LA-GEM: imputation of gene expression with incorporation of Local Ancestry

Mrinal Mishra†, Layan Nahlawi†, Yizhen Zhong, Tanima De, Guang Yang, Cristina Alarcon and Minoli A. Perera

Department of Pharmacology, Center for Pharmacogenomics, Feinberg School of Medicine, Northwestern University
Chicago, Illinois, USA
Email: minoli.perera@northwestern.edu

Gene imputation and TWAS have become a staple in the genomics medicine discovery space; helping to identify genes whose regulation effects may contribute to disease susceptibility. However, the cohorts on which these methods are built are overwhelmingly of European Ancestry. This means that the unique regulatory variation that exist in non-European populations, specifically African Ancestry populations, may not be included in the current models. Moreover, African Americans are an admixed population, with a mix of European and African segments within their genome. No gene imputation model thus far has incorporated the effect of local ancestry (LA) on gene expression imputation. As such, we created LA-GEM which was trained and tested on a cohort of 60 African American hepatocyte primary cultures. Uniquely, LA-GEM include local ancestry inference in its prediction of gene expression. We compared the performance of LA-GEM to PrediXcan trained the same dataset (with no inclusion of local ancestry) We were able to reliably predict the expression of 2559 genes (1326 in LA-GEM and 1236 in PrediXcan). Of these, 546 genes were unique to LA-GEM, including the *CYP3A5* gene which is critical to drug metabolism. We conducted TWAS analysis on two African American clinical cohorts with pharmacogenomics phenotypic information to identity novel gene associations. In our IWPC warfarin cohort, we identified 17 transcriptome-wide significant hits. No gene reached are prespecified significance level in the clopidogrel cohort. We did see suggestive association with *RAS3A* to P2RY12 Reactivity Units (PRU), a clinical measure of response to anti-platelet therapy. This method demonstrated the need for the incorporation of LA into study in admixed populations.

Keywords: Local Ancestry, Gene Expression Model, LA-GEM, PrediXcan, Gene Imputation, Population-specific Genetic Variations, Admixed Populations, Ancestry-specific Gene Associations

1. Introduction

It is widely acknowledged that large-scale genetic studies investigating human diseases have often failed to encompass the extensive diversity seen in global populations, as they primarily focus on individuals of European descent.[1]This insufficiency of ethnic diversity in such studies limits our understanding of the genetic underpinnings of human diseases and intensifies health disparities. Moreover, the paucity of ethnic diversity in human genomics research could lead to a potentially hazardous deficiency, or even errors, in our capacity to apply genetic research findings to clinical procedures or public health policies.

† Contributed equally to the work.

PrediXcan is one of the first and most popular methods used to predict gene expression levels in different tissues or cell types for use in transcriptome-wide association studies (TWAS).[2] The method leverages large publicly available multi-omic datasets that includes paired single nucleotide polymorphism (SNP) data and gene expression data from multiple individuals and tissues.[2-3] By training a predictive model on these reference datasets, PrediXcan can predict the expression levels of a given gene in a new individual, based on that person's genetic variation. Outside data can be trained through various available methods.[4,5] There are various extensions to PrediXcan that have been developed which extend this method to multi-tissue TWAS and causal gene prioritization.[5-9]

In any association studies, undetected population stratification can lead to false-positive. Therefore, it is critical to implement appropriate correction to adjust these effects.[10] One such measure, used in genome-wide association studies (GWAS), is the inclusion of principal components (PCs), with the first few PCs estimating global ancestry (GA) in the cohort. GA is largely directed by demographic history of the population. However, for admixed population the effects of nearby SNPs or epigenetic changes has been shown to have a significant effect of gene expression[11]. Thus, local ancestry may be an important consideration in gene expression prediction. Here we have incorporated LA as predictor in PrediXcan framework to assess the if including this variable in the African American population resulting in the improved predictability of the gene models.

2. Methods

In this paper, we propose a modification to PrediXcan method titled LA-GEM (Local Ancestry based Gene Expression prediction Model) to incorporate local ancestry predictors (LA) along with cis region genetic variants in the development of gene expression prediction models. We have used our African American multi-omic hepatocyte dataset (N = 60) to create gene expression prediction models, however this method can be used on any multi-omic data from an admixed cohort in which local ancestry inference is available.

2.1. *Primary Hepatocyte Cohort*

Sixty-three African Ancestry (AA) primary human hepatocyte (PHHs) cultures were acquired. AA PHHs were either purchased from commercial companies (BioIVT, TRL/Lonza, Life technologies, Corning, and Xenotech), or isolated in-house from cadaveric livers. Livers with active cancer or a history of hepatocarcinoma were excluded from the study. To account for differences in PHH sourcing, transcriptomic data went through additional QC measures (i.e., PC visualization, batch correction) to ensure any differences from source and isolation method were corrected. PHHs were isolated from cadaveric livers using a modified two-step collagenase perfusion procedure previously described in Park et. al.[12] Only hepatocyte cultures with RNA Integrity Number (RIN) over 8 and with sufficient RNA to conduct NGS were used in the study.

2.2. *Genotyping, quality control and imputation*

DNA was obtained from around 1 million cells of each PHH culture using the Gentra Puregene (Qiagen) kit following manufacturer's protocol. All extracted DNA samples were barcoded for genotyping. Illumina Infinium Multi-Ethnic Global Kit was used for SNP genotyping and standard genotyping protocol was followed. SNPs were filtered out before imputation based on following criterion: (1) SNPs present on the sex and mitochondrial chromosomes. They were filtered out as they could alter the minor allele frequency (MAF) values (2) SNPs having A/T or C/G as it may introduce flip-strand issues. (3) SNPs with low genotype quality (call rate < 0.95).

Using PLINK[9], individuals with discordant sex information were identified using the sex check function and duplicates or related individuals were identified using the identity-by-descent (IBD) method. An IBD cutoff score of 0.125 was used, indicating third-degree relatedness or closer. No samples were removed after these QC steps. SNPs with MAF<0.05 were removed. Patient ancestries were confirmed using a principal component analysis (PCA) plot of linkage disequilibrium (LD) pruned genotype data. LD pruning was conducted to identify the principal dimensions of genetic variation between samples. Samples that did not cluster along the spectrum for AA within this PCA plot of raw genotype data were removed.[11] One individual was excluded after sample and genotyping QC analysis, leaving 62 individuals.

Genotypes were imputed by the TOPMed imputation server (version 1.6.6)12-14 using the TOPMed r2 reference panel, GRCh38/hg38 array build, and 0.3 estimated r2 (rsq) filter threshold. Post-imputation QC includes removal of SNPs with poor imputation quality scores (<0.8), failed Hardy-Weinberg equilibrium tests ($p < 0.00001$), and low MAFs (<0.05). This resulted in a total of 5,189,820 SNPs included for model building.

2.3. *Local ancestry inference*

LA was inferred using RFMix (v.1.5.4). RFMix takes as input a set of reference panels (populations with known ancestry) and a set of test individuals, and uses a hidden Markov model to infer the most likely ancestry of each segment of the test individuals' genomes. The output of RFMix is a set of probabilities for each test individual, indicating the likelihood that a specific haplotype segment comes from one of the reference populations.[13] In this analysis we use Yoruba (African Ancestry) and American white (CEU – European Ancestry) as our refences populations.

2.4. *RNA-sequencing and Quality Control*

Total RNA was extracted from each PHH culture three days after plating using the Qiagen RNeasy Plus mini kit. Samples with an RNA integrity number (RIN) less than 8 were removed from analysis. This resulted in the removal of 2 samples leaving 60 individuals at the end. Libraries were prepared for sequencing using the TruSeq RNA Sample Prep Kit, Set A (Illumina) per manufacturer's protocol. The cDNA libraries were prepared and sequenced using either HiSeq2500 (Illumina) or HiSeq4000 (Illumina) instruments by the University of Chicago's Functional Genomics core, producing single-end 50bp reads with approximately 50 million reads per sample. As two

instruments were used in this study, we were cognizant of potential batch effect and incorporated methods for correction as previously described.[14]

2.5. *Gene Expression Quantification*

Gene expression was quantified using a collapsed gene model following the GTEx isoform collapsing procedure[15]. To evaluate gene-level expression, reads were mapped to genes referenced with GENCODE(v.25) using RNA-SeQC. HTSeq supplied raw counts for gene expression analysis using Bioconductor package DESeq2(v1.20.0). Counts were normalized by regularized log transformation, batch correction was performed using ComBat-Seq[14], and PCA was performed using DESeq2.

Gene expression was normalized by trimmed means of M-values normalization method (TMM) implemented in edgeR.[16] Transcripts per million (TPM) was calculated by first normalizing counts by gene length and then by read depth.[17] Gene expression values were filtered based on expression thresholds < 0.1 TPM in at least 20% of samples and ≤ 6 reads in at least 20% of samples. The expression values for each gene were normalized across samples with inverse normal transformation. To account for unmeasured confounding variables in transcriptome data, we used probabilistic estimation of expression residuals (PEER).[18]

2.6. *LA-GEM Framework*

LA-GEM consists of mainly three steps:

For gene expression prediction, a linear model was trained using reference panel that includes genotype, LA predictor, interaction predictor (interaction between genotype and LA predictor) and corresponding expression data[2,19] using the following training model equation[19]:

$$y_g \sim \sum_{a,b,c} \boldsymbol{w_a S_a} + \boldsymbol{w_b A_b} + \boldsymbol{w_c I_c} + \boldsymbol{\varepsilon} \tag{1}$$

where w_a, w_b and w_c are the regression parameter needed to be trained, S = (S_1, , S_2, …, S_a) is the genotype data in the cis region of interest, A = (A_1, A_2, …, A_b) is the local ancestry predictors for all SNP positions in the cis region and I = (I_1, I_2, …, I_c) is the Interaction predictor (I = S x A).

Genetically regulated gene expressions are then determined using the above model for new dataset that include combination of genotype and local ancestry information using the following equation:

$$\hat{y}_g \sim \sum_{d,e,f} \boldsymbol{w_d S_d} + \boldsymbol{w_e A_e} + \boldsymbol{w_f I_f} \tag{2}$$

Estimated genetically regulated gene expressions ŷ_g is then associated to the phenotype using the following equation:

$$Z \sim \hat{y}_g + \boldsymbol{\varepsilon} \tag{3}$$

LA-GEM prediction models were trained on 60 African American PHH samples followed by 5-fold cross-validation. Gene models with an average correlation $\rho \geq 0.1$ and $P < 0.05$ between predicted and observed Expression were deemed well predicted.

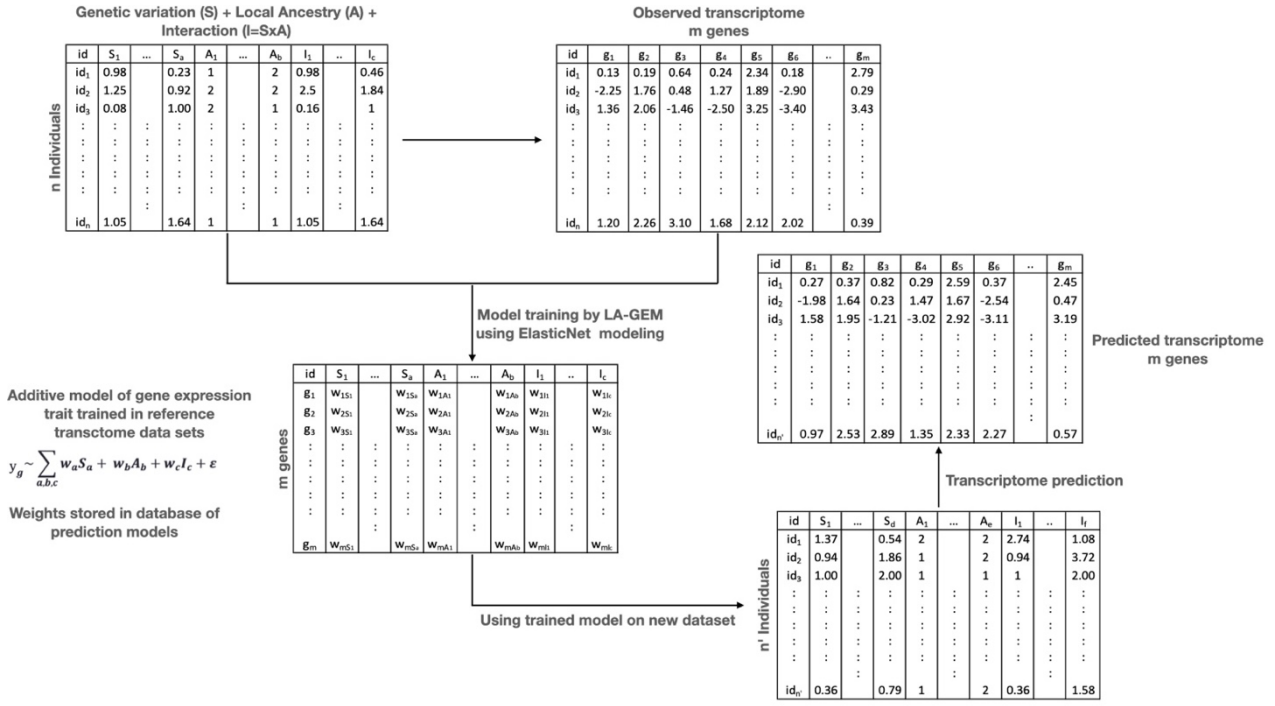

Fig. 1. Flowchart showing LA-GEM workflow.

2.7. *TWAS association using LA-GEM gene imputation.*

As a proof of concept, we use LA-GEM to impute hepatic gene expression in two clinical cohort to identify novel gene associations to drug response. As the expression of hepatic genes are especially important in platelet function and drug metabolism, we imputed gene expression of 1323 genes which were then used in the TWAS conducted using PrediXcan.[2] We prespecified a TWAS p-value of 3.8×10^{-5} as significant (0.05/1323).

2.7.1. *African American warfarin Cohort*

Through the International Warfarin Pharmacogenomics consortium (IWPC) we collect information from 340 African American patients on warfarin as well as 199 African Americans who were part of the University of Alabama Birmingham Warfarin cohort assess though dbGAP (phs000708.v1.p1). Briefly, clinical and demographic data on stable warfarin dose was collected, defined as the dose of warfarin needed to elicit and INR within therapeutic rage (2-3) for three consecutive clinical visits as previously described.[20]

2.7.2. *ACCOuNT Clopidogrel cohort*

Through the ACCOuNT Consortium[21] we recruited 180 African Americans on the anti-platelet drug, clopidogrel. All subjects included in the TWAS had a biomarker measure of clopidogrel

response, P2Y12 Reactivity Units (PRU). All subjects were on 75 mg of clopidogrel for at least 15 days with inclusion and exclusion criteria as described previously.[21]

2.8. *Log Ratio of Interaction Predictors*

To quantify the relative influence of interaction predictors in our LA-GEM model, we calculated a Log Ratio for each gene using the formula:

$$\text{Log Ratio} = \log_2(\text{Count of Interaction Predictors} + 1) - \log_2(\text{Count of SNP Dosage Predictors} + 1)$$

A positive Log Ratio indicates that a gene relies more heavily on interaction predictors, while a negative value suggests greater reliance on genetic dosage predictors.

2.9. *Code Availability*

The LA-GEM model was implemented in R and employs SNP-based local ancestry calculated using RFMix version 1.5.4. The source code is publicly available and can be accessed at https://github.com/pereralab/LA-GEM.

3. Results

We built two gene expression prediction models, LA-GEM and PrediXcan (using AA PHH as training). We assessed predictive performance using five-fold cross-validation (R2 of model performance). We found that LA-GEM was able to impute 1323 genes at a rho>0.1, p-value ≤ 0.05 (Average rho = 0.397) as compared to 1236 genes imputed well using the PrediXcan model (Average rho = 0.403) in the same dataset without LA (Fig. 2). The average number of predictors for LA-GEM is shown in Table 1.

Table 1 – Summary table showing total number of Predictable genes and number of different Predictors used to train the model.

	LA-GEM
Number of Predictable genes	1323
Number of Predictors	71702
Number of SNP Dosage Predictor	46028
Number of Interaction Predictors (L.A X SNP Dosage)	25674

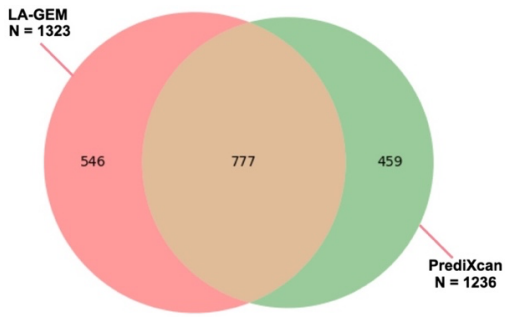

Fig. 2. Venn diagram showing number of predictable genes in each of the model.

3.1. *Gene list enrichment analysis of predictable genes*

KEGG Pathway enrichment analysis (Statistical overrepresentation test) was performed using g:Profiler[28] for predictable genes (1323 genes) obtained from LA-GEM. The analysis yielded significant enrichments for several pathways as shown in Fig, 3, notably those linked to pharmacogenomics. Among these, three pathways were found to be prominently enriched: "Metabolism of xenobiotics by cytochrome P450" (KEGG:00980) with a fold enrichment of 3.37 and an adjusted p-value of 0.00285, "Drug metabolism - cytochrome P450" (KEGG:00982) with a fold enrichment of 3.18 and an adjusted p-value of 0.01097, and "Drug metabolism - other enzymes" (KEGG:00983) with a fold enrichment of 2.74 and an adjusted p-value of 0.04196.

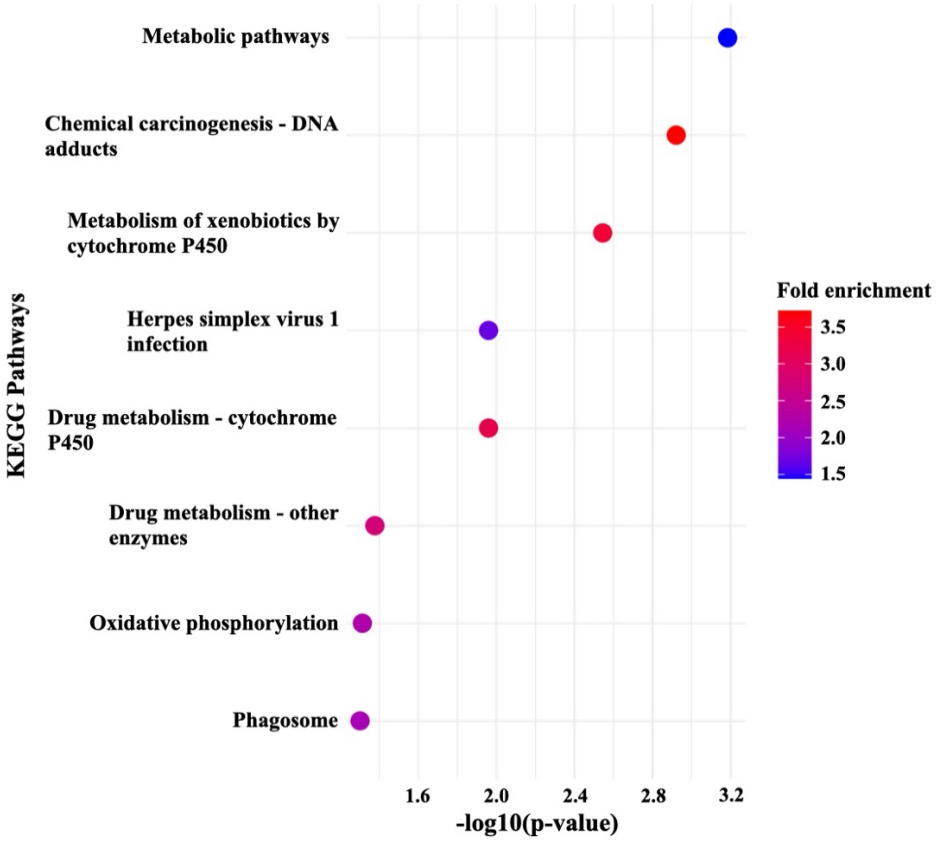

Fig. 3. Gene set enrichment of 1323 predictable genes obtained from LA-GEM. Y-axis show categories with their corresponding -$\log_{10}$(p-value) in the X-axis. Color shows the fold enrichment value for each of the processes.

3.2. *Genes unique to LA-GEM*

Among the 1323 predictable genes, 546 genes were found to be unique to LA-GEM model which were not reported by PrediXcan model as significant (Fig. 2). Out of the 546 unique genes, 2 genes (*MME* and *LRRC37A2*) were found to be strongly associated with global West African ancestry as previously reported[12]. In addition, *CYP3A5 CYP1A1, CYP4F2, CBR1,* and *UGT2A1* was also among

the genes unique to LA-GEM which is known to show significant variability in level of expression between population of different ancestry and are important to drug response[22].

3.3. *Genes unique to PrediXcan*

Among the 1323 predictable genes, 459 genes were found to only in the PrediXcan model (Fig. 2). Out of the 459 genes, 6 genes (*DHODH*, *SNAI1*, *RBBP9*, *ENSG00000271239*, *NPR2*, and *SLC39A11*) were found to be strongly associated with global West African ancestry as previously reported.[12]

3.4. *Genes common to LA-GEM and PrediXcan*

Among the 1323 predictable genes, 777 genes were found to be well imputed by both models. Out of these 777 genes, 4 genes (*CDK18*, *GREM2*, *COL26A1* and *MMP20-AS1*) were found to be strongly associated with West African ancestry as previously reported.[12] The rho average for *CDK18* and *GREM2* were higher in LA-GEM (0.48 versus 0.33 and 0.28 versus 0.26, respectively) but the inverse was true for *COL26A1* and *MMP20-AS* (0.39 versus 0.44 and 0.32 versus 0.54 respectively) The rho average for these genes were evenly distributed around the diagonal (Fig. 4), suggesting one model did not outperform the other in these commonly imputed genes. For genes that were unique to the PrediXcan model, the average difference in rho between models was 0.42. For those gene that were uniquely to LA-GEM the average difference in Rho was 0.46. However, these differences in prediction accuracy were not a significant difference between the two groups of genes ($p = 0.07$).

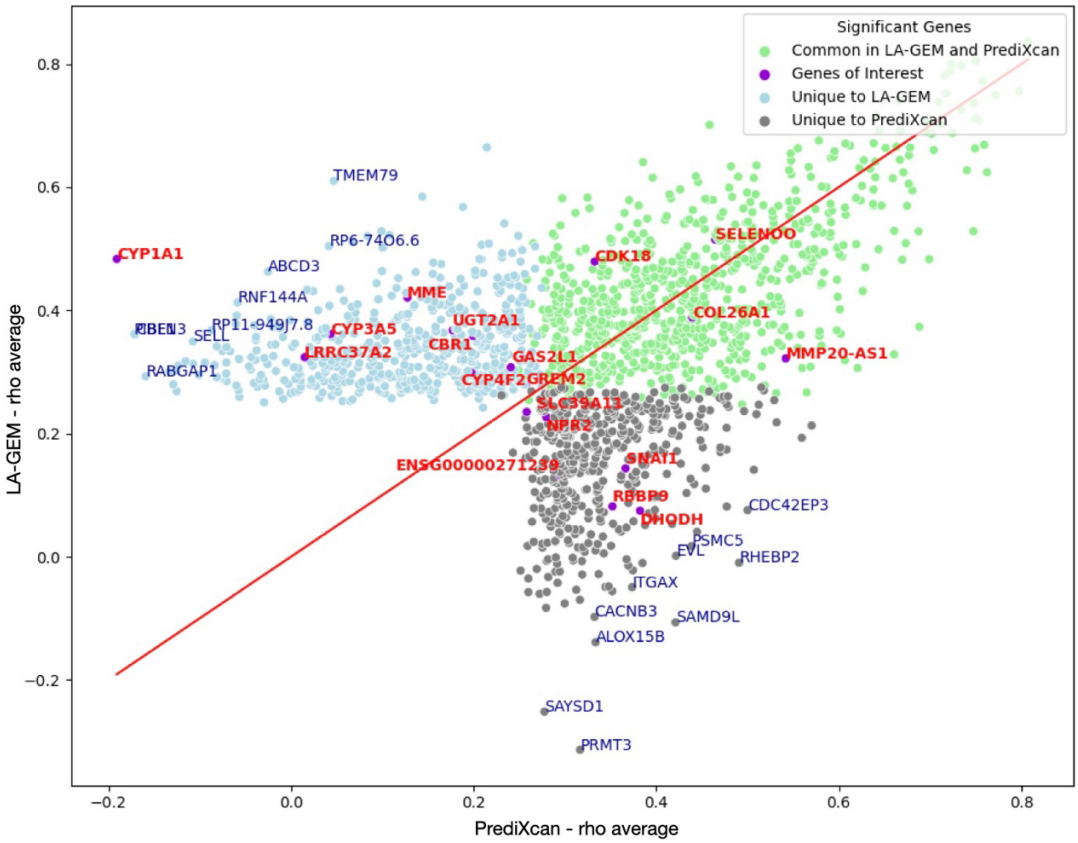

Fig. 4. Correlation plot between rho-averages of gene well predicted with LA-GEM and PrediXcan models. Top 10 genes showing the maximum rho-average difference between methods are labelled in dark blue color. Red line shows perfect correlation. Well predicted genes unique to LA-GEM model are shown in light blue. Well predicted genes unique to PrediXcan model are shown in grey. Well predicted genes common between LA-GEM and PrediXcan model are shown in light green. Genes of interest with pharmacogenomic relevance or which are associated with West African ancestry are shown in violet and are labelled in red.

3.5. *Differential Role of Interaction Predictors in LA-GEM and PrediXcan Models*

In the process of model training for LA-GEM, we observed differences in the role played by the type of predictors, especially interaction predictors, in model efficacy. Among the 546 genes uniquely imputed by the LA-GEM model, 137 genes (or approximately 25% of these significant genes) exhibited a positive Log Ratio of the Count of Interaction predictors. This observation underscores the relevance of interaction predictors as significant contributors in the unique imputation capability of the LA-GEM model.

In contrast, among the 777 genes that were common between LA-GEM and PrediXcan, only 119 genes (approximately 15% of these significant genes) had a positive Log Ratio of the Count of Interaction predictors. This relatively lower proportion suggests that the common genes might rely less on interaction predictors in the LA-GEM model than the genes unique to it.

The detailed distribution of the Log Ratio of the Count of Interaction predictors for these gene sets is depicted in Fig 5. This difference in the involvement of interaction predictors between genes unique to LA-GEM and those common with PrediXcan provides further insight into the distinguishing features of these models.

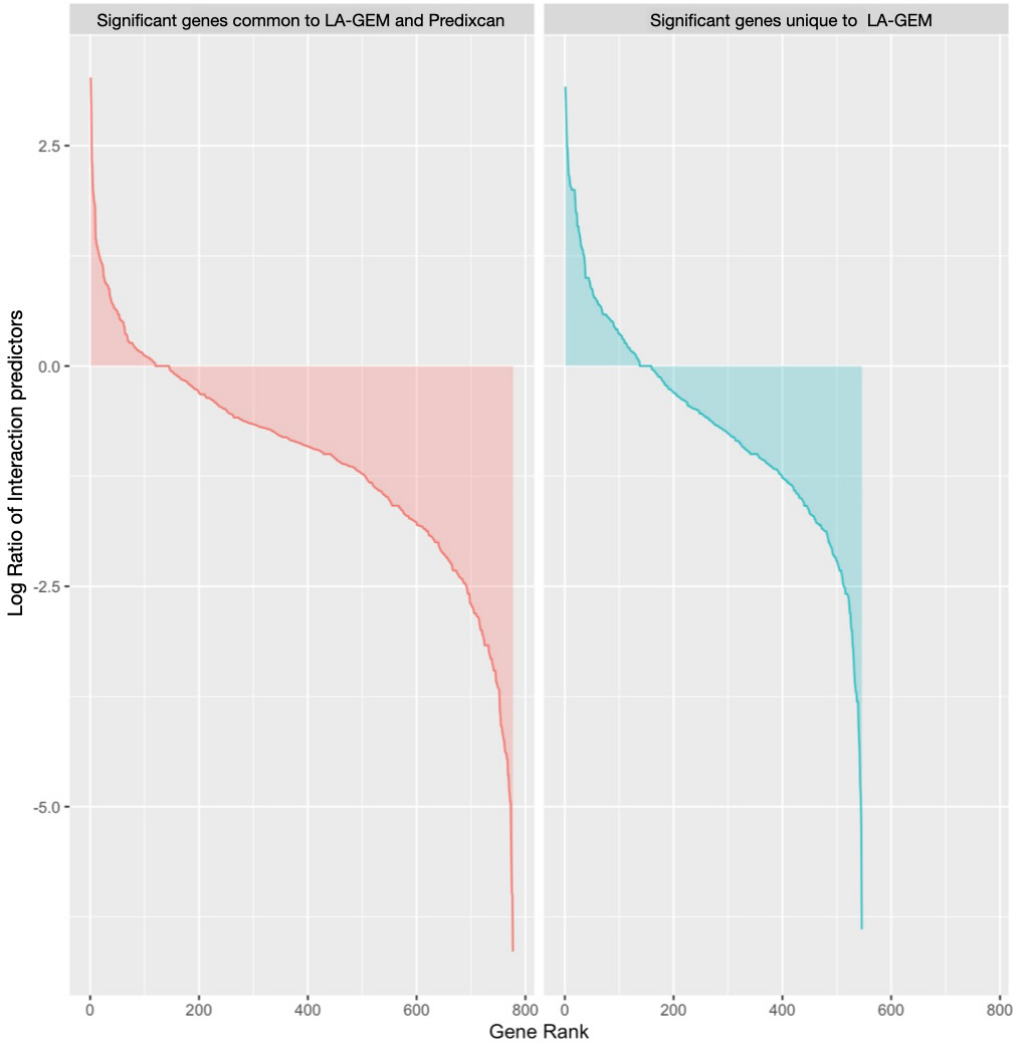

Fig. 5. Distribution of Positive Log Ratios of Count of Interaction Predictors in Genes Unique to LA-GEM and Common to LA-GEM and PrediXcan

3.6. *TWAS association to warfarin dose*

Using the IWPC warfarin cohort we imputed hepatocyte gene expression (restricted to those genes that were well imputed – N = 1325) and conducted a TWAS. The top associations are shown in the Manhattan plot (Fig. 6). Bonferroni corrected significant associations were found with 17 genes. No association was seen with known warfarin genes *VKORC1*, or *CYP2C9* as these gene were not well imputed in our models.

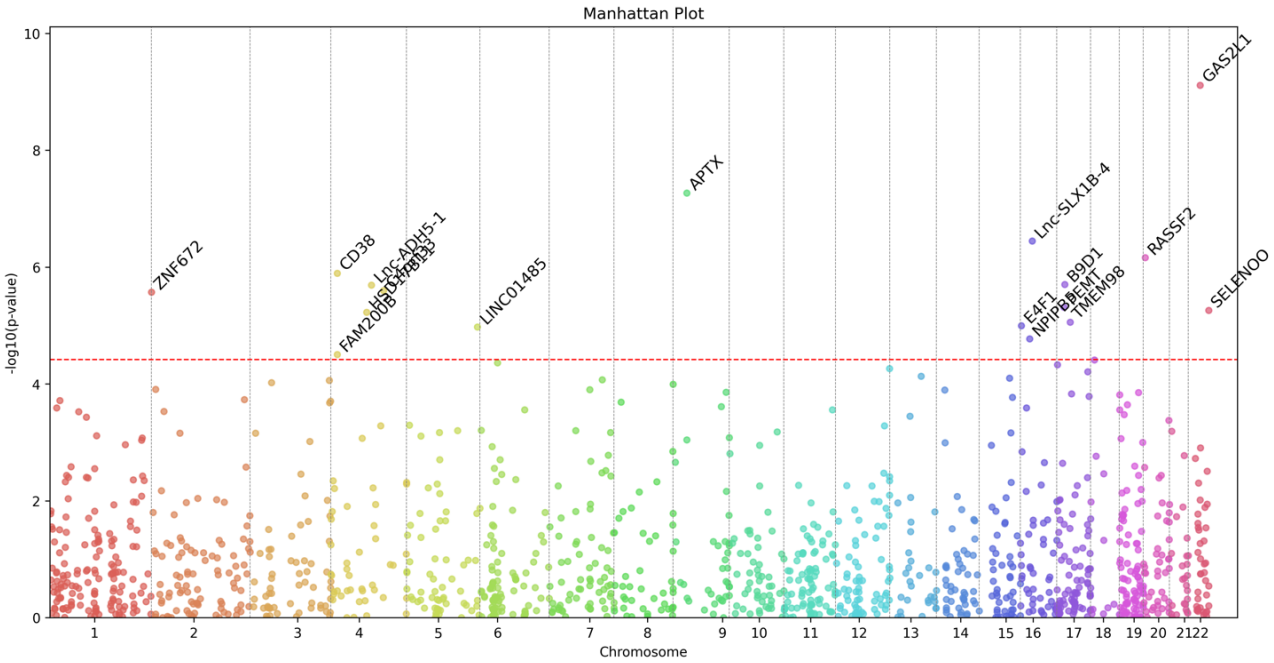

Fig. 6. Manhattan plot of TWAS results. The figure shows the association of imputed gene expression to stable warfarin dose in the IWPC cohort. The x-axis show the relative genomic position of each gene tested (N = 1323) and the y-axis show the Log(10) p-value. The red dashed line marked the threshold of significance for this study.

3.7. *TWAS association to PRU in patient taking clopidogrel.*

Using the ACCOuNT cohort, we imputed the hepatic gene expression in 180 African American patients on clopidogrel. We found no transcriptome-wide significant associations. However, one top association showed *RASA3* gene expression associated with increased PRU (p = 0.0014, Beta = 0.61). This gene has known association to platelet aggregation.[29]

4. Discussion

This study introduces a novel computational model, LA-GEM, designed to enhance gene expression prediction by integrating local ancestry (LA) predictors with cis-regional genetic variants. The development and deployment of such a model emerge from the understanding that complex trait prediction may be augmented by considering population-specific genetic variations. In many traditional models, such as PrediXcan, the unique genetic contributions of LA are not considered, potentially leading to overlooked associations.[2]

Our findings revealed that LA-GEM improved gene expression prediction compared to PrediXcan in some genes, suggesting that the inclusion of LA predictors can effectively supplement traditional cis-regional genetic variants. This improvement was demonstrated by the imputation of 1323 genes at a rho>0.1, p-value ≤ 0.05 by LA-GEM, compared to 1236 genes imputed by PrediXcan without considering LA.

Beyond these numbers, our study unveiled a set of 546 genes uniquely predicted by LA-GEM and 777 genes common in both LA-GEM and PrediXcan AA model. Out of 1323, 6 genes (*MME* and *LRRC37A2*, *CDK18*, *GREM2*, *COL26A1* and *ENSG00000281655*) were previously found to be associated with global West African ancestry and exhibited significant differential expression when compared to individuals of European descent.[12] These genes are not only statistically significant but also relevant to pharmacogenomics. For instance, *GREM2*, a gene involved in developmental processes[23,] is also associated with allopurinol efficacy[24]. and *MME*, implicated in neuropeptide degradation[25] and associated with ACE inhibitor-induced cough[26], were amongst the uniquely predicted genes. Lastly variants in *COL26A1* have been associated to Aspirin-intolerant asthma.[27]

Importantly, this study highlights the valuable implications of integrating LA predictors in gene expression models for drug response studies. By significantly predicting genes such as *CYP3A5*, *CYP1A1, CYP4F2, CBR1,* and *UGT2A1* - well-known contributors to drug metabolism and disease progression[30-33] - our model may aid in TWAS studies of inter-individual variations in drug responses and adverse drug reactions in African Americans. A particular emphasis should be placed on *CYP3A5*. This gene has been widely recognized for variability between different ethnic groups. The splice variant *CYP3A5*3*, associated with reduced enzyme activity, is less frequent in African populations, resulting in a functional enzyme in African populations. As most European carry the *CYP3A5*3,* the effect of this enzyme on drug response is not well accounted for in studies of European individuals. *CYP3A5* is thought to contribute to drug efficacy and toxicity, including responses to immunosuppressants such as tacrolimus.[34-35]

We applied LA-GEM to the African American warfarin and clopidogrel cohorts, demonstrating its utility in clinical studies. The warfarin cohort revealed 17 genes with significant associations with warfarin dose requirement, providing novel potential genetic influencers of warfarin dosage response beyond the well-known *VKORC1* and CYP2C9 genes[36-37]. The most significate TWAS hit was *GAS2L1 (*associated with increased warfarin dose requirement*,* $p = 7.7x10^{-10}$*),* which has previously been associated with thrombocytopenia in women.[38] Also, the gene *SELENOO* on chromosome 22 showed association to decrease warfarin dose requirement ($p = 5.5x10^{-6}$). A previous study in Sub-Saharan Africans found variants near this gene associated to increase R-6 Hydroxy-warfarin metabolite measurement.[39]

In the ACCOuNT clopidogrel cohort, we discovered an association between *RASA3* gene expression and increase P2Y12 Reactivity Units (PRU) level. While the most notable role of *RASA3* involves platelet function and hemostasis[29], this gene's function is not limited to platelets and the bloodstream. It is broadly expressed in many tissues, including the brain, lungs, and kidneys, suggesting it might have additional roles outside of platelets. In cancer biology, the Ras and RAP GTPases regulated by *RASA3* are often involved in tumorigenesis. For instance, inactivation of GAPs (like *RASA3*) can lead to overactive Ras signaling, which can contribute to the development of cancer.[40] This gene has also been associated to pulmonary hypertension in Sickle Cell Disease.[41]

In terms of computational efficiency, LA-GEM and PrediXcan showed similar performance during the model training phase. Specifically, for our limited dataset of 60 hepatocyte samples, both models completed the training within a time frame of approximately 2 to 3 hours. It's worth noting that the computational time is expected to scale linearly with the size of the sample pool, thus offering scalability as more comprehensive datasets become available.

Several innovative methods have set the stage in ancestry inform gene expression prediction. Notable among these are METRO[42], which enhances transcriptome-wide association studies (TWAS) through a likelihood-based inference framework, and MATS[43], which jointly analyzes samples from multiple populations to account for ancestral heterogeneity in gene expression effects. Additionally, a study by Lauren et al.[44] addressed the genetic architecture of gene expression across diverse populations, emphasizing the necessity for diverse population sampling in genomics. Despite their valuable contributions, none of these methods utilize SNP-based local ancestry as an intrinsic part of their predictive models. Our approach, LA-GEM, distinctively integrates SNP-based local ancestry predictors along with cis-regional variants to make more nuanced gene expression predictions. This unique aspect of LA-GEM not only adds a new layer of granularity to the existing methodologies but also paves the way for future explorations in this growing field.

While our findings are promising, there are several limitations to our study. First, we constructed the LA-GEM models with a limited cohort of 60 hepatocyte cultures. This is reflective of the overall lack of comprehensive multi-omics data in the African American population. With greater amounts of data on which to build these models, we will be better able to predict tissue specific patterns in the under-represented populations. This is also evident by the much greater number of well imputed gene available for the GTEx liver model (N = 3356) which is built on 153 liver samples. It should be noted that only 12 of these sample have any African Ancestry. Second, it is clear that there are still genes that are better predicted without the addition of LA. This suggests that to comprehensively use TWAS in African American population may require both LA-aware as well as traditional gene imputation methods. Lastly, the validation of LA-GEM in other tissues and larger cohorts remains a crucial next step. Ultimately, the incorporation of LA predictors can contribute significantly to personalized medicine, paving the way for treatments and interventions more attuned to a unique admixed genetic background of African Americans.

In conclusion, our study underscores the need for inclusion of LA in genomic methods. LA-GEM serves as a valuable tool in this endeavor, providing novel insights into the genomic architecture of complex traits in multiethnic populations, and highlighting the importance of considering local ancestry when predicting gene expression. The potential to uncover novel ancestry-specific gene associations can revolutionize our understanding of the interplay between genetics, disease, and therapeutic responses.

5. Acknowledgment

This work was made possible for through the following grants R01MD009217 (NIH, NIMHD), and R21HG011695 (NIH, NHGRI)

References

1. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. Nature, 538(7624), 161–164. https://doi.org/10.1038/538161a
2. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC; GTEx Consortium; Nicolae DL, Cox NJ, Im HK. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015 Sep;47(9):1091-8. doi: 10.1038/ng.3367. Epub 2015 Aug 10. PMID: 26258848; PMCID: PMC4552594.
3. Mikhaylova AV, Thornton TA. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. Front Genet. 2019 Apr 3;10:261. doi: 10.3389/fgene.2019.00261. PMID: 31001318; PMCID: PMC6456650.
4. Xu Z, Wu C, Wei P, Pan W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. Genetics. 2017 Nov;207(3):893-902. doi: 10.1534/genetics.117.300270. Epub 2017 Sep 11. PMID: 28893853; PMCID: PMC5676241.
5. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016 Mar;48(3):245-52. doi: 10.1038/ng.3506. Epub 2016 Feb 8. PMID: 26854917; PMCID: PMC4767558.
6. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S, Shi Y, Kunkle BW, Mukherjee S, Natarajan P, Naj A, Kuzma A, Zhao Y, Crane PK; Alzheimer's Disease Genetics Consortium,; Lu H, Zhao H. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019 Mar;51(3):568-576. doi: 10.1038/s41588-019-0345-7. Epub 2019 Feb 25. PMID: 30804563; PMCID: PMC6788740.
7. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, Björkegren JLM, Im HK, Pasaniuc B, Rivas MA, Kundaje A. Opportunities, and challenges for transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):592-599. doi: 10.1038/s41588-019-0385-z. Epub 2019 Mar 29. PMID: 30926968; PMCID: PMC6777347.
8. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am J Hum Genet. 2017 Mar 2;100(3):473-487. doi: 10.1016/j.ajhg.2017.01.031. Epub 2017 Feb 23. PMID: 28238358; PMCID: PMC5339290.
9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.
10. Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T, Zhu X. Assessing the impact of global versus local ancestry in association studies. BMC Proc. 2009 Dec 15;3 Suppl 7(Suppl 7):S107. doi: 10.1186/1753-6561-3-s7-s107. PMID: 20017971; PMCID: PMC2795878.

11. Zhong Y, De T, Alarcon C, Park CS, Lec B, Perera MA. Discovery of novel hepatocyte eQTLs in African Americans. PLoS Genet. 2020 Apr 20;16(4):e1008662. doi: 10.1371/journal.pgen.1008662. PMID: 32310939; PMCID: PMC7192504.
12. Park CS, De T, Xu Y, Zhong Y, Smithberger E, Alarcon C, Gamazon ER, Perera MA. Hepatocyte gene expression and DNA methylation as ancestry-dependent mechanisms in African Americans. NPJ Genom Med. 2019 Nov 25;4:29. doi: 10.1038/s41525-019-0102-y. PMID: 31798965; PMCID: PMC6877651.
13. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013 Aug 8;93(2):278-88. doi: 10.1016/j.ajhg.2013.06.020. Epub 2013 Aug 1. PMID: 23910464; PMCID: PMC3738819.
14. Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics, 2(3), lqaa078. https://doi.org/10.1093/nargab/lqaa078
15. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, … Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. Nature, 550(7675), 204–213. https://doi.org/10.1038/nature24277.1993).
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818.
17. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020 Aug;26(8):903-909. doi: 10.1261/rna.074922.120. Epub 2020 Apr 13. PMID: 32284352; PMCID: PMC7373998.
18. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012 Feb 16;7(3):500-7. doi: 10.1038/nprot.2011.457. PMID: 22343431; PMCID: PMC3398141.
19. Zou H, & Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B (Statistical Methodology). 2005 67(2), 301–320. http://www.jstor.org/stable/3647580
20. Perera MA, Cavallari LH, Limdi NA, Gamazon ER, Konkashbaev A, Daneshjou R, Pluzhnikov A, Crawford DC, Wang J, Liu N, Tatonetti N, Bourgeois S, Takahashi H, Bradford Y, Burkley BM, Desnick RJ, Halperin JL, Khalifa SI, Langaee TY, Lubitz SA, Nutescu EA, Oetjens M, Shahin MH, Patel SR, Sagreiya H, Tector M, Weck KE, Rieder MJ, Scott SA, Wu AH, Burmester JK, Wadelius M, Deloukas P, Wagner MJ, Mushiroda T, Kubo M, Roden DM, Cox NJ, Altman RB, Klein TE, Nakamura Y, Johnson JA. Genetic

variants associated with warfarin dose in African-American individuals: a genome-wide association study. Lancet. 2013 Aug 31;382(9894):790-6. doi: 10.1016/S0140-6736(13)60681-9. Epub 2013 Jun 5. PMID: 23755828; PMCID: PMC3759580.
21. Friedman PN, Shaazuddin M, Gong L, Grossman RL, Harralson AF, Klein TE, Lee NH, Miller DC, Nutescu EA, O'Brien TJ, O'Donnell PH, O'Leary KJ, Tuck M, Meltzer DO, Perera MA. The ACCOuNT Consortium: A Model for the Discovery, Translation, and Implementation of Precision Medicine in African Americans. Clin Transl Sci. 2019 May;12(3):209-217. doi: 10.1111/cts.12608. Epub 2019 Feb 12. PMID: 30592548; PMCID: PMC6510376.
22. Galaviz-Hernández C, Lazalde-Ramos BP, Lares-Assef I, Macías-Salas A, Ortega-Chavez MA, Rangel-Villalobos H, Sosa-Macías M. Influence of Genetic Admixture Components on CYP3A5*3 Allele-Associated Hypertension in Amerindian Populations From Northwest Mexico. Front Pharmacol. 2020 May 11;11:638. doi: 10.3389/fphar.2020.00638. PMID: 32477124; PMCID: PMC7232668.
23. Kosinski C, Li VS, Chan AS, Zhang J, Ho C, Tsui WY, Chan TL, Mifflin RC, Powell DW, Yuen ST, Leung SY, Chen X. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proc Natl Acad Sci U S A. 2007 Sep 25;104(39):15418-23. doi: 10.1073/pnas.0707210104. Epub 2007 Sep 19. PMID: 17881565; PMCID: PMC2000506.
24. Brackman DJ, Yee SW, Enogieru OJ, Shaffer C, Ranatunga D, Denny JC, Wei WQ, Kamatani Y, Kubo M, Roden DM, Jorgenson E, Giacomini KM. Genome-Wide Association and Functional Studies Reveal Novel Pharmacological Mechanisms for Allopurinol. Clin Pharmacol Ther. 2019 Sep;106(3):623-631. doi: 10.1002/cpt.1439. Epub 2019 May 23. PMID: 30924126; PMCID: PMC6941886.
25. Roques BP, Noble F, Daugé V, Fournié-Zaluski MC, Beaumont A. Neutral endopeptidase 24.11: structure, inhibition, and experimental and clinical pharmacology. Pharmacol Rev. 1993 Mar;45(1):87-146. PMID: 8475170.
26. Morice AH, Fontana GA, Sovijarvi AR, Pistolesi M, Chung KF, Widdicombe J, O'Connell F, Geppetti P, Gronke L, De Jongste J, Belvisi M, Dicpinigaitis P, Fischer A, McGarvey L, Fokkens WJ, Kastelik J; ERS Task Force. The diagnosis and management of chronic cough. Eur Respir J. 2004 Sep;24(3):481-92. doi: 10.1183/09031936.04.00027804. PMID: 15358710.
27. Pasaje CF, Kim JH, Park BL, Cheong HS, Kim MK, Choi IS, Cho SH, Hong CS, Lee YW, Lee JY, Koh IS, Park TJ, Lee JS, Kim Y, Bae JS, Park CS, Shin HD. A possible association of EMID2 polymorphisms with aspirin hypersensitivity in asthma. Immunogenetics. 2011 Jan;63(1):13-21. doi: 10.1007/s00251-010-0490-8. Epub 2010 Nov 18. PMID: 21086123.
28. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019 Jul 2;47(W1):W191-W198. doi: 10.1093/nar/gkz369. PMID: 31066453; PMCID: PMC6602461.
29. Stefanini L, Paul DS, Robledo RF, Chan ER, Getz TM, Campbell RA, Kechele DO, Casari C, Piatt R, Caron KM, Mackman N, Weyrich AS, Parrott MC, Boulaftali Y, Adams MD, Peters LL, Bergmeier W. RASA3 is a critical inhibitor of RAP1-dependent platelet activation. J Clin Invest. 2015 Apr;125(4):1419-32. doi: 10.1172/JCI77993. Epub 2015 Feb 23. PMID: 25705885; PMCID: PMC4396462.

30. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013 Apr;138(1):103-41. doi: 10.1016/j.pharmthera.2012.12.007. Epub 2013 Jan 16. PMID: 23333322.
31. Caldwell MD, Awad T, Johnson JA, Gage BF, Falkowski M, Gardina P, Hubbard J, Turpaz Y, Langaee TY, Eby C, King CR, Brower A, Schmelzer JR, Glurich I, Vidaillet HJ, Yale SH, Qi Zhang K, Berg RL, Burmester JK. CYP4F2 genetic variant alters required warfarin dose. Blood. 2008 Apr 15;111(8):4106-12. doi: 10.1182/blood-2007-11-122010. Epub 2008 Feb 4. PMID: 18250228; PMCID: PMC2288721.
32. Lal S, Sandanaraj E, Wong ZW, Ang PC, Wong NS, Lee EJ, Chowbay B. CBR1 and CBR3 pharmacogenetics and their influence on doxorubicin disposition in Asian breast cancer patients. Cancer Sci. 2008 Oct;99(10):2045-54. doi: 10.1111/j.1349-7006.2008.00903.x. PMID: 19016765.
33. Court MH, Hao Q, Krishnaswamy S, Bekaii-Saab T, Al-Rohaimi A, von Moltke LL, Greenblatt DJ. UDP-glucuronosyltransferase (UGT) 2B15 pharmacogenetics: UGT2B15 D85Y genotype and gender are major determinants of oxazepam glucuronidation by human liver. J Pharmacol Exp Ther. 2004 Aug;310(2):656-65. doi: 10.1124/jpet.104.067660. Epub 2004 Mar 25. PMID: 15044558.
34. Staatz CE, Tett SE. Clinical pharmacokinetics and pharmacodynamics of tacrolimus in solid organ transplantation. Clin Pharmacokinet. 2004;43(10):623-53. doi: 10.2165/00003088-200443100-00001. PMID: 15244495.
35. Birdwell KA, Decker B, Barbarino JM, Peterson JF, Stein CM, Sadee W, Wang D, Vinks AA, He Y, Swen JJ, Leeder JS, van Schaik R, Thummel KE, Klein TE, Caudle KE, MacPhee IA. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guidelines for CYP3A5 Genotype and Tacrolimus Dosing. Clin Pharmacol Ther. 2015 Jul;98(1):19-24. doi: 10.1002/cpt.113. Epub 2015 Jun 3. PMID: 25801146; PMCID: PMC4481158.
36. Johnson JA, Gong L, Whirl-Carrillo M, Gage BF, Scott SA, Stein CM, Anderson JL, Kimmel SE, Lee MT, Pirmohamed M, Wadelius M, Klein TE, Altman RB; Clinical Pharmacogenetics Implementation Consortium. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. Clin Pharmacol Ther. 2011 Oct;90(4):625-9. doi: 10.1038/clpt.2011.185. Epub 2011 Sep 7. PMID: 21900891; PMCID: PMC3187550.
37. Wadelius M, Chen LY, Downes K, Ghori J, Hunt S, Eriksson N, Wallerman O, Melhus H, Wadelius C, Bentley D, Deloukas P. Common VKORC1 and GGCX polymorphisms associated with warfarin dose. Pharmacogenomics J. 2005;5(4):262-70. doi: 10.1038/sj.tpj.6500313. PMID: 15883587.
38. Gnatenko DV, Zhu W, Xu X, Samuel ET, Monaghan M, Zarrabi MH, Kim C, Dhundale A, Bahou WF. Class prediction models of thrombocytosis using genetic biomarkers. Blood. 2010 Jan 7;115(1):7-14. doi: 10.1182/blood-2009-05-224477. Epub 2009 Sep 22. PMID: 19773543; PMCID: PMC2803693.
39. Asiimwe IG, Blockman M, Cohen K, Cupido C, Hutchinson C, Jacobson B, Lamorde M, Morgan J, Mouton JP, Nakagaayi D, Okello E, Schapkaitz E, Sekaggya-Wiltshire C, Semakula JR, Waitt C, Zhang EJ, Jorgensen AL, Pirmohamed M. A genome-wide association study of plasma concentrations of warfarin enantiomers and metabolites in sub-

Saharan black-African patients. Front Pharmacol. 2022 Sep 23;13:967082. doi: 10.3389/fphar.2022.967082. PMID: 36210801; PMCID: PMC9537548.
40. Vigil D, Cherfils J, Rossman KL, Der CJ. Ras superfamily GEFs and GAPs: validated and tractable targets for cancer therapy? Nat Rev Cancer. 2010 Dec;10(12):842-57. doi: 10.1038/nrc2960. Epub 2010 Nov 24. PMID: 21102635; PMCID: PMC3124093.
41. Prohaska CC, Zhang X, Schwantes-An TL, Stearman RS, Hooker S, Kittles RA, Aldred MA, Lutz KA, Pauciulo MW, Nichols WC, Desai AA, Gordeuk VR, Machado RF. RASA3 is a candidate gene in sickle cell disease-associated pulmonary hypertension and pulmonary arterial hypertension. Pulm Circ. 2023 Apr 1;13(2):e12227. doi: 10.1002/pul2.12227. PMID: 37101805; PMCID: PMC10124178.
42. Li Z, Zhao W, Shang L, Mosley TH, Kardia SLR, Smith JA, Zhou X. METRO: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. Am J Hum Genet. 2022 May 5;109(5):783-801. doi: 10.1016/j.ajhg.2022.03.003. Epub 2022 Mar 24. PMID: 35334221; PMCID: PMC9118130.
43. Knutson KA, Pan W. MATS: a novel multi-ancestry transcriptome-wide association study to account for heterogeneity in the effects of cis-regulated gene expression on complex traits. Hum Mol Genet. 2023 Apr 6;32(8):1237-1251. doi: 10.1093/hmg/ddac247. PMID: 36179104; PMCID: PMC10077507.
44. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, Johnson WC, Im HK, Liu Y, Wheeler HE. Genetic architecture of gene expression traits across diverse populations. PLoS Genet. 2018 Aug 10;14(8):e1007586. doi: 10.1371/journal.pgen.1007586. PMID: 30096133; PMCID: PMC6105030.