2020

Optimizing Gene Expression Prediction and Omics Integration in Populations of African Ancestry

Paul Chukwuebuka Okoro

LOYOLA UNIVERSITY CHICAGO

OPTIMIZING GENE EXPRESSION PREDICTION AND OMICS INTEGRATION IN POPULATIONS OF AFRICAN ANCESTRY

A THESIS SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY

PAUL CHUKWUEBUKA OKORO

CHICAGO, IL

AUGUST 2020

ACKNOWLEDGMENTS

I would like to thank in a very special way Dr. Heather Wheeler for giving me the opportunity to join in the amazing works of her laboratory in diversifying genomics research to include African ancestries as well as many other minority populations, and consequently shaping me to be a better scientist. I am very thankful to everyone who made this thesis possible, starting with the former and current members of the Wheeler lab – special thanks to both Angela Andaleon and Ryan Schubert for their bioinformatics analytic supports and debugging of my codes whenever I get stuck.

I would also like to thank all my class professors in Loyola. Special thanks to Dr. Dmitry Dligach for giving me the foundation in machine learning which became a big chunk of my thesis. Special thanks to Dr. Heather Wheeler once again for making my first taste of actual bioinformatics and application of my coding skills in bioinformatics analysis easy. Special thanks to Dr. Catherine Putonti for taking me through the rough path of taking both Advanced Bioinformatics and Computational Biology courses in Spring 2019 and coming out unbroken. I also thank all the members of my thesis committee – special thanks to Dr. Lara Dugas for providing me with data that formed the basis of my thesis. Additionally, I would like to thank the Bioinformatics program in Loyola for awarding me the MS Bioinformatics fellowships 2018-2019 and 2019-2020, which enabled me to pursue my research and complete this thesis.

Finally, I would like to thank my family and friends back home in Nigeria, as well as the friends I have made here, for all their support.

To the Glory of God, the Almighty.

It stands to the everlasting credit of science that by acting on the human mind, it has overcome man’s insecurity before himself and before nature.

—Albert Einstein

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFA	MESA African American
AFHI	Combination of AFA & HIS
ALL	Combination of AFA, CAU, & HIS
CAU	MESA European American
CV	Cross Validation
EN	Elastic Net
eQTL	Expression Quantitative Trait Loci
GWAS	Genome Wide Association Study
HDL	High-Density Lipoprotein
HIS	MESA Hispanic American
HWE	Hardy-Weinberg Equilibrium
IBD	Identity By Descent
KNN	K Nearest Neighbor
MAF	Minor Allele Frequency
MESA	Multiethnic Study of Atherosclerosis
METS	Modeling the Epidemiological Transition Study
ML	Machine Learning
PCA	Principal Component Analysis
RF	Random Forests

SVR	Support Vector Regression
TPM	Transcript Per Million
TWAS	Transcriptome Wide Association Study

ABSTRACT

Popular transcriptome imputation methods such as PrediXcan and FUSION use parametric linear assumptions, and thus are unable to flexibly model the complex genetic architecture of the transcriptome. Although non-linear modeling has been shown to improve imputation performance, replicability and potential cross-population differences have not been adequately studied. Therefore, to optimize imputation performance across global populations, we used the non-linear machine learning (ML) models random forest (RF), support vector regression (SVR), and K nearest neighbor (KNN) to build transcriptome imputation models, and evaluated their performance in comparison to elastic net (EN). We trained gene expression prediction models using genotype and blood monocyte transcriptome data from the Multi-Ethnic Study of Atherosclerosis (MESA) comprising individuals of African, Hispanic, and European ancestries and tested them using genotype and whole blood transcriptome data from the Modeling the Epidemiology Transition Study (METS) comprising individuals of African ancestries. We show that the prediction performance is highest when the training and the testing population share similar ancestries regardless of the prediction algorithm used. While EN generally outperformed RF, SVR, and KNN, we found that RF outperforms EN for some genes, particularly between disparate ancestries, suggesting potential robustness and reduced variability of RF imputation performance across global populations. When applied to a high-density lipoprotein (HDL) phenotype, we show including RF prediction models in PrediXcan reveals potential gene associations missed by EN models. Therefore, by integrating non-linear modeling into

PrediXcan and diversifying our training populations to include more global ancestries, we may uncover new genes associated with complex traits. We did not find any significant associations when the prediction models were applied to obesity status and microbiome diversity.

CHAPTER ONE

INTRODUCTION

Genome Wide Association Study (GWAS)

The human genome consists of approximately 3 billion nucleotide base pairs with 99.9% of the DNA sequence similar across humans (Chial, 2008). Despite the high degree of similarity of the genomic sequences across people and populations, there are different levels of variation in the DNA that contribute to the phenotypic manifestation that make us look different from one another as well as lead to different susceptibilities to diseases. The most common form of genetic variations in the DNA is the single nucleotide polymorphism (SNP), where at a single base pair in the genome, the individuals in a population have varying nucleotide sequence.

In recent years, advancements in high-throughput genotyping and sequencing technologies have assayed hundreds of thousands of SNPs leading to an explosion in the amount of genetic data publicly available (Visscher, Brown, McCarthy, & Yang, 2012). Consequently, researchers have leveraged strong statistical analysis to probe single nucleotide genetic variations through genome wide association study (GWAS) of traits of interest (Christensen & Murray, 2007).

Specifically, GWAS involves interrogating the entire genome by conducting multiple statistical association tests between SNPs and traits. Additionally, according to the National Institutes of Health, GWAS is defined as a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits

(Mitchell, Ferguson, & Ferguson, 2007). GWAS has been used by various researchers to successfully identify genetic associations for many complex diseases (MacArthur et al., 2017).

INTEGRATING THE TRANSCRIPTOME INTO GWAS

Expression Quantitative Trait Loci (eQTL)

Although GWAS has been remarkable in identifying disease susceptibility loci for complex traits, there are still many challenges associated with interpreting the results, one of which, is that the functional significance of some of these identified loci are unclear. Simply put, in most cases, knowing that a SNP is statistically associated with a complex trait does not fully shed light into the biological mechanism and regulation of the trait. Thus, GWAS successes are still many steps removed from clinical application, and subsequently, precision medicine. In fact, majority of the discovered significant GWAS disease associated loci have only explained a small portion of the variance in disease risk (Manolio et al., 2009). Indeed, most of these variants identified through GWAS are usually found in the noncoding region of the genome, thereby complicating identification of their functional importance in understanding the biology of complex traits (Huang, 2015; MacArthur et al., 2017; M. I. McCarthy et al., 2008; Visscher et al., 2012).

In a bid to incorporate functional genomics into GWAS – in order to further elucidate the mechanisms behind identified complex disease associations – increased research attention has been paid to the study of regulatory elements that can influence a gene's transcriptional activities and consequently alter phenotypes (Li et al., 2018). One important class of such regulatory elements are called the expression quantitative trait loci (eQTLs) (Albert & Kruglyak, 2015). Indeed, many studies have shown that the noncoding regions of the genome are particularly

enriched for gene regulatory variants such as eQTLs. This suggests that genetically regulated gene expression might play a critical role in explaining the phenotypic variability in a wide range of complex traits (Aguet et al., 2017; Gamazon, Huang, Cox, & Dolan, 2010; Gamazon et al., 2013; Nicolae et al., 2010). In fact, the Genotype-Tissue Expression (GTEx) Project aimed to collect a comprehensive set of eQTLs from different human tissues and to provide the scientific community a database of genetic associations with molecular traits such as mRNA levels (Aguet et al., 2017; GTEx Consortium, 2015; Lonsdale et al., 2013). More so, given that a handful of SNPs have large effect associations that can explain most of the heritable component of gene expression traits, mathematical modeling of the relationship between genotype and gene expression is achievable using moderate sample sizes (Wheeler et al., 2016). Indeed, this has led to the development of transcriptome methods such as PrediXcan (Gamazon et al., 2015) and FUSION (Gusev et al., 2016) which integrate cis-eQTL genotype and transcriptome datasets in order to predict the transcriptome from GWAS data, and subsequently test for association between the predicted transcriptome and trait of interest. cis-eQTLs are SNPs located near the target gene, usually within 1 mega base, and tend to have larger effect sizes than trans-eQTLs, which are farther away or on different chromosomes. Because most GWAS lack corresponding transcriptome data, these methods may identify gene regulatory mechanisms underlying complex traits as well as better interpretability of the direction of effect of uncovered genetic associations.

Optimizing Transcriptome Prediction

The mathematical model used in PrediXcan is Elastic Net (EN) (Zou & Hastie, 2005) while FUSION uses Bayesian Sparse Linear Mixed Model (BSLMM) (Zhou, Carbonetto, & Stephens, 2013). The EN model used by PrediXcan is a combination of L1 (LASSO) (Tibshirani,

1996) and L2 (Rigde) (Hoerl & Kennard, 1970) regularization of the cis-eQTLs effect sizes, thus assuming a parametric prior for the cis-eQTLs. The same parametric assumption is made by FUSION since BSLMM assumes a normal mixture prior, combining Bayesian Variable Selection Regression (BVSR) (Guan & Stephens, 2011) and Linear Mixed Modeling (LMM) (Yu et al., 2006). Given their parametric and linear assumptions, these tools fail to flexibly model the distributions of the cis-eQTL genotypes and their relationship with gene expression (Nagpal et al., 2019). Studies have shown that some cis-eQTL relationships can be best modeled mathematically with non-linear and non-parametric assumptions (Manor & Segal, 2013; Nagpal et al., 2019). Manor and Segal showed that by using very simple non-linear modeling with the *K* Nearest Neighbor (KNN) (Cover & Hart, 1967) algorithm, robust gene expression prediction can be achieved using just cis-eQTLs. Wang et al. 2016, found that a mixed model based random forest (Breiman, 2001) (a non-linear model) has the potential to capture the non-linear relationships of cis-eQTLs and gene expression, and thus improve imputation performance. Most recently, a method called TIGAR (Nagpal et al., 2019), which is based on a non-parametric Bayesian method called Dirichlet process regression (Zeng & Zhou, 2017), was shown to achieve better imputation R^2 than PrediXcan on simulation data where at least 1% of the cis-eQTLs are causal and true expression heritability is at most 0.2. TIGAR (Nagpal et al., 2019) was also shown to impute expression for more genes than PrediXcan in a real dataset, thus corroborating the potential of using non-parametric and non-linear modeling of gene expression prediction in order to uncover more gene associations with complex traits.

Although several studies have shown that non-linear modeling of cis-eQTLs and gene expression can improve imputation performance R^2 (Manor & Segal, 2013; Nagpal et al., 2019;

J. Wang et al., 2016), we sought to further explore the cross-population portability of non-linear modeling of cis-eQTLs and gene expression with new cohorts. Generally, a large UK Biobank based study has acknowledged the discrepancy in genetic prediction due to lack of diversity in training cohorts (Martin et al., 2017). More specifically, the importance of genetic ancestries diversity in gene expression prediction has also been corroborated by many recent studies (Fryett, Morris, & Cordell, 2020; Keys et al., 2019; Mikhaylova & Thornton, 2019; Mogil et al., 2018). Using parametric and linear modeling, these studies have shown that similarity in ancestries between the training and testing population improves gene expression prediction (Fryett et al., 2020; Keys et al., 2019; Mikhaylova & Thornton, 2019; Mogil et al., 2018). However, the replicability of these observations and the potential cross-population differences with non-linear machine learning modeling have not been adequately studied.

Microbiome

While some variants in the DNA, through GWAS and eQTL studies, have been discovered to be associated with many complex diseases, many other aspects of the human ecosystem contribute to diseases or influence morbidity. For centuries, microscopic living organisms known collectively as microbes have been studied and identified by scientists as the cause of many diseases in humans. The human gastrointestinal tract is inhabited by these microbes such as bacteria, viruses, fungi, archaea, and protozoa, all of which are collectively referred to as the gut microbiota (Davis, 2016). In fact, the human gut microbiota contains about 100 trillion microbes, all of which when combined, have about 100 times more genes (the microbiome) than are found in the entire human genome (Qin et al., 2010).

These microbes in humans have been found to play important roles such as breaking down soluble fiber and non-digestible nutrients, producing vitamins, metabolizing xenobiotics, preventing colonization by pathogens, and supporting development of a mature immune system (Bergman, 1990; Davis, 2016; den Besten et al., 2013; Krebs et al., 2002). In fact, microbiome studies have shown that there is a relationship between nutrients, gut microbiota, and human diseases such as obesity (Davis, 2016). Specifically, by assisting in breaking down fiber and non-digestible nutrients, the microbes directly and indirectly regulate adiposity and energy homeostasis through a genetic pathway and potential eQTL associations (GTEx Consortium, 2015; Hong et al., 2005; Kimura, Inoue, Hirano, & Tsujimoto, 2014). Thus, host microbiome compositional differences provide biomarkers that could be tested for risk or presence of diseases (Chassaing, Aitken, Gewirtz, & Vijay-Kumar, 2012; Clemente, Ursell, Parfrey, & Knight, 2012; Karlsson, Tremaroli, Nielsen, & Bäckhed, 2013). Microbiome diversity is measured in terms of diversity within (Alpha Diversity) and between (Beta Diversity) study samples (Kuczynski et al., 2010; Lozupone & Knight, 2008).

Genetics of Lipid and Obesity

Many studies have associated cardiovascular diseases with obesity and lipid measurements such as total cholesterol, high-density lipoprotein (HDL), triglycerides, as well as low-density lipoprotein (LDL) (Akil & Ahmad, 2011; Carbone et al., 2019; Rader & Hovingh, 2014; Stone et al., 2014). Focusing on lipids, HDL is considered the good cholesterol, while LDL is the bad one (Mozaffarian et al., 2016). Indeed, studies have shown that decreased HDL levels and increased LDL levels are associated with heart attacks and strokes (Stone et al., 2014). The “goodness” of HDL is due to its inherent property of being less prone to oxidation, and its

role in carrying cholesterol from tissues back to the liver as well as transporting lipid molecules out of arterial walls, thereby reducing the amount of cholesterol in circulation (Feingold & Grunfeld, 2018). Unchecked excessive accumulation of fat molecules in the artery causes blockage of blood flow, thereby causing stroke if the blockage occurs in the brain or heart attack if the blockage is in the coronary artery (Scott, 2004).

Obesity is one of the leading causes of cardiovascular disease mortality and morbidity (Akil & Ahmad, 2011). Obesity has been defined by the World Health Organization as abnormal or excessive fats that accumulate and present a risk to health (Mamat et al., 2011). Obesity is measured in terms of body mass index (BMI). BMI is calculated by dividing the body weight (kilograms) with the square of the body height (meters) such that a person with BMI score of 30 or above is considered obese (Mamat et al., 2011; Wyatt, Winters, & Dubbert, 2006). Obesity is a major risk factor for the development of diseases such as type-2 diabetes, hypertension, and coronary artery disease (Poirier et al., 2006; Ritchie & Connell, 2007). Indeed, obesity has been found to increase cardiovascular disease mortality and morbidity (Van Gaal, Mertens, & Christophe, 2006).

In recent years, scientists have sought to leverage the advances in next generation sequencing techniques to identify and understand the genetic variations underlying complex traits such as obesity and cardiovascular diseases. Through GWAS, some SNPs have been found to associate with lipid and obesity phenotypes. A GWAS using Framingham Heart Study (FHS) data identified twenty-nine genome-wide significant ($P < 5 \times 10^{-8}$) SNPs associated with total cholesterol and HDL-cholesterol (Ma et al., 2010). Indeed, GWAS has uncovered about 100 loci associated with lipid traits and experimental follow-up on the GWAS loci has identified the

functional relevance of genes *GALNT2, TRIB1, PNPLA3, SUGP1, SOCS2, RAMP3, APOB, CETP, ZPR1, FAAH* and *SORT1* in lipid traits study (Andaleon, Mogil, & Wheeler, 2019; DiStefano et al., 2015; Ma et al., 2010; Willer & Mohlke, 2012; Wood et al., 2013). In the same vein, a GWAS on obesity and related traits identified seventeen SNPs significantly associated with obesity status and waist to hip ratio (K. Wang et al., 2011). These SNPs were found within the *FTO* gene as well as *NRXN3* gene (K. Wang et al., 2011). SNP variations in the *FTO* gene region has been found to significantly associate with BMI and risk of obesity across multiple study populations (Fawcett & Barroso, 2010).

Lack of Diversity in Genetic Studies

While GWAS has been applied to shed light on many complex traits such as obesity and cardiovascular disease, majority of the studies were carried out largely on populations of European ancestries (Martin et al., 2019). In fact, the largest GWAS and meta-analysis to understand obesity biology was carried out largely in populations of European descent (Locke et al., 2015). Similarly, many lipid trait GWAS have been performed in predominately European individuals, including one from the Global Lipids Genetics Consortium of over 100,000 people (DiStefano et al., 2015; Teslovich et al., 2010; Wood et al., 2013). Generally, a study has shown that predicting disease risk based on European GWAS is skewed in African populations (Martin et al., 2017). Strikingly, the burden of obesity is disproportionately higher in US based adults of recent African origin when compared to populations of other ancestries (Dugas et al., 2017; Flegal, Kruszon-Moran, Carroll, Fryar, & Ogden, 2016). While the observed disproportionate burden of obesity is true, genetic differences alone cannot account for this disparate prevalence

because environment, social behavior and culture, diet, and consequent microbiota composition collectively play a big role (Archie & Tung, 2015; Singh et al., 2017).

Summary

In this thesis, we sought to optimize other machine learning models such as random forests (RF), support vector regression (SVR), and k nearest neighbor (KNN) for transcriptome prediction within and across populations, in comparison to the standard transcriptome prediction tool – PrediXcan – built on elastic net (EN). Additionally, we performed integrative transcriptome and gut microbiome studies to explore the possibility of discovering gene associations with HDL and obesity through a transcriptome wide association study (TWAS). In the machine learning comparisons, we found that gene prediction models were generally best in EN and closely followed by RF. Additionally, we corroborated previous findings that similarity in ancestry improves gene expression prediction accuracy. In the integration of the predicted transcriptome and microbiome to TWAS of HDL and obesity, we found a gene association reported in previous studies.

Next, we describe our methods in chapter two and present results in chapter three. We end with a discussion of our findings and directions for future research in chapter four.

CHAPTER TWO

METHODS

GENOMIC AND TRANSCRIPTOMIC TRAINING DATA PREPARATION

The Multi-Ethnic Study of Atherosclerosis (MESA)

The MESA cohort is made up of 6814 individuals recruited from 6 sites across the USA (Baltimore, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; northern Manhattan, NY; St. Paul, MN) and consists of 53% female and 47% male individuals between the ages of 45-84 (Bild et al., 2002) with the demographics approximately distributed as 38% European American (CAU), 23% Hispanic American (HIS), 28% African American (AFA), and 11% Chinese American (CHN). From the whole cohort, RNA was extracted from CD14+ monocytes from 1264 individuals across the three subpopulations (AFA, HIS, CAU) and quantified on the Illumina Ref-8 BeadChip (Liu et al., 2013). Individuals with both genotype (dbGaP: phs000209.v13.p3) and expression data (GEO: GSE56045) included 234 AFA, 386 HIS, and 582 CAU. Illumina IDs were converted to Ensembl IDs using the RefSeq IDs from MESA and GENCODE (Harrow et al., 2012) version 18 (gtf and metadata files) to match Illumina IDs to Ensembl IDs. If there were multiple Illumina IDs corresponding to an Ensembl ID, the average of those values was used as the expression level.

MESA Genotype Data Analysis and Quality Control

Genotype quality control and imputation were performed as previously described (Mogil et al., 2018). To summarize, all MESA population genotypes were in genome build

GRCh37/hg19. We used the Michigan Imputation Server and 1000 genomes phase 3 v5 reference panel and Eagle v2.3 to impute genotypes in each of the MESA subpopulation. The imputation reference populations were EUR for CAU and mixed population for AFA and HIS (Das et al., 2016; GTEx Consortium, 2015; Loh et al., 2016). Imputation results were first filtered by $R^2 < 0.8$, minor allele frequency (MAF) > 0.01, and ambiguous strand SNPs were removed. After filtering, 9,352,383 SNPs in AFA, 7,201,805 SNPs in HIS, and 5,559,636 SNPs in CAU were remaining for further analysis. PLINK (https://www.cog-genomics.org/plink2) (Chang et al., 2015) was used for quality control and cleaning of the genotype data. We removed SNPs with call rate < 99%, and LD pruned the resulting SNPs by removing 1 SNP in a 50 SNP window if $r^2 > 0.3$. We conducted identity by descent (IBD) analysis on the genotype data and removed one pair of related individuals (IBD > 0.05). The cleaned genotypes were merged with HapMAP populations (Yoruba in Ibadan, Nigeria (YRI), Utah residents with Northern and Western European descent (CEU), and East Asians from Beijing, China and Tokyo, Japan (ASN)) and principal component analysis was done both across and within populations using EIGENSTRAT (Price et al., 2006). After quality control, the final sample sizes used for the gene expression prediction model training are AFA = 233, HIS = 352, and CAU = 578. The final sample sizes used for downstream TWAS analysis are AFA=1188, HIS=952, and CAU=1716.

MESA Transcriptome Data and Quality Control

PEER factor analysis was performed on the expression data of each subpopulation using the PEER R package (Stegle, Parts, Piipari, Winn, & Durbin, 2012). Mogil et al. showed that the true positive replication rate was similar for 10, 20, and 30 PEER factors. As such, we used 10 PEER factors to adjust for potential batch effects and experimental confounders in the measured

gene expression data. Then, we quantile normalized adjusted expression levels for use in model building.

GENOMIC AND TRANSCRIPTOMIC TEST DATA PREPARATION

The Modeling the Epidemiology Transition Study (METS)

The METS cohort comprises of 2506 healthy individuals of African origin between the ages of 25-45 years, with approximately 500 (~50% male) from each of sites; Ghana, South Africa, Seychelles, Jamaica and Chicago, US (Luke et al., 2011). Out of these cohort, 77 female individuals (38 Ghana and 39 Chicago, US) underwent genome-wide genotyping on the Illumina Infinium Multi-Ethnic AMR/AFR BeadChip and RNA sequencing (RNA-seq) from whole blood using the NuGEN mRNA-Seq with AnyDeplete Globin library preparation kit. Single-end 50bp RNA-seq was performed by the Duke University Sequencing and Genomic Technologies Shared Resource. (Loyola IRB #210260091217).

METS Genotype Data Analysis and Quality Control

The METS genotype data is in genome build GRCh38/hg38. We performed all quality control using PLINK v1.90b4.4 (Chang et al., 2015). We removed SNPs on non-autosomal chromosomes, below a call rate threshold of 0.01, or not in Hardy-Weinberg equilibrium (HWE) ($P< 0.00001$). Prior to identity by descent (IBD) and principal component analysis (PCA), we LD-pruned variants using PLINK's --indep-pairwise option at thresholds 50 5 0.3. Due to small sample size, we did not remove individuals based on cryptic relatedness. Prior to PCA, we merged METS genotypes with HapMap reference populations and filtered the merged population for missingness (--geno 0.01) and minor allele frequency (MAF) (--maf 0.01) and performed LD-pruning (--indep-pairwise 50 5 0.3). We performed METS genotype imputation on the

Sanger Imputation service (Loh et al., 2016; S. McCarthy et al., 2016) using the African Genome Resources reference panel. After imputation, non-ambiguous strand SNPs in HWE ($P > 0.05$), with MAF > 0.05 and imputation $R^2 > 0.8$ were retained and the cleaned genotypes were lifted over to genome build GRCh37/hg19 for gene expression prediction analyses.

METS Transcriptome Data Analysis and Quality Control

We used FASTQC (Andrews et al., 2012) to analyze RNA-seq quality and found 50 high fidelity bases with no primers or over-represented sequences. We quantified gene expression using Salmon pseudoalignment (Patro, Duggal, Love, Irizarry, & Kingsford, 2017), which estimates the transcripts per million (TPM) for each gene using a reference transcriptome without performing the time-consuming process of an actual alignment. We used only protein-coding genes as defined by GENCODE (Harrow et al., 2012) version 28 and removed genes with mean TPM < 0.01. The resulting expression data were quantile normalized and PEER factor analyzed (Stegle et al., 2012). Since the study population originates from two divergent country populations (Ghana and USA), the Ghana individuals and USA individuals were subsequently corrected separately using 10 PEER factors to adjust for potential batch effects and experimental confounders. Then, adjusted expression levels were quantile normalized for use in model building.

MODEL BUILDING

Prediction Models

We used MESA expression values for protein coding genes and genotypes of SNPs within 1 Mb of each gene, i.e. in cis, to fit the models. We used the fitted model to predict expression in METS. Model performance were evaluated by Spearman correlation (ρ) of the

METS predicted and measured gene expression values defined by GENCODE (Harrow et al., 2012) version 28. Like prior studies, we considered $\rho > 0.1$ as significant (Gamazon et al., 2015; Mogil et al., 2018).

Elastic Net

We used the glmnet R package (Friedman, Hastie, & Tibshirani, 2010) to implement elastic net (EN) with the alpha parameter set at 0.5, which has previously been shown to perform optimally for predicting gene expression (Wheeler et al., 2016). For every single gene, we carried out nested cross-validation of the EN model as follows: firstly, training data was split into roughly five equal parts, secondly for each held-out fold, ten-fold cross-validation was performed on the remaining four folds to minimize the lambda parameter, and the model with the minimal lambda was used to predict on the held-out fold to determine the coefficient of determination (R^2). After going through each of the five folds, we used the average R^2 as our measure of model performance. The trained models with minimal lambda were used to predict expression in the test data (Mogil et al., 2018).

Random Forest

We used the scikit-learn Python package version 0.21.2 (Pedregosa et al., 2011) (Python version 3.7.3) to implement random forest (RF) regression and all the hyperparameters in the regressor were set to default except for the *n_estimators* hyperparameter (which is the number of trees in the forest). For every single gene, via five-fold cross-validation, we conducted a grid search of the best *n_estimators* hyperparameter ranging from 50 to 500 inclusive that yields the highest cross-validated regression coefficient of determination (R^2). The range of trees used in the grid search were informed by our preliminary analysis result as shown in Figure 1.

Subsequently, for every gene, we used the resulting best *n_estimators* hyperparameter to fit the random forest regressor model and predict expression in the test data.

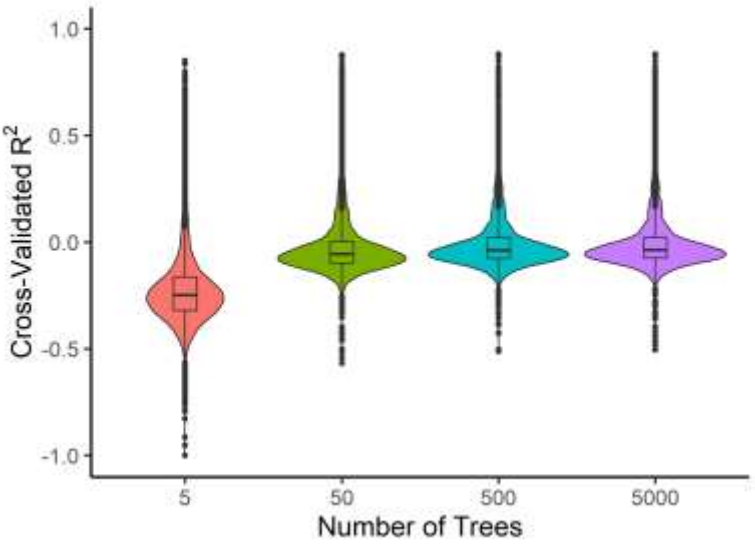

Figure 1. Random Forests Trees Performance**.** We compared the distribution of the cross-validated (CV) R^2 of all genes at different random forest number of trees (5, 50, 500, 5000). This informed the range of trees we used in the random forest model building hyperparameter tuning. In this plot, gene models with CV $R^2 < -1$ were filtered out.

K Nearest Neighbor

We used the scikit-learn Python package version 0.21.2 (Pedregosa et al., 2011) (Python version 3.7.3) to implement K nearest neighbor (KNN) regression. The hyperparameters were set to default except for *n_neighbours* (which is the number of neighbors (k) to use), *weights* (which is a weight function used in the prediction), and *P* (which is the power parameter for the Minkowski metric). We used two of the *weights* function parameters namely 'uniform' (wherein all points in each neighborhood are weighted equally) and 'distance' (wherein all points in each neighborhood are weighted by the inverse of their distance). For every gene, via five-fold cross-validation, we conducted a grid search of the best three hyperparameter combinations that yield the highest cross-validated regression coefficient of determination (R^2). The three

hyperparameter combinations were drawn from *k* (odd numbers between 3 and 31 inclusive), *weights* (uniform and distance), and *P* (1,2,3). Subsequently, for every gene, we used the resulting best hyperparameter combination to fit the KNN regressor model and predict expression in test data.

Support Vector Machine

We used the scikit-learn Python package version 0.21.2 (Pedregosa et al., 2011) (Python version 3.7.3) to implement support vector regression (SVR). We set all parameters to default except for the following: *gamma* (which was set to 'scale'), *kernel* (which is the type of kernel to use in the model), *degree* (which is specifically for the degree of the polynomial kernel function), and *C* (which is the penalty for error term). For every gene, via five-fold cross-validation, we conducted a grid search of the best three hyperparameter combinations that yield the highest cross-validated regression coefficient of determination (R^2). The three hyperparameter combinations were drawn from *kernel* ('linear', 'poly', 'rbf', 'sigmoid'), *degree* (2,3,4,5,6,7) and *C* (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 2.0). Subsequently, for every gene, we used the resulting best hyperparameter combination to fit the SVR regressor model and predict expression in test data.

METS GUT MICROBIOME

Microbiome Analysis

Microbial genomic DNA was extracted from the stool of 61 female individuals divided into 25 African Americans residing in Chicago USA, and 36 Ghana natives residing in Ghana (Individuals in the METS cohort). The V4 region of the 16S rRNA gene was amplified and paired end multiplex sequencing performed on Illumina Miseq platform (Dugas et al., 2018). The

resulting raw sequences were processed with DADA2 R package version 1.16.0 (Callahan et al., 2016). Quality control and filtering of low quality regions of the sequences (Figures 2 and 3) were performed with the DADA2 filterAndTrim function using the following parameters: truncLen = c(275, 175), maxEE = c(2,2), truncQ = 2. The DADA2 denoised forward and reverse reads were merged (Parameters: mergePairs(justConcatenate = TRUE)), and chimera sequences were removed, yielding the final Amplicon Sequence Variant (ASV) table. The ASVs were classified to species level using the DADA2 formatted training set of the SILVA reference database (Callahan et al., 2016; Quast et al., 2012). Alpha diversity indices such as Shannon, Fisher, and Inverse Simpson (Peet, 1974) were calculated using Phyloseq R package version 1.30.0 (McMurdie & Holmes, 2013).

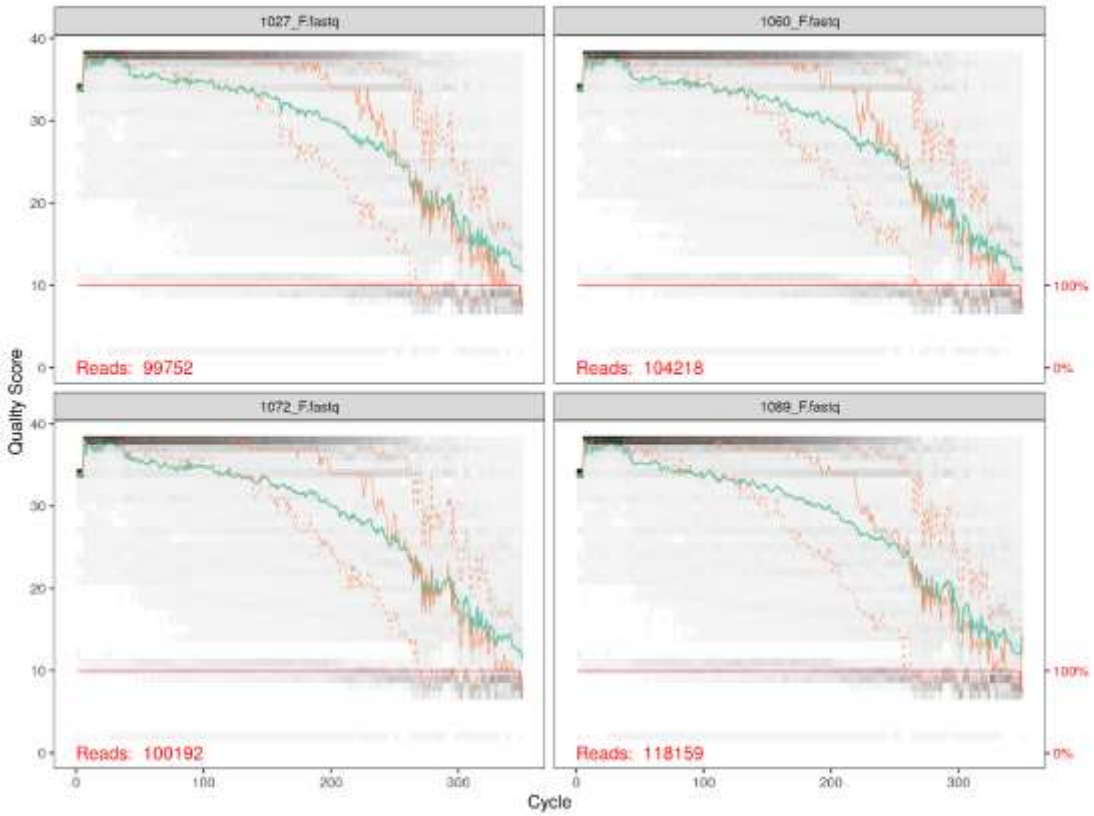

Figure 2. Sample of raw forward sequences before filtering. Quality profiles of the forward reads wherein Phred quality score is on the y axis and base position is on the x axis. The grey scale in each plot is a heat map of the frequency of each quality score at each base position. The green line represents mean quality score at each position while the orange line represents the quartiles of the quality score distribution. The red line shows the scaled proportion of reads that extend to at least that position.

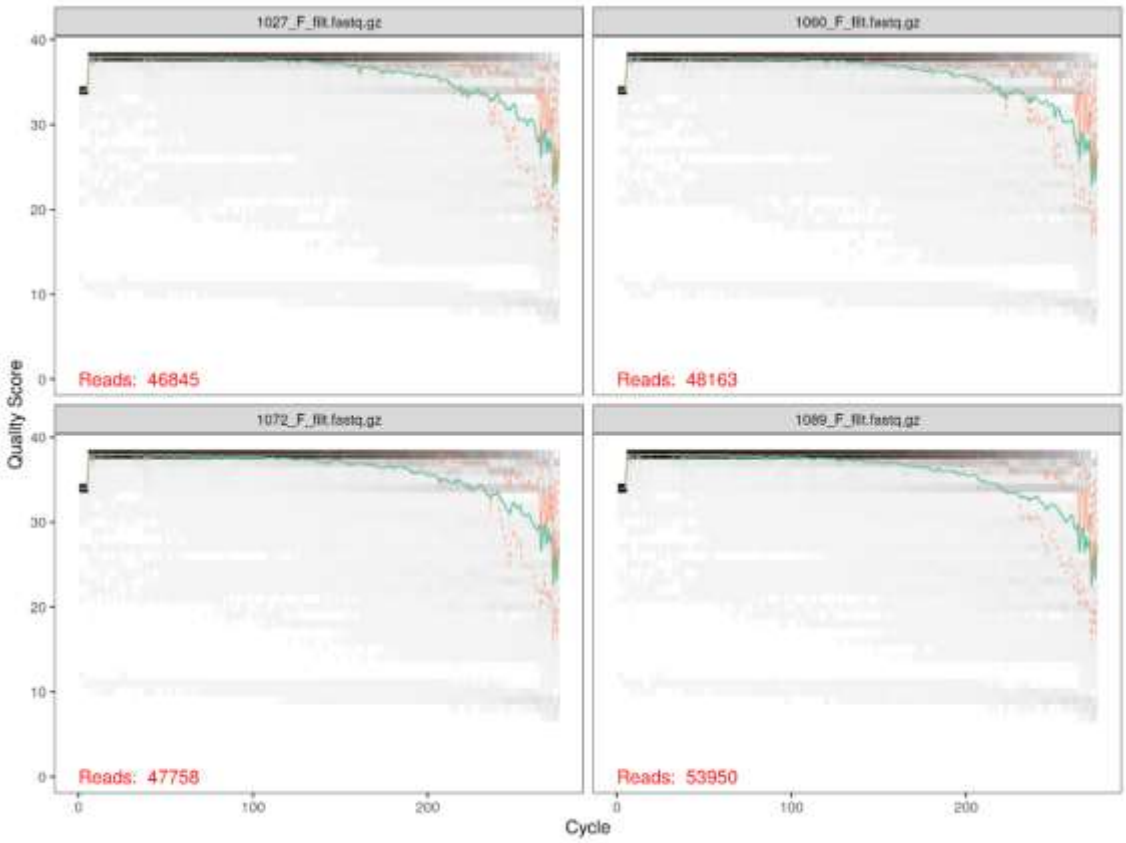

Figure 3. Sample of raw forward sequences after filtering. Quality profiles of the forward reads wherein Phred quality score is on the y axis and base position is on the x axis. The grey scale in each plot is a heat map of the frequency of each quality score at each base position. The green line represents mean quality score at each position while the orange line represents the quartiles of the quality score distribution. The red line shows the scaled proportion of reads that extend to at least that position.

CHAPTER THREE

RESULTS

Elastic Net Outperforms Machine Learning Models for Cross-Validated Gene Expression Prediction

We sought to determine if non-parametric machine learning models could improve SNP-based imputation of the transcriptome across populations compared to the parametric elastic net models currently used in PrediXcan (Gamazon et al., 2015). We trained each of the machine learning algorithms, random forest (RF), support vector regression (SVR), and K nearest neighbor (KNN), using genotype and blood monocyte transcriptome data from each subpopulation in the Multi-Ethnic Study of Atherosclerosis (MESA). The training samples in the MESA subpopulations are distributed as African Americans (AFA, n=233), European Americans (CAU, n=578), and Hispanic Americans (HIS, n=352). To have a larger sample size, we also combined the genotype and transcriptome of the MESA subpopulations (AFA, HIS, CAU) into the ALL cohort (n=1163). Standard quality control analysis was done on the genotype data. We also adjusted for potential batch effects and experimental confounders in the transcriptome data using PEER factor analysis (see Methods) and for population structure using the first 3 genotypic principal components. Using each of the MESA subpopulations and ALL, we then performed model training through 5-fold cross-validation of RF, SVR, and KNN, and nested cross validation of EN by using SNPs within 1 Mb of each gene to predict its expression level. We used the coefficient of determination (R^2) between predicted and observed expression as our measure of model performance. We found that across all the subpopulations and prediction

algorithms, *ERAP2, HLA-C, HLA-DRB1, CHURC1, RAD51*, and S*NAP29* have $R^2 > 0.5$, thus suggesting their SNP predictors are conserved across global populations. We also found that EN usually outperformed the ML models, but RF outperformed EN on many gene models, especially those trained in HIS and CAU (Figures 4 and 5). This suggests that different prediction algorithms may be potentially more robust for different training populations.

Focusing only on the model training built in the ALL cohort, the model building converged and completed for 9623 genes in RF, SVR, KNN, and 9622 in EN. The 9622 genes in EN models are also in SVR and KNN, while 9621 are in RF. The average R^2 for each of the prediction algorithms are EN=0.0733, SVR=0.0476, RF=0.0409, and KNN=0.0103. *TACSTD2, RNF150, HLA-DRB5, HLA-DRB1, CHURC1* genes have $R^2 > 0.8$ across EN, RF and SVR models while all genes in the KNN model have $R^2 < 0.8$. Overall, EN outperformed all ML models (Figure 4). Focusing on the overlapping genes with $R^2 > 0.01$ (EN vs SVR =3736, EN vs RF =3635, EN vs KNN = 2598), EN performed better on approximately 99%, 97%, and 93% of the overlapping genes than KNN, SVR, and RF, respectively. Table 1 shows the number of genes that have models in each of the prediction algorithm at different R^2 thresholds. EN had the most genes with models compared to the ML methods across all thresholds. However, at $R^2 > 0.5$, RF has almost same number of gene models as EN (RF=194, EN=222), distantly followed by SVR, while KNN has just 28 genes. This clearly shows that EN, RF, and SVR models have generally good performance for most of the highly predictable genes. The same comparison trend is generally observed in the imputation models trained with AFA, CAU, and HIS (Tables 2, 3, and 4). However, unlike ALL and AFA, we observed that RF outperformed EN on HIS and CAU trained data (Figure 4). This suggests integrating both EN and RF models into

transcriptome prediction may be useful. Next, we sought to determine how our models performed in an independent test cohort.

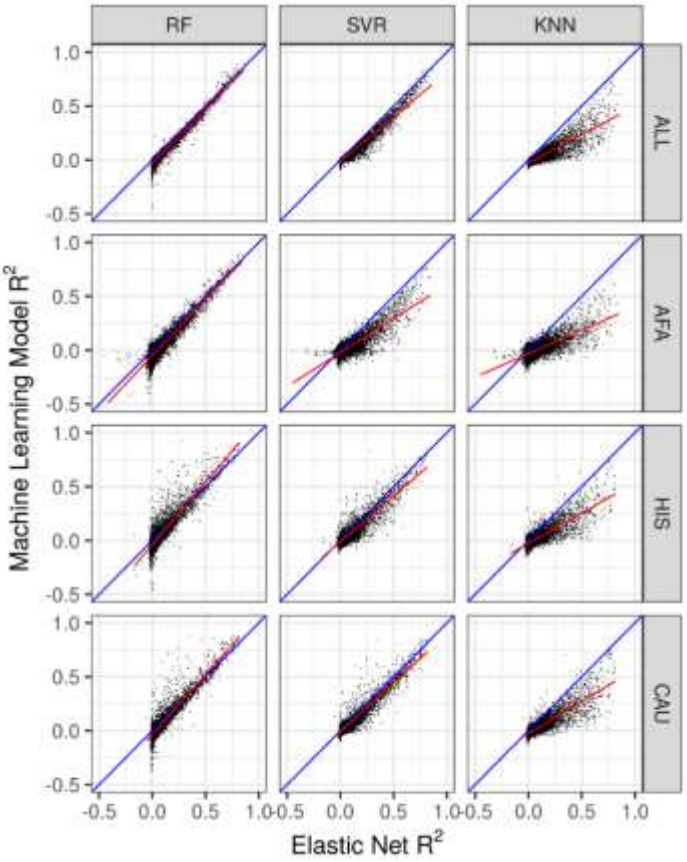

Figure 4. Comparison of the Cross-Validated Gene Expression Prediction Performance in the MESA Cohort. Machine learning (ML) models prediction R^2 compared to elastic net across MESA subpopulations wherein each point on the plot is a gene. The linear regression fit is shown by the red line and the identity line (slope=1) is blue in each plot. In the ALL cohort (combination of AFA, HIS, CAU subpopulations), RF model have 9621 genes while SVR and KNN models have 9622 genes in common with Elastic Net. Pearson correlations (R) between Elastic Net (EN) performance and Random Forest (RF), Support Vector Regression (SVR), and K nearest neighbor (KNN) are 0.98, 0.97, and 0.89, respectively. In the AFA cohort, the overlapping genes between models are RF vs EN = 9608, SVR and KNN vs EN = 9609 while the R are 0.93, 0.86, and 0.75, respectively. In the HIS cohort, ML models have 9499 genes in common with EN, and the R between EN and RF, SVR, and KNN are 0.91, 0.92, and 0.84, respectively. In the CAU cohort, ML models have 9499 genes in common with EN, and the R between EN and RF, SVR, and KNN are 0.94, 0.96, and 0.88, respectively. EN generally outperformed RF, SVR, and KNN, except for some genes where RF outperforms EN, particularly in the HIS and CAU subpopulations.

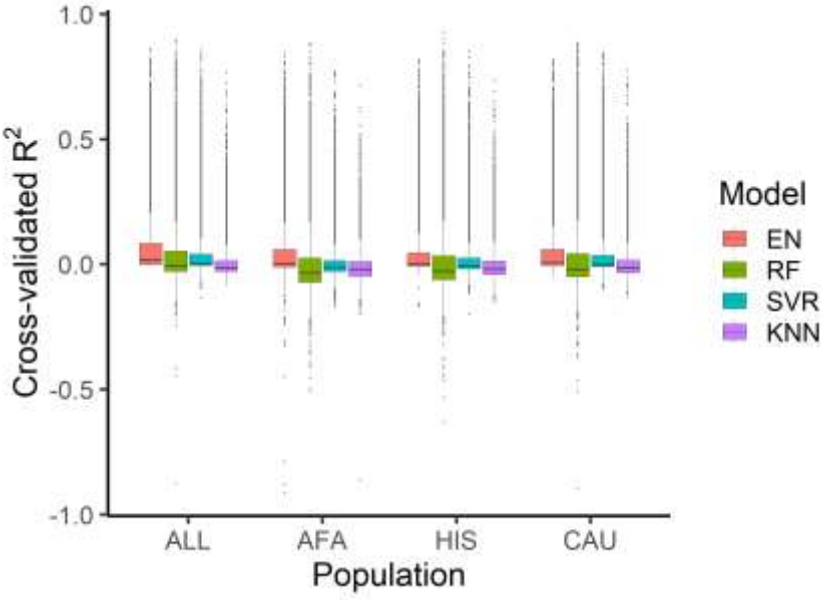

Figure 5. Distribution of the Cross-Validated Gene Expression Prediction Performance in the MESA Cohort. The distribution of gene models with CV $R^2 > -1$ in the ALL (EN=9622, RF=9623, SVR=9623, KNN=9623), AFA (EN=9609, RF=9622, SVR=9623, KNN=9623), HIS (EN=9621, RF=9501, SVR=9501, KNN=9501), and CAU (EN=9621, RF=9501, SVR=9501, KNN=9501) cohorts. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), K Nearest Neighbor (KNN).

Table 1. Numbers of genes with expression prediction models for each method after filtering by cross-validated R^2 in the ALL cohort. Total gene models before filtering; EN=9622, RF=9623, SVR=9623, KNN=9623. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbor (KNN).

Method	**$R^2 > -0.1$**	**$R^2 > -0.01$**	**$R^2 > 0$**	**$R^2 > 0.01$**	**$R^2 > 0.05$**	**$R^2 > 0.1$**	**$R^2 > 0.5$**
EN	9622	9621	6823	5729	3176	2108	222
RF	9544	4924	4158	3651	2449	1687	194
SVR	9622	8929	5355	3772	2185	1454	141
KNN	9263	4193	3206	2601	1422	839	28

Table 2. Number of genes with expression prediction models for each method after filtering by cross-validated R^2 in the AFA cohort. Total gene models before filtering; EN=9623, RF=9623, SVR=9623, KNN=9623. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbor (KNN).

Method	**R^2 > -0.1**	**R^2 > -0.01**	**R^2 > 0**	**R^2 > 0.01**	**R^2 > 0.05**	**R^2 > 0.1**	**R^2 > 0.5**
EN	9589	6641	4860	4051	2601	1814	181
RF	8538	3608	3165	2841	1970	1398	157
SVR	9574	4492	3258	2648	1462	917	52
KNN	9361	3864	3093	2473	1163	581	10

Table 3. Number of genes with expression prediction models for each method after filtering by cross-validated by R^2 in the HIS cohort. Total gene models before filtering EN=9621, RF=9501, SVR=9501, KNN=9501. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbor (KNN).

Method	**R^2 > -0.1**	**R^2 > -0.01**	**R^2 > 0**	**R^2 > 0.01**	**R^2 > 0.05**	**R^2 > 0.1**	**R^2 > 0.5**
EN	9618	8009	5038	3959	2288	1532	147
RF	8858	3701	3295	2976	2101	1530	187
SVR	9497	5630	3841	3056	1784	1153	95
KNN	9460	3914	3135	2529	1317	716	17

Table 4. Number of genes with expression prediction models for each method after filtering by cross-validated R^2 in the CAU cohort. Total gene models before filtering EN=9621, RF=9501, SVR=9501, KNN=9501. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbor (KNN).

Method	**R^2 > -0.1**	**R^2 > -0.01**	**R^2 > 0**	**R^2 > 0.01**	**R^2 > 0.05**	**R^2 > 0.1**	**R^2 > 0.5**
EN	9621	9405	5758	4314	2619	1753	221
RF	9210	4025	3527	3108	2214	1577	241
SVR	9501	7084	4402	3387	2059	1396	178
KNN	9496	4089	3202	2606	1481	878	38

Similarity in Ancestry Improves Prediction Performance Across Prediction Models

Recent studies using EN have observed that similarity in training and testing population improves prediction performance (Fryett et al., 2020; Keys et al., 2019; Mikhaylova & Thornton, 2019; Mogil et al., 2018). Consequently, in order to see if the same observation will be replicated using nonlinear ML algorithms, we used new genotype and whole blood transcriptome data from 77 African American individuals in Chicago, USA and Africans in Ghana enrolled in the Modelling the Epidemiology Transition study (METS) as a replication cohort (Luke et al., 2011, 2014). We performed standard quality control and adjusted for potential confounders in the METS genotype and transcriptome data (see Methods). We predicted gene expression in the METS cohort using only gene models with CV $R^2 > 0.01$ in each of the prediction algorithms trained with the MESA cohort. Specifically, we tested models trained in each of the MESA subpopulations; AFA=233, HIS=352, CAU=578, and the combined population; ALL=1163. To accommodate for any effect sample size may have in our study, we also used the combination of AFA and HIS subpopulations (AFHI=585), which is a similar sample size as CAU, to train the prediction algorithms. Both AFA and HIS contain recent African admixture and thus share more genetic ancestries with our test cohort (METS) than CAU (Figure 6). To determine how accurate the prediction algorithms trained in MESA are in METS, we computed the Spearman correlation (ρ) between the METS predicted expression values and METS measured expression values.

To evaluate the prediction performance of the training MESA subpopulation in METS, for each of the prediction algorithm methods, we calculated the mean ρ for common predicted genes across the subpopulations (Table 5). Across the training subpopulations, the mean ρ in METS is highest when using AFHI-trained models for all the prediction algorithms. As shown in

Tables 5 and 6, across all the tested prediction algorithms, the training subpopulations comprising individuals of recent African ancestries (AFA, HIS, AFHI, ALL) significantly outperformed the training subpopulation comprising only individuals of European descent (CAU). This shows that prediction performance is highest when the genetic distance between the training population and testing population are closest regardless of the prediction algorithm used. Also, larger sample size improves prediction performance but not as much as when majority of the individuals in the training set share similar ancestries with those in the test set, i.e. AFHI-trained models perform the same as ALL-trained models (Table 5). If larger sample size were the main factor to improve prediction performance, we would expect the average ρ to be significantly highest in ALL. However, we see that average ρ in ALL is less than in AFHI, even though AFHI has lower sample size. More so, the ALL-trained models average ρ were not significantly better than AFA-trained models (Welch test p-values, EN=0.3369, RF=0.8892, SVR=0.1916, KNN=0.3382) (AFA has the lowest sample size and closest ancestry similarity to METS across the training MESA subpopulations). Thus, this highlights the importance of similarity in ancestry at improving prediction performance.

Table 5. Mean prediction performance of MESA-trained models in METS. We focused on the common predicted genes across the training subpopulations for each prediction method. The number of common genes across the training subpopulation for each prediction method are EN=2221, RF=1589, SVR=1435, and KNN=1078. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), K Nearest Neighbor (KNN).

Model	**AFA**	**HIS**	**CAU**	**AFHI**	**ALL**
EN	0.1243	0.0975	0.0767	0.1301	0.1297
RF	0.1152	0.1096	0.0853	0.1197	0.1161
SVR	0.0932	0.0925	0.0758	0.1058	0.1015
KNN	0.0823	0.0788	0.0634	0.0893	0.0888

Table 6. Welch two Sample t-test of the prediction performance of MESA-trained models in METS. The t-test was carried out between training subpopulations comprising individuals of African ancestries (AFA, HIS, AFHI, ALL) and subpopulation comprising only individuals of European ancestries (CAU). Only the P-values from the t-test are recorded in the table.

Model	**AFA vs CAU**	**HIS vs CAU**	**AFHI vs CAU**	**ALL vs CAU**
EN	2.200×10^{-16}	5.043×10^{-5}	2.200×10^{-16}	2.200×10^{-16}
RF	2.860×10^{-6}	1.267×10^{-4}	8.648×10^{-8}	1.436×10^{-6}
SVR	5.005×10^{-3}	6.334×10^{-3}	1.998×10^{-6}	4.760×10^{-5}
KNN	4.286×10^{-3}	1.671×10^{-2}	8.751×10^{-5}	1.430×10^{-4}

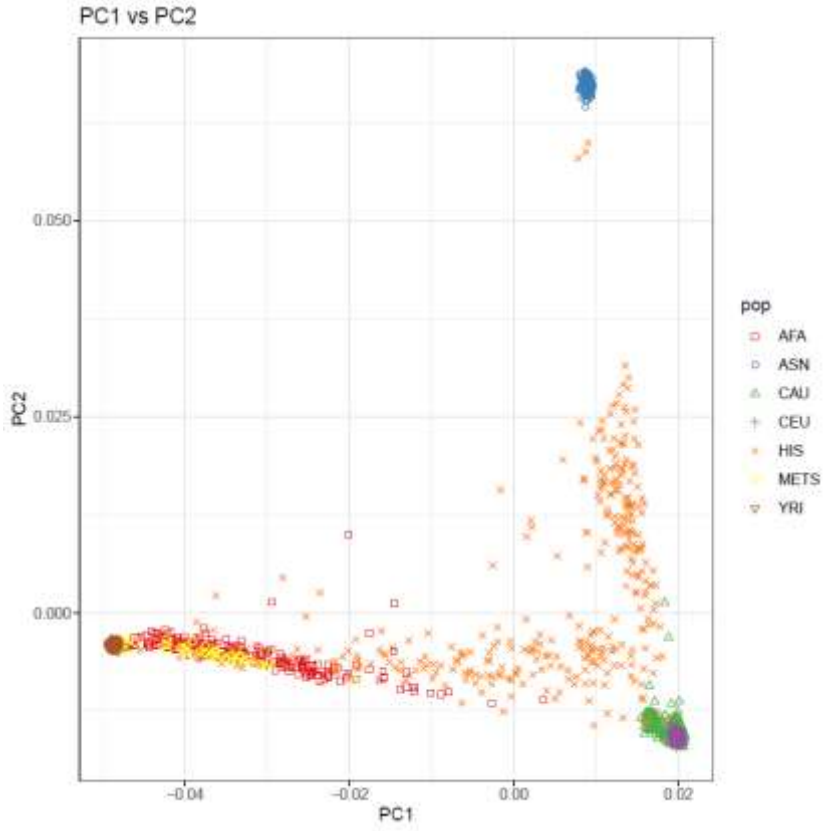

Figure 6. Principal Component Analysis of METS. The genotypic principal component plot of the METS (Modeling the Epidemiological Transition Study) and MESA (Multi-ethnic Study of Atherosclerosis) populations analyzed with HapMap populations. The abbreviations are MESA African Americans (AFA), East Asians from Beijing, China and Tokyo, Japan (ASN), MESA European Americans (CAU), European ancestry from Utah (CEU), MESA Hispanic Americans (HIS), Yoruba from Ibadan, Nigeria (YRI).

When we examine all prediction results in METS, the numbers of genes we were able to predict gene expression values for varied across algorithms and populations (Figure 7). The gene models trained with the ALL cohort predicted gene expression values for more genes than the rest of the other training populations across all prediction algorithms. This is probably because, ALL cohort has the largest sample size. In fact, the number of genes captured decreases from ALL to AFA as the sample size decreases. Interestingly though, when we filter by $\rho > 0.1$, AFA captures more genes (EN=1545, RF=1167, SVR=961, KNN=824) than HIS and CAU, again showing the importance of similarity in ancestry between training and testing population for gene expression prediction regardless of prediction algorithm. The models trained with AFHI and ALL cohorts capture more genes than AFA most probably because of their larger sample size and the fact that they also contain the AFA cohort. Therefore, although larger sample size is important in prediction performance, it is paramount that individuals in the training population have similar ancestry with the testing population.

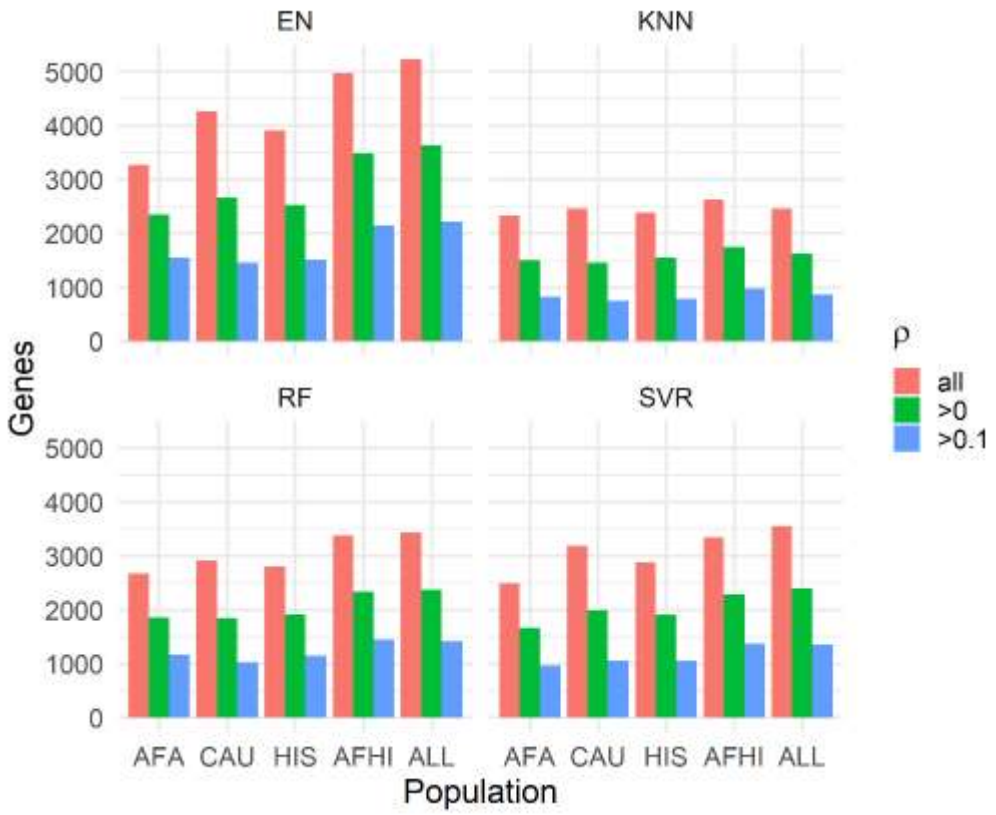

Figure 7. Number of Predicted Genes in METS after Filtering by ρ. The MESA population used to train each set of models is shown on the x-axis and the number of genes with predicted expression values in METS is shown on the y-axis. ρ is the Spearman correlation between predicted and observed gene expression in METS.

Elastic Net Trained Models Outperform Machine Learning Models in Test Cohort

The elastic net models predict gene expression values in METS for more genes than Random Forest (RF), Support Vector Regression (SVR), and K Nearest Neighbor (KNN) (Figure 7). When all genes predicted in METS are compared, prediction performance is highest for RF-trained models in the HIS and CAU populations, while performance is highest for EN-trained models in the AFA, AFHI, and ALL populations (Figure 8). However, when we compare test prediction performance of the machine learning algorithms against EN on the genes they both can predict (intersection), EN performs best regardless of training population with RF being the

closest to EN (Figures 9 and 10). In our comparison of ALL-trained models, the number of overlapping genes between EN and the other algorithms are RF=1126, SVR=1063, and KNN=654. Although EN generally outperforms the other algorithms, we observe that all the genes in each of the algorithms did not overlap with those in EN even though they captured fewer genes than EN (Table 7). That is, these algorithms have significant performance ($\rho > 0.1$) on some genes that EN does not, and vice versa. To probe further into the algorithm pairs, we counted the genes unique to each algorithm (Table 7). Expectedly, EN captures over 1000 unique genes, however, the few unique genes (<300) captured by each of RF, SVR, and KNN suggests that prediction performance in test cohorts can be improved by combining gene models from EN and these other algorithms.

Table 7. Number of ALL-trained Predicted Genes in METS in Algorithm Pairs. Abbreviations are Elastic Net (EN), Random Forest (RF), Support Vector Regression (SVR), K Nearest Neighbor (KNN).

Genes	**EN vs RF**		**EN vs SVR**		**EN vs KNN**	
Overlap	1126		1063		654	
Unique	1092	292	1155	296	1564	211

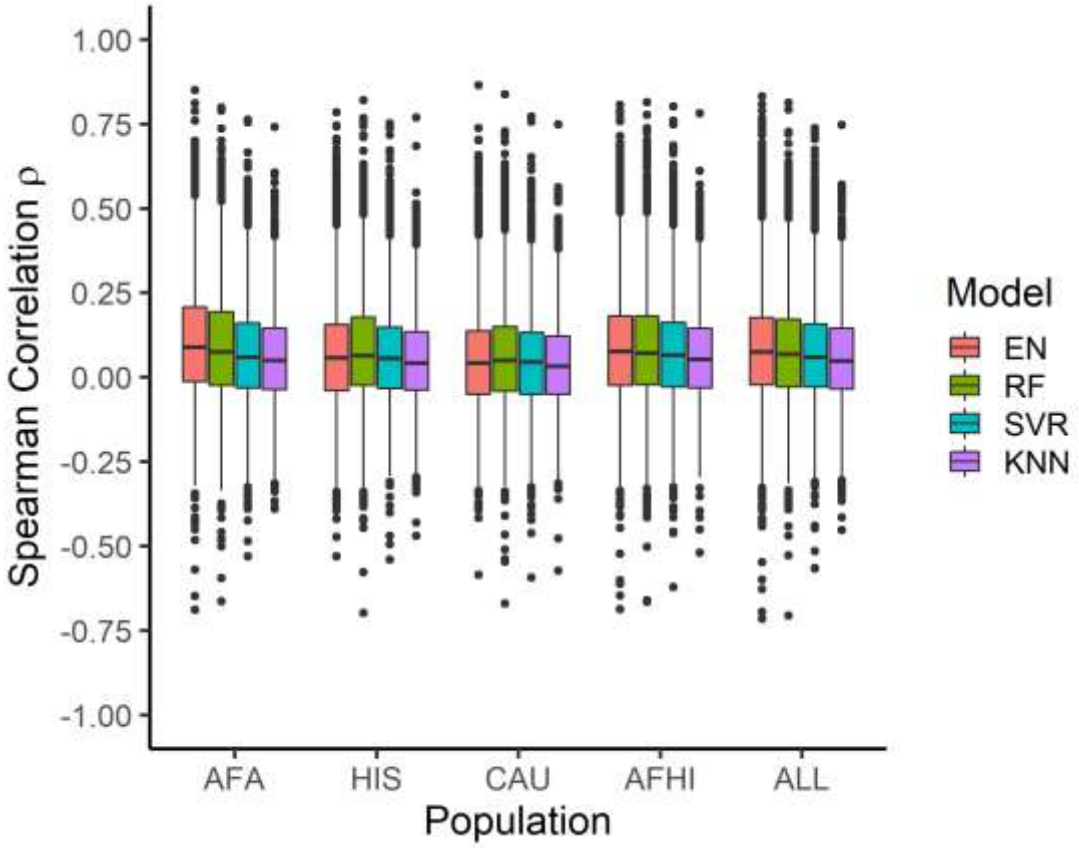

Figure 8. Prediction performance of models trained in MESA populations and tested in METS. We predicted expression in METS using only gene models with $R^2 > 0.01$. The MESA population used to train each set of models is shown on the x-axis and the Spearman correlation between predicted and observed gene expression in METS is shown on the y-axis. All METS predicted genes are shown in the plot.

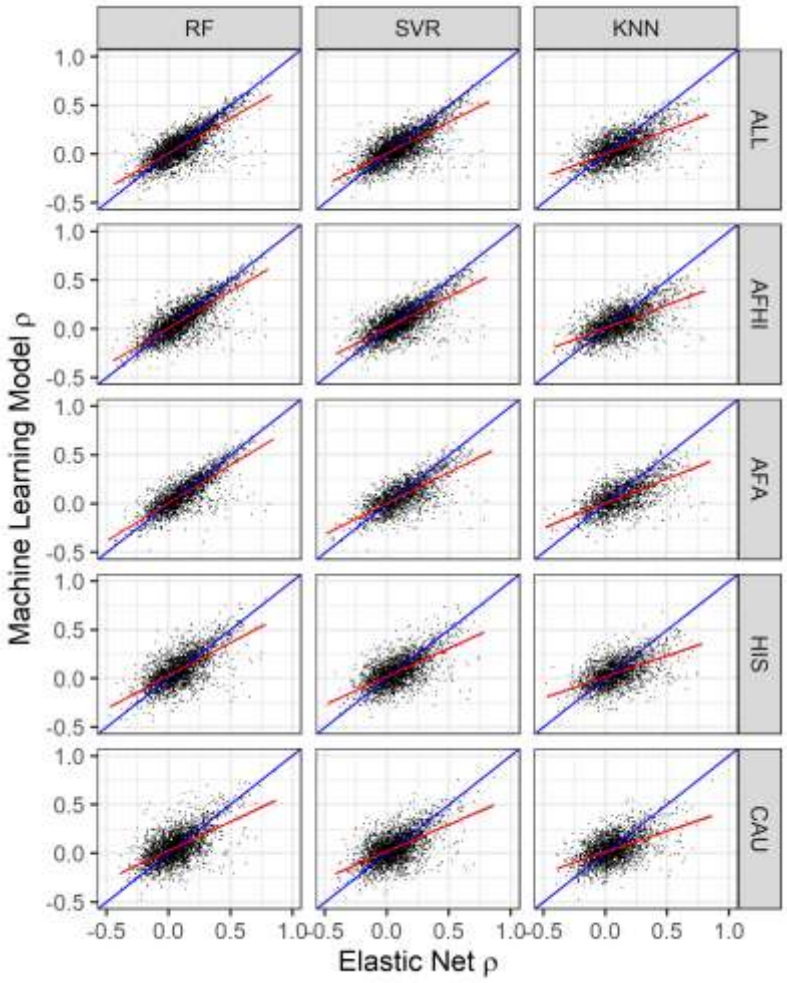

Figure 9. Comparison of Algorithm Test Prediction Performance in METS from Models Trained in MESA. Prediction performance ρ (Spearman correlation between predicted and observed gene expression in METS) for each gene in each machine learning (ML) model vs. elastic net (EN) is shown. Only genes with $\rho > 0.1$ are plotted. The linear regression fit is shown by the red line and identity line (slope=1) is blue in each plot. In the ALL cohort, the numbers of genes that overlap are EN vs RF = 1126, EN vs SVR = 1063, EN vs KNN = 654, and their Pearson correlations (R) are 0.8121, 0.7699, and 0.6199, respectively. In the AFHI cohort, the numbers of genes that overlap are EN vs RF = 1182, EN vs SVR = 1055, EN vs KNN = 717, and their Pearson correlations (R) are 0.8212, 0.7547, and 0.6150, respectively. In the AFA cohort, the numbers of genes that overlap are EN vs RF = 922, EN vs SVR = 683, EN vs KNN = 554, and their Pearson correlations (R) are 0.8260, 0.7339, and 0.5753, respectively. In the HIS cohort, the numbers of genes that overlap are EN vs RF = 762, EN vs SVR = 663, EN vs KNN = 496, and their Pearson correlations (R) are 0.6289, 0.6179, and 0.5371, respectively. In the CAU cohort, the numbers of genes that overlap are EN vs RF = 614, EN vs SVR = 623, EN vs KNN = 434, and their Pearson correlations (R) are 0.6336, 0.6096, and 0.4701, respectively.

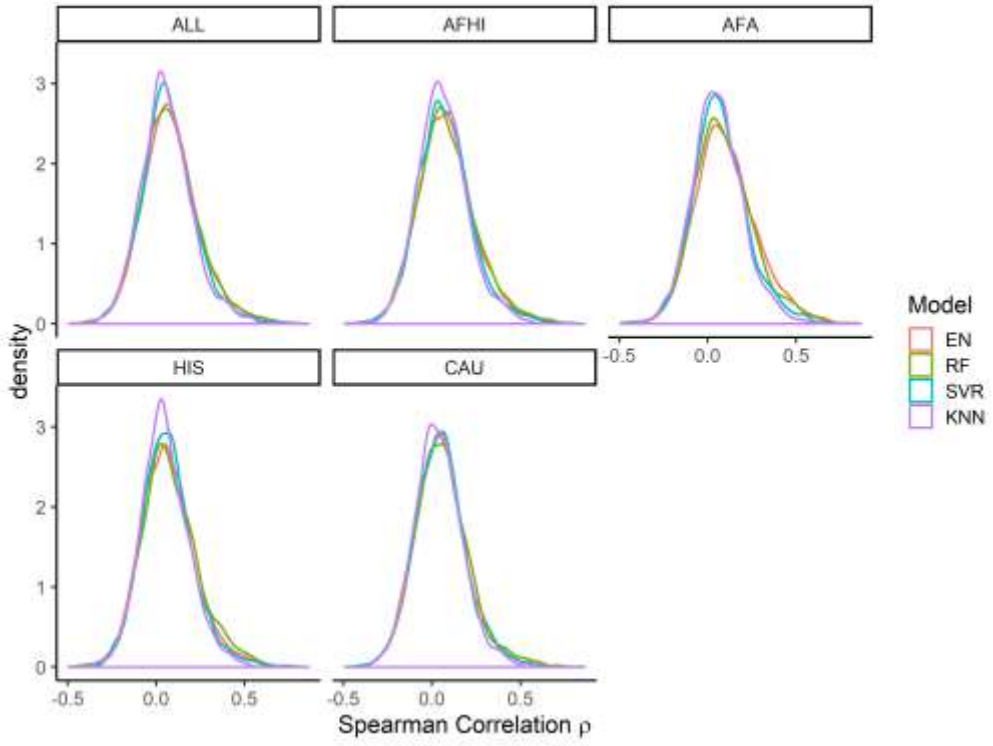

Figure 10. Distribution of Prediction Performance in METS from Models Trained in MESA cohort. Distributions of prediction performance (Spearman's ρ) for genes with ρ > -0.5 in each algorithm. Note, EN and RF models have similar distributions and are shifted to the right compared to SVR and KNN.

Elastic Net and Machine Learning Models Identify the Same Gene in Lipid TWAS

To evaluate the biological importance of the prediction algorithms in identifying significant genes associated with traits, we carried out transcriptome wide association studies (TWAS) on high-density lipoprotein (HDL) levels. In our analysis, we used a genotype dataset from the MESA cohort (n=3856), comprising individuals from the subpopulations that were not used in building any of the imputation models and in which we have corresponding lipid phenotype data (AFA=1188, HIS=952, and CAU=1716). The genotype data were cleaned using standard quality control procedures (See Methods). We used the ALL-trained imputation gene models (genes with CV $R^2 > 0.01$) from each algorithm to impute transcriptome levels from the

MESA genotypes. We adjusted the predicted transcriptome levels for population structure using the first 10 genotype principal components and rank normalized the HDL levels. Using the adjusted predicted transcriptome levels and normalized HDL data, we conducted association tests using linear regression. Interestingly, all tested prediction algorithms except KNN identified a significant association ($P<5 \times 10^{-8}$) for the gene *CETP* (Figures 11 and 12). The lack of association with HDL for all gene expression values predicted from KNN trained models is consistent with our earlier results in this paper that KNN is worse at imputing transcriptome levels compared to the other algorithms. The directions of effect of *CETP* transcriptome levels as predicted by EN, RF, and SVR are the same (Figure 13). An increase in predicted *CETP* expression is associated with decreased HDL levels across EN, RF, and SVR. The ability of the three algorithms to identify the same significant hit underscores their effectiveness at imputing gene expression (*CETP* R^2, EN=0.0917, RF=0.0772, SVR=0.0539). More so, the p-value of HDL association with *CETP* predicted transcriptome levels was most significant in RF ($p=7.933 \times 10^{-14}$), and highest in SVR ($p=2.278 \times 10^{-8}$), thus showing that RF outperformed EN ($p=6.869 \times 10^{-11}$) in this instance.

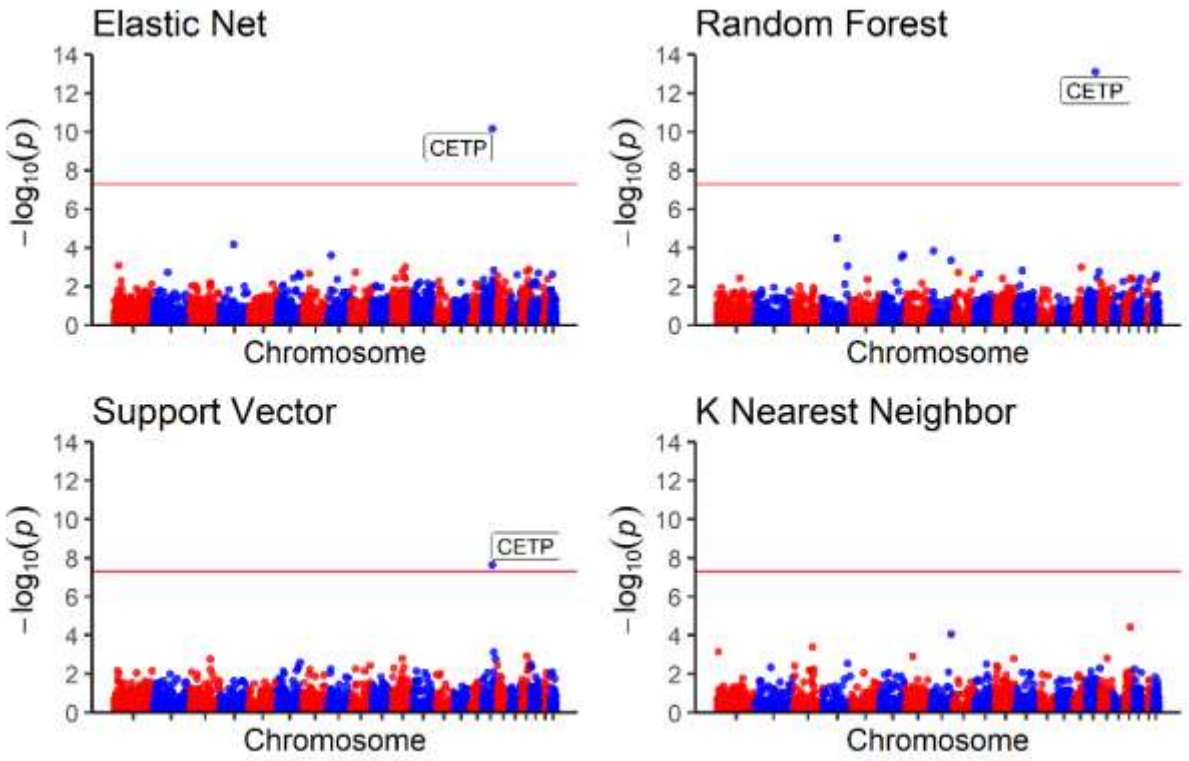

Figure 11. HDL Transcriptome-Wide Association Studies Results. Manhattan plot of the gene P-values from the TWAS between HDL (high density lipoprotein) values and predicted gene expression. Using models trained in MESA ALL cohort, we predicted gene expression in MESA (n=3856) genotype data comprising individuals not used in the model training with HDL phenotype data and then carried out TWAS. Genome-wide significance ($P < 5 \times 10\text{-}8$) is shown by the red line in the plots. The X axis are ordered from chromosomes 1 to 22 (left to right).

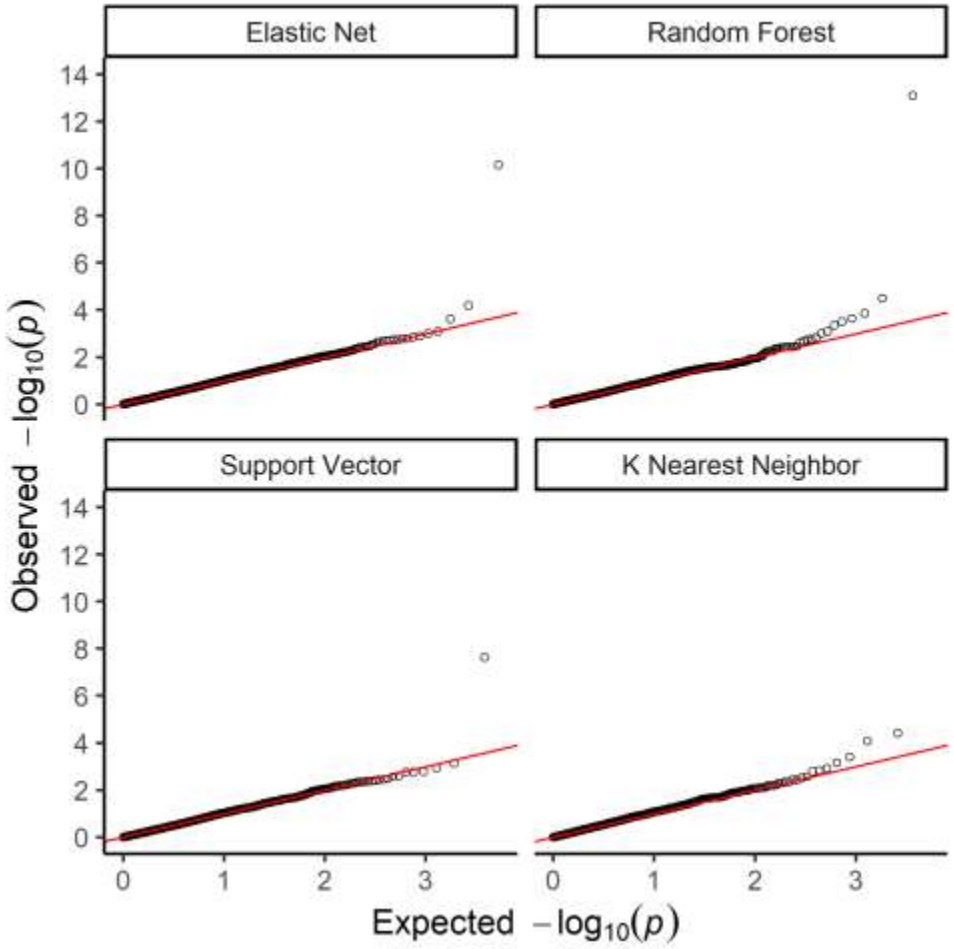

Figure 12. Q-Q Plot of Association Tests P-Values. Q-Q plot of the P-values from the TWAS between HDL (high density lipoprotein) values and predicted gene expression. Using models trained in MESA ALL cohort, we predicted gene expression in MESA (n=3856) genotype data comprising of individuals not used in the model training and that equally has HDL phenotype data and then carried out TWAS. The red line in each plot show the null expected distribution of the P-values.

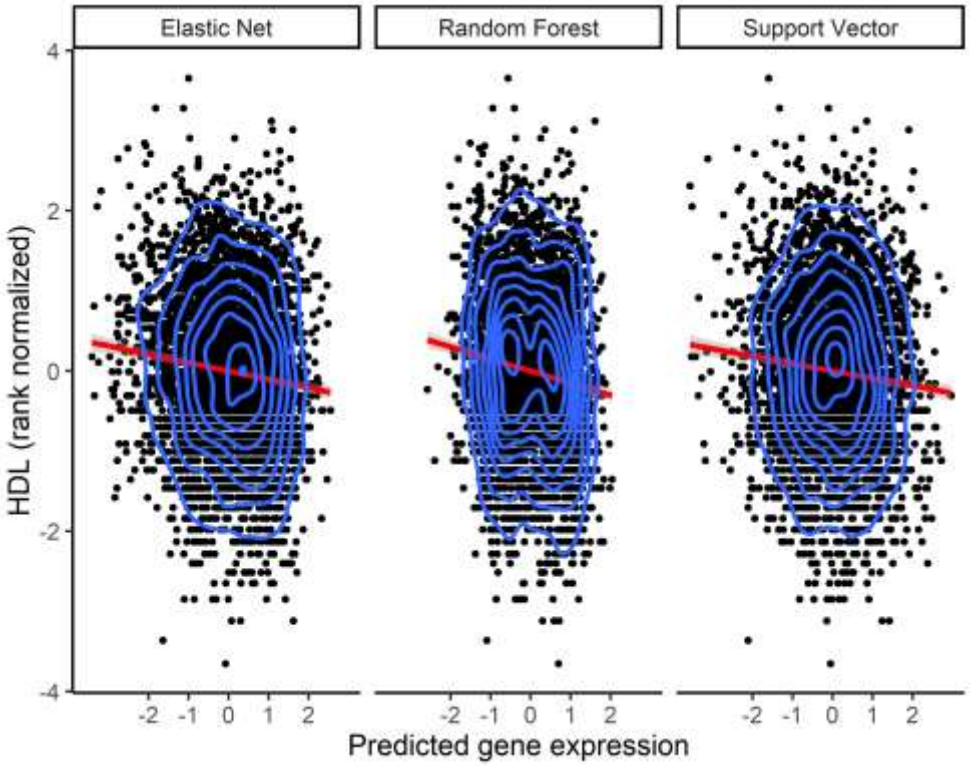

Figure 13. Increased HDL Levels correlate with decreased CETP Predicted Expression. Direction of effect of the *CETP* gene on HDL levels. Using models trained in the MESA ALL cohort, we predicted gene expression in MESA (n=3856) genotype data comprising individuals not used in the model training with HDL phenotype data. Each point in the plot represents an individual. The linear regression fit is shown by the red line in each plot. The blue contour lines from two-dimensional kernel density estimation help visualize where the points are concentrated.

Consequently, we carried out comparison of EN and RF on their t-statistic values from the association tests between HDL and predicted gene expression. We found that both EN and RF t-statistic values were almost parallel for the genes they have in common thus corroborating the observed similar performance on their common genes from our previous results (Figures 4 and 8). In the EN TWAS, 5279 genes were tested for association with HDL. In the RF TWAS, 16 unique genes that were not present in EN TWAS were tested for association with HDL (Figure 14). Among the RF unique genes, we found a gene, *ST8SIA4*, that may potentially be associated with normalized HDL ($p=4.288 \times 10^{-3}$) but was missed by EN (*ST8SIA4* R^2, EN=-

0.0005, RF=0.0100) (Figure 14). This discovery is consistent with our previous results wherein we found that although EN has many genes in common with RF in their imputation models, the RF algorithm generated some unique gene models (Table 7). Thus, by combining EN and RF models in gene expression imputation and subsequent TWAS analysis, we may uncover more and new significant gene-trait associations. Note however, that by combining EN and RF models, we are not significantly changing the number of tests performed. Depending on predictive performance inclusion threshold, adding RF expression prediction models may increase the number of tests by up to 13% (Table 7), which does not dramatically change the Bonferroni correction threshold.

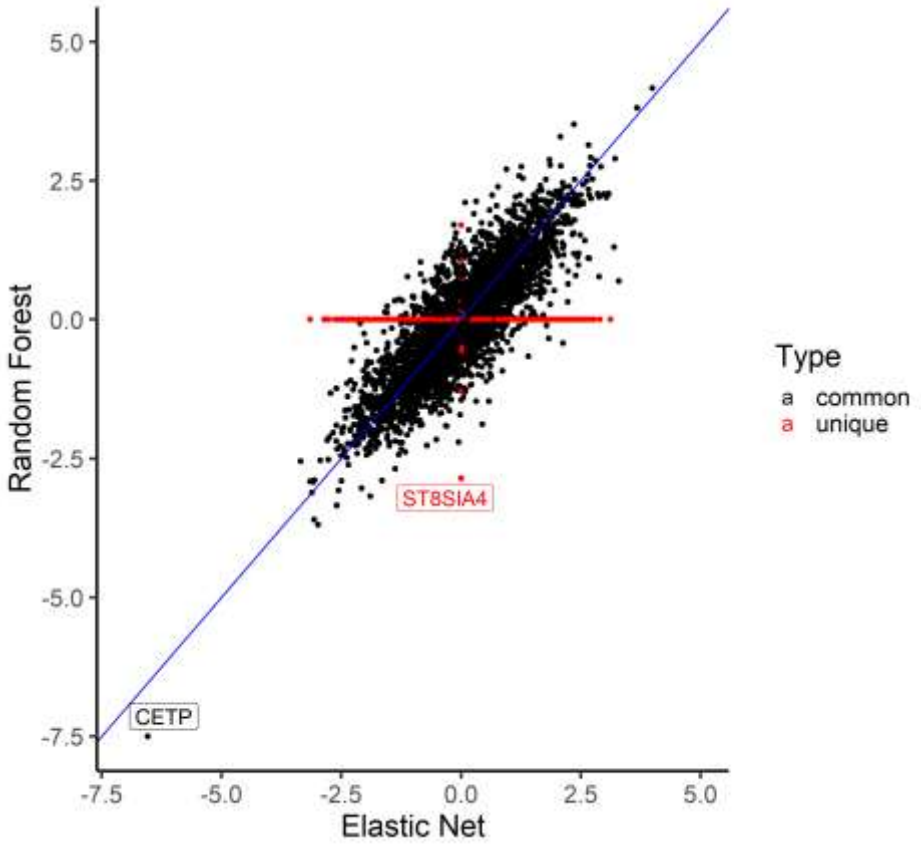

Figure 14. Comparison of the HDL Association t-statistics from RF and EN models trained in the MESA ALL cohort. Comparison of RF and EN t-statistics from the TWAS of HDL and predicted transcriptome in MESA individuals not used for imputation model building. Each dot in plot represents the t-statistic values of a gene from the HDL TWAS while the identity line (slope=1) is shown in blue. We see that the t-statistic values are similar between RF and EN except for genes that are unique in each algorithm shown as red dots in the plot. *CETP* is strongly associated with HDL levels using both EN- and RF-trained models. RF-trained models revealed the unique gene *ST8SIA4* (no prediction model in EN) maybe potentially associated with HDL levels ($p=4.288 \times 10^{-03}$).

Microbiome Diversity Differs Between Ghanaians and African Americans

We compared gut microbiome profiles (See Methods) of 36 Ghanaians in Ghana and 25 African Americans in USA from the METS cohort (Figure 15). We found that Ghanaians have higher alpha diversity than Americans (Figure 16). Since Ghanaians are more in our sample, we randomly removed 11 Ghana individuals from the analysis to match the Americans sample size.

Even though the samples sizes were equal, alpha diversity was still highest in the individuals from Ghana (Figure 17), thus suggesting clear microbiome differences between the two groups. However, when we analyzed the microbiome profile based on obesity status of the study individuals (obesity was classified as having body mass index BMI ≥ 30), we found no significant difference between the alpha diversity of obesity status (Figure 18).

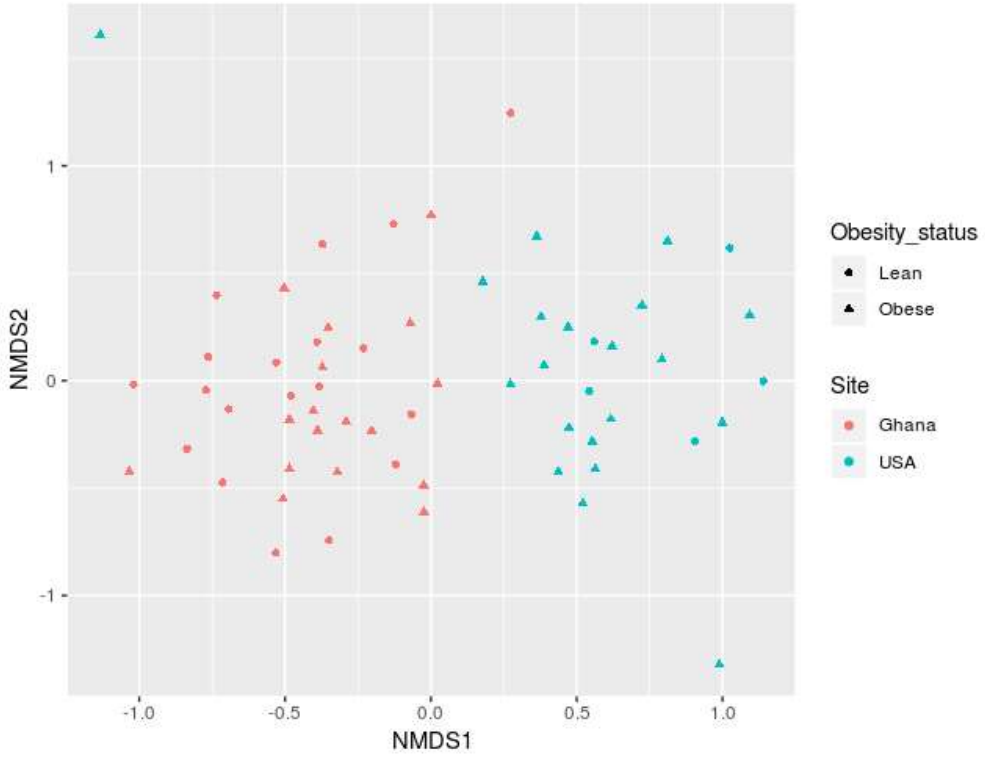

Figure 15. Bray Curtis Dissimilarity plot of the METS cohort. The microbiome species composition of the cohorts is clearly separated by site. Also, there is no marked difference between the microbial compositions of obese and lean individuals. Obesity was classified as BMI ≥ 30.0

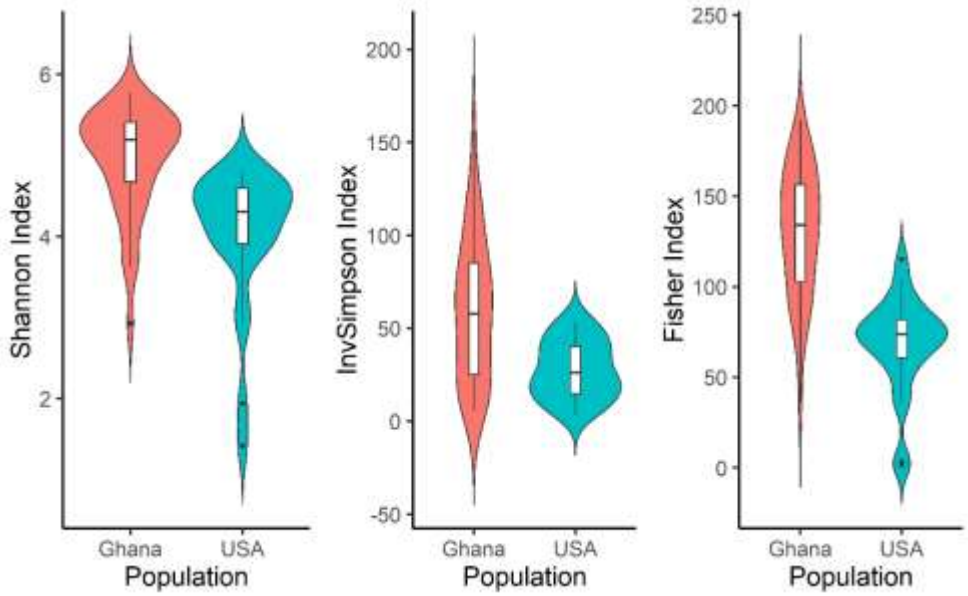

Figure 16. Alpha diversity distribution of all 61 METS sample by population. As seen in the plot, Africans in Ghana have higher alpha diversity than African Americans in the USA across all tested diversity indices.

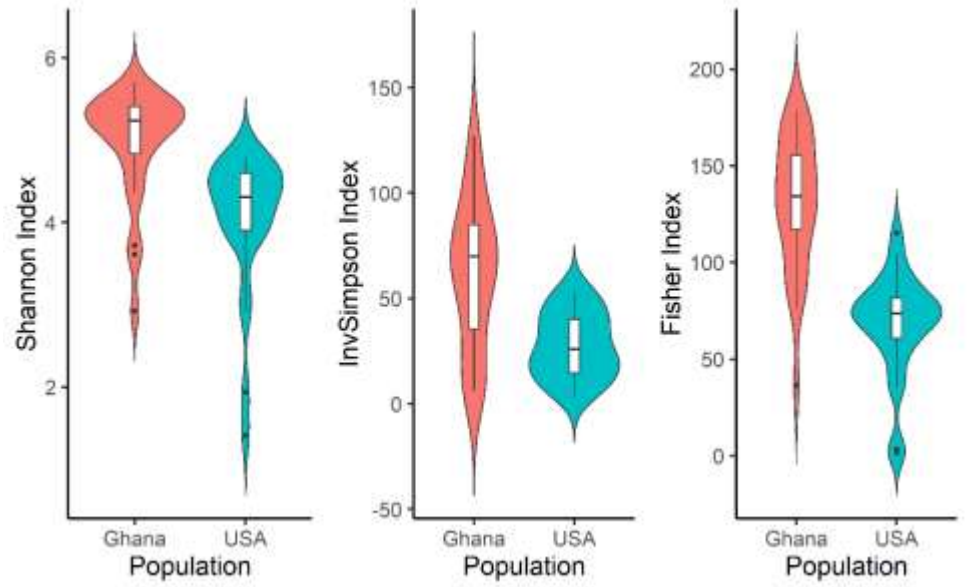

Figure 17. Alpha diversity distribution of METS with equal number of Ghanaians and Americans (25 each). As shown in the plot, microbial alpha diversity is highest in Ghanaians than in Americans (Welch t-test p-values; Shannon Index=9.777×10^{-05}, InvSimpson Index=8.504×10^{-05}, Fisher Index=1.033×10^{-08}).

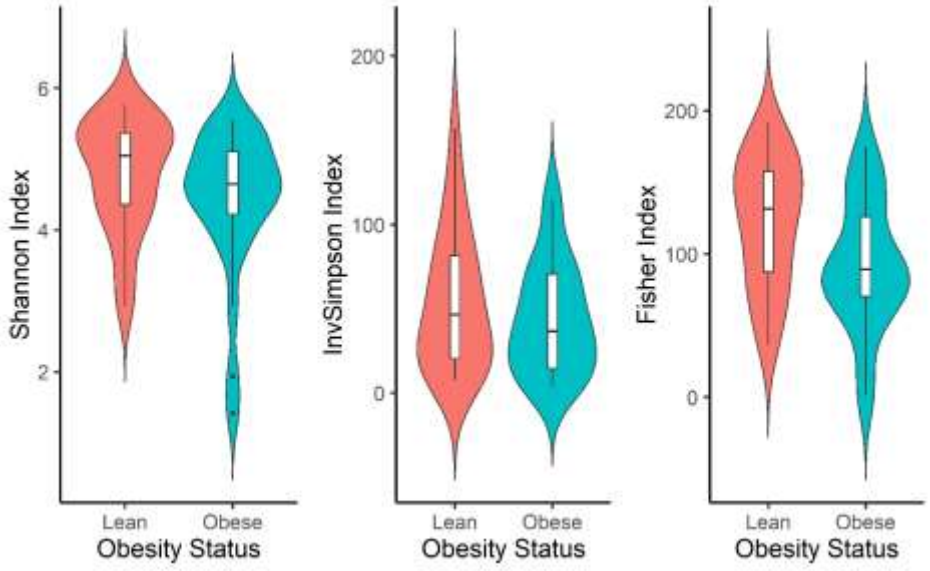

Figure 18. Alpha diversity distribution of METS with equal number of obese and lean individuals (24 each). As shown in the plot, there is no significant difference in the microbial alpha diversity of obese and lean individuals (Welch t-test p-values; Shannon Index=0.1634, InvSimpson Index=0.3309, Fisher Index=0.0230). Obesity was classified as BMI $\geq$ 30.0

No Associations Found in Limited Sample Transcriptome-wide Association Studies of Microbiome Diversity and Obesity

We sought to explore potential genetic relationship between gene expression and microbiome diversity index as a quantitative heritable trait as well as obesity. To achieve this goal, we predicted gene expression on the 61 individuals in the METS cohort whom we have genotype, transcriptome, and microbiome data using the prediction models trained with the ALL cohort. The predicted expression profiles were adjusted for population structures and confounders (See Methods). We subsequently carried out association tests between the adjusted predicted expression and obesity status using logistic regression (Figures 19 and 20). We found no genome wide significant association between predicted expression and obesity status across all the predictive algorithms. We also conducted association test between the adjusted predicted expression and microbiome alpha diversity index (Shannon Index) using linear regression

(Figures 21 and 22). Again, we found no significant associations, thus suggesting a very low effect sizes of gene expression regulations due to microbial composition, like other complex traits where thousands of individuals are required for genome-wide significance.

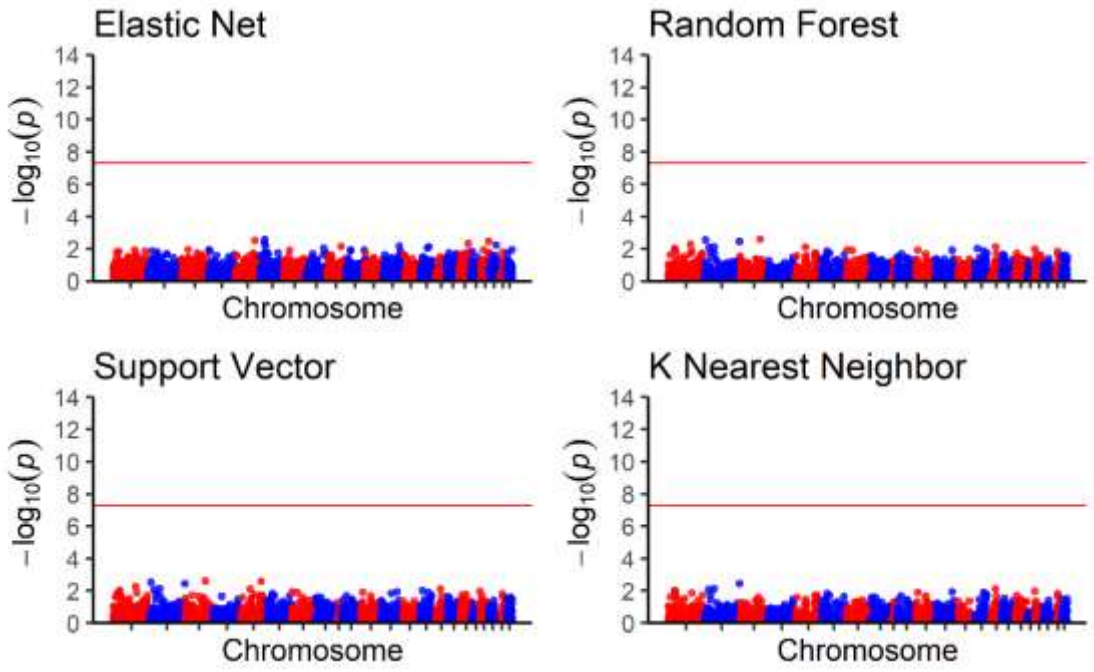

Figure 19. Transcriptome-Wide Association with Obesity. We carried out logistic regression between predicted transcriptome and obesity status from 61 individuals in the METS cohort. The predicted transcriptome profiles were generated using the ALL-trained imputation models. Genome-wide significance ($P < 5 \times 10\text{-}8$) is shown by the red line in the plots. The X axis are ordered from chromosomes 1 to 22 (left to right). Obesity was classified as BMI ≥ 30.0.

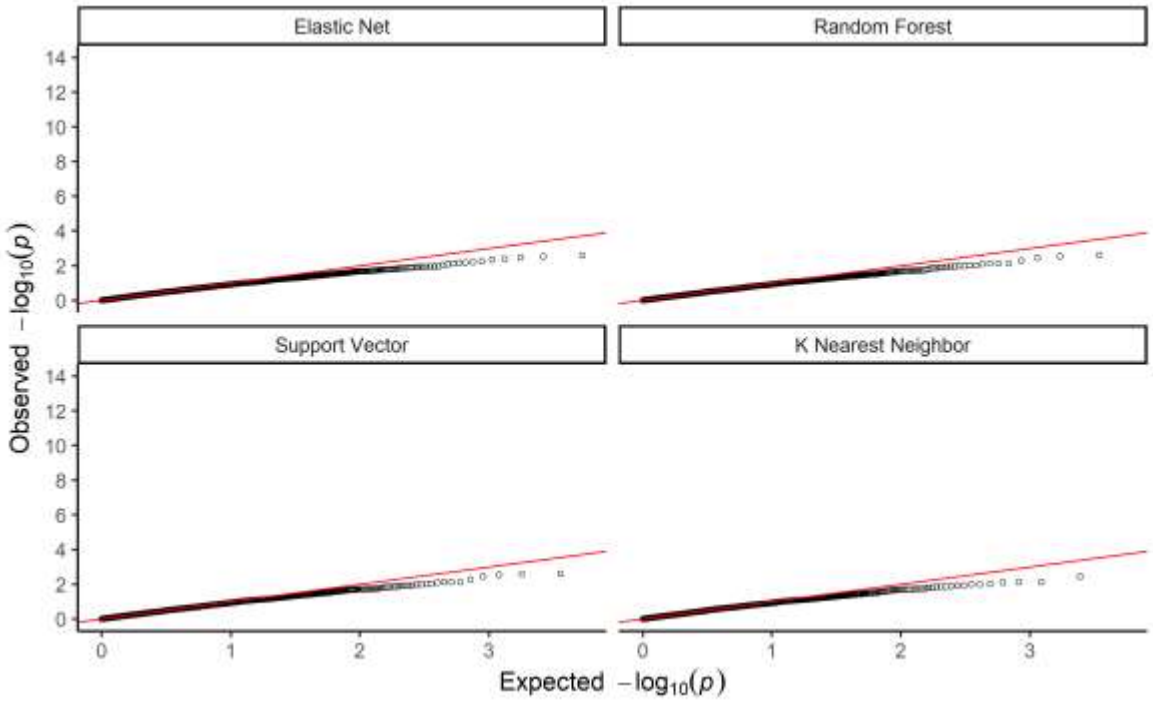

Figure 20. Q-Q plot of METS obesity association tests p-values. Q-Q plot of the P-values from the TWAS between obesity status and predicted gene expression in METS cohort (n=61). The red line in each plot show the null expected distribution of the P-values. Obesity was classified as BMI ≥ 30.0.

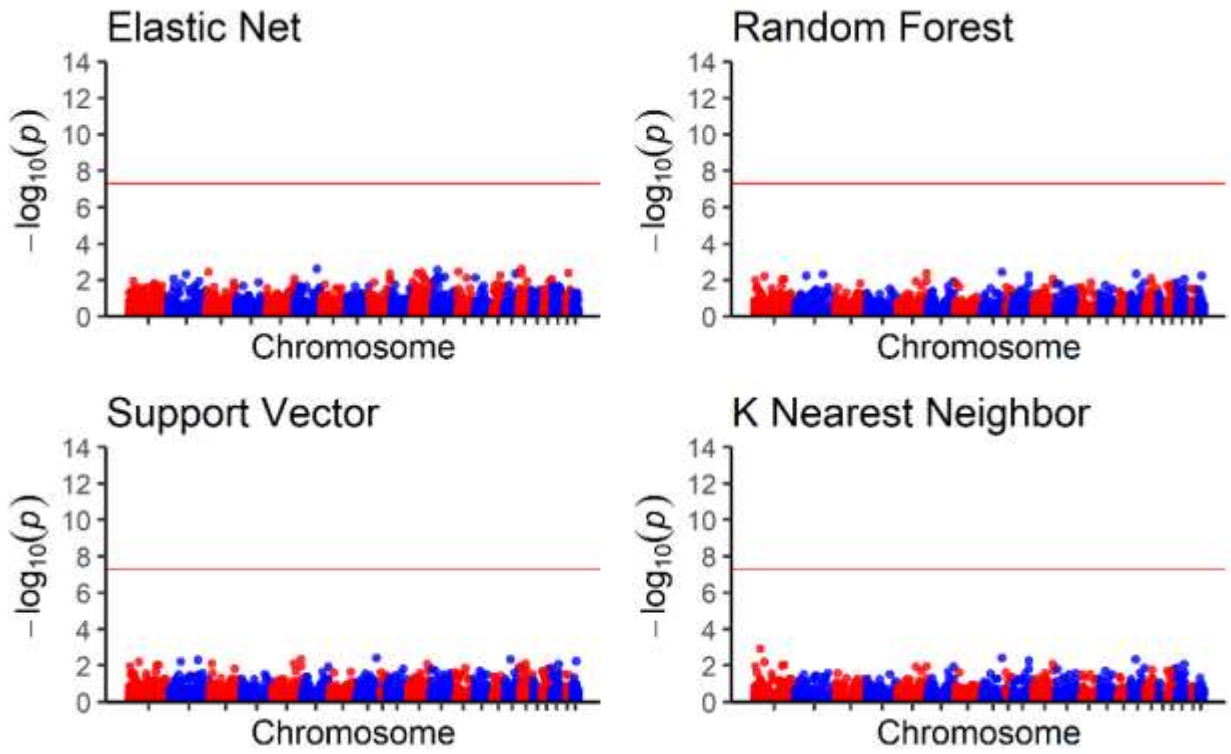

Figure 21. Transcriptome-Wide Association with microbiome alpha diversity index – Shannon index. We carried out linear regression between predicted transcriptome and Shannon index values from 61 individuals in the METS cohort. The predicted transcriptome profiles were generated using the ALL-trained imputation models. Genome-wide significance ($P < 5 \times 10\text{-}8$) is shown by the red line in the plots. The X axis are ordered from chromosomes 1 to 22 (left to right).

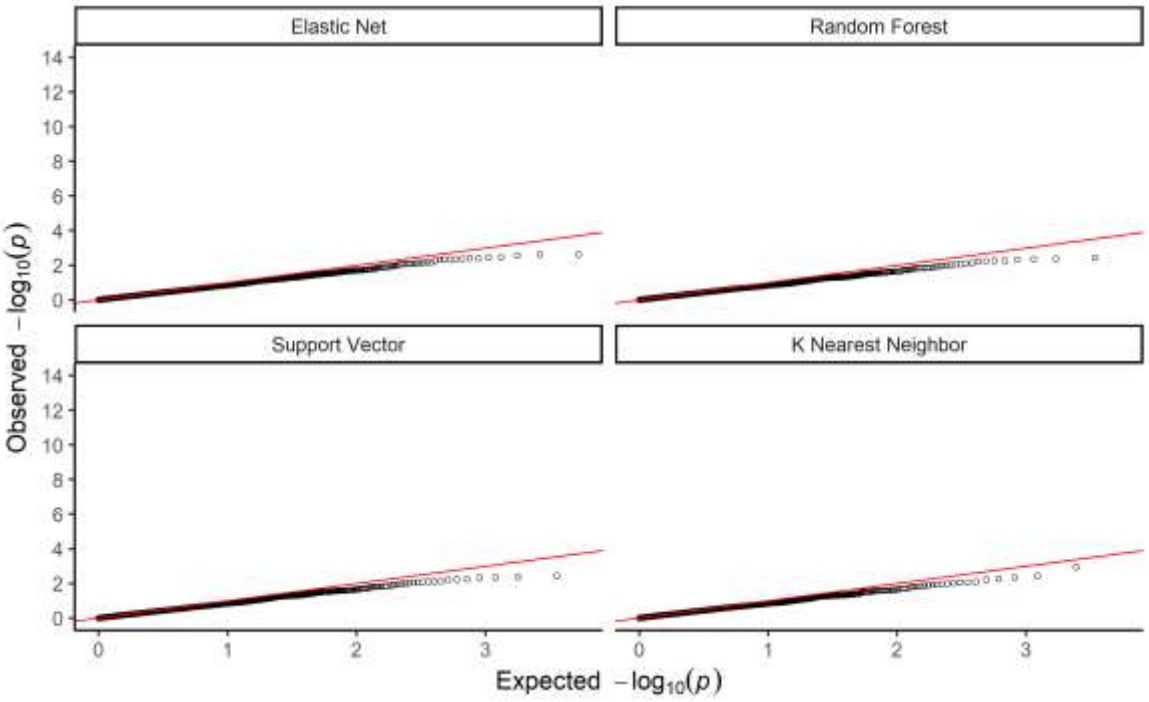

Figure 22. Q-Q plot of METS microbiome alpha diversity (Shannon Index) association tests p-values. Q-Q plot of the P-values from the TWAS between Shannon index values and predicted gene expression in METS cohort (n=61). The red line in each plot show the null expected distribution of the P-values.

CHAPTER FOUR

DISCUSSION AND CONCLUSION

We explored the potential of using non-linear machine learning modeling, including random forest (RF), support vector regression (SVR), and *K* nearest neighbor regression (KNN), to further improve gene expression prediction performance across global populations in comparison to parametric linear elastic net (EN) modeling, which is currently used in PrediXcan (Gamazon et al., 2015). To accomplish this, we trained each of the prediction models with genotype and transcriptome data from the MESA cohort on 9623 protein coding genes and compared their cross-validated imputation performance (R^2). Although almost paralleled by RF, we found EN generally outperformed the tested non-parametric machine learning models. This is consistent with a recent study where it was shown that the genome wide polygenic risk score method based on simple linear additive effects of genetic factors outperformed non-linear machine learning models in genetic prediction of cardiovascular disease risk (Gola, Erdmann, Müller-Myhsok, Schunkert, & König, 2020). However, in our study, we found that when the prediction models are trained within each of the MESA subpopulations, RF sometimes outperformed EN, specifically on HIS and CAU data (Figure 4, 8). This suggests potential robustness and reduced variability of RF imputation performance across global populations. In addition, we also found genes *ERAP2, HLA-C, HLA-DRB1, CHURC1, RAD51*, and *SNAP29* have $R^2 > 0.5$ for all the training subpopulations across all prediction models, indicating the high heritability (Gamazon et al., 2015; Mogil et al., 2018; Wheeler et al., 2016) and suggesting commonality of their cis-eQTL predictors across global populations.

We further tested the MESA trained models on genotype and transcriptome data from African-origin individuals in the METS cohort. We show that models trained with the cohorts (AFA, HIS, AFHI, ALL) comprising individuals similar in ancestries with METS have better prediction performance than the models trained with individuals (CAU) of no recent African ancestries (Table 5, Figure 8). Thus, as seen in recent studies (Fryett et al., 2020; Keys et al., 2019; Mikhaylova & Thornton, 2019; Mogil et al., 2018), we show similarity in ancestries between training and testing populations improves prediction performance. Notably, we found that the improvement in prediction due to ancestries similarity is consistent within all tested prediction algorithms, further underscoring the huge importance of diverse ancestries in genetic studies.

We applied the trained models on out-of-sample MESA genotype data with corresponding high-density lipoprotein (HDL) phenotype values. All tested prediction models except for KNN identified the gene CETP to be significantly associated with HDL (Figure 11). As seen in a recent study on lipids traits (Andaleon et al., 2019), we show that increased CETP expression is significantly associated with lower HDL levels and the direction of effect are the same for EN, RF, and SVR models (Figure 13). Thus, we computationally corroborate the biological importance of CETP gene in HDL associated diseases. In many studies, the CETP gene has been experimentally associated with HDL levels in humans, and currently stands as a potential drug target for the treatment of atherosclerosis (Barter et al., 2003; de Grooth et al., 2004; Kosmas, Dejesus, Rosario, & Vittorio, 2016; Tall & Rader, 2018; Thompson et al., 2003). Thus, our analysis in a relatively small TWAS (n=3856) identified a known drug target that has been studied extensively in the context of preventing cardiovascular disease. However, because

of the inability of nonlinear models to use GWAS summary statistics as training data, applicability of the nonlinear machine learning models in TWAS is limited to only GWAS with genotypes and phenotypes available.

We analyzed microbiome data of our test cohort (METS) to understand the geographical and phenotypical microbial composition differences. We found that Africans in Ghana have higher gut microbiome diversity than African Americans in the USA (Figures 16, 17, and 18) as shown in a previous study (Dugas et al., 2018). The differences in the two groups microbiota composition can be attributed to their different social behaviors and nutrition due to culture and socioeconomic realities. Indeed, many studies have shown that social interactions and diet can alter microbiota composition (Archie & Tung, 2015; Singh et al., 2017). Focusing on obesity status, we found no significant differences in the gut microbiome diversity of obese and lean individuals in our study which contrasts with many other studies (Kalliomäki, Carmen Collado, Salminen, & Isolauri, 2008; Le Chatelier et al., 2013; Ley et al., 2005). Nonetheless, our finding of no microbiota diversity differences in the obese and lean individuals are not unfounded as some studies have also shown inconsistencies in the microbial composition of these two groups (Duncan et al., 2008; Zhang et al., 2009). These inconsistencies can be attributed to confounding factors such as fasting, diet, and use antibiotics. We also explored the possibility of integrating microbiome into our transcriptome prediction model. We found no significant association between the predicted transcriptome and obesity status, as well as microbiome alpha diversity (Shannon Index) across all the prediction algorithms. Thus, suggesting that any potential gene expression regulations due to microbiota composition maybe too small to detect with the few sample size (61) used in our study.

Overall, although linear modeling of cis-eQTLs and gene expression is generally good at imputing expression for new data, linear models fail to accurately predict expression for some genes. Interestingly, our study shows the imputation performances for some genes are comparatively better with non-linear machine learning modeling like random forest (Figure 9) than linear modeling like elastic net. Therefore, by increasing ancestries diversity and sample sizes of study populations, optimizing prediction performance on these genes with machine learning modeling, and incorporating the models into the existing PrediXcan tool, we may further increase the probability of uncovering new gene-trait associations in downstream transcriptome-phenotype analyses.

BIBLIOGRAPHY

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., … Site—NDRI, B. C. S. (2017). Genetic effects on gene expression across human tissues. Nature, 550(7675), 204–213. https://doi.org/10.1038/nature24277

Akil, L., & Ahmad, H. A. (2011). Relationships between obesity and cardiovascular diseases in four southern states and Colorado. Journal of Health Care for the Poor and Underserved, 22(4 Suppl), 61.

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature Reviews Genetics, 16(4), 197–212.

Andaleon, A., Mogil, L. S., & Wheeler, H. E. (2019). Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. PloS One, 14(8).

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012, January). {FastQC}. Babraham, UK.

Archie, E. A., & Tung, J. (2015). Social behavior and the microbiome. Current Opinion in Behavioral Sciences, 6, 28–34.

Barter, P. J., Brewer Jr, H. B., Chapman, M. J., Hennekens, C. H., Rader, D. J., & Tall, A. R. (2003). Cholesteryl ester transfer protein: a novel target for raising HDL and inhibiting atherosclerosis. Arteriosclerosis, Thrombosis, and Vascular Biology, 23(2), 160–167.

Bergman, E. N. (1990). Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. Physiological Reviews, 70(2), 567–590.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V, Folsom, A. R., … Tracy, R. P. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. American Journal of Epidemiology.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581.

Carbone, S., Canada, J. M., Billingsley, H. E., Siddiqui, M. S., Elagizi, A., & Lavie, C. J. (2019). Obesity paradox in cardiovascular disease: where do we stand? Vascular Health and Risk Management, 15, 89.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4(1), s13742--015.

Chassaing, B., Aitken, J. D., Gewirtz, A. T., & Vijay-Kumar, M. (2012). Gut microbiota drives metabolic disease in immunologically altered mice. In Advances in immunology (Vol. 116, pp. 93–112). Elsevier.

Chial, H. (2008). DNA sequencing technologies key to the Human Genome Project. Nature Education, 1(1), 219.

Christensen, K., & Murray, J. C. (2007). What genome-wide association studies can do for medicine. N Engl J Med, 356(11), 1094–1097.

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. Cell, 148(6), 1258–1270.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … others. (2016). Next-generation genotype imputation service and methods. Nature Genetics, 48(10), 1284–1287.

Davis, C. D. (2016). The gut microbiome and its role in obesity. Nutrition Today, 51(4), 167.

de Grooth, G. J., Klerkx, A. H. E. M., Stroes, E. S. G., Stalenhoef, A. F. H., Kastelein, J. J. P., & Kuivenhoven, J. A. (2004). A review of CETP and its relation to atherosclerosis. Journal of Lipid Research, 45(11), 1967–1974.

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. Journal of Lipid Research, 54(9), 2325–2340.

DiStefano, J. K., Kingsley, C., Wood, G. C., Chu, X., Argyropoulos, G., Still, C. D., … Gerhard, G. S. (2015). Genome-wide analysis of hepatic lipid content in extreme obesity. Acta Diabetologica, 52(2), 373–382.

Dugas, L. R., Bernabé, B. P., Priyadarshini, M., Fei, N., Park, S. J., Brown, L., … others. (2018). Decreased microbial co-occurrence network stability and SCFA receptor level correlates with obesity in African-origin women. Scientific Reports, 8(1), 1–17.

Dugas, L. R., Forrester, T. E., Plange-Rhule, J., Bovet, P., Lambert, E. V, Durazo-Arvizu, R. A., … others. (2017). Cardiovascular risk status of Afro-origin populations across the spectrum of economic development: findings from the Modeling the Epidemiologic Transition Study. BMC Public Health, 17(1), 438.

Duncan, S. H., Lobley, G. E., Holtrop, G., Ince, J., Johnstone, A. M., Louis, P., & Flint, H. J. (2008). Human colonic microbiota associated with diet, obesity and weight loss. International Journal of Obesity, 32(11), 1720–1724.

Fawcett, K. A., & Barroso, I. (2010). The genetics of obesity: FTO leads the way. Trends in Genetics, 26(6), 266–274.

Feingold, K. R., & Grunfeld, C. (2018). Introduction to lipids and lipoproteins. In Endotext [Internet]. MDText. com, Inc.

Flegal, K. M., Kruszon-Moran, D., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2016). Trends in obesity among adults in the United States, 2005 to 2014. Jama, 315(21), 2284–2291.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Fryett, J. J., Morris, A. P., & Cordell, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. Genetic Epidemiology, n/a(n/a). https://doi.org/10.1002/gepi.22290

Gamazon, E. R., Huang, R. S., Cox, N. J., & Dolan, M. E. (2010). Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proceedings of the National Academy of Sciences, 107(20), 9287–9292.

Gamazon, E. R., Im, H. K., Liu, C., Nicolae, D. L., Cox, N. J., & others. (2013). The convergence of eQTL mapping, heritability estimation and polygenic modeling: emerging spectrum of risk variation in bipolar disorder. ArXiv Preprint ArXiv:1303.6227.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., … others. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics, 47(9), 1091.

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., & König, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. Genetic Epidemiology.

GTEx Consortium. (2015). GTEx pilot analysis: multitissue gene regulation in humans. Science, 348(6235), 648–660. https://doi.org/10.1126/science.1262110

Guan, Y., & Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics, 1780–1815.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., … others. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics, 48(3), 245.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … others. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research, 22(9), 1760–1774.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

Hong, Y.-H., Nishimura, Y., Hishikawa, D., Tsuzuki, H., Miyahara, H., Gotoh, C., … others. (2005). Acetate and propionate short chain fatty acids stimulate adipogenesis via GPCR43. Endocrinology, 146(12), 5092–5099.

Huang, Q. (2015). Genetic study of complex diseases in the post-GWAS era. Journal of Genetics and Genomics, 42(3), 87–98.

Kalliomäki, M., Carmen Collado, M., Salminen, S., & Isolauri, E. (2008). Early differences in fecal microbiota composition in children may predict overweight. The American Journal of Clinical Nutrition, 87(3), 534–538.

Karlsson, F., Tremaroli, V., Nielsen, J., & Bäckhed, F. (2013). Assessing the human gut microbiota in metabolic diseases. Diabetes, 62(10), 3341–3349.

Keys, K. L., Mak, A. C. Y., White, M. J., Eckalbar, W. L., Dahl, A., Mefford, J., … Gignoux, C. R. (2019). On the cross-population portability of gene expression prediction models. BioRxiv. https://doi.org/10.1101/552042

Kimura, I., Inoue, D., Hirano, K., & Tsujimoto, G. (2014). The SCFA receptor GPR43 and energy metabolism. Frontiers in Endocrinology, 5, 85.

Kosmas, C. E., Dejesus, E., Rosario, D., & Vittorio, T. J. (2016). CETP inhibition: past failures and future hopes. Clinical Medicine Insights: Cardiology, 10, CMC--S32667.

Krebs, M., Krssak, M., Bernroider, E., Anderwald, C., Brehm, A., Meyerspeer, M., … Roden, M. (2002). Mechanism of amino acid-induced skeletal muscle insulin resistance in humans. Diabetes, 51(3), 599–605.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., & Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nature Methods, 7(10), 813.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., … others. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature, 500(7464), 541–546.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences, 102(31), 11070–11075.

Li, B., Verma, S. S., Veturi, Y., Verma, A., Bradford, Y., Haas, D. W., & Ritchie, M. D. (2018). Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. PSB, 448–459.

Liu, Y., Ding, J., Reynolds, L. M., Lohman, K., Register, T. C., De La Fuente, A., … others. (2013). Methylomics of gene expression in human monocytes. Human Molecular Genetics, 22(24), 5065–5074.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … others. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature, 518(7538), 197–206.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., … others. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics, 48(11), 1443.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … others. (2013). The genotype-tissue expression (GTEx) project. Nature Genetics, 45(6), 580.

Lozupone, C. A., & Knight, R. (2008). Species divergence and the measurement of microbial diversity. FEMS Microbiology Reviews, 32(4), 557–578.

Luke, A., Bovet, P., Forrester, T. E., Lambert, E. V, Plange-Rhule, J., Schoeller, D. A., … others. (2011). Protocol for the modeling the epidemiologic transition study: a longitudinal observational study of energy balance and change in body weight, diabetes and cardiovascular disease risk. BMC Public Health, 11(1), 927.

Luke, A., Bovet, P., Plange-Rhule, J., Forrester, T. E., Lambert, E. V, Schoeller, D. A., … Cooper, R. S. (2014). A mixed ecologic-cohort comparison of physical activity & weight among young adults from five populations of African origin. BMC Public Health, 14, 397. https://doi.org/10.1186/1471-2458-14-397

Ma, L., Yang, J., Runesha, H. B., Tanaka, T., Ferrucci, L., Bandinelli, S., & Da, Y. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. BMC Medical Genetics, 11(1), 55.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … others. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, 45(D1), D896--D901.

Mamat, M., Kamila Deraman, S., Maizura Mohamad Noor, N., Rokhayati, Y., Afoakwah, A. N., Owusu, W. B., … others. (2011). Health topics: Obesity. Asian Journal of Applied Sciences, 5(1), 1–13.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … others. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753.

Manor, O., & Segal, E. (2013). Robust prediction of expression differences among human individuals using only genotype information. PLoS Genetics, 9(3).

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., … Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics, 100(4), 635–649.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics, 51(4), 584.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics, 9(5), 356–369.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … others. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics, 48(10), 1279–1283.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One, 8(4).

Mikhaylova, A. V, & Thornton, T. A. (2019). Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations. Frontiers in Genetics, 10, 261.

Mitchell, J., Ferguson, S. M., & Ferguson, S. M. (2007). National Institutes of Health policy for sharing data obtained in NIH supported or conducted genome-wide association studies (GWAS). In Federal Register (Vol. 72). National Institutes of Health.

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., … Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genetics, 14(8), e1007586.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., … others. (2016). Heart disease and stroke statistics-2016 update: a report from the American Heart Association. Circulation, 133(4), e38.

Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., … others. (2019). Tigar: An improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. The American Journal of Human Genetics, 105(2), 258–266.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genetics, 6(4).

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. Journal of Machine Learning Research, 12, 2825–2830.

Peet, R. K. (1974). The measurement of species diversity. Annual Review of Ecology and Systematics, 5(1), 285–307.

Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X., & Eckel, R. H. (2006). Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 American Heart Association Scientific Statement on

Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical. Circulation, 113(6), 898–918.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), 904–909.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., … others. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285), 59–65.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(D1), D590--D596.

Rader, D. J., & Hovingh, G. K. (2014). HDL and cardiovascular disease. The Lancet, 384(9943), 618–625.

Ritchie, S. A., & Connell, J. M. C. (2007). The link between abdominal obesity, metabolic syndrome and cardiovascular disease. Nutrition, Metabolism and Cardiovascular Diseases, 17(4), 319–326.

Scott, J. (2004). Pathophysiology and biochemistry of cardiovascular disease. Current Opinion in Genetics & Development, 14(3), 271–279.

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., … others. (2017). Influence of diet on the gut microbiome and implications for human health. Journal of Translational Medicine, 15(1), 73.

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols, 7(3), 500.

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Merz, C. N. B., Blum, C. B., Eckel, R. H., … others. (2014). 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology, 63(25 Part B), 2889–2934.

Tall, A. R., & Rader, D. J. (2018). Trials and tribulations of CETP inhibitors. Circulation Research, 122(1), 106–112.

Teslovich, T. M., Musunuru, K., Smith, A. V, Edmondson, A. C., Stylianou, I. M., Koseki, M., … others. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature, 466(7307), 707–713.

Thompson, J. F., Lira, M. E., Durham, L. K., Clark, R. W., Bamberger, M. J., & Milos, P. M. (2003). Polymorphisms in the CETP gene and association with CETP mass and HDL levels. Atherosclerosis, 167(2), 195–204.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Van Gaal, L. F., Mertens, I. L., & Christophe, E. (2006). Mechanisms linking obesity with cardiovascular disease. Nature, 444(7121), 875–880.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. The American Journal of Human Genetics, 90(1), 7–24.

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., … Chen, L. S. (2016). Imputing gene expression in uncollected tissues within and beyond GTEx. The American Journal of Human Genetics, 98(4), 697–708.

Wang, K., Li, W.-D., Zhang, C. K., Wang, Z., Glessner, J. T., Grant, S. F. A., … Price, R. A. (2011). A genome-wide association study on obesity and obesity-related traits. PloS One, 6(4), e18939.

Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, Gte., … Im, H. K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS Genetics, 12(11), e1006423.

Willer, C. J., & Mohlke, K. L. (2012). Finding genes and variants for lipid levels after genome-wide association analysis. Current Opinion in Lipidology, 23(2), 98.

Wood, A. R., Perry, J. R. B., Tanaka, T., Hernandez, D. G., Zheng, H.-F., Melzer, D., … others. (2013). Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One, 8(5).

Wyatt, S. B., Winters, K. P., & Dubbert, P. M. (2006). Overweight and obesity: prevalence, consequences, and causes of a growing public health problem. The American Journal of the Medical Sciences, 331(4), 166–174.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., … others. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics, 38(2), 203–208.

Zeng, P., & Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nature Communications, 8(1), 1–11.

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., … others. (2009). Human gut microbiota in obesity and after gastric bypass. Proceedings of the National Academy of Sciences, 106(7), 2365–2370.

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics, 9(2).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.

VITA

Paul Chukwuebuka Okoro was born and raised in Onitsha, Anambra State, Nigeria. Before attending Loyola University Chicago, he attended the University of Lagos, Lagos Nigeria, where he earned a Bachelor of Science degree in Cell Biology and Genetics, with Second Class Honors (Upper Division), in 2015. Mr. Okoro was awarded Loyola's MS Bioinformatics Fellowship to study for a Master of Science degree in Bioinformatics and conduct research in the Wheeler Lab under the mentorship of Dr. Heather Wheeler. Before moving to Loyola Chicago from Nigeria, Mr. Okoro enrolled into the University of Ibadan, Ibadan Nigeria, for a Master of Science degree in Epidemiology and Medical Statistics, which he completed in 2019 while still studying in Loyola Chicago. Going forward, Mr. Okoro desires to pursue doctoral degree study in computational biology.