Hurme[21], Marjo-Riitta Jarvelin[3,10,22,23], Mika Ala-Korpela[10,24,25,26], Jaspal S. Kooner[5,27,28,29], Annette Peters[2], Terho Lehtimäki[4], John C. Chambers[3,5,14,28,29], Christian Gieger[1,2], Johannes Kettunen[10,25,26] and Melanie Waldenberger[1,2,8*]

MCFA is a powerful tool for exploratory analysis

Observed = Shared + Private



- Unsupervised and exploratory
  - Interpret with caution
- MCFA remains in active development
  - Numerous additional applications!
- NHLBI TOPMed is among the most ambitious multi-omic data collection efforts
  - We look forward to future analyses with larger sample sizes

Thank you!

Development and analysis

- Collin Wang *
- Silva Kasela
- Francois Aguet
- Daniel Nachun
- Stephen Montgomery
- David Knowles *
- Tuuli Lappalainen *

MESA/TOPMED Collaborators

- Kent Taylor
- Russ Tracy
- Peter Durda
- Yongmei Liu
- W. Craig Johnson
- David Van Den Berg
- Namrata Gupta
- Stacy Gabriel
- Josh Smith
- Robert Gerzsten
- Clary Clish
- Quenna Wong
- George Papanicolau
- Jerome I. Rotter
- Stephen S. Rich

MESA NHLBI TOPMed NEW YORK GENOME CENTER® COLUMBIA UNIVERSITY