Large-scale imputation models for multi-ancestry proteome-wide association analysis

Chong Wu

Department of Biostatistics

The University of Texas MD Anderson Cancer Center

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

2024 Joint Statistical Meetings

August 6, 2024

Outline

- **Background**
- New method
- Results
- Extension

Causal inference in observational data

Identify causal biomarkers for a complex disease

Why:

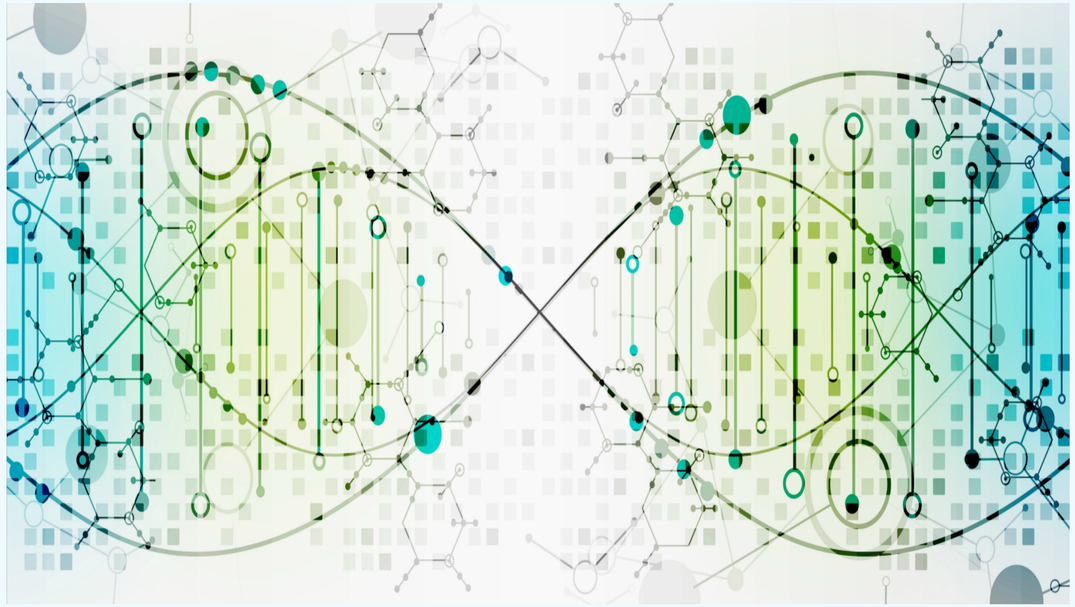
- understand the etiology
- drug development

Challenges:

- the number of biomarkers is large
- biomarkers are correlated

Goal:

identify likely causal biomarkers by using observational data



This figure is downloaded from Google Image

Identify likely causal gene expression

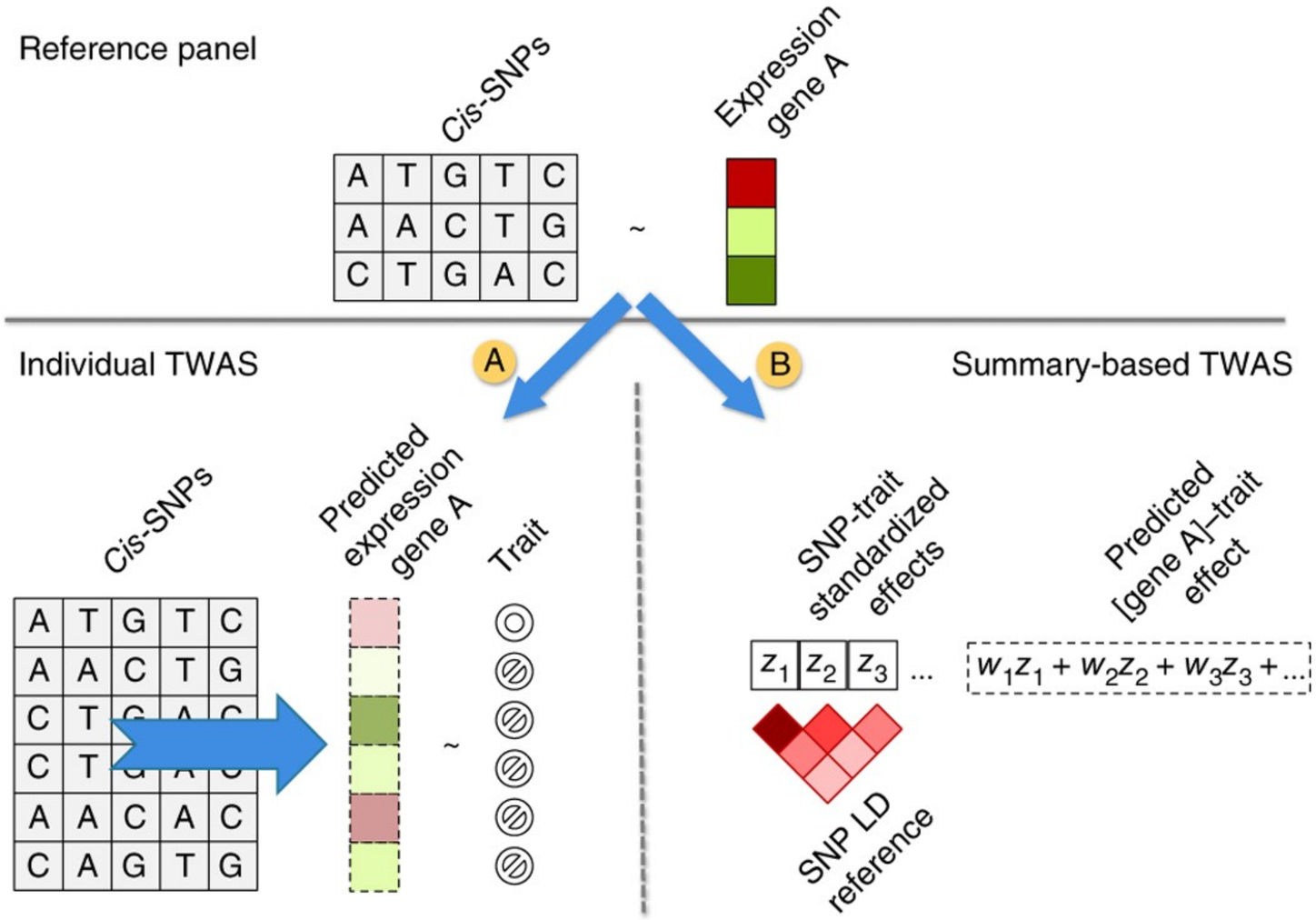
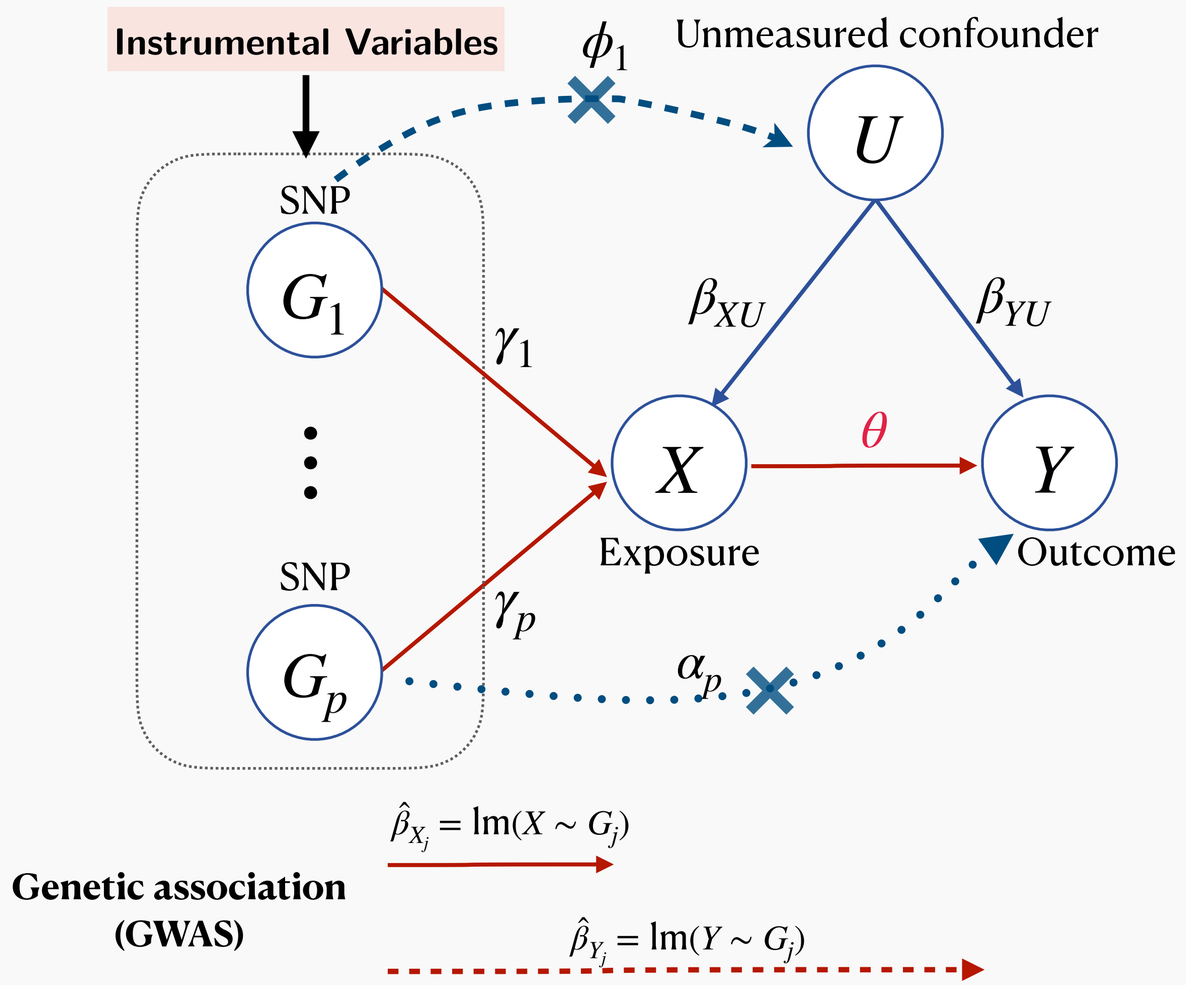


Figure: Workflow of TWAS[1]

1 Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." *Nature Genetics* 48.3 (2016): 245-252.

Mendelian randomization



Structure equation model:

$$\beta_{X_j} = \gamma_j + \phi_j \cdot \beta_{XU}$$

$$\beta_{Y_j} = \beta_{Y_j,\mathrm{M}} + \beta_{Y_j,\mathrm{D}} = \theta \cdot \beta_{X_j} + (\alpha_j + \phi_j \cdot \beta_{YU})$$

SNP j is a valid instrumental variable (IV) if

- **Relevance:** $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$

For a valid IV SNP j:

$$\beta_{X_j} = \gamma_j$$

$$\beta_{Y_j} = \theta \cdot \beta_{X_j}$$

Outline

- Background
- **New method**
- Results
- Extension

Three main challenges

- We only have the access to summary-level pQTL data for many large cohorts (e.g., deCODE and ARIC)
- It is hard to find the exact match independent validation/tuning dataset for protein prediction model building
- The sample size of non-European ancestry is currently relatively small

Build protein prediction model with summary-level data

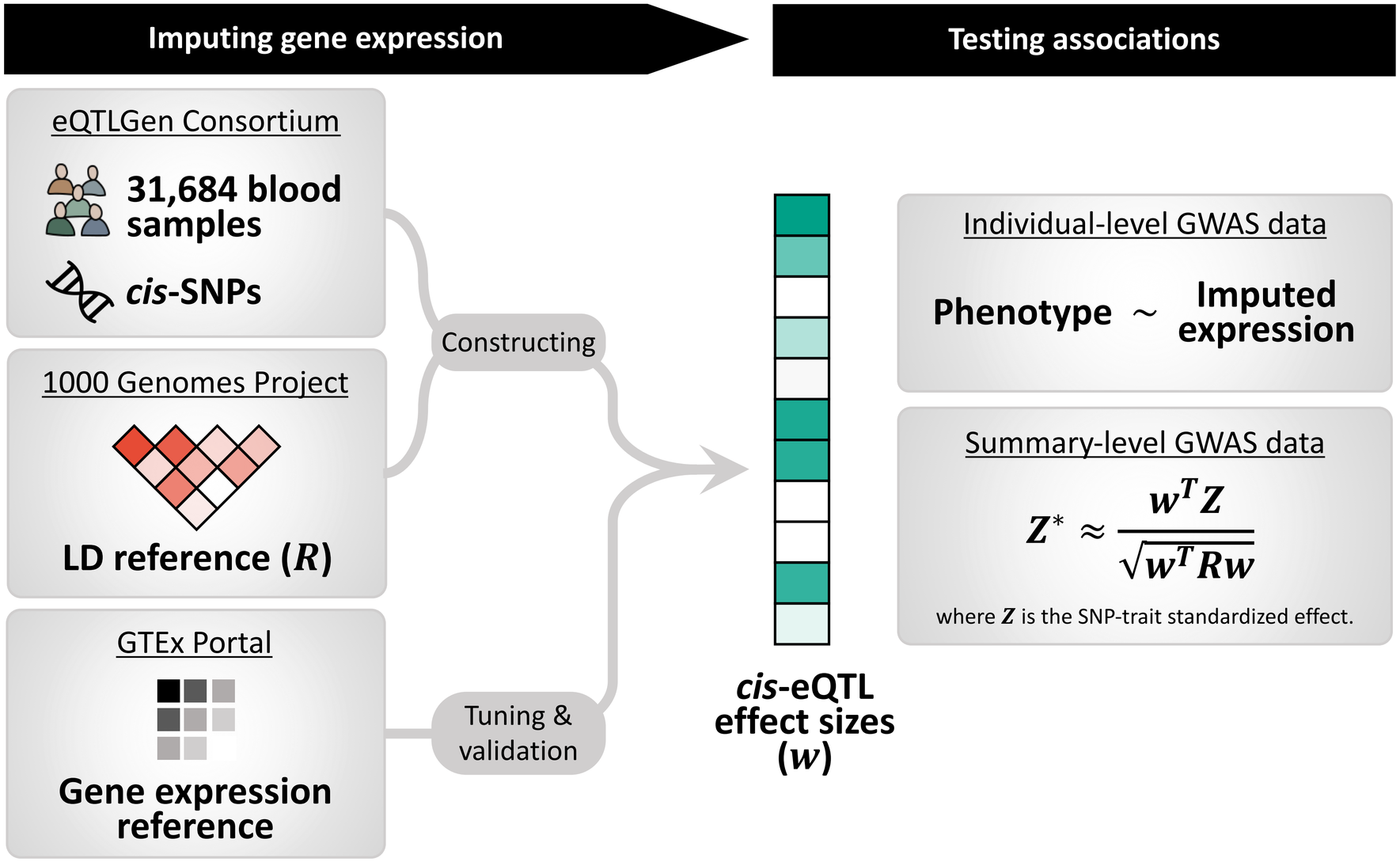


Figure: Workflow of SUMMIT

SUMMIT

Notation and model setup

$$\mathbf{Y} = \sum_{j=1}^{p} w_j \mathbf{X}_j + \epsilon$$

- $\mathbf{Y}$ is the gene expression levels; $\mathbf{X} = (\mathrm{X}_1^{'}, \cdots, \mathrm{X}_p^{'})'$ is the $N \times p$ standardized genotype matrix of p cis-SNPs around the gene; $\mathbf{w} = (w_1, \cdots, w_p)'$ is the cis-eQTL effect size, which can be estimated by

$$f(\mathbf{w}) = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{w})'(\mathbf{Y} - \mathbf{X}\mathbf{w})}{N} + J_\lambda(\mathbf{w}) = \frac{\mathbf{Y}'\mathbf{Y}}{N} + \mathbf{w}'\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)\mathbf{w} - 2\mathbf{w}'\frac{\mathbf{X}'\mathbf{Y}}{N} + J_\lambda(\mathbf{w})$$

SUMMIT

Notation and model setup

$$f(\mathbf{w}) = \frac{\mathbf{Y}'\mathbf{Y}}{N} + \mathbf{w}'\mathbf{R}\mathbf{w} - 2\mathbf{w}'\mathbf{r} + J_\lambda(\mathbf{w}),$$

Not depend on w

- $J_\lambda(\cdot)$ is a penalty term; such as LASSO, elastic net, MCP, SCAD, and MNet
- $\mathbf{r} = \mathbf{X}'\mathbf{Y}/N = (r_1, \cdots, r_p)'$ is p-dimensional vector of standardized marginal effect size for cis-SNPs (i.e., correlation between cis-SNPs and gene expression levels)
- $\mathbf{R} = \mathbf{X}'\mathbf{X}/N$ is the linkage disequilibrium (covariance) matrix of the cis-SNPs.
- The objective function is

$$\tilde{f}(\mathbf{w}) = \mathbf{w}'\tilde{\mathbf{R}}\mathbf{w} - 2\mathbf{w}'\tilde{\mathbf{r}} + \theta\mathbf{w}'\mathbf{w} + J_\lambda(\mathrm{w})$$

Ensure a unique solution upon optimization

Limitation in SUMMIT

1. We require a matched individual-level data to select the tuning parameters in SUMMIT, which are often hard to obtain

Solution: "**self-training**" of pQTL summary statistics: we generate independent pseudo-training and validation datasets for selecting tuning parameters

2. In Stage 2 test, standard TWAS/PWAS assumes that LD matrix estimated from the reference panel precisely matched that from the GWAS data

Solution: We explicitly consider the difference and use a slightly different formula to estimate the effect size

$$\hat{\gamma} = \frac{\hat{w}'Z/\sqrt{n_s}}{\sigma_r} \text{ and } \widehat{\text{Var}}\left(\hat{\gamma}\right) = \left(\frac{1}{n_s} + \frac{1}{n_r}\right)\hat{\gamma}^2 + \frac{\zeta^2}{n_s\sigma_r^2}$$

BLISS (Biomarker expression Level Imputation using Summary-level Statistics)

"self-training" of pQTL summary statistics: generate independent pseudo-training and validation datasets for selecting tuning parameter

- key idea was to sample marginal association statistics of pQTL data for a subset of individuals conditional on the complete summary-level pQTL data.
- We generated the pseudo-training data

$$G'_{(tr)}X_{(tr)} \mid G'X \sim \mathcal{N}\left(\frac{N-n}{N}G'X, \frac{N-n}{N}\Sigma\right)$$

- Obtain the pseudo-validation data

$$G'_{(v)}X_{(v)} = G'X - G'_{(tr)}X_{(tr)}$$

- We calculated the summary-level predictive R^2, which was the squared Pearson correlation coefficient between genetically predicted and directly measured protein expression levels, using the pseudo-validation data

Build non-European PWAS models with transfer learning

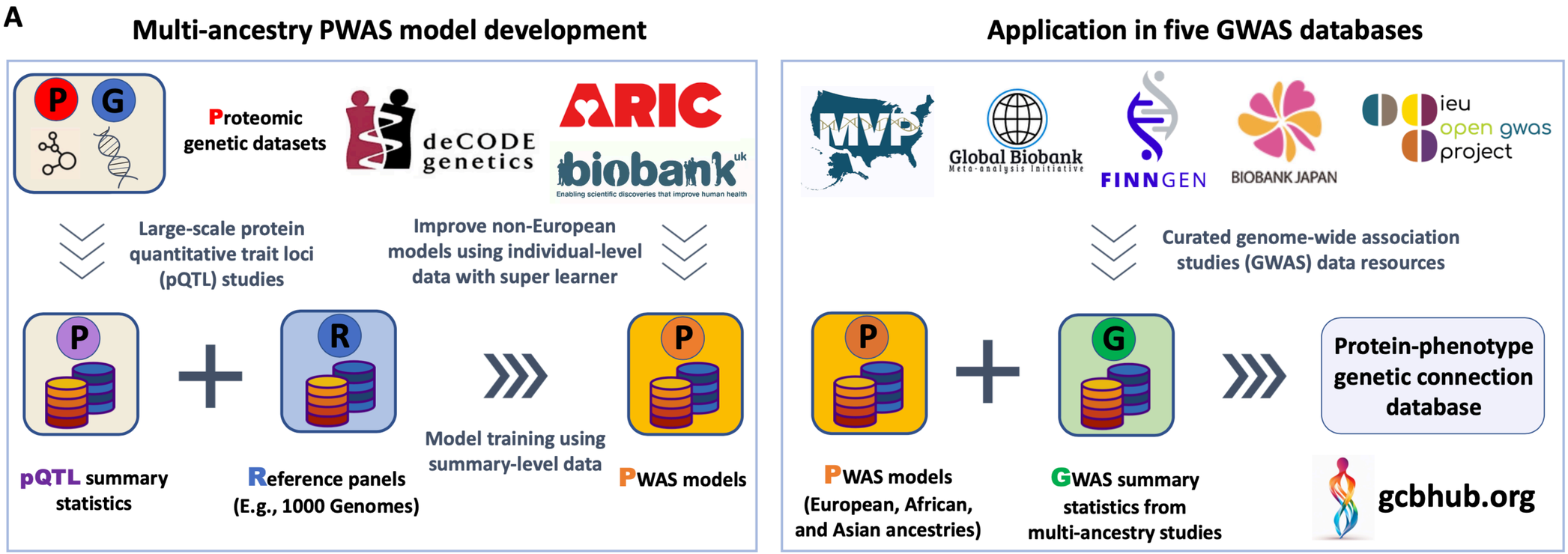
1. **Data: the individual-level UKB data of African and Asian ancestries**
2. **Methods: Super Learner Integration**

- We built the protein imputation model for each protein by Elastic-net using *cis*-SNPs
- Recognizing that the PWAS models for Europeans were built on much larger sample sizes and could potentially improve the prediction accuracy of Asian models, we applied super learner to combine the standard (Elastic-net) Asian models and BLISS-based European models
- We use non-negative least squares (NNLS) to produce a weighted sum of predictions from standard and BLISS models, where the weights were learned from nested five-fold cross-validation

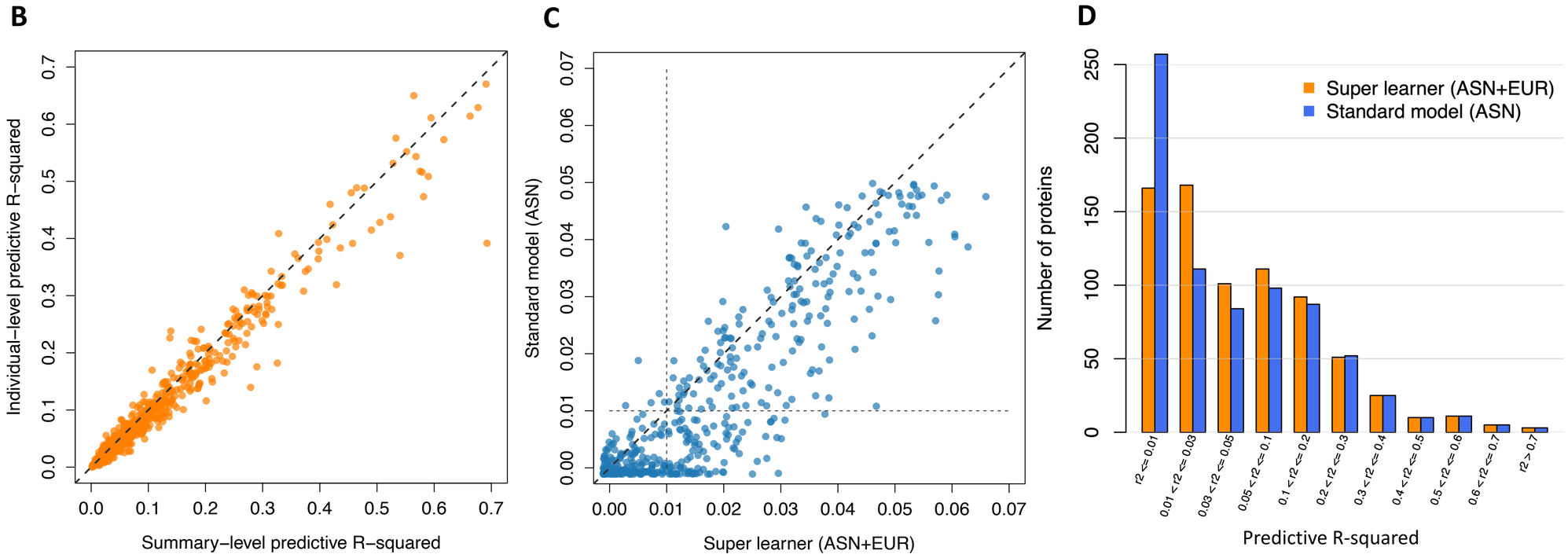
Outline

- Background
- New method
- **Results**
- Extension

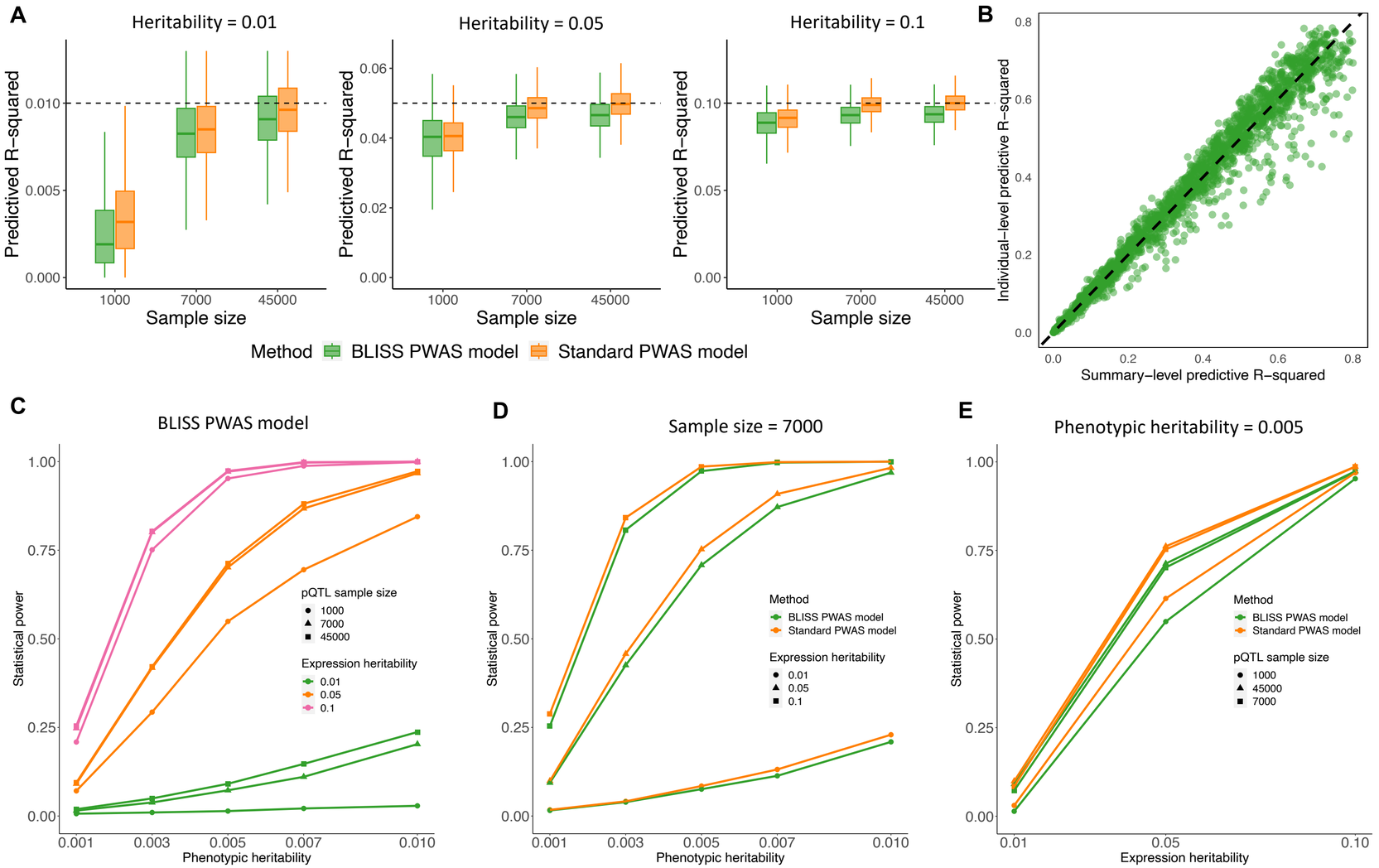
Overview of PWAS



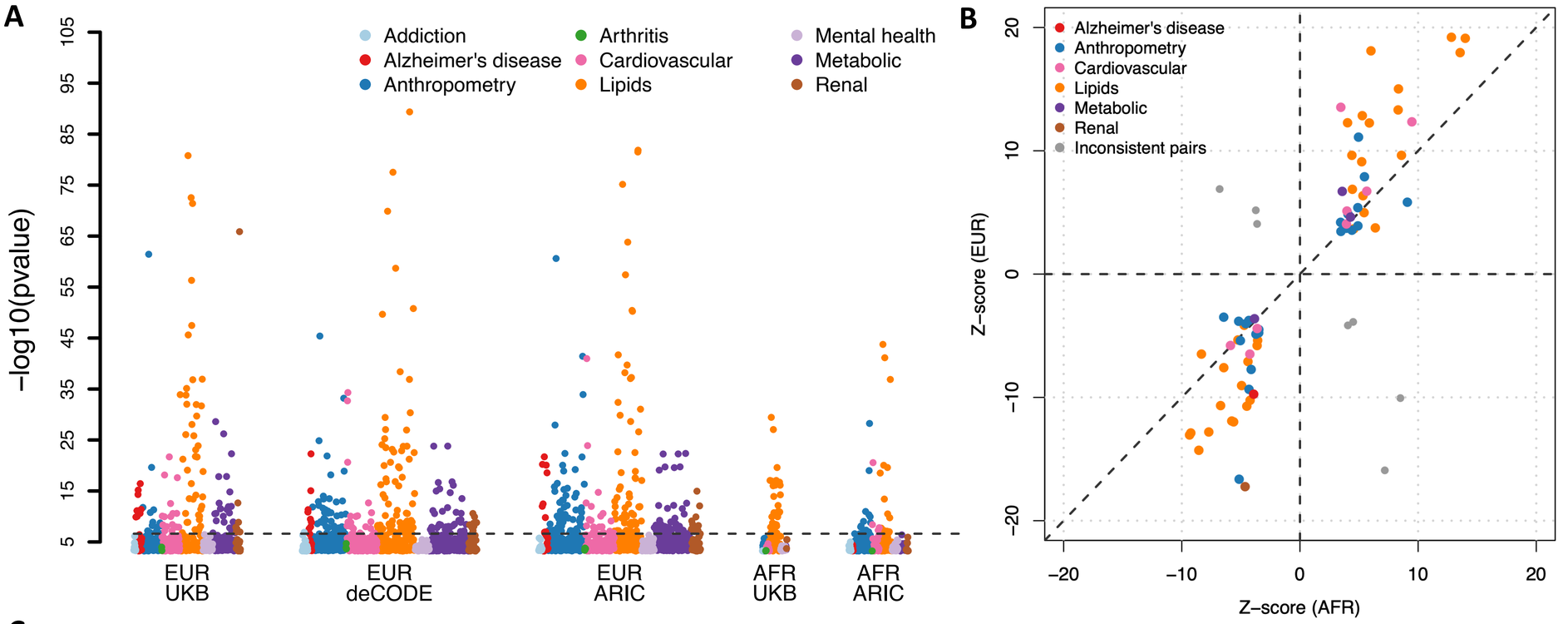
Results


B
Individual−level predictive R−squared
Summary−level predictive R−squared
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
C
Standard model (ASN)
Super learner (ASN+EUR)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
D
Number of proteins
Predictive R-squared
Super learner (ASN+EUR)
Standard model (ASN)
0
50
100
150
200
250
r2 <= 0.01
0.01 < r2 <= 0.03
0.03 < r2 <= 0.05
0.05 < r2 <= 0.1
0.1 < r2 <= 0.2
0.2 < r2 <= 0.3
0.3 < r2 <= 0.4
0.4 < r2 <= 0.5
0.5 < r2 <= 0.6
0.6 < r2 <= 0.7
r2 > 0.7

Simulation results



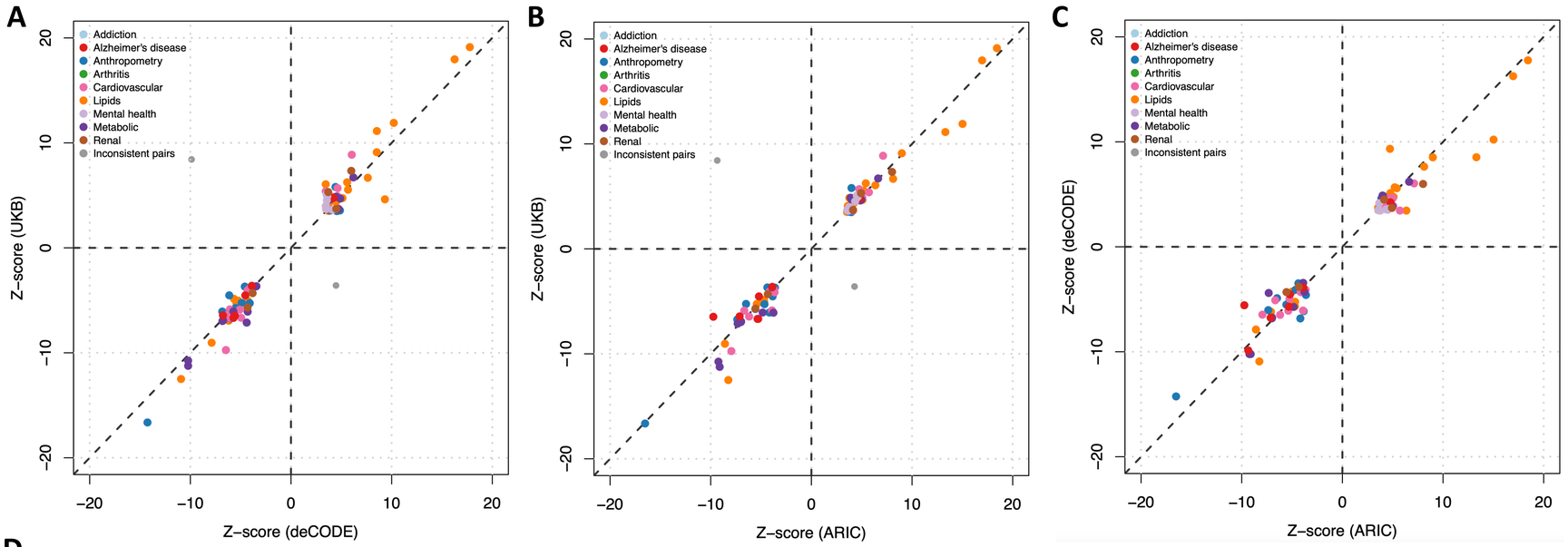
Transferability between African and European PWAS results



C

Trait(s)	Gene(s)	Beta	*P* value	Beta	*P* value	GWAS (GWAS Catalog)	Colocalization (PPH4 > 0.8)	pQTL MR (FDR < 0.05)
Alzheimer's disease	*TREM2*	β < -0.013	$P < 2.57\times10^{-8}$	β < -0.017	$P < 3.67\times10^{-4}$	*TREM2*	*TREM2*	*TREM2*
Venous thromboembolism (VTE)	*ABO, F11, PROC, PROS1, THBD*	|β| > 0.012	$P < 1.01\times10^{-5}$	|β| > 0.020	$P < 5.45\times10^{-4}$	*ABO, F11, THBD*	*ABO, F11*	*ABO, F11*
Varicose veins	*ABO*	β > 0.002	$P < 1.23\times10^{-4}$	β = 0.011	$P = 1.59\times10^{-8}$	*ABO*	*ABO*	*ABO*
Cardiovascular ideal health score (IHS)	*ERBB4, HP*	β > 0.005	$P < 4.87\times10^{-5}$	β > 0.020	$P < 8.40\times10^{-5}$		*HP*	*ERBB4, HP*
Non-alcoholic fatty liver disease (NAFLD)	*IL1RN*	β > 0.025	$P < 5.47\times10^{-10}$	β = 0.076	$P = 3.46\times10^{-4}$			*IL1RN*
Type 2 diabetes (T2D)	*MSR1, TREML2*	|β| > 0.008	$P < 2.92\times10^{-4}$	|β| > 0.011	$P < 6.08\times10^{-4}$	*TREML2*		*MSR1, TREML2*

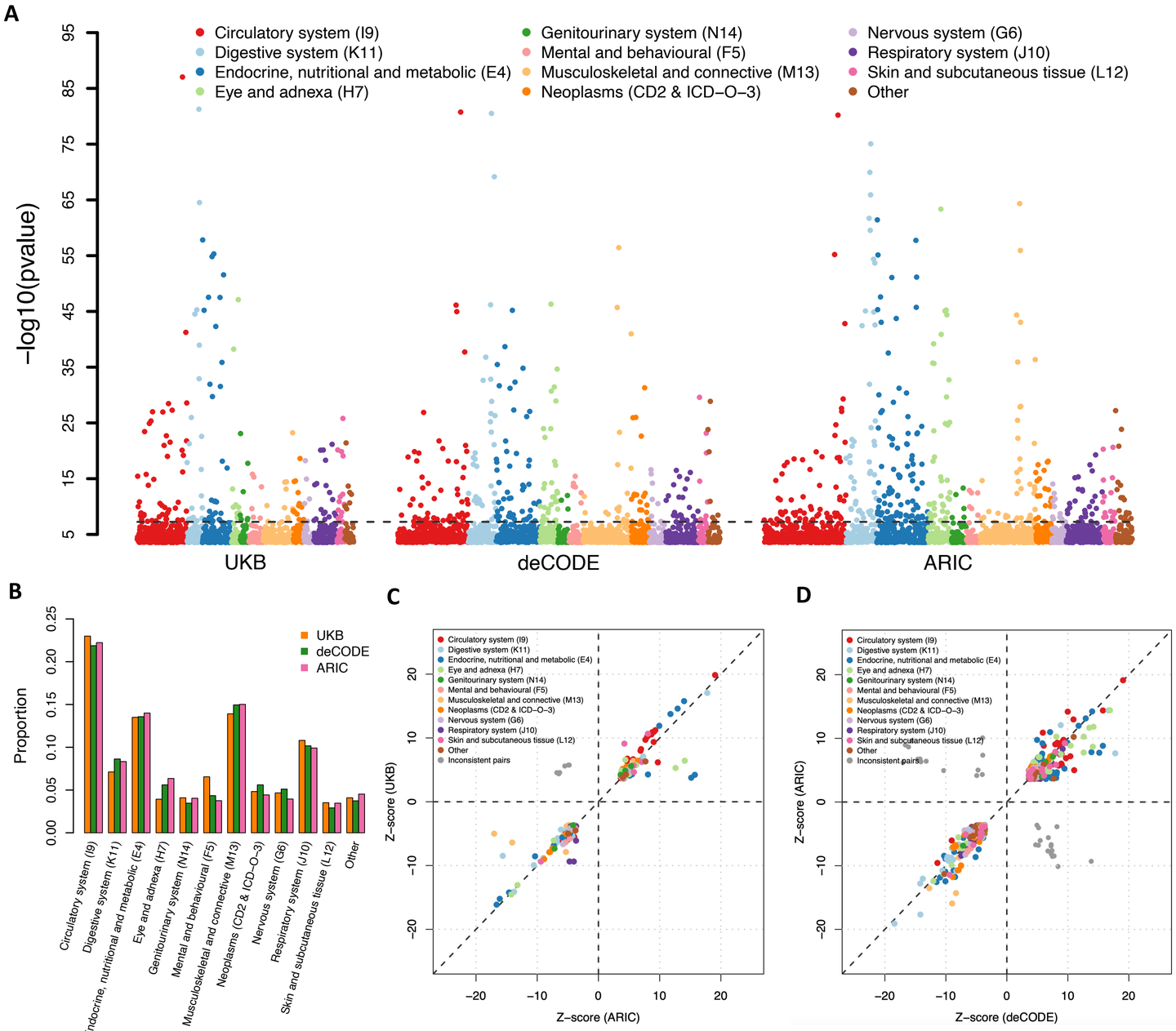
Consistent PWAS findings across reference proteomic datasets



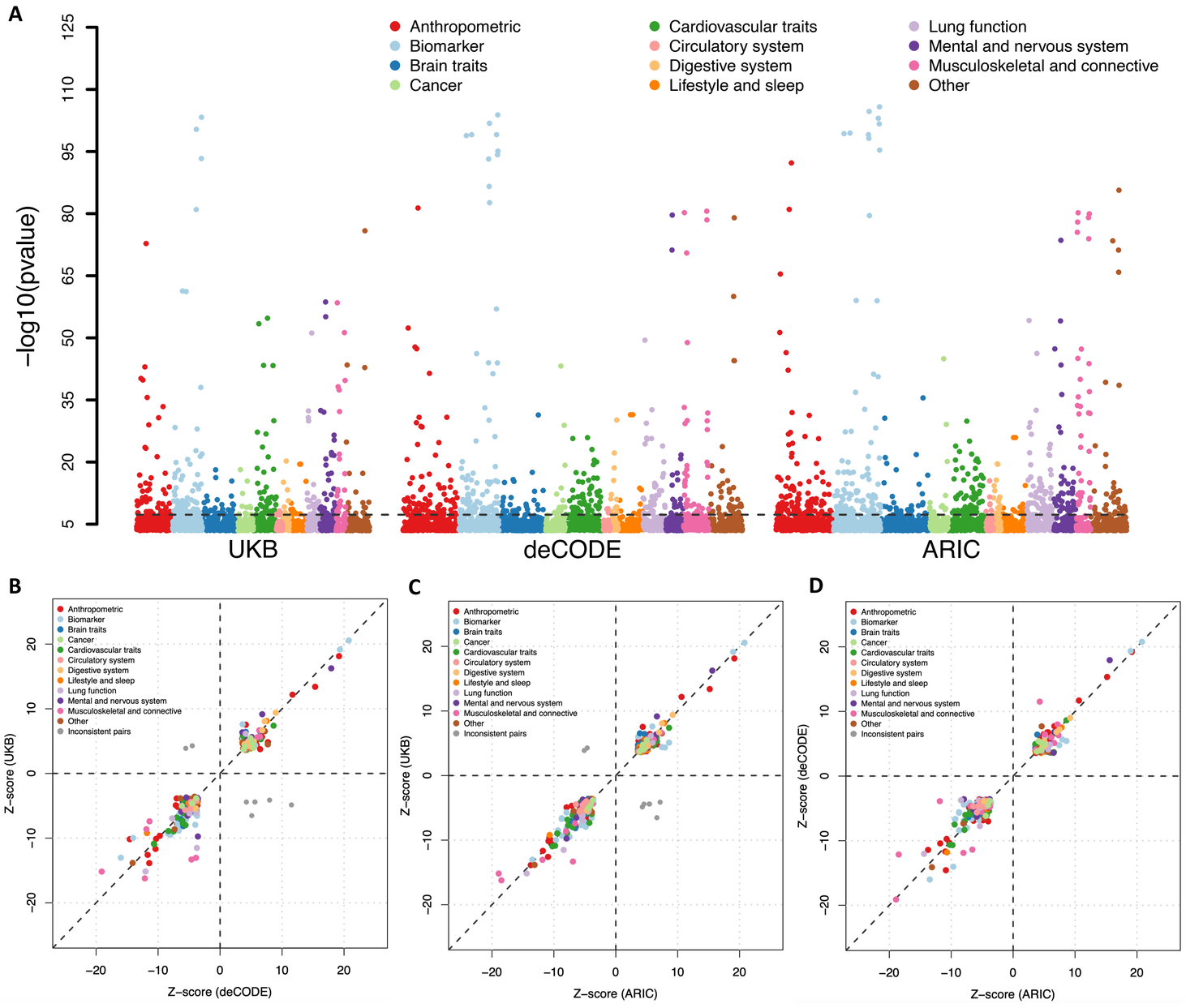
D

Trait(s)	Gene(s)	Beta	*P* value	Beta	*P* value	Beta	*P* value	GWAS (GWAS Catalog)	Colocalization (PPH4 > 0.8)	pQTL MR (FDR < 0.05)
Alzheimer's disease	*BCAM, CD55, EPHB4, GRN, LILRB1, SIRPA, TREM2*	|β| > 0.005	$P < 2.95\times10^{-4}$	|β| > 0.005	$P < 1.03\times10^{-4}$	|β| > 0.002	$P < 1.00\times10^{-4}$	*BCAM, CD55, EPHB4, GRN, TREM2*	*GRN, TREM2, SIRPA*	*BCAM, CD55, EPHB4, GRN, LILRB1, SIRPA, TREM2*
Post-traumatic stress disorder (PTSD)	*CTSF, CTSV, CD14*	β > 0.036	$P < 2.83\times10^{-4}$	β > 0.022	$P < 3.64\times10^{-4}$	β > 0.014	$P < 5.59\times10^{-4}$		*CTSF*	*CTSF, CTSV, CD14*
Coronary artery disease (and angiographic burden)	*IL6R, PCSK9, SPARCL1*	|β| > 0.016	$P < 1.09\times10^{-4}$	|β| > 0.023	$P < 2.56\times10^{-4}$	|β| > 0.008	$P < 4.15\times10^{-5}$	*IL6R, PCSK9*	*IL6R, PCSK9*	*IL6R, PCSK9*
Peripheral artery disease	*C2, MMP12*	|β| > 0.006	$P < 7.05\times10^{-8}$	|β| > 0.012	$P < 1.09\times10^{-8}$	|β| > 0.007	$P < 5.57\times10^{-4}$		*MMP12*	*C2, MMP12*
Venous thromboembolism (VTE)	*GP6, NPPB, OBP2B*	|β| > 0.018	$P < 1.01\times10^{-6}$	|β| > 0.012	$P < 9.28\times10^{-5}$	|β| > 0.019	$P < 9.85\times10^{-5}$		*GP6, NPPB*	*GP6, NPPB*
Varicose veins	*FABP2, RSPO3, TNFSF12*	|β| > 0.026	$P < 2.82\times10^{-4}$	|β| > 0.008	$P < 3.63\times10^{-5}$	|β| > 0.002	$P < 5.85\times10^{-5}$	*RSPO3*	*RSPO3*	*FABP2, RSPO3, TNFSF12*
Cardiovascular ideal health score (IHS)	*PCSK9, ENTPD6*	β < -0.025	$P < 5.16\times10^{-6}$	β < -0.041	$P < 3.42\times10^{-5}$	β < -0.024	$P < 1.55\times10^{-5}$		*PCSK9, ENTPD6*	*PCSK9, ENTPD6*
Non-alcoholic fatty liver disease (NAFLD)	*APOH, FCRLB, IL1RN, RSPO3, SPON1**	|β| > 0.009	$P < 2.68\times10^{-6}$	|β| > 0.017	$P < 9.48\times10^{-5}$	|β| > 0.009	$P < 1.08\times10^{-5}$	*RSPO3*	*APOH*	*APOH, FCRLB, IL1RN, RSPO3*
Type 2 diabetes (T2D)	*MLN, NCAN, NELL1, PAM, AGER, BST1*	|β| > 0.005	$P < 2.53\times10^{-4}$	|β| > 0.004	$P < 1.65\times10^{-4}$	|β| > 0.002	$P < 5.87\times10^{-4}$	*NELL1, PAM*	*NELL1, PAM*	*MLN, NCAN, NELL1, PAM, AGER, BST1*

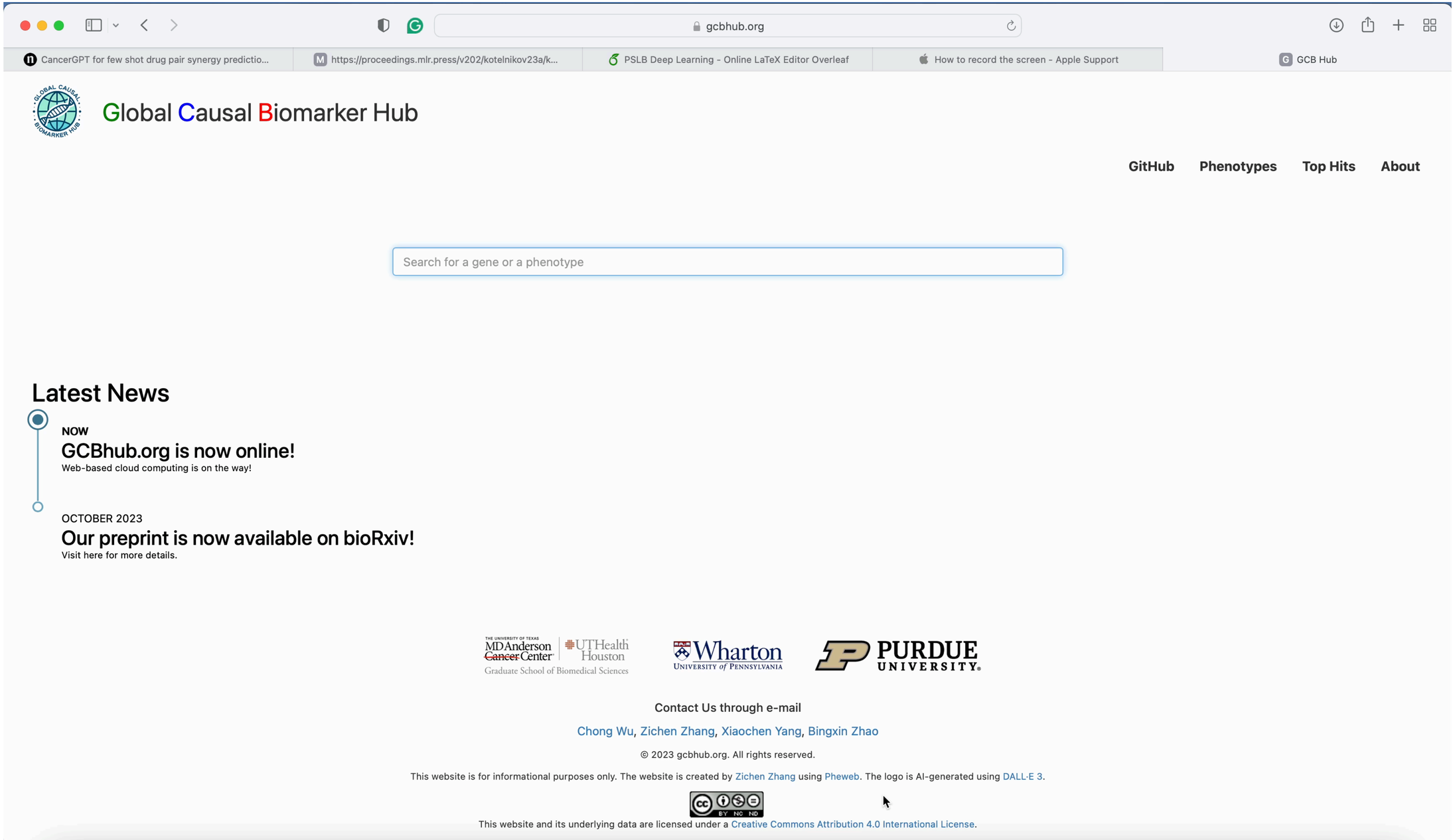
FinnGen data analysis



MRC IEU OpenGWAS data analysis



https://www.gcbhub.org



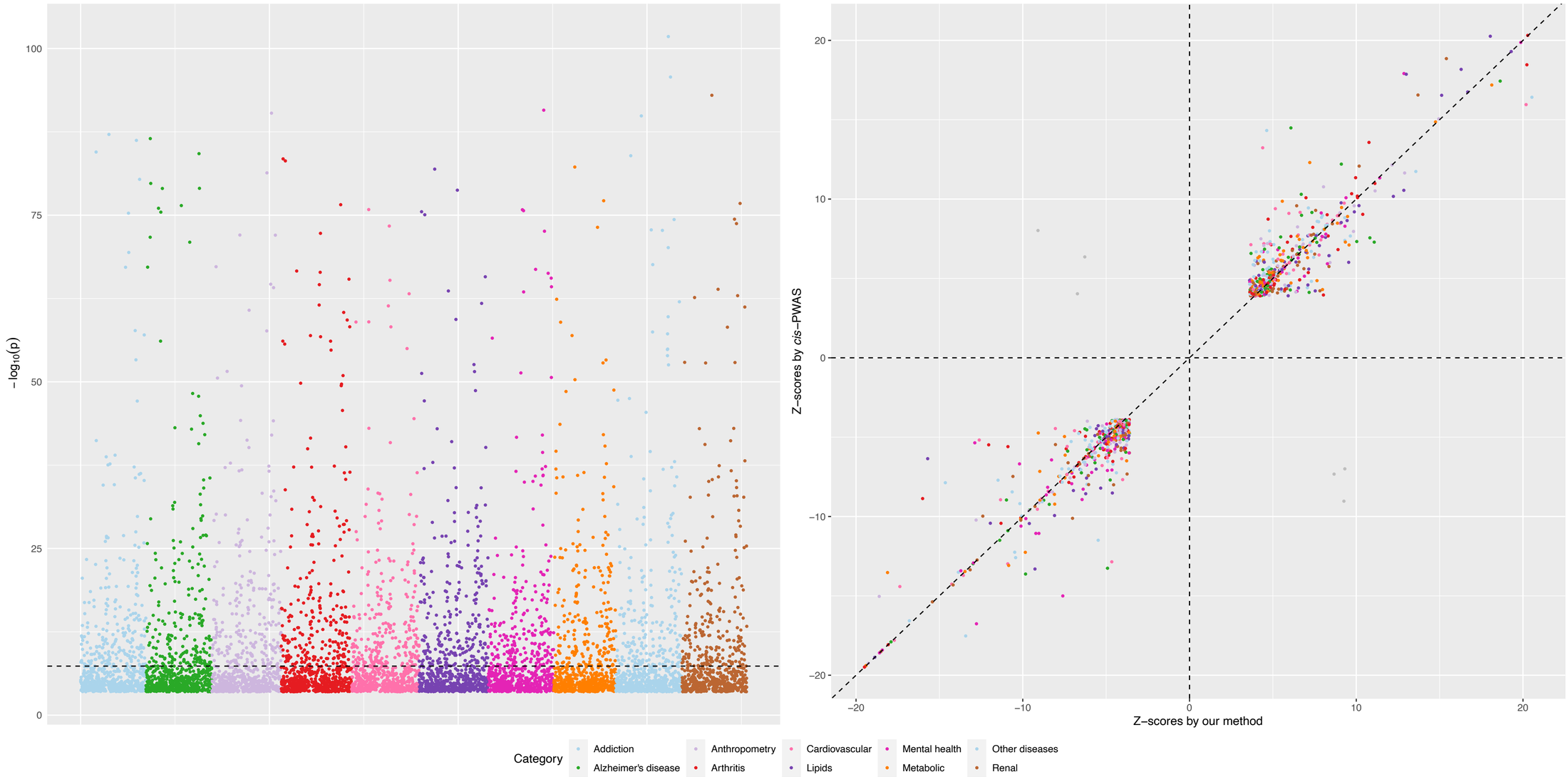
Future and Ongoing work

- BLISS can be extended to other omics data: single-cell TWAS
- TWAS/PWAS methods, including SUMMIT/BLISS, can be viewed as one type of gene-based Mendelian randomization (MR) and can provide valid causal interpretations only when all genetic variants used in the expression prediction models are valid instrumental variables **(Strong and uncheckable assumption)**
- Non-linearity: deep learning model
- **Trans-acting elements:** how to incorporate information from trans regions (many challenges, including weak signals, pleiotropy effects, etc.)
- Multi-ethnicity: Improve the robustness and performance (transfer learning)

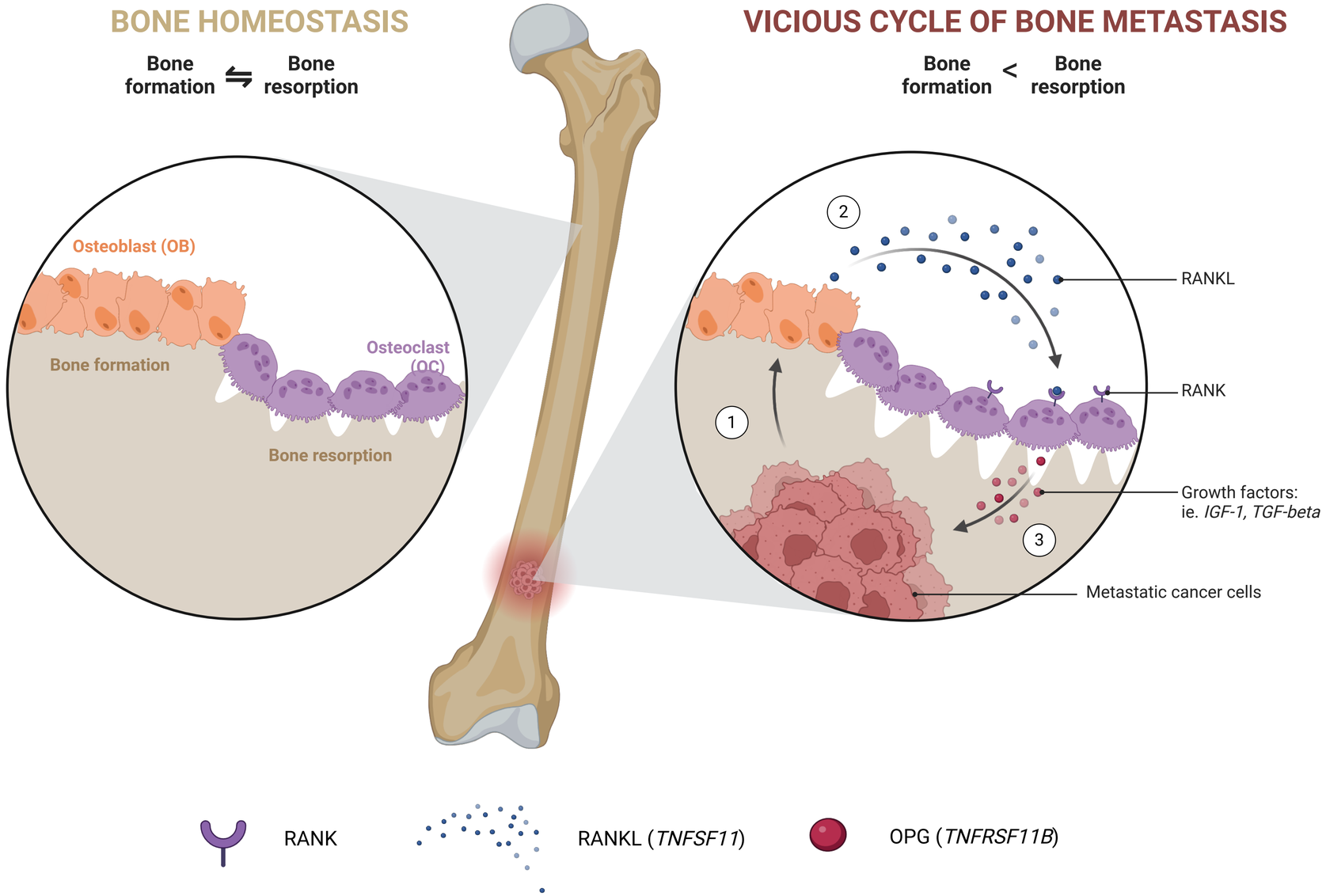
Future and Ongoing work

- BLISS can be extended to other omics data: single-cell TWAS
- TWAS/PWAS methods, including SUMMIT/BLISS, can be viewed as one type of gene-based Mendelian randomization (MR) and can provide valid causal interpretations only when all genetic variants used in the expression prediction models are valid instrumental variables **(Strong and uncheckable assumption)**
- Non-linearity: deep learning model
- **Trans-acting elements:** how to incorporate information from trans regions (many challenges, including weak signals, pleiotropy effects, etc.)
- Multi-ethnicity: Improve the robustness and performance (transfer learning)

Trans results



Trans results



References

SUMMIT:

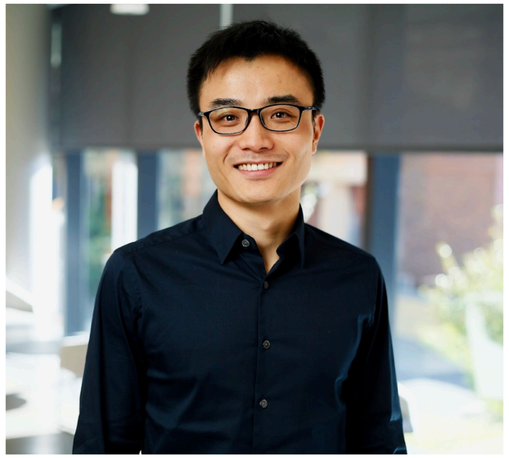
Zhang, Zichen, Ye Eun Bae, Jonathan R. Bradley, Lang Wu, and Chong Wu. "SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification." *Nature Communications* **13**, 6336 (2022).

BLISS:

Wu, Chong, Zichen Zhang, Xiaochen Yang, and Bingxin Zhao. "Large-scale imputation models for multi-ancestry proteome-wide association analysis." *bioRxiv* (2023): 2023-10.

GUB-Hub: https://www.gcbhub.org/

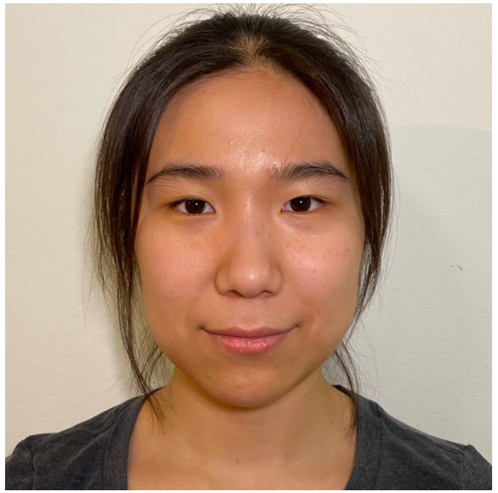
Acknowledgements



Bingxin Zhao @ Upenn

Lang Wu @ Hawaii

Jon Bradley @ FSU



Xiaochen Yang @ Purdue

Zichen Zhang @MDA

Ye Eun Bae @ FSU

Thank Aditya and Chongliang for invitation!

NIH

biobank uk

Thank you!

Chong Wu

Email: cwu18@mdanderson.org

Website: https://wuchong.org