Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Database for ancestry-specific cross-tissue gene expression models: AGEMdb

Inti Pagnuco

inti.pagnuco@manchester.ac.uk

University of Manchester

Francisco Herrerías-Azcué

University of Manchester

Stephen Eyre

University of Manchester

Magnus Rattray

University of Manchester

Andrew P Morris

University of Manchester

---

database

Keywords:

Posted Date: June 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10068853/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Bioinformatics on November 6th, 2026. See the published version at https://doi.org/10.1186/s12859-026-06743-2.

Abstract

Background

Transcriptome-wide association studies (TWAS) integrate genetically regulated gene expression with genome-wide association study (GWAS) data to identify gene–trait associations. These analyses rely on accurate gene expression prediction models trained using genotype and transcriptomic data from the same set of individuals. However, most publicly available TWAS models are based on resources that predominantly consist of participants of European ancestry. This European ancestry bias limits the transferability and generalizability of gene expression models across diverse populations.

Description

Using genotype and gene expression data from the Genotype-Tissue Expression (GTEx) Project, we curated datasets comprising 689 European ancestry individuals and 111 African American individuals after quality control and confounder adjustment. We then developed ancestry-specific cross-tissue gene expression imputation models using the UTMOST framework. These models are accessible through AGEMdb, a web-based database that allows users to browse, search, and download gene expression imputation models with flexible filtered options (e.g. by tissue and gene). By incorporating African American-specific models, AGEMdb addresses the European ancestry bias present in existing TWAS resources and supports more equitable gene–trait association analyses.

Conclusion

AGEMdb provides a valuable resource for ancestry-specific TWAS, which promotes more inclusive studies that support the advancement of equitable genomic research.

1. BACKGROUND

Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with complex human traits and diseases. However, association signals at most of these loci (~ 90%) are driven by variants that map to non-coding regions of the genome [1], making it difficult to interpret their functional consequences or determine the effector genes through which their effects are mediated.

To address this challenge, transcriptome-wide association studies (TWAS) have emerged as a powerful strategy to prioritize candidate genes underlying GWAS signals. TWAS integrates GWAS data with gene expression imputation models trained on reference datasets containing both genotype and transcriptomic information. These models capture the regulatory effects of expression quantitative trait loci (eQTLs), enabling researchers to estimate genetically-regulated gene expression levels in independent cohorts based only on genotype data. TWAS then tests for associations between predicted

gene expression and complex traits, offering gene-level insights that complement traditional GWAS findings.

The performance of TWAS critically depends on the accuracy of the gene expression imputation models. Several statistical methods have been developed to train these models using eQTLs [2–4], primarily focusing on evaluating the additive effects of cis-eQTLs —variants that are located within 1Mb of the gene they regulate. While many TWAS methods build prediction models separately for each tissue, several cross-tissue approaches have been developed to improve gene expression prediction by leveraging shared genetic regulation [4–6]. These methods are motivated by the observation that many regulatory mechanisms are shared across tissues [7], allowing information from correlated tissues to enhance model performance. Among these approaches, UTMOST ("Unified Test for MOlecular SignaTures" [4]) jointly models gene expression across tissues, resulting in improved imputation performance and enhanced detection of shared regulatory signals compared with single-tissue approaches [4, 8, 9].

Despite these methodological advancements, a significant limitation of TWAS lies in its reliance on reference transcriptomic datasets that lack population diversity. For example, the widely used Genotype-Tissue Expression (GTEx) Project [7] consists predominantly of individuals of European ancestry (84.9%). As a result, gene expression models built from GTEx—such as those provided by UTMOST [4], PrediXcan [2], and Fusion [3]—are primarily tailored to European ancestry populations. While these models have substantially advanced genetic research, their utility in non-European ancestry groups remains limited.

Several studies have demonstrated that gene expression prediction models trained exclusively on European ancestry data perform poorly in other populations [9–14]. This challenge reflects similar issues in GWAS portability, where findings from European ancestry populations often fail to replicate in other ancestry groups due to differences in allele frequencies and linkage disequilibrium (LD) patterns [15–18]. This ancestry-specific genetic architecture similarly impacts the generalisability of eQTL models and downstream TWAS results.

Although African Americans are the second most represented ancestry group in GTEx (12.6%), dedicated pre-trained gene expression models for this population remain largely unavailable. To address this gap, we developed the database for Ancestry-specific cross-tissue Gene Expression Models (AGEMdb). Built using the UTMOST framework and high-quality GTEx data, AGEMdb provides cross-tissue gene expression imputation models for the two most represented ancestry groups—European and African American. These ancestry-specific models enable more accurate and equitable TWAS [9], particularly for African American GWAS. Featuring broad gene coverage, customizable filtering by gene and tissue, and a user-friendly web interface, AGEMdb represents a critical step toward reducing disparities in genomic research.

2. CONSTRUCTION AND CONTENT

AGEMdb is a specialized resource that provides gene expression imputation models tailored for individuals of European and African American ancestry—the two largest ancestry groups represented in the GTEx dataset. It is designed to support TWAS analyses in these ancestry groups (Fig. 1). TWAS typically involves two main steps: (1) training models to impute gene expression from genotype data, and (2) identifying gene–trait associations by integrating gene expression imputation models with individual level genotype/phenotype data or GWAS summary statistics and SNP dosage covariance matrices.

2.1. Genotype and Gene Expression Data Processing

Gene expression imputation models were trained using paired genotype and gene expression data collected from the same individuals in the GTEx Project, accessed via dbGaP (accession number phs000424.v9.p2). Detailed quality control procedures are described in our previous work [9]. Briefly, genotype data from 953 individuals were analyzed after excluding samples with a genotype call rate below 99.9%. The distribution of autosomal heterozygosity was bimodal, consistent with the presence of two major ancestral groups in the cohort. Heterozygosity thresholds were selected based on the observed distribution and validated against self-reported ancestry information, identifying a European ancestry group (heterozygosity 0.045–0.050) and an African American group (0.060–0.080).

Within each ancestry group, we removed single nucleotide variants (SNVs) with a call rate below 90% or Hardy-Weinberg equilibrium exact P-values less than 10^{-6}. Principal component analysis (PCA) was performed using PLINK v1.9, following the exclusion of SNVs with minor allele frequency (MAF) below 5%, removal of regions with high LD, and LD pruning ($r^2 < 0.05$). Consistent with the quality-control framework described previously [9], ancestry outliers were identified and excluded based on the first two principal components, which accounted for the major axes of genetic variation within each ancestry group. This process resulted in a final dataset comprising 689 individuals of European ancestry and 111 individuals of African American ancestry.

The GTEx dataset includes normalized gene expression data from 49 tissues, which were adjusted to account for potential confounding effects, including sex, sequencing platform (HiSeq2000 or HiSeqX), library construction protocol (PCR-based or PCR-free), the first two genotype principal components, and the top probabilistic estimation of expression residuals (PEER) factors [19]. Following previous recommendations [7], the number of PEER factors included varied according to tissue sample size: 15 for tissues with fewer than 150 samples, 30 for tissues with 150–249 samples, 45 for tissues with 250–349 samples, and 60 for tissues with at least 350 samples.

2.2. Cross-Tissue Gene Expression Prediction Model training

We trained cross-tissue expression imputation models using UTMOST [4] separately for each ancestry group. Models were trained for 38,539 autosomal genes using data from 49 GTEx tissues. For each

gene, SNVs located within 1 Mb of the transcription start and stop sites were included as predictors; excluding variants with ambiguous alleles (AT or GC) or a MAF below 1%.

In the European ancestry cohort, sample sizes per tissue ranged from 58 (kidney cortex) to 514 (skeletal muscle) (Sup. Table 1), resulting in 35,090 gene expression imputation models across 49 tissues, each containing at least one SNP with a non-zero effect size in at least one tissue. For the African American cohort, sample sizes ranged from 7 to 77 across 44 tissues (Sup. Table 1), resulting in 34,918 models meeting the same criterion. No models were generated for five tissues in the African American dataset (brain amygdala, brain anterior cingulate cortex BA24, brain hypothalamus, brain spinal cord cervical C1, and kidney cortex), due to insufficient sample size (Sup. Table 1). Model training was performed using the default UTMOST pipeline (https://github.com/yiminghu/CTIMP).

2.3. AGEMdb Infrastructure and Implementation

AGEMdb integrates these ancestry-specific models into a publicly accessible, cross-tissue TWAS database. The main data processing pipeline was implemented in Python, which parses UTMOST output matrices, extracts SNP effect sizes, counts non-zero predictors, and maps variant identifiers to gene names and rsIDs. R was used for summary statistics generation and model performance evaluation. Data are stored in a PostgreSQL database via SQLAlchemy, with schema validation handled by Pydantic.

The web interface was built using Flask (Python), HTML with Jinja templating, and styled with Tailwind CSS. Minimal JavaScript enables user interactivity. AGEMdb is hosted by the University of Manchester, providing a secure and scalable platform for public access. Separate portals are available for different ancestries: European ancestry data can be accessed at https://agemdb-eur.manchester.ac.uk/ and African American ancestry data at https://agemdb-afa.manchester.ac.uk/.

2.4. AGEMdb Database Schema: Gene Expression Model Outputs and Performance Metadata

The database contains three main tables that store gene expression model outputs along with related metadata.

The “Weights” table contains parameter estimates for the pre-trained models, reporting the effect sizes of significant SNPs on gene expression across different tissues. Each entry details how a specific genetic variant influences gene expression, including the SNP identifier, the corresponding gene, the effect size (weight), and the tissue to which that effect size applies. This table facilitates the investigation of how SNVs drive tissue-specific gene regulation.

Building on this, the “Gene Level Metadata” table provides a summary of key characteristics for each gene model. It records the number of non-zero SNPs (i.e., SNPs with non-zero weights that significantly affect gene expression) included in the model, alongside the average number of samples from the tissues used during training. It also reports performance metrics derived from 5-fold cross-validation, including the average R^2—calculated as the square of the Pearson correlation coefficient between

predicted and observed expression levels—which reflects the proportion of variance explained by the model, and the average Mean Squared Error (MSE) across tissues. In cases where a model trained for a given fold results in all-zero weights—indicating no detectable predictive signal—these metrics may be reported as NaN. A detailed benchmarking and cross-ancestry comparison of model performance has been previously reported in [9]. Additionally, this table includes links to the “Gene Tissue Metadata” table, which provides detailed information for each gene-tissue pair, reporting the number of training samples and evaluation metrics specific to each tissue. This enables a refined analysis of gene expression modelling across diverse biological contexts.

2.5. Covariance Resources for Single- and Cross-Tissue TWAS

AGEMdb provides pre-computed covariance matrices of the SNP dosages required for downstream TWAS analyses using UTMOST or similar methods that rely on GWAS summary statistic information to test gene–trait associations. These matrices, generated from the ancestry-specific genotype data used for model training, cover all 49 European ancestry tissues and 44 African American tissues with trained gene expression imputation models. Alternatively, users can generate custom covariance matrices using reference genotype panels tailored to their specific study populations or analytical frameworks.

3. UTILITY AND DISCUSSION

AGEMdb is a web-based resource that provides ancestry-specific gene expression imputation models for use in TWAS. It addresses the long-standing underrepresentation of African American individuals in existing imputation frameworks by generating models specifically trained on this ancestry group, alongside those for European ancestry. Using a cross-tissue training approach, AGEMdb enables tissue-specific gene expression imputation from genotype data, supporting gene-trait association analyses when direct expression measurements are unavailable.

3.1. Gene-Trait Association Analysis using AGEMdb

TWAS improves the interpretation of GWAS results by refining association signals, prioritizing candidate causal genes, and elucidating tissue-specific regulatory mechanisms underlying complex traits. AGEMdb supports TWAS analysis by providing pre-trained, ancestry-specific models that can be applied to impute tissue-specific gene expression using individual-level genotype data or be directly integrated with GWAS summary statistics to uncover expression–trait associations.

The models in AGEMdb were developed using the UTMOST framework, which employs GWAS summary statistics, cross-tissue derived gene expression imputation models and SNP dosage covariance matrices to improve TWAS power. Despite this, the models are fully compatible with other popular TWAS tools, such as S-PrediXcan (also requiring GWAS summary statistics) and PrediXcan (which requires individual-level genotype/phenotype data), enabling straightforward integration into existing analysis pipelines.

By reducing the need for computationally intensive model training and offering a user-friendly interface, AGEMdb increases accessibility—especially for researchers studying ancestrally diverse populations or those without access to tissue-specific expression data.

3.2. Download Options and Flexible Filtering Tools

AGEMdb offers an intuitive web interface that allows researchers to easily access and download data. Users can obtain the full database in .db (SQLite) format, which includes both the “Weights” and “Gene Level Metadata” tables, or download filtered datasets based on specific criteria such as tissue, gene, or SNP attributes. Additionally, individual tables are available for download in .csv or .json formats.

By default, the “Weights” table is organized by tissue, producing one file per tissue. However, this tissue-based grouping can be disabled to enable downloading of the complete dataset as a single file without tissue separation.

All AGEMdb tables offer flexible filtering through the web interface, allowing users to customize queries with exact matches, exclusions, numeric ranges, and combined multiple conditions. For example, filters in the “Weights” table include SNP ID (rsID), gene name, tissue, weight, performance metrics, and sample count, enabling selections such as specific tissues or value ranges. Numeric filters support thresholds and intervals (e.g., values between 0.5 and 0.7). Advanced users can utilize URL-based queries to access all columns, including gene IDs, alleles, and chromosome positions, facilitating tailored data extraction for diverse research purposes.

3.3. Discussion and Implications for TWAS Research

TWAS have enhanced our ability to link genetic variants to complex traits by integrating gene expression predictions with GWAS data. However, most TWAS rely on reference datasets that predominantly consist of individuals of European ancestry. This lack of diversity limits the accuracy and generalizability of gene expression models across populations with distinct genetic architectures and linkage disequilibrium patterns. Recognizing that African Americans represent the second largest ancestry group in GTEx, AGEMdb addresses this critical gap by developing ancestry-specific gene expression models. Leveraging the UTMOST framework, we constructed cross-tissue imputation models tailored to both European and African American individuals, providing a valuable resource to enable more equitable and precise genetic analyses. The importance of this strategy is underscored by previous evaluations of TWAS performance in under-represented populations, which demonstrated substantially improved transferability of gene expression prediction models within the same ancestry group than between ancestry groups [14]. In addition, ancestry-specific models outperform those trained on multi-ancestry data in ancestry-matched evaluations, despite the larger sample sizes achieved by combining data across ancestry groups.

Despite these advances, the relatively small sample size of African American individuals in GTEx limits the statistical power and stability of gene expression prediction models. This limitation is compounded by substantial variation in sample sizes across tissues, which range from 7 to 77 individuals, with many tissues containing fewer than 20 samples. Such small sample sizes may reduce the precision of eQTL

effect estimates, decrease prediction accuracy, and limit the biological interpretability of tissue-specific models. The cross-tissue framework implemented in UTMOST partially mitigates these challenges by leveraging shared genetic regulation across tissues and borrowing information from genetically correlated tissues. For this reason, we retained all available African American tissues rather than imposing a minimum sample-size threshold, which would have substantially reduced tissue coverage and further constrained the development of ancestry-specific models. Nevertheless, tissues with limited sample sizes are expected to yield less stable predictions and should therefore be interpreted with appropriate caution. Furthermore, because cross-tissue models integrate information across tissues, tissues with larger sample sizes may contribute disproportionately to model estimation, potentially reducing sensitivity to tissue-specific regulatory effects in underrepresented tissues.

Despite these limitations, the integration of African American-specific models with existing European ancestry models enables investigations into the cross-ancestry transferability of TWAS findings. For example, examining whether gene-trait associations identified in European ancestry populations replicate or are enriched in African American cohorts can offer valuable insights into the robustness and universality of genetic signals. Future work could also explore the integration of rare-variant resources (such as RAVAR [20]) and rare-variant-aware prediction frameworks to better capture ancestry-specific regulatory effects that may not be adequately represented by models based primarily on common cis-eQTLs.

While AGEMdb marks a meaningful advancement for African American representation in TWAS, many other ancestry groups remain severely underrepresented. Although GTEx is a valuable resource, it includes relatively few individuals from these other ancestry groups, resulting in sample sizes too small to support training of gene expression imputation models in these populations. Future efforts should focus on expanding genotypic and transcriptomic data collection to include more diverse populations and larger sample sizes. Doing so will improve our ability to detect population-specific genetic effects, explore allelic heterogeneity, and ensure that the benefits of genetic research are equitably distributed. Promoting diversity in genetic studies is not only scientifically essential but also critical for advancing health equity worldwide.

4. CONCLUSION

AGEMdb provides ancestry-specific gene expression models for both European ancestry and African American populations. The inclusion of African American models represents a novel and critical advancement beyond existing European ancestry-centric resources. AGEMdb supports ancestry-specific multi-tissue TWAS analyses, enhancing the power and precision of gene–trait association studies and facilitating the clinical translation of GWAS findings for complex diseases. Despite this progress, many other ancestry groups remain underrepresented, underscoring the urgent need to expand genotypic and transcriptomic data collection across diverse populations and larger sample sizes. Such efforts are vital to reveal population-specific genetic effects and allelic heterogeneity, ensuring that genomic research benefits are equitably distributed worldwide.

Declarations

Funding

We acknowledge funding from the Medical Research Council (MR/V020749/1) and support from the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (NIHR203308). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Acknowledgments

We are grateful for the assistance given by The University of Manchester IT Services and for the use of the Computational Shared Facility.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used in this research were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v9.p2.

Ethics approval and consent to participate

Not applicable. This study used only publicly available, de-identified human data from the GTEx Project (dbGaP accession number phs000424.v9.p2). Therefore, ethics approval and consent to participate were not required. No new data involving human participants were collected.

Consent for publication

Not applicable.

References

1. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. Epub; 2012;7;337(6099):1190-5. https://doi.org/10.1126/science.1222794.
2. Gamazon E, Wheeler H, Shah K, Mozaffari S, Aquino-Michaels K, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8. https://doi.org/10.1038/ng.3367.
3. Gusev A, Ko A, Shi H, Bhatia G, Chung W, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52. https://doi.org/10.1038/ng.3506.
4. Hu Y, Lu Q, Weng H, Wang J, Zekavat S, Yu Z, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019;51:568–76. https://doi.org/10.1038/s41588-019-0345-7.
5. Barbeira A, Pividori M, Zheng J, Wheeler H, Nicolae D, Im H. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. 2019;15. https://doi.org/10.1371/journal.pgen.1007889
6. Zhou D, Jiang Y, Zhong X, Cox N, Liu C, Gamazon E. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020;52. https://doi.org/10.1038/s41588-020-0706-2
7. GTEX Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. 2015. p. 648–60. https://doi.org/10.1126/science.1262110.
8. Barbeira AN, Melia O, Liang Y, Bonazzola R, Wang G, Wheeler H, et al. Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. Genet Epidemiol. 2020;44:854–67. https://doi.org/10.1002/gepi.22346.
9. Pagnuco I, Eyre S, Rattray M, Morris A. Transferability of Single- and Cross-Tissue Transcriptome Imputation Models Across Ancestry Groups. Genet Epidemiol. 2025;49. https://doi.org/10.1002/gepi.22611.
10. Gottlieb A, Daneshjou R, DeGorter M, Bourgeois S, Svensson P, et al. Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. Genome Med. 2017;9. https://doi.org/10.1186/s13073-017-0495-0.
11. Li B, Verma S, Veturi Y, Verma A, Bradford Y, et al. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. Pac Symp Biocomput. 2018;23:448–59.
12. Mogil LS, Andaleon A, Badalamenti A, Dickinson S, Guo X, et al. Genetic architecture of gene expression traits across diverse populations. PLOS Genet. 2018;14:1007586. https://doi.org/10.1371/journal.pgen.1007586.
13. Mikhaylova AV, Thornton T. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. Front Genet. 2019;10:1664–8021. https://doi.org/10.3389/fgene.2019.00261.

14. Bhattacharya A, Hirbo J, Zhou D, Zhou W, Zheng J, Kanai M, et al. Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. Cell Genomics. 2022;2. https://doi.org/10.1016/j.xgen.2022.100180
15. Adeyemo A, Rotimi C. Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics. 2010;13:72–9. https://doi.org/10.1159/000218711.
16. Carlson CS, Matise T, North K, Haiman C, Fesinmeyer M, et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. 2013;11:1001661. https://doi.org/10.1371/journal.pbio.1001661.
17. Li YR, Keating B. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. 2014;6. https://doi.org/10.1186/s13073-014-0091-5.
18. Kelly DE, Hansen M, Tishkoff S. Global variation in gene expression and the value of diverse sampling. Curr Opin Syst Biol. 2017;1:102-108,. https://doi.org/10.1016/j.coisb.2016.12.018.
19. Stegle O, Parts L, Durbin R, Winn J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. PLOS Comput Biol. 2010;6:1000770. https://doi.org/10.1371/journal.pcbi.1.
20. Cao C, Shao M, Zuo C, Kwok D, Liu L, Ge Y, et al. RAVAR: a curated repository for rare variant-trait associations. Nucleic Acids Res. 2024;52:990–7. https://doi.org/10.1093/nar/gkad876

Figures

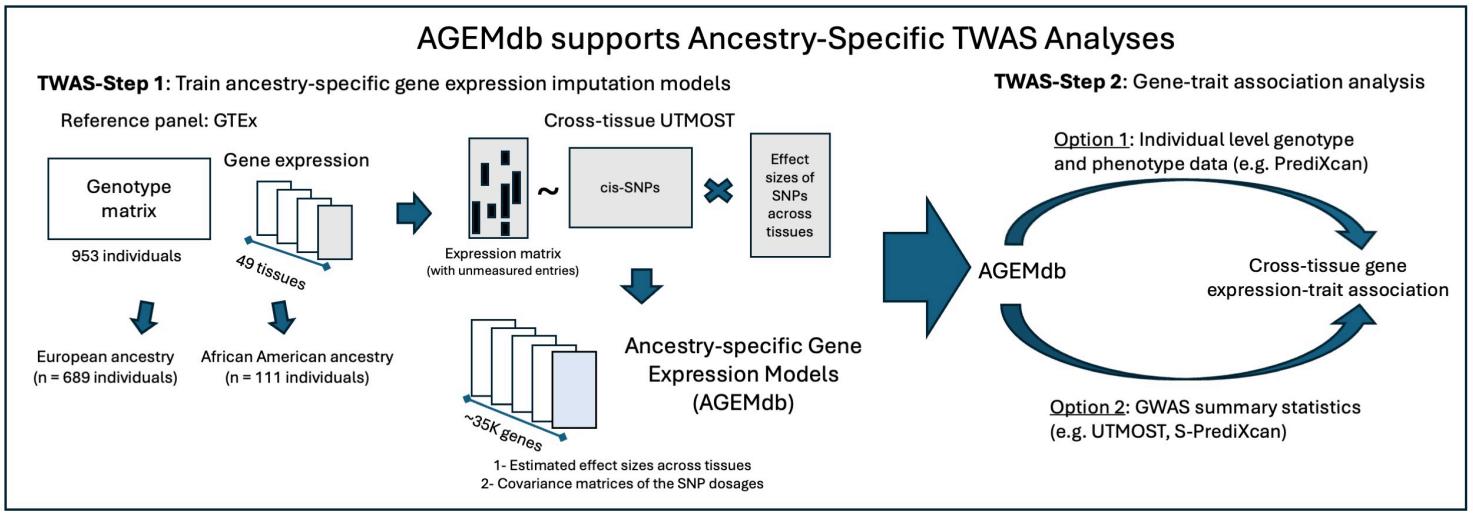


Figure 1

Overview of the AGEMdb resource and its application in ancestry-specific transcriptome-wide association studies (TWAS).

AGEMdb provides gene expression imputation models trained separately for European and African American ancestry groups, using GTEx as the reference panel. Sample sizes for each ancestry group are shown after genotype quality control. In TWAS Step 1, models are trained using the cross-tissue UTMOST method to predict gene expression from genotype data. AGEMdb includes SNP effect sizes on gene expression, estimated via the imputation models across tissues for each gene, along with SNP dosage covariance matrices. In Step 2, the models are applied to identify gene–trait associations using either individual-level genotype/phenotype data or GWAS summary statistics combined with SNP dosage covariance matrices.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppTable1.xlsx