The Pennsylvania State University
The Graduate School

LEVERAGING BIG GENETIC DATA FOR PREDICTION IN MULTI-ETHNIC STUDIES: APPLICATIONS TO TOBACCO USE PHENOTYPES

A Dissertation in
Biostatistics
by
Lina Yang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2021

The dissertation of Lina Yang was reviewed and approved by the following:

Dajiang J. Liu
Associate Professor, Department of Public Health Sciences
Co-Director of Bioinformatics and Genomics Graduate Program
Dissertation Adviser
Co-Chair of Committee

Vernon M. Chinchilli
Distinguished Professor, Department of Public Health Sciences
Chair of Public Health Sciences
Co-Chair of Committee

David T. Mauger
Professor, Department of Public Health Sciences

Bibo Jiang
Assistant Professor, Department of Public Health Sciences

Ian M. Paul
Professor, Department of Pediatrics

Arthur Berg
Associate Professor, Department of Public Health Sciences
Director of Biostatistics PhD Program

Abstract

Large-scale genetic datasets have revolutionized human genetic research. As the cost of sequencing decreases dramatically, biobank scale datasets with millions of individuals and hundreds of millions of genetic variants have emerged. Given the scale of the sequence datasets, query and retrieval of information from them have become a central problem that precedes genetic analyses. We developed an R-package *seqminer2* for efficient querying and retrieving genetic variants in biobank scale datasets. It implements a variant-based index and substantially improves the speed of querying sequence datasets by several magnitudes compared to the other state-of-the-art tools. It also requires much smaller memory to run making it feasible to directly read genetic data into R program. It supports popular file formats for statistical genetic analysis, including VCF/BCF, BGEN, and PLINK formats. The improved efficiency and comprehensive support for various file formats has greatly facilitated our method development for risk prediction in multi-ethnic populations and will facilitate others' research in the genetic and genomic field as well.

With the help of *seqminer2*, we developed a novel meta-analysis approach to predict polygenic risk score (PRS) in multi-ethnic samples. It is currently challenging to construct PRS in the diverse US and worldwide populations because the majority of the available training data are from the European population. If using European samples as training data to predict PRS in non-European samples, the prediction accuracy can be low due to the different patterns of linkage disequilibrium (LD) and heterogeneity of genetic effects in diverse ethnic populations. An alternative way is to train the model with the same population as the target population. However, the sample size for the target population other than European can be much reduced and results in worse prediction performance. Our method integrates multi-ethnic studies as training dataset while accommodating heterogeneity in genetic effects and linkage disequilibrium patterns. It decomposes genetic effect heterogeneity into a fixed effect and top principal components (PCs) of genetic variation. It integrates the heterogeneous genetic effect estimates across ancestries to improve the PRS prediction for individuals from diverse ancestries. We showed our method improved the prediction accuracy for individuals from different ancestries in the simulation comparisons over various scenarios for heterogeneity across diverse populations. Applying our method to GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) dataset improved prediction for tobacco use phenotypes. Our approach facilitates stratifying the risk of smoking behaviors across ancestries and would contribute to quantifying nicotine dependence risk in diverse populations.

Keywords: GWAS, meta-analysis, linkage disequilibrium, polygenic risk score, multi-ethnic population, tobacco usage.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

TOPMed	Trans-Omics Precision Medicine
VCF	Variant Call Format
BCF	Binary Call Format
Vbi	Variant-based Index
LD	Linkage Disequilibrium
GWAS	Genome-wide Association Studies
PRS	Polygenic Risk Score
CHD	Coronary Heart Disease
CRP	C-reactive Protein
LASSO	Least Absolute Shrinkage and Selection Operator
FDR	False Discovery Rates
CS	Continuous Shrinkage
EUR	European
LAT	Latino
EAS	East Asian
SAS	South Asian
AMR	Native American
PC	Principal Component
MI	Myocardial Infarction
BMI	Body Mass Index
PAINTOR	Probabilistic Annotation INTegratOR
MANTRA	Meta-ANalysis of TRansethnic Association studies

MR-MEGA	Meta-Regression of Multi-Ethnic Genetic Association
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion
BLUE	Best Linear Unbiased Estimator
CigDay	Cigarettes per Day
AgeInit	Age of smoking Initiation
GSCAN	Sequencing Consortium of Alcohol and Nicotine use
MAF	Minor Allele Frequency

Acknowledgements

I would like to express my deepest gratitude to my advisor Dr. Dajiang Liu for his invaluable guidance, inspiration, patience, and support throughout my Ph.D. study. His immense knowledge and insightful feedback always pushed me to sharpen my thinking and brought my research work to a higher level. I would also like to thank my committee members Dr. Vernon M. Chinchilli, Dr. David T. Mauger, Dr. Bibo Jiang, and Dr. Ian M. Paul for their suggestions and comments through each stage of my dissertation research.

I would like to extend my thanks to all the faculty members in the Department of Public Health Sciences for their support during my coursework and research. I am grateful to the Division of Biostatistics & Bioinformatics and the Department of Public Health Sciences for offering me a professional training platform to develop my skills and critical thinking. In addition, I am thankful to my collaborator Dr. Xiaowei Zhan at the University of Texas Southwestern Medical Center for his help. I would like to thank my outstanding lab members for their friendship and support throughout this process. I would also like to thank my friends for their companionship and encouragement especially during the pandemic.

Last but not the least, I sincerely thank my parents and grandmother whose love and supports are with me in whatever I pursue.

Chapter 1 Introduction

With the advances in sequencing and genotyping technologies, genome-wide genotype data for large-scale biobanks have been collected. For example, in the Trans-Omics Precision Medicine (TOPMed) sequencing project, over 460 million variants have been identified from over 100,000 whole genome deep sequenced individuals [1]; in the UK biobank dataset, ~500,000 individuals were genotyped [2], and after phasing and genotype imputation using the Haplotype Reference Consortium panel, plus the UK10K and 1000 Genomes reference panel, the number of variants in UK biobank can be over 80 million. These enormous genomic data resources provide opportunities for discovering novel genetic associations of complex diseases and traits and offer comprehensive insights into human genetics research and drug development.

It is not surprising that analyzing biobank scale datasets encounters new computational challenges. Reading large-scale human genetics data directly into statistical analysis programs such as the R program is infeasible. Efficient query and retrieval of information from these datasets have become a central problem that precedes virtually all population genomic analyses and method development. Index for genetics datasets, which is conceptually analogous to the index for dictionaries, allows us to narrow the searches to the genomic intervals of interest. An effective index can be critical for efficient query and retrieval of sequence data. Most of the existing tools rely on *tabix* index [3], the default choice for query VCF/BCF format files. However, because of the design limitations of the *tabix* index, these tools can become very slow for the biobank scale datasets. Other efforts to improve the query of large sequence datasets involve defining more compact file formats, but these file formats have not been adopted by many downstream analysis softwares. A tool that is highly efficient for querying and retrieving biobank scale sequence datasets and at the same time comprehensive support for commonly used file formats is in demand for modern statistical genetics analysis.

Tobacco use is the single leading preventable cause of death globally. Previous family and twin-based studies have shown the heritability of smoking behaviors. Genome-wide association studies (GWAS) also showed significant associations between smoking behaviors and hundreds of genetic variants which linked to biological mechanisms in substance use, including nicotine dependence. Identifying people at high risk of nicotine dependence could facilitate early prevention which could be a critical strategy in reducing tobacco usage. The polygenic risk score (PRS) is a commonly used estimate to gauge an individual's risk for a complex disease or trait compared to others. It reflects an aggregated genetic variants' effect on the likelihood of a complex disease or trait for an individual. To date, the majority of genome-wide studies examined European ancestry individuals. Because of this lack of diversity, the predictive validity of PRS in individuals of other ancestries is many folds less accurate than in individuals of European ancestry. Thus, it is widespread concern about this disparity in PRS accuracy for diverse populations. Methods and tools to make PRS useful for other populations are needed before PRS can be routinely used in clinical practice.

In this dissertation, I address challenges in prediction in multi-ethnic populations and present tool and methods to leverage big genetic data for prediction in multi-ethnic studies. In Chapter 2, we develop a tool to efficiently inquire and retrieve sequencing data from ultra-large biobank scale datasets. In Chapter 3, we propose a meta-analysis method for risk prediction in multi-ethnic populations and apply it to tobacco usage.

Chapter 2 Tool for Efficient Retrieval of Genetic Variants in Biobank Datasets

2.1 Introduction

Human genetic studies are revolutionized by the cost-effective sequencing and genotyping technologies. As the cost for deep whole genome sequencing drops below $1,000 USD and that for genome- wide genotyping drops below $100 USD, many ultra-large biobank scale datasets with millions of individuals begin to emerge. For sequence datasets with many individuals, the number of discovered variations is also increasing rapidly, as more rare variants can be uncovered as more samples are sequenced. For example, in the Trans-Omics Precision Medicine (TOPMed) sequencing project, there are >460 million variants identified from >100,000 whole genome deep sequenced individuals [1]. In the UK biobank dataset, approximately half a million individuals were genotyped [2]. After genotype imputation using the Haplotype Reference Consortium panel, plus the UK10K and 1000 Genomes reference panel, there are over 80 million genetic variants. Given the scale of these sequence datasets, it is infeasible to directly read them into RAM. Efficient query and retrieval of information from these datasets have become a central problem that precedes virtually all genetic analyses.

An effective index is critical for efficient query and retrieval for sequence datasets. Conceptually similar to the index used in books and dictionaries, index for computer files allows us to narrow down the searches to the genomic intervals that may overlap the variants of interest. For most sequence data formats, such as VCF/BCF format, *tabix* [3] index is the default choice for query which combines binning and linear index to query the files. *Tabix* works by clustering chromosomal positions into bins of fixed sizes. Therefore, the index file size is small and does not vary with the number of variants. For densely genotyped datasets with many individuals, each bin in the *tabix* index contains numerous variants. Even when querying and retrieving a single variant, the entire bin that contains the variant will be uncompressed and extracted, which greatly reduces the efficiency of retrieving variants. Existing R packages such as *VariantAnnotation* [4] and *PopGenome* [5] integrate *tabix* index and can become very slow for biobank scale dataset.

There have been efforts to improve the query of large sequence datasets. One approach is to define more compact file formats as BGT [6], GQT [7], so that indexing and queries can be more efficiently performed. A major limitation for these approaches is that they only support limited variant types. Multi-allelic variants or imputed genotypes may not be supported well. Also these new file formats have not been adopted by many downstream analysis softwares, such as genetic association analysis. These limitations prevent them from being widely used in statistical genetics research.

Alternatively, to improve the query of sequence datasets, we seek to develop a novel variant-based index (*vbi*) that can work with VCF/BCF format, and provide implementations both in R and in command line tools for querying genomic sequence datasets. With the novel *vbi* index, *seqminer2* greatly improved existing R packages for querying VCF/BCF files. Compared to *VariantAnnotation* and *PopGenome*, the query speed for *seqminer2* can be magnitude faster. The companion *seqminer2* command line tool for querying VCF files outperformed *bcftools* [8] in speed by 5-folds when extracting single range from VCF file and by more than 200-folds when extracting multiple randomly chosen genomic ranges. It slightly outperformed *GIGGLE* [9] in speed (*seqminer2* supported many other features). Such speed improvement makes it possible to retrieve variant genotypes on the fly for many applications such as the calculation of linkage disequilibrium (LD) coefficients, the calculation of LD scores, gene-level association analysis and transcriptomic wide association analysis. In addition to improved speed, *seqminer2* also requires much smaller memory to run compared to other tools when querying or retrieving sequence variants.

Another uniqueness of *seqminer2* is its comprehensive support for popular file formats. To make *seqminer2* a convenient and comprehensive tool, we also reimplemented support for a few other file formats that are commonly used in statistical genetics analysis, including BGEN [10] format, which was developed to store imputed genotypes for large datasets and binary PLINK format, which is a state-of-art file format for storing array genotypes. Our implementation was considerably faster than the *rbgen* package for querying BGEN files, and of comparable speed as *BEDMatrix* for querying binary PLINK files. To our knowledge, *seqminer2* is the only package that supports all commonly used file formats in statistical genetics analysis (Table 2.1). Its comprehensive feature and improved efficiency make it a valuable tool for data analysis and method development in R.

***Table 2.1*: Feature Comparison for seqminer2, VariantAnnotation, PopGenome, giggle, bcftools, rbgen, snpStat, and BEDMatrix.**

Software	Query File Types	Integration with R
seqminer2	VCF, BCF, BGEN, PLINK	Yes
VariantAnnotation	VCF	Yes
PopGenome	VCF	Yes
giggle	VCF, PLINK	No
bcftools	VCF, BCF	No
rbgen	BGEN	Yes
snpStat	PLINK	Yes
BEDMatrix	PLINK	Yes

2.2 Methods

In this section, I first illustrate the motivation for the development of the variant based index (*vbi*) which is implemented in *seqminer2*. I then describe *vbi* in details and the algorithms *seqminer2* uses. I also describe the reimplementation of BGEN file support and how it improves the speed of query and retrieval.

2.2.1 Motivation for the Development of VBI

Tabix is widely used tool to randomly query and retrieve sequence variants from large-scale genomic datasets. It uses a hybrid of binning and linear index to query bgzip compressed VCF/BCF or generic tab-delimited files. Binning index is small in size and efficient for querying moderate-sized VCF files with a few thousand individuals and millions of genetic variants. Yet, for densely genotyped biobank scale datasets, *tabix* index become very inefficient owing to the design of binning index.

For binning index, the genetic variants in each chromosome are clustered hierarchically into bins of different sizes. When querying a genetic interval, the smallest bin that contains the queried interval will be uncompressed and processed. The bin size is preset and not dependent on the density of the variants across the chromosome. For modern sequence datasets with many individuals, there can be one genetic mutation observed per 20 basepair [1]. In this case, for a sequence dataset with 100,000 individuals, the smallest bin of 16,384 bp used by *tabix* contains 80 MB data. So to query 100 randomly chosen genetic variants, 100 bins with 80,000 genetic variants and 8 GB data will need to be uncompressed, even though most of the uncompressed data is irrelevant to the queried variants.

The limitation for binning index motivated us to develop an alternative method to query VCF/BCF index.

2.2.2 Build an Index for bgzipped VCF/BCF File

As *tabix* index, *vbi* is also based upon bgzip file. In *vbi* index file, we store the genomic position and the offset for each variant in the index file. So for a file with M variants, all the indices form a M by two matrix. For a chromosome with 10 million markers, the index file size is 10^7*2* (8 byte) = 160 MB. The size of the *vbi* index file does not depend on the number of samples, and remains small enough to be loaded to computer RAM in entirety. With *vbi*, we can directly locate the location of the queried genetic variants, and minimize the amount of redundant data that needs to be uncompressed.

2.2.3 Query bgzip Compressed VCF/BCF by Positions

To locate any marker in a large VCF/BCF file by position, we make use of binary search, which takes no more than log2(number of marker) comparisons. It means that for a dataset with 10 million genetic variants, it takes only up to 24 comparisons. In this case, to query 100 random selected genomic ranges from a VCF file, this index strategy takes less than 20 s, while the alternative *tabix* index takes 4 min, 12 times slower. To query 100 randomly chosen non-consecutive ranges from a BCF file, this index strategy takes less than 1 s.

2.2.4 Reimplementation of BGEN Support

BGEN format was developed to store genotype probabilities from genotype imputation. It has become a popular file format in statistical genetic analysis. For example, imputed genotypes from UK Biobank were released in BGEN format. We followed the BGEN file format specification and reimplemented the query and retrieval in C++. This reimplementation is in fact considerably faster than the official implementation. We attribute the speed improvement to several software engineering advances:

First, we seek to minimize disk I/O, which is a major bottleneck for query and retrieval speed. For example, to read multiple genomic ranges, we first cluster these ranges, and merge ranges in proximity and read the whole block of the variants of all samples into the memory. Second, our implementation improves by optimizing the most common cases. BGEN format, as developed is very general, and capable of handling multiploidy and multiallelic variants. Yet, human germline genome is diploid, and a majority of the variants are biallelic. We improve the performance by simplifying the implementation on the most common case, and then separately considering the more special cases. Compared to the original implementation that treats common and uncommon cases indifferently, *seqminer2* gains considerable efficiency. Lastly, we used the dictionary-based decompression algorithm in ZSTD, which has faster decompression speed compared to the system-default decompression algorithm for large files. Together these software engineering works considerably improved the speed of *seqminer2* over *rbgen* package.

2.3 Results

We extensively evaluated *seqminer2* for querying VCF/BCF, BGEN and PLINK files using both the R package and the command line tool. We considered scenarios with very large number of individuals (N=487,409) as motivated by UK Biobank datasets [2]. We compared the query of various numbers of genomic ranges from largest chromosome (chr2 with number of markers M=8,129,063) and smallest chromosome (chr21 with number of markers M=1,261,158). Comparisons were conducted on a computer server with

Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz CPU, 128 GB RAM and 7,200 rpm hard drive. We also conducted comparisons requesting only 4 GB of memory in each node.

The time used for querying and parsing the output was recorded. For each scenario, the query was repeated 100 times to ensure stability of the results. The median time used was reported.

2.3.1 Evaluation of Reading VCFs with Variant-based Index

First, we evaluated *seqminer2* against *VariantAnnotation* and *PopGenome* R packages to query *tabix*-indexed VCF files. For a file containing 487,409 individuals and 8,129,063 genetic variants, the function *readSingleChromosomeVCFToMatrixByRange* took 22.35 s to read a single range with 100 genetic variants, and 16.57 s to read 100 randomly chosen ranges. The time used for querying multiple randomly selected ranges is often comparable and sometimes even less than the time used for querying one single range, which represents a unique advantage for *vbi* index.

On the other hand, the R packages based upon *tabix* index have greatly reduced speed for querying multiple randomly selected ranges. This is because randomly chosen ranges tend to fall into multiple distinct bins, each of which contains a large amount of data. Retrieval of variants in these ranges requires uncompressing all overlapping bins, which constitutes a severe bottle neck.

In our comparison, we used *readGT* function in *VariantAnnotation* package Version 1.28.11 to query variants. It took 99.86 s to read a single range with 100 genetic-variants and 355.40 s to read 100 randomly selected ranges from the same file as we used for *seqminer2*. We also compared with *readVCF* function in *PopGenome* package. It took 830.70 s to read a single range with 100 genetic variants. To the best of our knowledge, readVCF does not support the query of multiple *tabix* ranges. If simply looped through all 100 randomly chosen ranges, the query took more than 21h. The advantage of *seqminer2* over *VariantAnnotation* and *PopGenome* increased as more ranges were queried and retrieved (Table 2.2 A). For instance, *seqminer2* took 187.62 s to read a single range with 1,000 genetic variants, whereas *VariantAnnotation* took 1,006.10 s which was more than 5 times of *seqminer2*, and *PopGenome* took 802.90 s. When reading 1,000 randomly chosen ranges, *seqminer2* took 166.60 s, but *VariantAnnotation* and *PopGenome* respectively took 4,035.86 s and >1 d, which was substantially slower than *seqminer2*.

We next evaluated *seqminer2* command line tool against two other command line tools, *GIGGLE* and *bcftools*. *Seqminer2* outperformed GIGGLE when reading a single range or randomly selected ranges. *Bcftools* was far less efficient than *seqminer2*. It took up to 4 times more time when reading a single range, and 200 times more time than *seqminer2* when reading randomly selected ranges (Table 2.2 B).

2.3.2 Evaluation of Reading BCF Files

Seqminer2 is one of the few tools that support BCF format, which is a binary version of the VCF format. BCF format was developed as an improvement to VCF files for more efficient storage and query. Both *seqminer2* and *bcftools* performed efficiently while reading a single region. They took less than 1 s to extract 100 variants and less than 5 s to extract 1000 variants, which was indeed much faster compared to reading VCF files. However, when reading multiple randomly selected ranges, there was a sharp increase in time for *bcftools*. *Bcftools* took up to 23 min to extract 1,000 randomly chosen ranges, a disadvantage owing to the *tabix* index, while *seqminer2* still only took less than 5 s (Table 2.2 C).

2.3.3 Evaluation of Reading Binary PLINK Files

We used the function *readPlinkToMatrixByIndex* in *seqminer2* to query binary PLINK files. *Seqminer2* and *BEDMatrix* [11] performed about equally well when read in a single range or multiple randomly selected ranges. The function *read.plink* in *snpStats* [12] failed to read in variants, which revealed its limitations for handling biobank scale datasets.

Unlike *seqminer2*, *PLINK2* can only output extracted sequence variants from BED files into a text file, which needs to be read into R separately. This appears to be less convenient. In our evaluation, *seqminer2* is also faster than *PLINK2* in almost all scenarios for querying and reading variants into R (Table 2.2 D).

2.3.4 Evaluation of Reading BGEN Files

We used *seqminer2* and an R package specifically designed to load BGEN format files called *rbgen* to query *bgenix*-indexed bgen files. The function *readBGENToMatrixByRange* in *seqminer2* was >10 times faster than *bgen.load* in *rbgen* to read a single range. When reading randomly chosen ranges, the advantage for *seqminer2* was even bigger: *seqminer2* was >200 times faster than *rbgen* when reading randomly chosen ranges from largest chromosome in UK Biobank, and remained >10 times faster when reading from smallest chromosome in UK Biobank (Table 2.2 E).

***Table 2.2*: Comparison of Query Speed of seqminer2 with Alternative Software Packages.** For each comparison, we used files with the largest chromosome (chr2) and the smallest chromosome (chr21) in the UK Biobank dataset. Sample size is N=487,409. We considered scenarios 1) for querying a single genomic range that contains 100 or 1000 variants and 2) for querying 100 or 1000 randomly chosen genomic ranges with 1 variant in each range. For each scenario, we repeat the comparison for 100 times and median time was recorded. The queried results were also compared and verified for the correctness.

A) Comparison for Querying VCF Files.

Datasets	Total Number of Queried Variants	Single Range			Multiple Random Genomic Ranges		
		seqminer2	VariantAnnotation	PopGenome	seqminer2	VariantAnnotation	PopGenome
chr2	100	22.35	99.86	830.70	16.57	355.40	78,253.79
	1000	187.62	1,006.10	802.90	166.60	4,035.86	> 1d
chr21	100	20.48	98.22	860.15	21.15	395.70	81,948.25
	1000	193.92	979.51	903.74	250.93	5,401.41	> 1d

B) Comparison for Querying VCF Files Using Command Line Tool.

Datasets	Total Number of Queried Variants	Single Range			Multiple Random Genomic Ranges		
		seqminer2	giggle	bcftools	seqminer2	giggle	bcftools
chr2	100	1.47	1.83	7.59	1.72	2.26	311.86
	1000	13.29	14.58	69.21	16.38	18.91	3,202.17
chr21	100	1.59	2.00	8.39	1.39	1.92	329.53
	1000	15.93	17.37	83.28	16.03	18.64	3,404.56

C) Comparison for Querying BCF Files.

Datasets	Total Number of Queried Variants	Single Range		Multiple Random Genomic Ranges	
		seqminer2	bcftools	seqminer2	bcftools
chr2	100	0.75	0.51	0.99	128.35
	1000	2.73	4.65	4.30	1,396.00
chr21	100	0.25	0.18	0.63	125.04
	1000	1.21	2.02	2.80	1,296.98

D) Comparison for Querying PLINK Files.

Datasets	Total Number of Queried Variants	Single Range				Multiple Random Genomic Ranges			
		seqminer2	snpStat	BEDMatrix	PLINK2	seqminer2	snpStat	BEDMatrix	PLINK2
chr2	100	13.47	ERROR	32.03	6.13	16.89	ERROR	16.42	30.71
	1000	15.78	ERROR	34.60	39.62	18.16	ERROR	17.02	81.04
chr21	100	3.16	ERROR	3.27	4.02	3.46	ERROR	2.58	8.66
	1000	4.37	ERROR	5.69	38.84	4.84	ERROR	5.13	52.02

E) Comparison for Querying BGEN Files.

Datasets	Total Number of Queried Variants	Single Range		Random Ranges	
		seqminer2	rbgen	seqminer2	rbgen
chr2	100	2.55	25.31	3.17	75.42
	1000	23.15	233.23	27.38	736.72
chr21	100	2.35	22.73	2.78	30.37
	1000	21.50	212.88	22.53	293.12

2.3.5 Evaluation of memory usage

We recorded the maximum memory each software used when querying files with the largest chromosome (chromosome 2) in the UK Biobank dataset. *Seqminer2* almost always required less memory than other tools (Table 2.3).

Table 2.3: **Comparison of Memory Usage of Query.** We recorded the maximum memory each software used to query files with the largest chromosome (chr2) in the UK Biobank dataset. Sample size is N=487,409.

Tools	Datasets	Total Number of Queried Variants	Single Range	Multiple Random Genomic Ranges
seqminer2 R package	VCF file	100	1.0 GB	2.7 GB
	BCF file	100	1.3 GB	2.1 GB
	BGEN file	100	0.1 GB	2.0 GB
	PLINK file	100	1.0 GB	1.9 GB
seqminer2 command line tool	VCF file	100	0.4 GB	2.0 GB
giggle	VCF file	100	0.4 GB	2.1 GB
VariantAnnotation	VCF file	100	10.9 GB	12.5 GB
PopGenome	VCF file	100	17.3 GB	18.5 GB
bcftools	VCF file	100	0.4 GB	2.1 GB
	BCF file	100	0.1 GB	0.1 GB
PLINK2	PLINK file	100	1.8 GB	1.8 GB
BEDMatrix	PLINK file	100	1.6 GB	1.8 GB
rbgen	BGEN file	100	0.2 GB	2.1 GB
seqminer2 R package	VCF file	1000	7.9 GB	9.4 GB
	BCF file	1000	3.5 GB	4.3 GB
	BGEN file	1000	0.1 GB	2.0 GB
	PLINK file	1000	4.4 GB	5.3 GB
seqminer2 command line tool	VCF file	1000	3.0 GB	4.1 GB
giggle	VCF file	1000	3.0 GB	4.4 GB
VariantAnnotation	VCF file	1000	27.5 GB	11.1 GB
PopGenome	VCF file	1000	17.5 GB	18.9 GB
bcftools	VCF file	1000	3.0 GB	3.0 GB
	BCF file	1000	0.7 GB	0.7 GB
PLINK2	PLINK file	1000	6.7 GB	1.9 GB
BEDMatrix	PLINK file	1000	3.4 GB	3.7 GB
rbgen	BGEN file	1000	0.2 GB	2.1 GB

We also benchmarked the performance of each software using computers with small memories. Specifically, we requested only 4 GB of memory in each node in the cluster when testing each software. In this setting, *seqminer2* still remained the fastest tool. When extracting variants from VCF files, *VariantAnnotation* and *PopGenome* failed to run, because they required much more RAM than 4 GB to retrieve 1,000 variants. *Seqminer2* was the only R package that extracted variants successfully (Table 2.4).

***Table 2.4*: Comparison of Query Speed of *seqminer2* with Alternative Software Packages using Computers with 4GB of Memory.** For each comparison, we used files with the largest chromosome (chr2) in the UK Biobank dataset. Sample size is N=487,409. We considered scenarios for 1) querying a single genomic range that contains 100 or 1000 variants and 2) for querying 100 or 1000 randomly chosen genomic ranges with 1 variant in each range. For each scenario, we repeated the comparison for 100 times and median time was recorded. The queried results were also compared and verified for the correctness.

A) **Comparison for Querying VCF Files Using R packages.**

Total Number of Queried Variants	Single Range			Multiple Random Genomic Ranges		
	seqminer2	VariantAnnotation	PopGenome	seqminer2	VariantAnnotation	PopGenome
100	24.91	Fail to run	Fail to run	23.34	Fail to run	Fail to run
1000	363.43	Fail to run	Fail to run	509.54	Fail to run	Fail to run

B) **Comparison for Querying VCF Files Using Command Line Tool.**

Total Number of Queried Variants	Single Range			Multiple Random Genomic Ranges		
	seqminer2	giggle	bcftools	seqminer2	giggle	bcftools
100	1.80	2.40	11.86	1.72	2.44	328.28
1000	22.47	22.69	111.70	16.68	19.26	3396.07

C) **Comparison for Quering BCF Files.**

Total Number of Queried Variants	Single Range		Multiple Random Genomic Ranges	
	seqminer2	bcftools	seqminer2	bcftools
100	0.79	0.59	1.56	140.69
1000	3.71	5.42	30.45	1493.62

D) Comparison for Quering PLINK Files.

Total Number of Queried Variants	Single Range			Multiple Random Genomic Ranges		
	seqminer2	BEDMatrix	PLINK2	seqminer2	BEDMatrix	PLINK2
100	15.53	12.47	7.66	16.24	12.61	37.06
1000	55.71	58.11	181.37	49.61	60.56	231.71

* PLINK2 has some failed jobs when randomly extracting 1000 non-consecutive regions. BEDMatrix have some failed jobs when randomly extracting 100, or 1000 non-consecutive regions. Time was calculated based on the successful jobs.

E) Comparison for Quering BGEN Files.

Total Number of Queried Variants	Single Range		Multiple Random Genomic Ranges	
	seqminer2	rbgen	seqminer2	rbgen
100	0.19	4.20	0.18	35.81
1000	0.18	4.21	0.28	322.37

2.4 Conclusions and Discussions

Here we showed that *seqminer2* is a very efficient software tool optimized for indexing and querying VCF/BCF. It also accommodates other commonly used file format in statistical genetic analysis, such as BGEN and PLINK files. It can scale well to biobank scale sequence datasets with hundreds of millions of variants.

While *seqminer2* greatly improves the efficiency over other tools, a few practical considerations in data analysis can also make huge differences in speed and warrants discussions. Disk I/O is a major bottleneck for large scale data analysis. Minimizing disk I/O is key to improving the speed for data analysis. If multiple ranges of data (e.g. multiple genes) need to be analyzed, we recommend reading in multiple ranges in one batch using *seqminer2*, instead of reading each range separately. If the goal is to analyze the entire chromosome, as long as the system memory allows, reading in all variants on the chromosome can be more efficient.

The current implementation focuses widely used file formats VCF/BCF, BGEN and PLINK. We noted that the indexing strategy can be readily extended to support generic file formats, e.g. tab- delimited files that include columns of chromosomal positions. Generic file formats have been broadly used in store annotation information, GWAS summary statistics. Implementing *vbi* index could be extremely useful to accelerate the query of files of these files.

Taken together, the improved efficiency and comprehensive features of *seqminer2* have already greatly facilitated method development and large data analysis in our research. We expect it to be a very useful tool for biobank scale data analysis for others as well.

2.5 Reference

[1] Taliun D. et al. (2019) *Sequencing of 53 831 diverse genomes from the NHLBI TOPMed Program*. bioRxiv. doi:10.1101/563866v1.
[2] Bycroft C. et al. (2018) *The UK Biobank resource with deep phenotyping and genomic data*. Nature, 562, 203–209.
[3] Li H. (2011) *Tabix: fast retrieval of sequence features from generic TAB-delimited files*. Bioinformatics, 27, 718–719.
[4] Obenchain V. et al. (2014) *VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants*. Bioinformatics, 30, 2076–2078.
[5] Pfeifer B. et al. (2014) *PopGenome: an efficient Swiss army knife for population genomic analyses in R*. Mol. Biol. Evol., 31, 1929–1936.
[6] Li H. (2016) *BGT: efficient and flexible genotype query across many samples*. Bioinformatics, 32, 590–592.
[7] Layer R.M. et al. (2016) *Efficient genotype compression and analysis of large genetic-variation data sets.* Nat. Methods, 13, 63–65.
[8] Li H. et al. (2009) *The sequence alignment/map format and SAMtools*. Bioinformatics, 25, 2078–2079.
[9] Layer R.M. et al. (2018) *GIGGLE: a search engine for large-scale integrated genome analysis*. Nat. Methods, 15, 123–126.
[10] Band G. , Marchini J. (2018) *BGEN: a binary file format for imputed genotype and haplotype data*. (preprint) https://www.biorxiv.org/content/10.1101/308296v2.
[11] Grueneberg A. (2019) *BEDMatrix: extract genotypes from a PLINK.bed file*. https://rdrr.io/cran/BEDMatrix/man/BEDMatrix-package.html
[12] Clayton D. (2019) *snpStats: SnpMatrix and XSnpMatrix classes and methods*. https://www.bioconductor.org/packages/release/bioc/html/snpStats.html

2.6 Funding

Research reported in this chapter was supported by grants R01HG008983 from NIH/NHGRI and R01GM126479 from NIH/NIGMS. The findings and conclusions do not necessarily reflect the view of the funding agency.

Chapter 3 Meta-analysis Model for Prediction in Multi-ethnic Studies: Applications to Tobacco Use Phenotypes

3.1 Introduction

Tobacco use is the single leading preventable cause of death globally. It caused 8.71 million deaths globally (15.4% of all deaths) in 2019 [1]. 7.69 million of those deaths are attributable to smoking tobacco use, and the rest are attributable to chewing and non-smokers being exposed to second-hand smoke [2, 3]. 20.2% of the deaths among males and 5.84% of the deaths among females result from smoking [2]. Tobacco use increases the risk of serious illness, including lung cancer, chronic obstructive pulmonary disease, heart disease, stroke, and diabetes [4]. Tobacco use behavior is heritable. From twin and family studies, smoking behaviors such as smoking initiation and cigarette per day are reported to have a heritability of 40%-60% [5-7]. Biobank study of smoking behaviors estimated heritability of 18% for smoking initiation and 12% for smoking cessation [6]. Genome-wide association studies (GWAS) showed significant associations between smoking behaviors and genetic variants. Work from our group analyzed up to 1.2 million individuals and found 566 genetic variants associated with tobacco use and alcohol use [8]. These variants affect multiple chemical functions related to glutamatergic and dopaminergic transmission, which is linked to neuronal communication and tied to reward-related memories and learning; acetylcholine nicotinic receptors, activating of which exerts rewarding effects and increase drug-seeking behavior.

Comprehensive information on genetic liability to smoking behaviors could contribute to quantifying nicotine dependence risk. Identifying people at high risk of nicotine dependence could facilitate early prevention, which could be a critical strategy in reducing tobacco usage. Polygenic risk score (PRS) reflects a mathematical aggregation of an individual's genetic variants' effect on the likelihood of a complex disease or trait. Calculation of PRS is based on GWAS data and individual genetic profiles. It has been increasingly used in clinical care and applied in personalized medicine. It is more attractive than traditional clinical risk factors, such as high cholesterol for coronary heart disease (CHD) and high level of C-reactive protein (CRP) for rheumatoid arthritis, because it is relatively affordable and only required to test once in a life time and available from birth. With the increasing number of available GWAS data and the emerging methods development for how PRS is constructed, the PRS is improving its accuracy and interpretability in clinical utility. For example, PRS for CHD has been demonstrated to be a more accurate prediction for potential adverse CHD events when added to established clinical risk factors, such as Framingham risk score [9], QRISK3 score [10] and ACC (American College of Cardiology)/AHA (American Heart Association) pooled cohort equation (PCE) [9, 10]. Especially when the later-life clinical risk factors are typically unknown at an early age, the prevalence of myocardial infarction is more than ten times in the top PRS percentile compared to the mean prevalence [11]. It suggests a potentially significant role of PRS in CHD risk stratification in early life. PRS for breast cancer has also been extensively researched and showed considerable improvement in predicting breast cancer incidence in carriers of established high-risk genes, such as *PALB2*, *CHEK2* [12], *BRCA1*, and *BRCA2* [13]. Models which integrate PRS with conventional non-genetic risks (family history, demographic information such as age and lifestyle) also showed improved performance in stratifying risk of breast cancer and thus could benefit women from better risk management strategies [14-16]. Existing researches also investigated the association between PRS for tobacco consumptions and nicotine dependence and found tobacco consumption PRS can be a good genetic surrogate for nicotine dependence [17]. In a longitudinal study of a community-representative sample, researchers found a significant effect of PRS for cigarettes per day in predicting smoking behavior at age 20 and 24 [18]. Studies also showed a meaningful genetic sharing between the smoking initiation and other substance usages by analyzing the variation the PRS of smoking initiation explained in substance usages initiation [19]. These findings and applications demonstrate that studies on tobacco use could benefit from integrating PRS and further suggest the usage of PRS of smoking-related phenotypes in clinical practice.

There are a few different algorithms to choose from in terms of PRS calculation. The PRS is a weighted sum of an individual's genome-wide genotypes, with weights obtained from estimated effect size from GWAS. So basically, PRS algorithms differ by 1) the genetic variants got picked in the summation, and; 2) the weights assigned to each picked genetic variant. Calculation of PRS is challenging because of the presence of linkage disequilibrium (LD), i.e., variants that are close on genome tend to inherit together, so they are highly correlated. Also, GWAS is usually performed on samples drawn from particular ethnicity groups. So the generalizability of PRS can be poor to the target population if it is not from the same ethnic group. Following are some popular PRS calculation methods that address one or both issues.

1. Lassosum [20]

Lassosum integrates summary statistics from GWAS and an LD reference panel to account for the underlying genetic structure in a penalized regression framework. It selects genetic variants and shrinks effect size β by minimizing

$$f(\beta) = y^T y + (1-s)\beta^T X_r^T X_r \beta - 2\beta^T r + s\beta^T \beta + 2\lambda\left|\left|\beta\right|\right|_1^1$$

where $X_r^T X_r$ is the LD matrix estimated from the reference panel, and $r = X^T y$ is the correlation between variants and the phenotype that can be calculated from GWAS summary statistics. s and λ are tuning parameters selected from a predefined set. s=0.2, 0.5, 0.9 or 1, and λ are 20 values equally spaced on the log-scale from 0.001 and 0.1. When the validation dataset is available, the tuning parameters are optimized by maximizing the correlation of the PRS with the phenotype data in the validation dataset. When the validation dataset is not available, tuning parameters can be optimized by a procedure referred to as “pseudovalidation” by the authors. In pseudovalidation, the correlation of the PRS with the validation phenotypes is substituted by a shrunken *r* that is estimated from a local false discovery rate (FDR), and the local FDR can be calculated using a procedure of Strimmer [21].

2. LDPred2 and LDPred2-auto [22]

LDPred2 also uses summary statistics from GWAS and the external LD reference panel. It is a Bayesian approach and assumes the prior for the variant effects sizes is a point-normal distribution such that only a fraction of variants are causal. The prior has the following form:

$$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

where p is the fraction of causal variants, M is the number of markers and h_g^2 is the heritability of phenotype (the proportion of variance of a phenotype can be explained by genetic variants). The posterior mean of effect size can be derived as

$$E(\beta_j | \tilde{\beta}_j) = \frac{\bar{p}_j \tilde{\beta}_j}{1 + \frac{Mp}{nh^2}}$$

where $\tilde{\beta}_j$ is the residual marginal effect size by subtracting the effect of correlated variants in the LD region and $\bar{p}_j$ is the posterior probability that β_j is from Gaussian distribution.

There are two hyper-parameters, heritability h^2 and sparsity p (the fraction of causal variants). LDPred2 uses a validation set to tune them from a grid. LDPred2-auto, as its name suggested, automatically estimates h^2 and p, so it does not require a validation set. LDPred2-auto is an appealing method because the external validation set with individual level phenotypes may not be available in many applications. Especially when calculating PRS for multi-ethnic populations, the dataset that has individual phenotypes measured usually only includes people from a specific ethnic group. Therefore, tuning hyper-parameters for calculating PRS for other ethnicity groups can be invalid.

3. SBayesR [23]

SBayesR compromises a Bayesian multiple regression with summary statistics from GWAS. For each standardized variant, it assumes the effect is drawn from one of four possible distributions with different probabilities as follows:

$$\beta_j|\pi,\sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1 \\ \sim N(0, 0.01\sigma_\beta^2) & \text{with probability } \pi_2 \\ \sim N(0, 0.1\sigma_\beta^2) & \text{with probability } \pi_3 \\ \sim N(0, \sigma_\beta^2) & \text{with probability } 1 - \sum_{c=1}^{3} \pi_c \end{cases}$$

Since one of the distributions is 0, it indicates that there is π_1 fraction of variants non-causal with 0 effects to the phenotype. For the rest of the variants, their contribution to the phenotype varies because they can fall into a normal distribution with different variances.

4. PRS-CS [24]

PRS-CS also utilizes a Bayesian regression framework. It imposes a continuous shrinkage prior on the variant effects with the following form:

$$\beta_i|\psi_i \sim N(0, \phi\psi_j), \qquad \psi_j \sim g$$

where ϕ is a global scaling parameter common to all variants and determines the model sparsity in general, ψ_i is a local variant specific parameter and g is a continuous density function designed to shrink more on noise and less on truly causal effects. In their paper, the authors used $\psi_i \sim Gamma(1, \delta_j)$ and $\delta_i \sim Gamma(0.5,1)$ and showed good performance in the simulation study and the real data analysis. The strength of continuous shrinkage prior is that it allows adaptive shrinkage for each variant; thus effect for variants that have weak association signals in GWAS would be shrunk more, whereas variants with strong association signals would be shrunk less. Another strength of continuous shrinkage is that it enables effects for variants in the LD block to get updated jointly in posterior inference. Therefore, the model can be appropriately adjusted by the actual LD pattern. The parameter ϕ can be learned from GWAS summary statistics, and no validation data is required.

5. Meta-GRS [25]

Meta-GRS combines multiple PRSs that are calculated using summary statistics from different GWAS and builds a new meta-score. It is a weighted average of multiple standardized scores:

$$\mathrm{GRS}_i^{\mathrm{meta}} = \frac{\beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3}}{\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2\rho_{1,2} + 2\beta_1\beta_3\rho_{1,3} + 2\beta_2\beta_3\rho_{2,3}}}$$

Z_{i1}, Z_{i2}, Z_{i3} are the PRSs for the ith individual using different summary statistics. For instance, Z_{i1} is calculated by summing up genome-wide risk alleles for the ith individual with weights obtained from estimated effects from a specific GWAS-1. It is derived using the basic pruning (thinning variants in high LD with a threshold r^2) and thresholding (filtering out variants above an association p-value P_t) approach. β_1 , β_2, β_3 are the coefficients for each PRS and $\rho_{\mathrm{j,k}}$ are the correlation between jth and kth PRS. Meta-GRS requires a training dataset to train the model.

$\mathrm{GRS}_i^{\mathrm{meta}}$ can also be written in the following form in terms of a per-variant score:

$$\mathrm{GRS}_i^{\mathrm{meta}} \propto \sum_{j=1}^{m} x_{ij} \left(\frac{\beta_1}{\sigma_1} \alpha_{j1} + \frac{\beta_2}{\sigma_2} \alpha_{j2} + \frac{\beta_3}{\sigma_3} \alpha_{j3} \right)$$

where x_{ij} is the genotype for the ith individual's jth variant, $\sigma_1, \sigma_2, \sigma_3$ are the empirical standard deviations of the scores obtained from the training data, $\alpha_{\mathrm{j1}}, \alpha_{\mathrm{j2}}, \alpha_{\mathrm{j3}}$ are the variant effect sizes for the jth variant in each of the scores, $\beta_1, \beta_2, \beta_3$ are the coefficients for each PRS score and m is the total number of variants.

6. Multi-ethnic PRS [26]

Similar to meta-GRS, multi-ethnic PRS also combines PRS from different GWAS, but it restricts each GWAS to samples from a distinct ethnic population. For example, a multi-ethnic PRS with two populations, European (EUR) and Latino (LAT), has the following form:

$$PRS_{EUR+LAT} = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{LAT}$$

where PRS_{EUR} denotes PRS built using the EUR sample and PRS_{LAT} denotes PRS built using the LAT sample. It also employs the pruning and thresholding approach for calculating PRS_{EUR} and PRS_{LAT}, but the pruning parameter r^2 and thresholding parameter P_t are optimized using training data with a single population. The weights α_1 and α_2 are estimated from separate validation data.

In addition, the authors also proposed another multi-ethnic PRS that include the top principal component (PC) of a target dataset:

$$PRS_{EUR+LAT} = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{LAT} + \alpha_3 PC$$

PC in the model can account for the ancestry information, but in their simulation studies and the real data applications, the accuracy of PRS did not gain much from this additional PC.

All these PRS approaches except multi-ethnic PRS are designed for calculating PRS for samples from homogeneous ancestry. Although the evidence showed that common causal variants for a complex trait are shared across diverse ancestries [27], and the directions of these variants' effects are also consistent across different ethnicities [28], due to the different patterns of LD and heterogeneity of genetic effects in different ethnic populations, directly applying these PRS methods to multi-ethnic samples is not appropriate. The majority of the training data that are currently exist are from European samples. During 2008-2017, 67% of PRS studies exclusively included European samples, 19% of studies included East Asian samples, and only 3.8% of studies investigated PRS in African, Hispanic, or Indigenous samples [29]. It is challenging to construct PRS in the diverse US and worldwide populations. If using phenotypes from European samples to predict PRS in non-European samples, the prediction accuracy can be low. For example, for PRS across a few phenotypes studied during 2008-2017 (phenotypes included schizophrenia, myocardial infarction (MI), body mass index (BMI), breast cancer, and type II diabetes), the median effect size of European-derived PRS in African samples was 42% that of matched European samples [29]. A more recent study that investigated 17 anthropometric and blood-panel traits in UK Biobank showed that, on average, the European-derived PRS accuracy was 4.5-fold lower in African samples, 2.0-fold lower in East Asians, and 1.6-fold lower in Hispanic Americans and South Asians compared to European samples [30]. Instead of relying on European training samples to predict PRS in other ancestry target samples, an alternative way for calculating PRS in a diverse population is training the model with the same population as the target population. However, the sample size for the target population other than European is usually much reduced and can result in worse prediction performance.

Properly modeling the genetic effects across multi-ethnic samples while accommodating the heterogeneity in the genetic architecture, especially LD patterns, and heterogeneity in genetic effects can be the key to improving the PRS prediction accuracy. Existing methods for aggregating summary statistics for multi-ethnic analysis (such as Probabilistic Annotation INTegratOR (PAINTOR) [31] and Meta-ANalysis of TRansethnic Association studies (MANTRA) [32]) often cluster studies into discrete ancestry groups and perform meta-analysis, which ignores the fact that study ancestries may vary continuously. Another multi-ethnic analysis approach, Meta-Regression of Multi-Ethnic Genetic Association (MR-MEGA), uses meta-regression to model genetic effects adjusting for the axes of genetic variation among diverse ancestry groups [33]. However, it does not model the non-ancestral heterogeneity of impact due to diet, geographic locations, or other environmental exposures. Better modeling the multi-ethnic samples will lead to improvements in the prediction accuracy of genetic effects and can further increase the PRS prediction accuracy.

To overcome the limitations of the methods that group studies into discrete ancestry groups and meta-regression method that only models the ancestral heterogeneity, we developed a novel meta-regression method which decomposed genetic effect heterogeneity into a fixed effect (an intercept to account for homogeneous effect across studies) and top principal components (PCs) of genetic variation (as derived from a matrix of genome-wide allele frequencies across multi-ethnic studies). We also innovatively integrated the heterogeneous genetic effect estimates across ancestries to improve the PRS prediction for individuals from diverse ancestries. It borrows strengths from predicting PRS with better estimated heterogeneous genetic effects and the existing PRS methods' capability of selecting and shrinking the genetic effects. We evaluated our method and compared it with fixed-effect method through extensive simulation studies over a range of scenarios for heterogeneity across diverse populations. We also presented the results from applying our proposed approach to ultra-large biobank scale datasets with 48 studies worldwide. We showed that our new meta-regression method produced estimated heterogeneous genetic effects with higher accuracy, especially when the heterogeneity is high among ancestry groups It further increased the PRS prediction accuracy and facilitated stratifying risk of smoking behaviors across ancestries.

3.2 Methods and Materials

3.2.1 Meta-Regression Model for Genetic Effect in Trans-Ethnic Meta-Analysis

We model the genetic effect estimates from GWAS as a function of an intercept to account for homogeneous effect across multi-ethnic studies and up to three top allele frequency PCs to account for heterogeneous effect across ancestries. We allow the number of PCs included in the model to vary, and the model with minimal Bayesian information criterion (BIC) value is selected as the final model. We define our model as the following linear regression model:

$$b_{jk} = \sum_{l=0}^{L_j} Z_{lk}\gamma_{jl} + \epsilon_{jk}$$

where b_{jk} is the effect size for variant j in the k^{th} study, Z_{lk} is the l^{th} PC in the k^{th} study, for notational convenience, we set Z_{0k} to 1, γ_{jl} is the effect size for Z_l in variant j, and ϵ_{jk} is the random error which follows $N\left(0, s_{jk}^2\right)$ where s_{jk}^2 is the variance for b_{jk}. L_j is the number of PCs in the model for variant j. We consider L_j=0,1,2 or 3, and the value of L_j is determined by minimal BIC value:

$$BIC_j = K * log\left(RSS_j/K\right) + L_j * log(K)$$

where K is the number of multi-ethnic studies, and RSS_j is the residual sum of squares defined as:

$$RSS_j = \sum_{k=1}^{K} \frac{\left(b_{jk} - \sum_{l=0}^{L} Z_{lk}\gamma_{jl}\right)^2}{s_{jk}^2} / \sum_{k=1}^{K} \frac{1}{s_{jk}^2}$$

When L=0, Z reduces to a column of 1, i.e. with only intercept in the model, the meta-regression model is equivalent to a fixed effect meta-analysis model. It is suitable for modeling genetic effects for variants whose effect size is consistent across studies. When at least one PC is included in the model, it can capture the heterogeneity of genetic effects across studies.

Using the weighted least square method, we can get the best linear unbiased estimator (BLUE) for $\boldsymbol{\gamma}_j$ from a set of training studies as:

$$\hat{\boldsymbol{\gamma}}_j = \left(\boldsymbol{Z}^T\boldsymbol{\Omega}_j\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{\Omega}_j\boldsymbol{b}_j$$

where $\boldsymbol{\Omega}_j$ is a diagonal matrix with diagonal entries being $\frac{1}{s_{j1}^2}, \frac{1}{s_{j2}^2}, \dots, \frac{1}{s_{jK}^2}$.

Based on the meta-regression coefficients estimated from training studies, we can estimate the genetic effect for a target study using:

$$\hat{b}_{jt} = \sum_{l=0}^{L_j} Z_{lt}\hat{\gamma}_{jl}$$

where $\hat{b}_{jt}$ is the estimated genetic effect for variant j in target study t and Z_{lt} is the l^{th} PC for the target study t (target study PCs are calculated from projecting the allele frequency vector from the target study onto the PC space of the training studies). The variance of $\hat{b}_{jt}$ can be derived as:

$$Var\left(\hat{b}_{jt}\right) = \boldsymbol{Z}_t\left(\boldsymbol{Z}^T\boldsymbol{\Omega}_j\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}_t^T$$

3.2.2 Phenotype Definition

We analyzed the cigarettes per day (CigDay) and age of smoking initiation (AgeInit) because they are critical measurements for studying nicotine dependence and are widely recorded in existing epidemiological studies. We adopted their definitions from Brazel et al.'s paper [34].

1. CigDay: A quantitative trait that measures current or former smokers' average number of cigarettes smoked per day. They are binned into 4 categories: 1=1-10, 2=11-20,3=21-30,4=31 or more.

2. AgeInit: A quantitative trait that measures the age that an ever smoker first started smoking regularly.

3.2.3 Dataset Description

For real data application, we used GWAS summary statistics from GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) which collaborates with ultra-large scale datasets, such as 23andMe and UK Biobank, over 50 independent studies worldwide with millions of participants [8]. More details about GSCAN can be found at: https://genome.psych.umn.edu/index.php/GSCAN.

3.3 Simulation Study

3.3.1 Simulation Design

We performed extensive simulation studies to evaluate our proposed method's performance. We used UK Biobank imputed exome genome data. We estimated the ancestry of individuals in UK Biobank using ADMIXTURE [35]. We identified 188,623 Europeans, 2,924 Africans, and 1,602 South Asian (with proportion for putative ancestry of each individual greater than 80%) as the major populations. We filtered variants and kept common variants in all three ancestries with minor allele frequency (MAF) >0.01. The variants were further restricted to HapMap3 variants to ease the computational burden from including too many variants that exceed the capacity of some of the PRS algorithms. We chose the HapMap3 panel because it is well imputed with high accuracy, captures variants from diverse populations, and is widely used in the literature. It resulted in 1,043,770 variants in the final simulation data. All genetic data manipulation in the UK Biobank was conducted using our *seqminer2* tool [36] and PLINK 1.9 software [37].

We fixed 1% variants to be causal variants and set the heritability h_g^2 for a trait at 0.5. The true effect size for variant j is sampled from the following point-normal distribution:

$$\beta_j \sim_{iid} \begin{cases} N\left(0, \dfrac{h_g^2}{Mp}\right) & with\, probability\; p = 1\% \\ 0 & with\, probability\; 1 - p = 99\% \end{cases}$$

where M is the number of variants.

To evaluate our proposed method in the presence of LD at large sample sizes, we simulated the normalized marginal effect estimates using the following distribution:

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta}\sim_{iid} N(\boldsymbol{D}\boldsymbol{\beta},\frac{\boldsymbol{D}}{N})$$

where $\boldsymbol{D}$ is the ancestry-specific LD matrix calculated from a random sample of 1,000 individuals from each ancestry population using PLINK 1.9 software [37], and *N* is the sample size for each study we set as 2,000. For computational efficiency, we simulated $\widehat{\boldsymbol{\beta}}$ by chromosomes, and for large chromosomes with more variants, we chunked the chromosomes into 2-3 blocks and simulated $\widehat{\boldsymbol{\beta}}$ by blocks. The breakpoints for blocks were chosen such that they would not break regions with strong LD. We used the pre-calculated breakpoints as provided by Berisa and Pickrell [38].

We set a proportion of causal variants to be shared across ancestries (European, South Asian, and African population) and have the same effects. The rest of causal variants to be ancestry specific. We performed simulations of genetic effect estimates under three different scenarios:

1. Homogenous Eurasian effect: genetic effects for ancestry specific causal variants present in European and South Asian population with identical effect size.
2. Modest heterogeneous Eurasian effect: genetic effects for ancestry specific causal variants present in European and South Asian population but the effect size in South Asian is twice as much as European.
3. Excessive heterogeneous Eurasian effect: similar as modest heterogeneous Eurasian effect, but the effect size for ancestry specific causal variants in South Asian is triple as much as European.

For each scenario, we also varied the proportions of causal variants that were shared across all three ancestries. We simulated ten replicates of $\widehat{\boldsymbol{\beta}}$ for each ancestry, which represented ten independent studies from each ancestry population.

We generated phenotype data based on the linear model:

$$Y_i = \sum_{j=1}^{M} X_{ij}\beta_j + \epsilon_i$$

where X_{ij} is the standardized genotype for variant j in individual i, and ϵ_i represents the environmental effect for individual *i* and follows $N(0, 1 - h_g^2)$.

3.3.2 Simulation Results

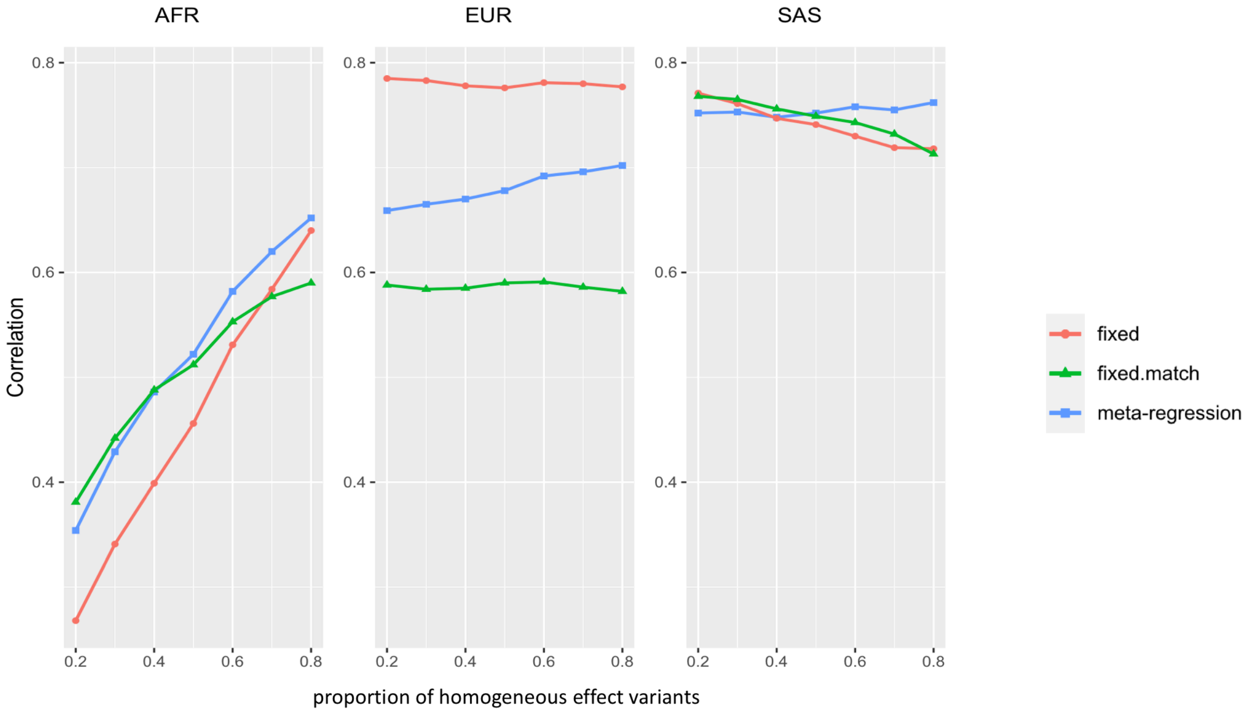
In this section, we first evaluated the performance of our proposed meta-regression method on estimating the genetic effects. For each simulation scenario, we performed leave-one-out cross-validation and compared the prediction accuracy with (i) fixed effect model using studies from all ancestries as training data; (ii) fixed effect model using studies whose ancestry match the target ancestry as training data. We then calculated PRS using the estimated genetic effects from our meta-regression method. To find the most suitable PRS algorithm to integrate with our meta-regression method, we assessed the performance of existing PRS algorithms, including lassosum, LDPred2-auto, SBayesR, and PRS-CS. We used our estimated genetic effects as input of the PRS algorithms. We also calculated the PRS using the multi-ethnic PRS algorithm. Since multi-ethnic PRS is specifically designed for calculating PRS using studies from different ethnicity populations, we directly used the simulated genetic effects as the input for multi-ethnic PRS and compared its PRS prediction accuracy with our meta-regression method integrated with the above described PRS algorithms.

I. Genetic Effects Estimation

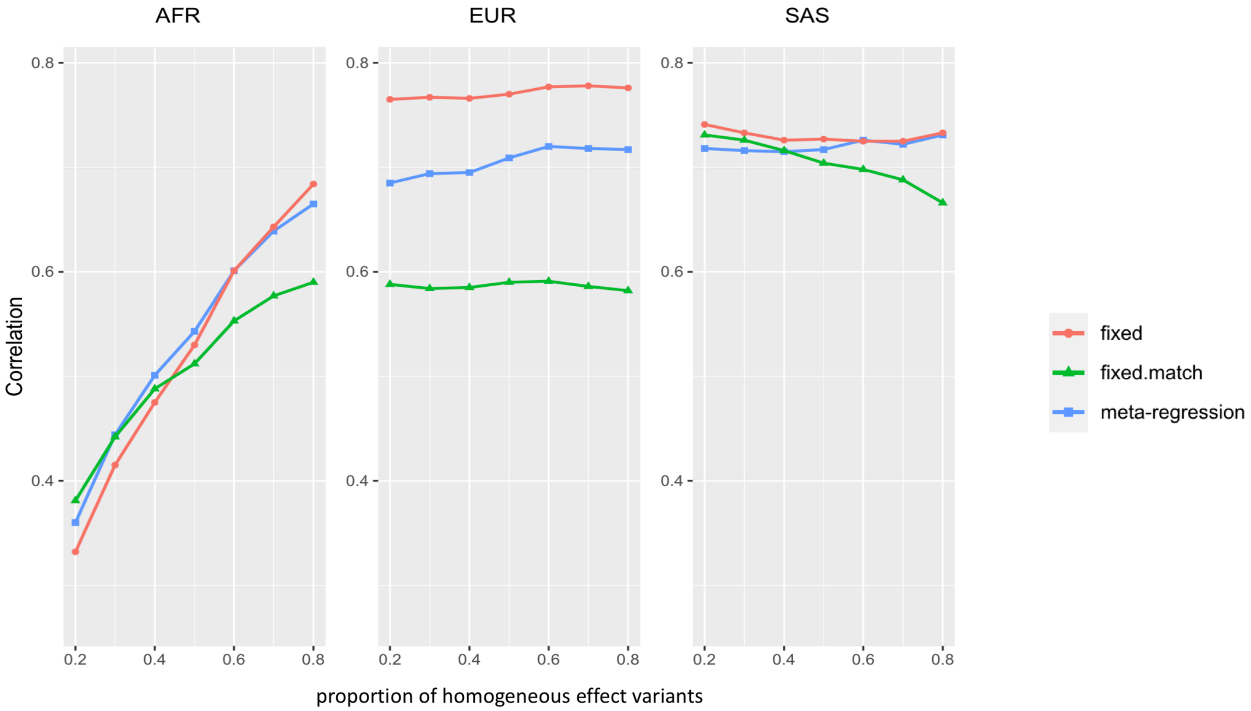
We observed that in heterogeneous Eurasian effect scenario, our meta-regression approach outperformed both the fixed effect model using studies from all ancestries as training data and the fixed effect model using studies whose ancestry match the target ancestry as training data when predicting genetic effects in African and South Asian populations in some ranges of proportions of homogeneous effect variants (Figure 3.1). In Figure 3.1A, we gave the median of Pearson correlation between estimated genetic effects and actual genetic effects of causal variants over cross-validations in excessive heterogeneous Eurasian effect scenario. When estimating genetic effects in the African and South Asian populations, our method has higher prediction accuracy than both fixed effect methods when at least 50% of causal variants have homogenous effects across ancestries. When predicting in the European population, fixed effect model using studies from all ancestries as training data performs the best. Our method is the second-best method, and the prediction accuracy increases as the proportion of homogenous effect variants increases. In the modest heterogeneous Eurasian effect scenario, our method has more advantages in estimating genetic effects in African populations. As shown in Figure 3.1B, our method still has the highest prediction accuracy when estimating genetic effects in African when 30%-60% of causal variants have homogenous effects across ancestries. As the proportion of homogeneous effect varies, the fixed effect model using studies from all ancestries as training data or fixed effect model using studies whose ancestry match the target ancestry as training data provides either the highest or the lowest prediction accuracy, but our meta-regression always performs close to the best method if it is not the best method. In homogenous Eurasian effect scenario, the fixed effect model using studies from all ancestries as training data has the highest prediction accuracy, and our method provides slightly lower prediction accuracy than it but much higher than the fixed effect model using studies whose ancestry match the target ancestry as training data (Figure 3.1C). We also investigated model performance in Afrasian scenario, where a proportion of causal variants only exist in the African and the South Asian populations. The results are displayed in Figure A.1 in Appendix A.

Figure 3.1: Median of the correlation of the estimated genetic effects with the true genetic effects in Eurasian scenario. We performed leave-one-out cross validation on 30 studies with 10 studies from each ancestry population. Each simulated study has 1,043,770 variants from presumably 2,000 individuals. We assessed the estimation accuracy of our meta-regression method versus fixed effect method using studies from all ancestries as training data (fixed) and fixed effect method using studies whose ancestry match the target ancestry as training data (fixed.match) in every ancestry across three simulation scenarios. (A) Estimation accuracy in excessive heterogeneous Eurasian effect scenario. A proportion of causal variants have homogeneous effects across all ancestries. For the rest of causal variants only European and South Asian populations have none-zero effects. The effect size in South Asian is triple as much as European. (B) Estimation accuracy in modest heterogeneous Eurasian effect scenario. Similar to the modest heterogeneous Eurasian effect, but the effect sizes for the ancestry specific causal variants in South Asian is twice as much as European. (C) Estimation accuracy in homogenous Eurasian effect scenario. A proportion of causal variants have homogeneous effects across all ancestries. For the rest of the ancestry specific causal variants, only European and South Asian population has non-zero identical genetic effects.

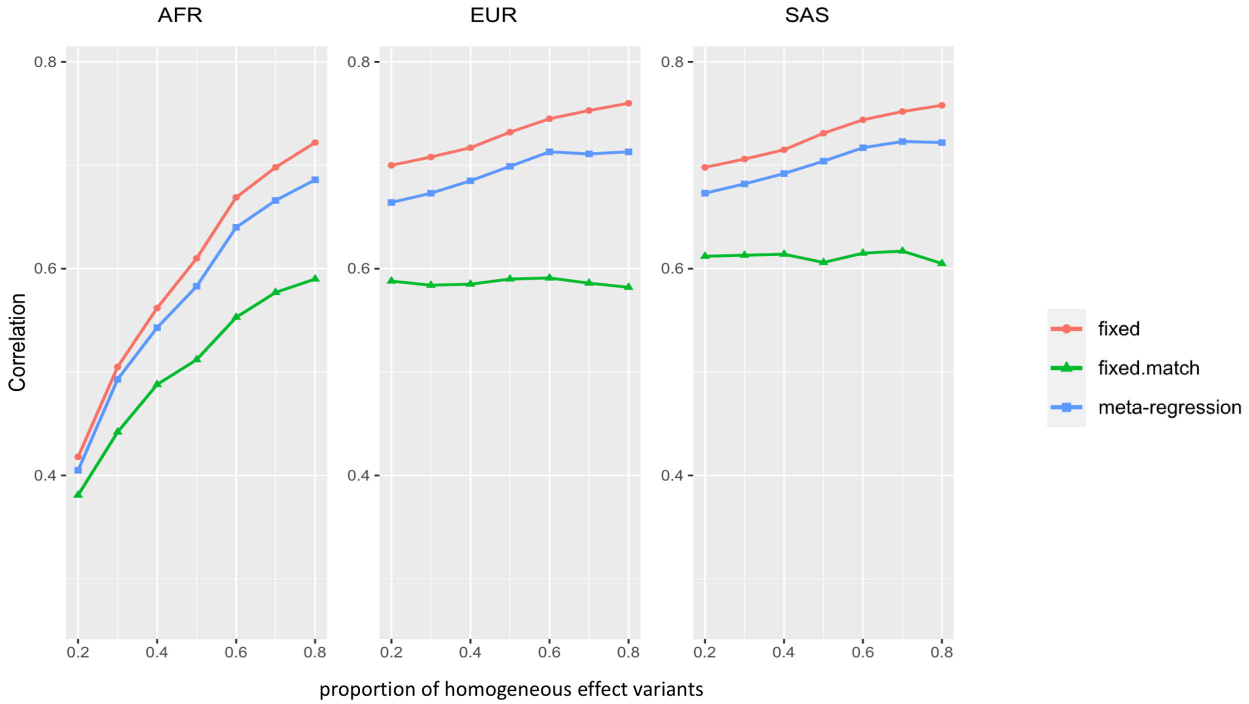
A.



B.


AFR
EUR
SAS
Correlation
0.8
0.6
0.4
0.2
0.4
0.6
0.8
proportion of homogeneous effect variants
fixed
fixed.match
meta-regression

C.


AFR
EUR
SAS
Correlation
0.8
0.6
0.4
0.2
0.4
0.6
0.8
proportion of homogeneous effect variants
fixed
fixed.match
meta-regression

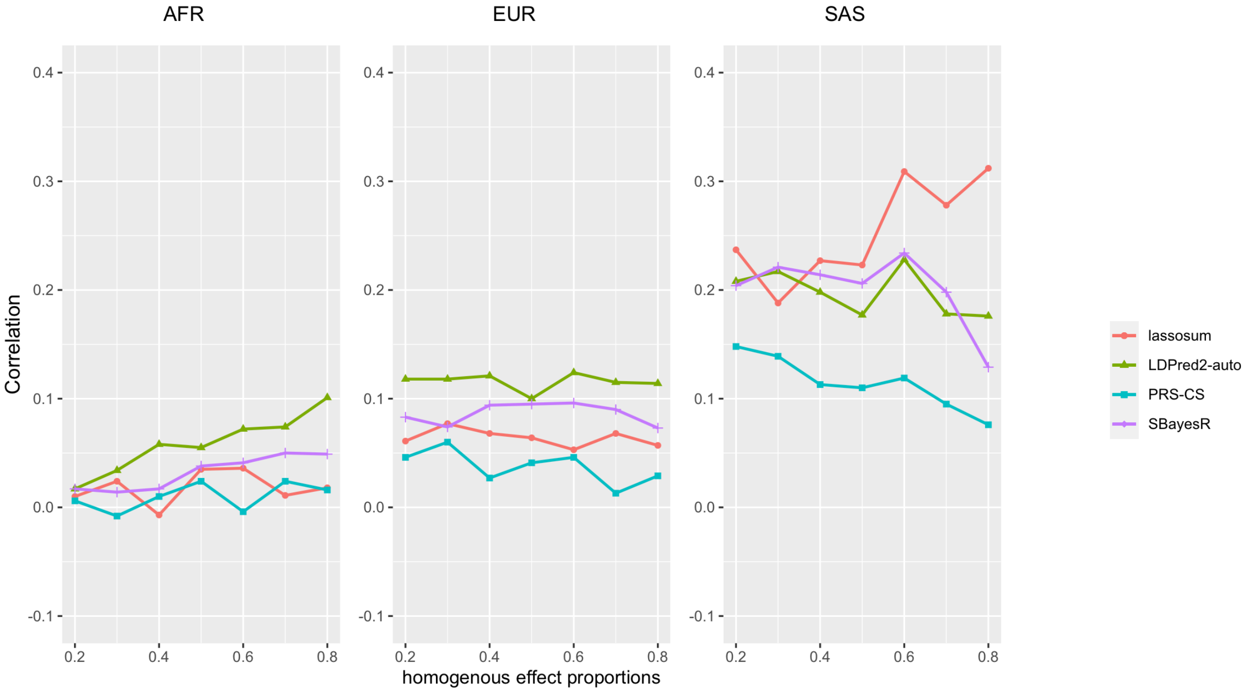
II. PRS Prediction

To find the most suitable PRS method to integrate with our meta-regression method, we investigated the performance of existing PRS methods using genetic effect estimates from our meta-regression method and two fixed effect methods in the excessive heterogeneous Eurasian effect scenario. We observe that the prediction accuracy in South Asian is higher than in African and European populations across all methods. It is consistent with what we observe from genetic effects estimation. It suggests the performance of PRS is associated with the quality of input genetic effects estimates. Lassosum tends to perform the best among all PRS methods in general. PRS-CS method gives the worst performence overall. Lassosum using genetic effects estimates from our meta-regression method as input provides the highest accuracy in a few situations. For example, when 20%, 60% or 80% of causal variants have homogenous effects across ancestries, the correlations between PRS.lassosum-meta-regression and true phenotypes in South Asian are 0.237, 0.309 and 0.312 respectively, whereas PRS.lassosum-fixed gives 0.236, 0.258 and 0.237 respectively, and PRS.lassosum-fixed.match gives 0.236, 0.298, 0.283 respectively. If we use LDPred2-auto to calculate PRS, PRS.LDPred2-auto-meta-regression has the highest prediction accuracy compared to PRS.LDPred2-auto-fixed and PRS.LDPred2-auto-fixed.match across all proportions of causal variants when predicting in South Asian population (Figure 3.2A-C).

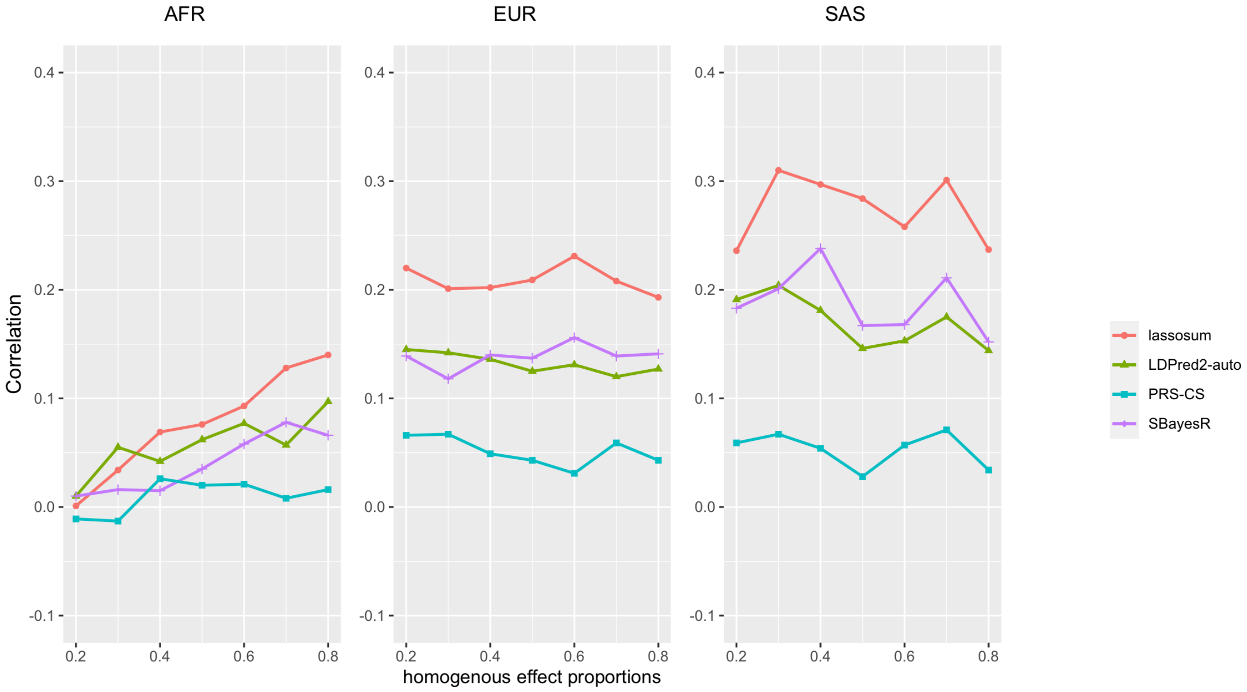
Multi-ethnic PRS calculated with simulated genetic effects gives lower prediction accuracy, especially in the South Asian population (Figure 3.2D). It shows the advantage of using estimated genetic effects over directly using the simulated genetic effects as input of PRS algorithms.

Figure 3.2: Median of the correlation of the PRS with the phenotype. We used estimated genetic effects as input for PRS algorithms lassosum, LDPred2-auto, SBayesR, and PRS-CS or directly used simulated genetic effects as input for PRS algorithm multi-ethnic PRS. We calculated the correlation of the estimated PRS with the phenotypes in 30 studies, 10 studies from each ancestry population. Each study has 1,043,770 variants from 1,000 individuals sampled from specific estimated ancestry groups in UK Biobank accordingly. (A) Prediction accuracy evaluation for PRS methods using estimated genetic effects from our meta-regression method as input. (B) Prediction accuracy evaluation for PRS methods using estimated genetic effects from the fixed method as input. (C) Prediction accuracy evaluation for PRS methods using estimated genetic effects from the fixed.match method as input. (D) Prediction accuracy evaluation for multi-ethnic PRS method using simulated genetic effects as input.

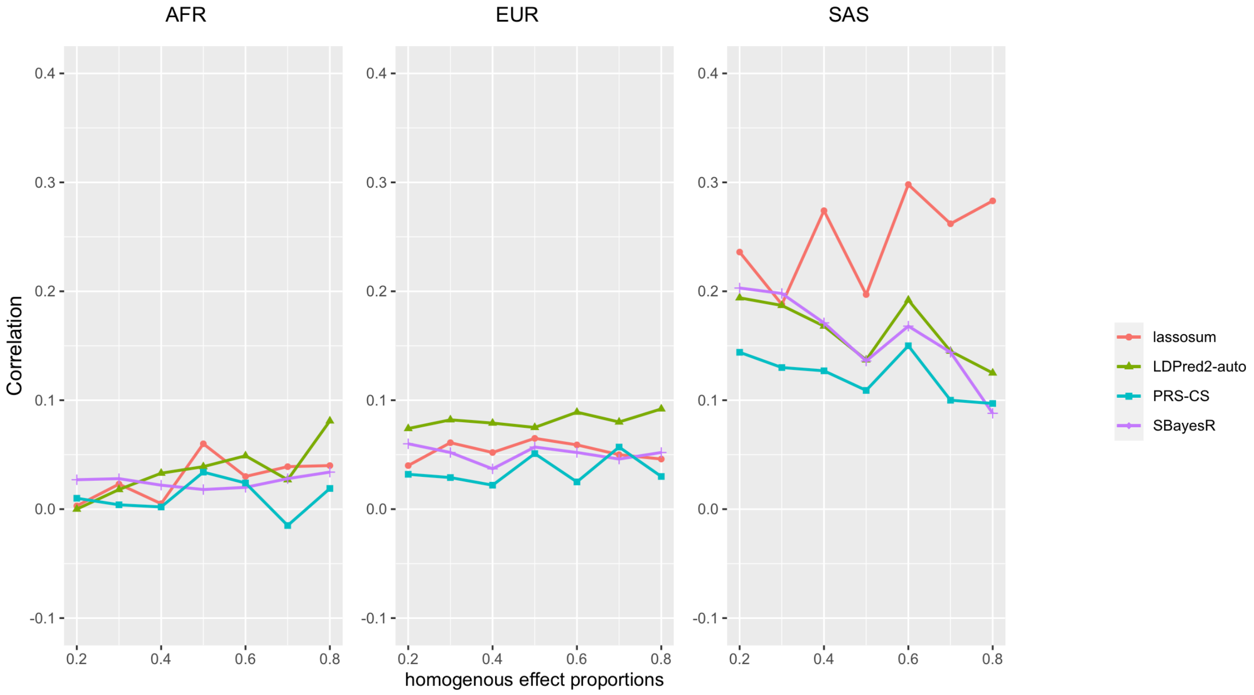
A. PRS-meta-regression method



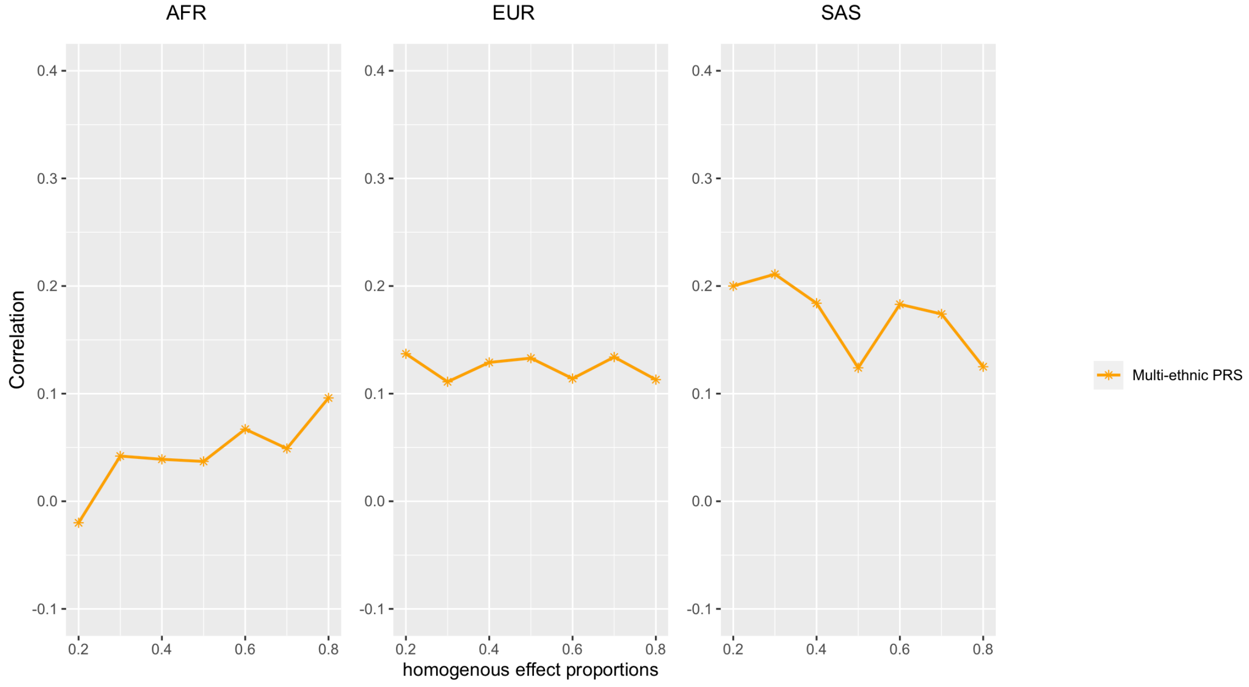
B. PRS-fixed method



C. PRS-fixed.match method



D. multi-ethnic PRS



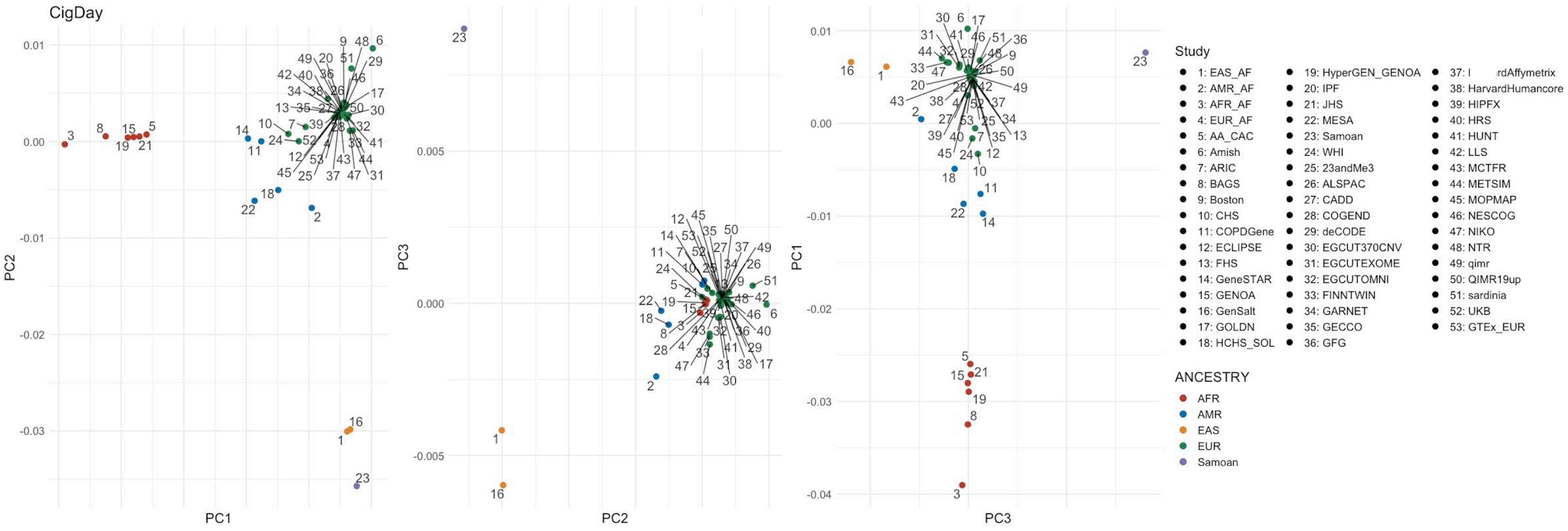
3.4 Real Data Application

To assess methodology in real data, we applied our meta-regression method on GSCAN. We used the summary statistics from all the studies excluding UK Biobank (total of 47 studies) from GSCAN to train our model and estimated the genetic effects for the whole genome of 3,483,748 genetic variants. We then predicted the PRS for smoking related phenotypes CigDay (UK Biobank fields: 2887 (number of cigarettes previously smoked daily), 3456 (number of cigarettes currently smoked daily (current cigarette smokers)), and 6183 (number of cigarettes previously smoked daily (current cigar/pipe smokers))), and AgeInit (UK Biobank fields: 3436 (age started smoking in current smokers) and 2867 (age started smoking in former smokers)) on individuals in UK Biobank with the estimated genetic effects. Lassosum was used for calculating PRS. We picked lassosum as our final PRS calculating method because most studies in GSCAN have restricted access to individual-level data; therefore, only summary statistics are accessible, and lassosum (when using the pseudovalidation option) does not require individual level data, such as genotypes or phenotypes to build the model. Another reason we decided to use lassosum is that in our simulation studies, lassosum shows the overall best performance in most circumstances. Lassosum is also easy to implement and requires substantially less computation time to process compared to other methods. This is important because there are over 3 million genetic variants in our real data application, whereas in the simulation studies we restricted the number of genetic variants to around 1 million. This dramatically increased number of variants made it challenging to calculate PRS with other methods such as PRS-CS.

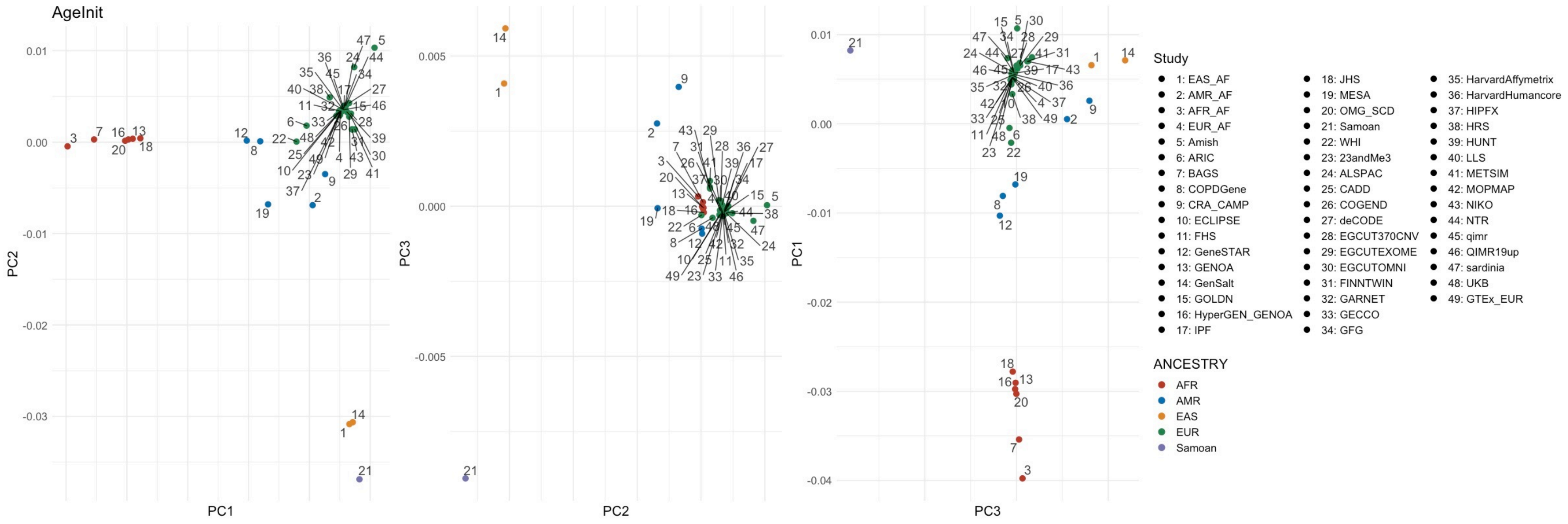
Figure 3.3 shows the first three PCs of genetic variations in allele frequencies separating ancestry groups from GSCAN studies. Figure 3A separates the GSCAN studies which have CigDay measured. Figure 3B separates the GSCAN studies which have AgeInit measured. Studies that are closer on plots are more likely from the same ancestry population. We also plotted the PCs from the 1000 Genome Projects as reference populations. GSCAN studies are labeled with their closest ancestry based on the Euclidean distance to reference population from 1000 Genome Projects. We observed that GSCAN studies' labeled ancestries are consistent with their studies' recruitment descriptions. It suggests the top 3 PCs are sufficient to separate ancestry groups of GSCAN studies. The estimated ancestries from CigDay studies are identical to the estimated ancestries from AgeInit studies for the studies which have both CigDay and AgeInit measured. The majority of studies are from the European population, but there are still 5 studies from the African population, 1 study from the East Asian population, and 4 studies from the Native American population.

Figure 3.3: PCs of genetic variations in allele frequencies separating ancestry groups from GSCAN studies and reference populations from 1000 Genomes Project. We identified 5 ancestry groups as African ancestry (AFR), Native American ancestry (AMR), East Asian ancestry (EAS), European ancestry (EUR), and Samoan, the majority of whom are isolated islanders. (A) The first 3 PCs of genetic variations from GSCAN studies which have CigDay measured. (B) The first 3 PCs of genetic variations from GSCAN studies which have AgeInit measured.

A.


CigDay
PC1
PC2
PC3
0.01
0.00
-0.01
-0.02
-0.03
-0.04
0.005
0.000
-0.005
Study
1: EAS_AF
2: AMR_AF
3: AFR_AF
4: EUR_AF
5: AA_CAC
6: Amish
7: ARIC
8: BAGS
9: Boston
10: CHS
11: COPDGene
12: ECLIPSE
13: FHS
14: GeneSTAR
15: GENOA
16: GenSalt
17: GOLDN
18: HCHS_SOL
19: HyperGEN_GENOA
20: IPF
21: JHS
22: MESA
23: Samoan
24: WHI
25: 23andMe3
26: ALSPAC
27: CADD
28: COGEND
29: deCODE
30: EGCUT370CNV
31: EGCUTEXOME
32: EGCUTOMNI
33: FINNTWIN
34: GARNET
35: GECCO
36: GFG
37: I rdAffymetrix
38: HarvardHumancore
39: HIPFX
40: HRS
41: HUNT
42: LLS
43: MCTFR
44: METSIM
45: MOPMAP
46: NESCOG
47: NIKO
48: NTR
49: qimr
50: QIMR19up
51: sardinia
52: UKB
53: GTEx_EUR
ANCESTRY
AFR
AMR
EAS
EUR
Samoan

B.


AgeInit
PC2
PC1
0.01
0.00
-0.01
-0.02
-0.03
PC3
PC2
0.005
0.000
-0.005
PC1
PC3
0.01
0.00
-0.01
-0.02
-0.03
-0.04
Study
● 1: EAS_AF
● 2: AMR_AF
● 3: AFR_AF
● 4: EUR_AF
● 5: Amish
● 6: ARIC
● 7: BAGS
● 8: COPDGene
● 9: CRA_CAMP
● 10: ECLIPSE
● 11: FHS
● 12: GeneSTAR
● 13: GENOA
● 14: GenSalt
● 15: GOLDN
● 16: HyperGEN_GENOA
● 17: IPF
● 18: JHS
● 19: MESA
● 20: OMG_SCD
● 21: Samoan
● 22: WHI
● 23: 23andMe3
● 24: ALSPAC
● 25: CADD
● 26: COGEND
● 27: deCODE
● 28: EGCUT370CNV
● 29: EGCUTEXOME
● 30: EGCUTOMNI
● 31: FINNTWIN
● 32: GARNET
● 33: GECCO
● 34: GFG
● 35: HarvardAffymetrix
● 36: HarvardHumancore
● 37: HIPFX
● 38: HRS
● 39: HUNT
● 40: LLS
● 41: METSIM
● 42: MOPMAP
● 43: NIKO
● 44: NTR
● 45: qimr
● 46: QIMR19up
● 47: sardinia
● 48: UKB
● 49: GTEx_EUR
ANCESTRY
● AFR
● AMR
● EAS
● EUR
● Samoan

The predicted PRSs for CigDay in ever smokers, using our meta-regression method integrated with lassosum, are normally distributed with means of 2.541 in the African population, -0.062 in the European population, and 0.617 in the South Asian population (Figure 3.4A). If using genetic effects estimated from fixed effect methods and integrating with lassosum, the predicted PRSs for CigDay are also normally distributed but are more centered around 0 (Figure 3.4B-C). We gained extra prediction accuracy from inputting lassosum with estimated genetic effects from our meta-regression method in the South Asian population: compared with PRSs using the genetic effects estimated from fixed effect method using studies from all ancestries and integrating with lassosum (r=0.014), and PRSs using the genetic effects estimated from fixed effect method using studies from European ancestries (since the majority of UK Biobank are from European population) and integrating with lassosum (r=0.031), our method (r=0.054) attained 286% and 74% improvement respectively in predicting PRS of CigDay (Figure 3.5).

Figure 3.4: PRS distributions for CigDay of smokers from UK Biobank. We calculated PRSs for smokers from African (N=450), European (N=55,715), and South Asian (N=215) populations in UK Biobank. (A) Distributions of PRSs calculated by using our meta-regression method integrated with lassosum. (B) Distributions of PRSs calculated by using the genetic effects estimated with fixed effect method using studies from all ancestries and integrated with lassosum. (C) Distributions of PRSs calculated using the genetic effects estimated with fixed effect method using European ancestries studies and combined with lassosum.

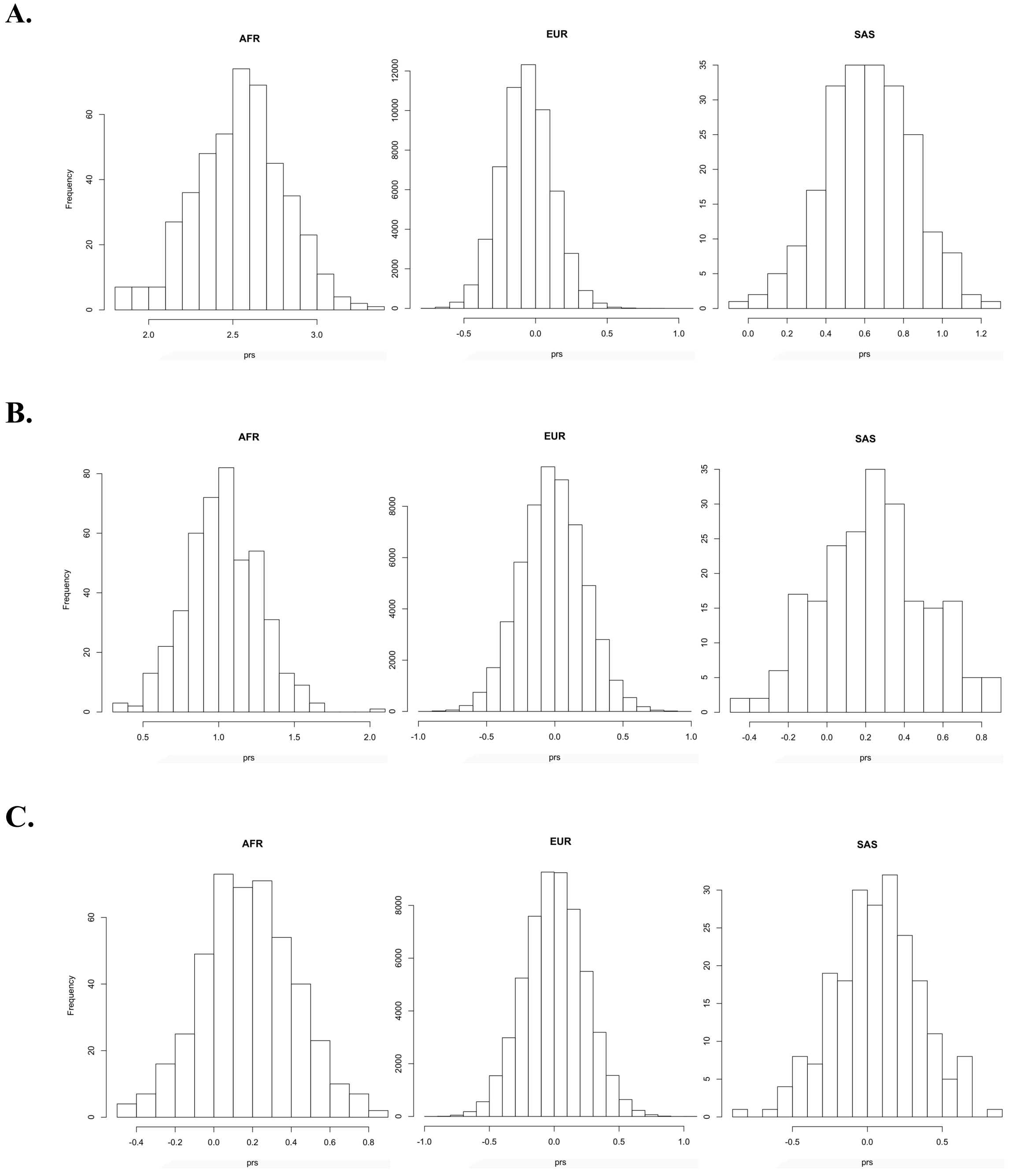
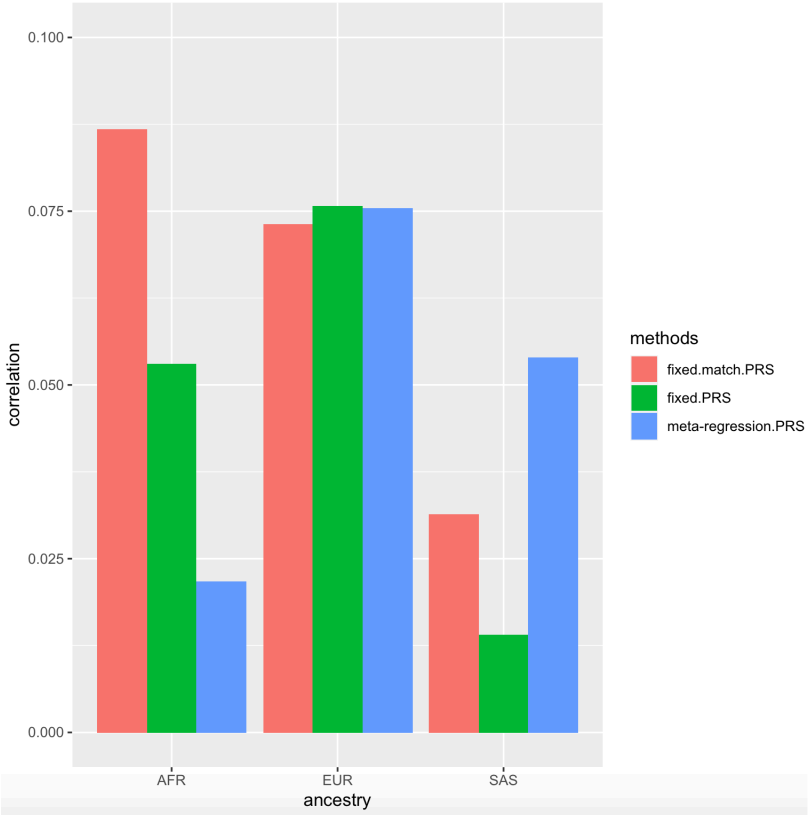


Figure 3.5: Prediction accuracy comparison for CigDay across PRSs calculated by using our meta-regression method integrated with lassosum, PRSs calculated by using the genetic effects estimated with fixed effect method using studies from all ancestries and integrating with lassosum, and PRSs calculated by using the genetic effects estimated with fixed effect method using European ancestries studies and integrating with lassosum. We calculated the correlation of PRSs of CigDay with the reported CigDay for smokers from African (N=450), European (N=55,715), and South Asian (N=215) populations in UK Biobank.



For another smoking related phenotype, AgeInit, the predicted PRSs for AgeInit in ever smokers are normally distributed as well (Figure 3.6). The prediction accuracy of PRSs for AgeInit is in general lower than that in CigDay for African and European populations. LikeCigDay, our meta-regression method shows the greatest strength in predicting PRS in the South Asian population in UK Biobank (Figure 3.7).

Figure 3.6: PRS distributions for AgeInit of smokers from UK Biobank. We calculated PRSs for smokers from African (N=460), European (N=58,285), and South Asian (N=217) populations in UK Biobank. (A) Distributions of PRSs calculated by using our meta-regression method integrated with lassosum. (B) Distributions of PRSs calculated by using the genetic effects estimated with fixed effect method using studies from all ancestries and integrated with lassosum. (C) Distributions of PRSs calculated by using the genetic effects estimated with fixed effect method using European ancestries studies and integrated with lassosum.

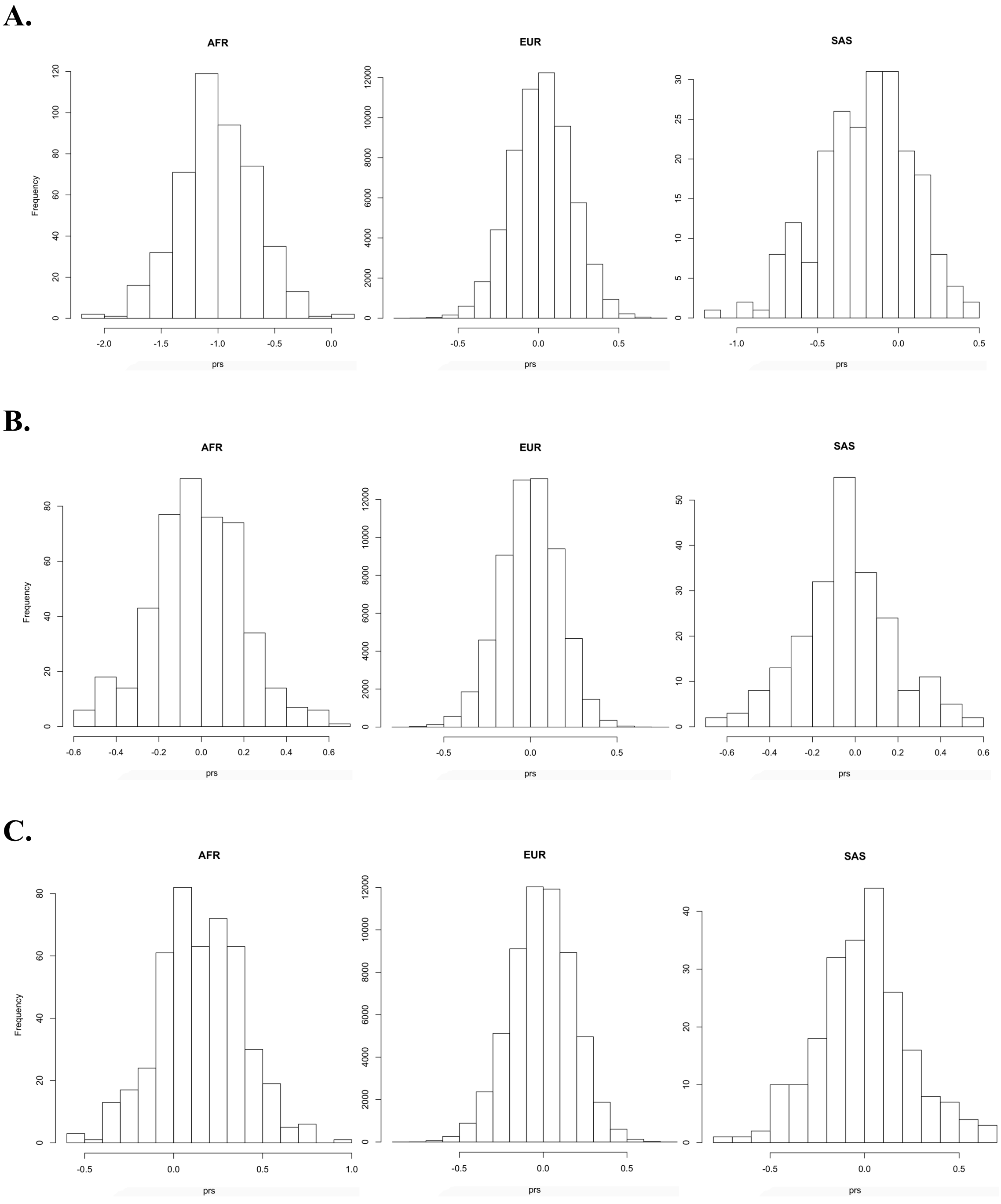
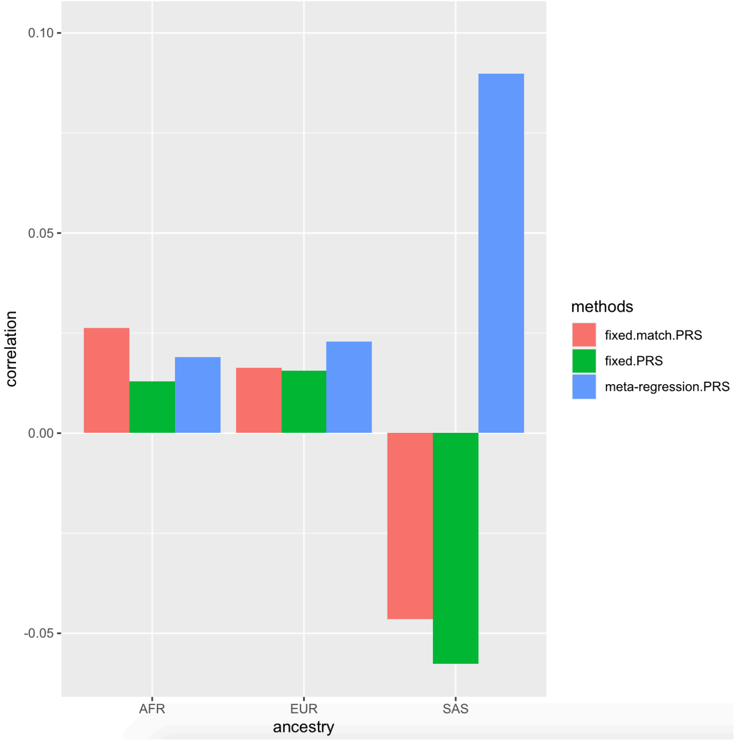


Figure 3.7: Prediction accuracy comparison for AgeInit across PRSs calculated by using our meta-regression method integrated with lassosum, PRSs calculated by using the genetic effects estimated with fixed effect method using studies from all ancestries and integrated with lassosum, and PRSs calculated by using the genetic effects estimated with fixed effect method using European ancestries studies and integrated with lassosum. We calculated the correlation of PRSs of AgeInit with the reported AgeInit for smokers from African (N=460), European (N=58,285), and South Asian (N=217) populations in UK Biobank.



To investigate the relationship between CigDay and AgeInit, we calculated the correlation between predicted PRSs for CigDay and predicted PRSs for AgeInit (Table 3.1). Overall, we noted that the correlation between PRSs for CigDay and AgeInit is negative, which indicates that smokers who start to smoke early tend to consume more cigarettes per day than those who start to smoke at a later age. It suggests people who have more severe nicotine dependence are more likely to initiate smoking earlier and become heavy smokers.

Table 3.1: Correlation between observed value for CigDay and observed value for AgeInit, and correlation between predicted PRSs for CigDay and predicted PRSs for AgeInit. For smokers who have both CigDay and AgeInit recorded, we calculated the correlation between observed CigDay and observed AgeInit, and the correlation between predicted PRSs for CigDay and predicted PRSs for AgeInit across different methods.

	observed values	meta-regression.PRS	fixed.PRS	fixed.match.PRS
All population in UKB (N=56,006)	-0.147	-0.397	-0.076	-0.077
African in UKB (N=442)	-0.072	-0.134	0.025	-0.021
European in UKB (N=55,354)	-0.146	-0.096	-0.079	-0.082
South Asian in UKB (N=210)	-0.164	-0.110	-0.218	-0.178

3.5 Discussion

All existing PRS methods, to the best of our knowledge, jointly model genetic variants across the genome using the estimated genetic effects directly from GWAS. We proposed a conceptually different approach for PRSs calculation which first uses meta-regression across ancestry studies to get the estimates of genetic effects with higher accuracy and then feed these better estimated genetic effects into existing PRS methods. Our meta-regression method gains extra prediction accuracy when there are GWAS results from multiple ancestries available. We showed through extensive simulation studies that our meta-regression method outperformed fixed effect methods in prediction accuracy when a fraction of causal variants for a trait are shared across ancestries with heterogeneous variant effects.

Our meta-regression method enables partitioning heterogeneity in variant effects between GWAS which is correlated with the ancestries by modeling the PCs of genetic variation. For variants whose effects are relatively homogeneous across ancestries, our method should still accommodate the model by picking the number of PCs adjusted in the model based on the smallest BIC value. BIC maximizes the likelihood while penalizing the number of parameters in the model. We also implemented minimal AIC and minimal p-value criteria, i.e., picking the number of PCs in the model based on the smallest AIC value or the smallest p-value. Overall, the minimal BIC criterion provides the highest prediction accuracy among all criteria (data not shown). Therefore, we decided to adhere to the minimal BIC criterion for all the analyses in this work.

In the simulation studies, we mainly focused on the heterogeneity scenario "Eurasian" where a proportion of causal variants only present in the European population and the Asian population, and we also differed the degree of effect sizes heterogeneity between European and Asian populations to the excessive heterogeneous scenario and the modest heterogeneous scenario. These scenarios incorporated heterogeneity between ancestry groups that are close to the real world. A previous study investigated the trans-ethnic genetic correlation in 31 complex diseases and complex traits, including body mass index, type 2 diabetes, rheumatoid arthritis, and schizophrenia using large studies in European and East Asian populations from Biobank Japan, UK Biobank, and CONVERGE consortium [39]. They found that the squared trans-ethnic genetic correlations were significantly depleted or enriched in functionally vital regions. They concluded that the causal effect sizes are population-specific for a wide range of diseases and traits. Unfortunately, our meta-regression method is not universally optimal across the broad range of simulation scenarios for heterogeneity in every ancestry population. For example, the fixed method using all ancestry studies as training data provides the highest prediction accuracy when predicting genetic effects in the European population. It may be because, in our simulation settings, the causal variants effect sizes in the European population are between the causal variants effect sizes in the African population and the Asian population. Mathematically as an averaging effect, the fixed effect method using all ancestry studies as training data would outperform both the meta-regression method and the fixed effect method using only European studies as training data. Our meta-regression method focuses on partitioning the variant effects on the top PCs of genetic variations, which distinguish populations from different ancestries. It shows higher prediction accuracy in the African or Asian populations which have more divergent variant effect sizes in our settings. We also noticed that our method almost consistently outperforms the fixed effect method using only European studies as training data. Our meta-regression method utilizes the information from all the studies, and the increased sample sizes would help with the prediction accuracy. The only scenario that fixed effect method using only target ancestry (except for European as target ancestry) studies as training data outperforms our meta-regression method is when less than 30% of causal variants are homogenous across ancestries.

When applying our meta-regression method to UK Biobank dataset, we observed that the PRS for CigDay is generally more accurate than the PRS for AgeInit. This is reasonable because the amount of cigarette consumption daily can reflect an individual's dependence on nicotine from genetic prospects well especially after an individual becomes a regular smoker, and it is consistent with other literature which found tobacco consumption PRS can be a good genetic surrogate for nicotine dependence. On the other hand, the age of smoking initiation is also associated with one's nicotine dependence. Still, compared with amount of daily cigarette consumptions, age of smoking initiation may be attributable more to the environment.

Our work still leaves several potential improvements for future exploration. First, since we modeled the genetic effects in a linear regression framework, it is relatively simple to include more factors that potentially contribute to the heterogeneity in variant effects across studies. For example, we can adjust for environmental factors in predicting the PRS for tobacco usages, such as geographic location, socioeconomic status, and educational attainment, which all showed significant associations with tobacco usage. Since tobacco usage showed a substantially difference in males and females, adding sex information can also potentially increase the prediction accuracy. Second, our current method restricted to modeling the genetic effects for continuous or ordering phenotypes, extending our meta-regression to categorical phenotypes using logistic regression can be addressed in the future work and would facilitate the risk prediction for other categorical tobacco usage related phenotypes such as ever been a regular smoker and ever attempt to quit smoking. Third, in this work, we focused on the variants that all GWAS have in common with allele frequency greater than 0.01. It is possible that some variants, although contribute to the phenotype of interest, got filtered out since they were not genotyped or their allele frequencies were less than 0.01 in some of the studies. For future improvement, we can keep those variants and estimate their effect sizes using the genotyped studies and have the allele frequency greater than the 0.01 threshold. Fourth, we limited our PRS calculation to lassosum in the real data application in UK Biobank. One of the main reasons was that lassosum was much faster than other PRS methods when calculating PRS using millions of variants. For example, in our simulation studies, it cost LDPred2 more than a day to calculate PRS for a single study, whereas lassosum only took a few hours. One of the bottlenecks for computational efficiency for LDPred2 is that it relies on the *snp_readBed* function from "bigsnpr" package to generate an rds file for the genotype data that can be later read into R for LD calculation. One solution to increase the computation efficiency is that instead of using *snp_readBed* function, we can implement the *readPlinkToMatrixByIndex* function from our "seqminer2" package. As we showed before, seqminer2 can read ultra large biobank scale data into R and it is faster than most state-of-art tools, implementing seqminer2 with PRS tools embedded with less efficient sequencing reading tools can potentially magnitude improve their computation efficiency.

In conclusion, our meta-regression method provides a way to estimate genetic variant effects with higher accuracy and further improve the PRS prediction in multi-ethnic studies. Although the PRS prediction accuracy is still not high enough to make it into clinical practice, we believe that with the increasing number of available GWAS from diverse populations, the PRS accuracy will keep improving and provides insights to the pathology of diseases and ultimately contribute to the personalized medicine.

3.6 Reference

1. Collaborators, G.B.D.R.F., *Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1223-1249.
2. Collaborators, G.B.D.T., *Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019.* Lancet, 2021. **397**(10292): p. 2337-2360.
3. Collaborators, G.B.D.C.T., *Spatial, temporal, and demographic patterns in prevalence of chewing tobacco use in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019.* Lancet Public Health, 2021. **6**(7): p. e482-e499.
4. Warren, G.W., et al., *The 2014 Surgeon General's report: "The health consequences of smoking--50 years of progress": a paradigm shift in cancer care.* Cancer, 2014. **120**(13): p. 1914-6.
5. Vink, J.M., G. Willemsen, and D.I. Boomsma, *Heritability of smoking initiation and nicotine dependence.* Behav Genet, 2005. **35**(4): p. 397-406.
6. Carmelli, D., et al., *Genetic influence on smoking--a study of male twins.* N Engl J Med, 1992. **327**(12): p. 829-33.
7. Kaprio, J., M. Koskenvuo, and S. Sarna, *Cigarette smoking, use of alcohol, and leisure-time physical activity among same-sexed adult male twins.* Prog Clin Biol Res, 1981. **69 Pt C**: p. 37-46.
8. Liu, M., et al., *Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use.* Nat Genet, 2019. **51**(2): p. 237-244.
9. Abraham, G., et al., *Genomic prediction of coronary heart disease.* Eur Heart J, 2016. **37**(43): p. 3267-3278.
10. Elliott, J., et al., *Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease.* JAMA, 2020. **323**(7): p. 636-645.
11. Isgut, M., et al., *Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later.* Genome Med, 2021. **13**(1): p. 13.
12. Mars, N., et al., *The role of polygenic risk and susceptibility genes in breast cancer over the course of life.* Nat Commun, 2020. **11**(1): p. 6383.
13. Kuchenbaecker, K.B., et al., *Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers.* J Natl Cancer Inst, 2017. **109**(7).
14. Lakeman, I.M.M., et al., *Addition of a 161-SNP polygenic risk score to family history-based risk prediction: impact on clinical management in non-BRCA1/2 breast cancer families.* J Med Genet, 2019. **56**(9): p. 581-589.
15. Hurson, A.N., et al., *Prospective evaluation of a breast-cancer risk model integrating classical risk factors and polygenic risk in 15 cohorts from six countries.* Int J Epidemiol, 2021.

16. Maas, P., et al., *Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States.* JAMA Oncol, 2016. **2**(10): p. 1295-1302.
17. Sanchez-Roige, S., et al., *Alcohol and cigarette smoking consumption as genetic proxies for alcohol misuse and nicotine dependence.* Drug Alcohol Depend, 2021. **221**: p. 108612.
18. Vrieze, S.I., M. McGue, and W.G. Iacono, *The interplay of genes and adolescent development in substance use disorders: leveraging findings from GWAS meta-analyses to test developmental hypotheses about nicotine consumption.* Hum Genet, 2012. **131**(6): p. 791-801.
19. Chang, L.H., et al., *Association between polygenic risk for tobacco or alcohol consumption and liability to licit and illicit substance use in young Australian adults.* Drug Alcohol Depend, 2019. **197**: p. 271-279.
20. Mak, T.S.H., et al., *Polygenic scores via penalized regression on summary statistics.* Genet Epidemiol, 2017. **41**(6): p. 469-480.
21. Strimmer, K., *A unified approach to false discovery rate estimation.* BMC Bioinformatics, 2008. **9**: p. 303.
22. Prive, F., J. Arbel, and B.J. Vilhjalmsson, *LDpred2: better, faster, stronger.* Bioinformatics, 2020.
23. Lloyd-Jones, L.R., et al., *Improved polygenic prediction by Bayesian multiple regression on summary statistics.* Nat Commun, 2019. **10**(1): p. 5086.
24. Ge, T., et al., *Polygenic prediction via Bayesian regression and continuous shrinkage priors.* Nat Commun, 2019. **10**(1): p. 1776.
25. Inouye, M., et al., *Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention.* J Am Coll Cardiol, 2018. **72**(16): p. 1883-1893.
26. Marquez-Luna, C., et al., *Multiethnic polygenic risk scores improve risk prediction in diverse populations.* Genet Epidemiol, 2017. **41**(8): p. 811-823.
27. Marigorta, U.M. and A. Navarro, *High trans-ethnic replicability of GWAS results implies common causal variants.* PLoS Genet, 2013. **9**(6): p. e1003566.
28. Ntzani, E.E., et al., *Consistency of genome-wide associations across major ancestral groups.* Hum Genet, 2012. **131**(7): p. 1057-71.
29. Duncan, L., et al., *Analysis of polygenic risk score usage and performance in diverse human populations.* Nat Commun, 2019. **10**(1): p. 3328.
30. Martin, A.R., et al., *Clinical use of current polygenic risk scores may exacerbate health disparities.* Nat Genet, 2019. **51**(4): p. 584-591.
31. Kichaev, G. and B. Pasaniuc, *Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies.* Am J Hum Genet, 2015. **97**(2): p. 260-71.
32. Morris, A.P., *Transethnic meta-analysis of genomewide association studies.* Genet Epidemiol, 2011. **35**(8): p. 809-22.
33. Magi, R., et al., *Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution.* Hum Mol Genet, 2017. **26**(18): p. 3639-3650.
34. Brazel, D.M., et al., *Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use.* Biol Psychiatry, 2019. **85**(11): p. 946-955.

35. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): p. 1655-64.

36. Yang, L., et al., *Seqminer2: an efficient tool to query and retrieve genotypes for statistical genetics analyses from biobank scale sequence dataset.* Bioinformatics, 2020. **36**(19): p. 4951-4954.

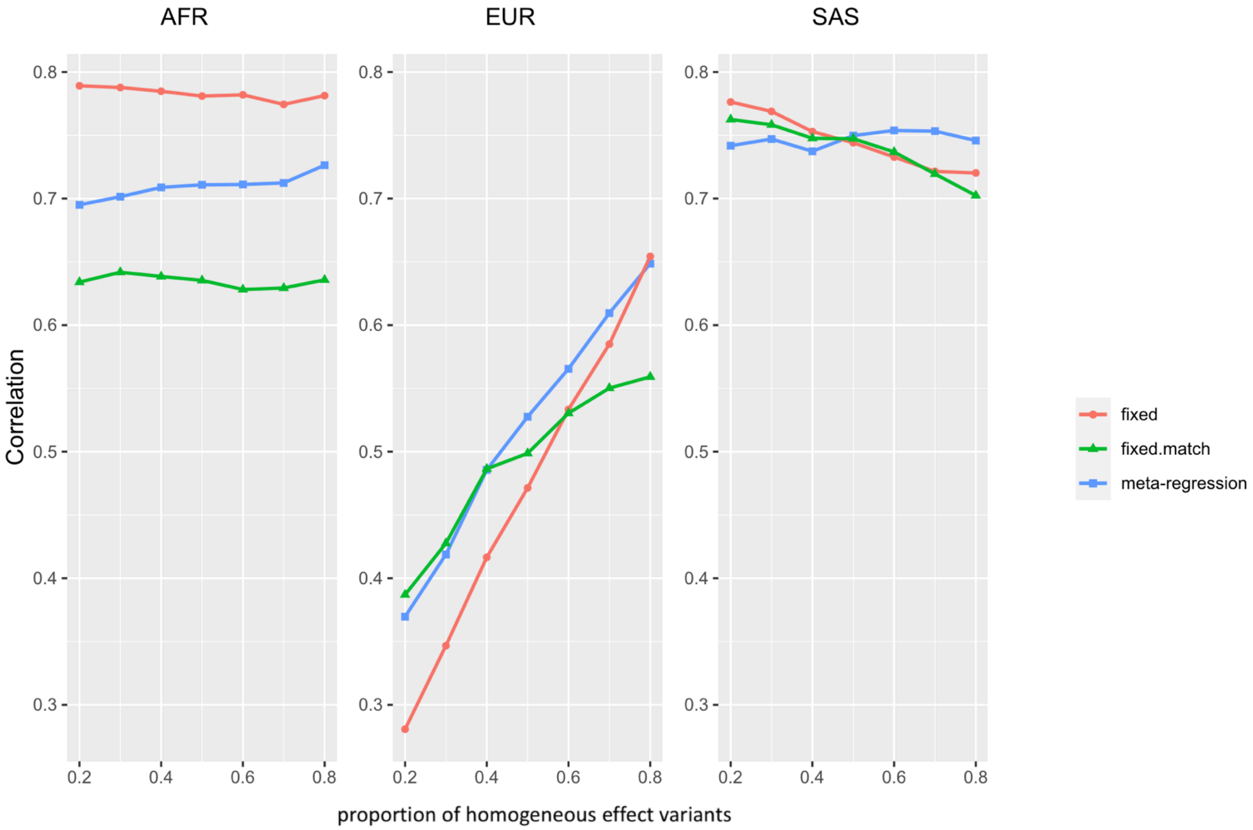
37. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.

38. Berisa, T. and J.K. Pickrell, *Approximately independent linkage disequilibrium blocks in human populations.* Bioinformatics, 2016. **32**(2): p. 283-5.

39. Shi, H., et al., *Population-specific causal disease effect sizes in functionally important regions impacted by selection.* Nat Commun, 2021. **12**(1): p. 1098.

Appendix A Chapter 3

Figure A.1: Median of the correlation of the estimated genetic effects with the true genetic effects in Afrasian scenario. We performed leave-one-out cross validation on 30 studies with 10 studies from each ancestry population. Each simulated study has 1,043,770 variants from presumably 2,000 individuals. We assessed the estimation accuracy of our meta-regression method versus fixed effect method using studies from all ancestries as training data (fixed) and fixed effect method using studies whose ancestry match the target ancestry as training data (fixed.match) in every ancestry. We evaluated in the scenario where a proportion of causal variants have homogeneous effects across all ancestries. For the rest of causal variants only the African and the South Asian populations have none-zero effects. The effect size in the South Asian is triple as much as the African.



VITA

Lina Yang

EDUCATION

Ph.D. in Biostatistics 2021
PENNSYLVANIA STATE UNIVERSITY, Hershey, PA, USA
Advisors: Prof. Dajiang Liu and Prof. Vernon Chinchilli

M.S. in Biostatistics 2017
DUKE UNIVERSITY, Durham NC, USA
Advisor: Prof. Terry Hyslop

M.S. in Physiology 2015
STATE UNIVERSITY OF NEW YORK UPSTATE MEDICAL UNIVERSITY, Syracuse, NY, USA

B.S. in Biological Science 2012
ZHEJIANG UNIVERSITY, Hangzhou, China

PUBLICATIONS

McGuire, D., Jiang, Y., Liu, M., Weissenkampen, J., Eckert, S., **Yang, L.**, …, Jiang, B., Li, Q. & Liu, DJ. Model-based assessment of replicability for genome-wide association meta-analysis, *Nature Communications* 2021

Masci, AM., White, S., Neely, B., Ardini-Polaske, M., Hill, CB., Misra, R., Aronow, B., Gaddis, N., **Yang, L.**, Wert, S., Palmer, S., Chan, C. & LungMAP Consortium. Ontology-guided Segmentation and Object Identification for Developmental Mouse Lung Immunofluorescent Images, *BMC Bioinformatics* 2021

Yang, L., Jiang, S., Jiang, B., Liu, D. J., & Zhan, X. Seqminer2: An Efficient Tool to Query and Retrieve Genotypes for Statistical Genetics Analyses from Biobank Scale Sequence Dataset, *Bioinformatics* 2020

Faraone, SV., Zhang-James, Y., Lloyd, d., James, ML., **Yang, L.**, & Richards, JB. Oral Methylphenidate Treatment of an Adolescent ADHD Rat Model Has Limited Effects on Cocaine-Conditioned Place Preference during Adulthood, J*ournal of Psychiatry and Brain Science* 2019

Yang, L., Faraone, SV. & Zhang-James, Y. Autism Spectrum Disorder Traits in *Slc9a9* Knock-out Mice, *Am J Med Genet B Neuropsychiatr Genet* 2016

Zhang-James, Y., Yang, Li., Middleton, F., **Yang, L.**, Patak, J. & Faraone, SV. Autism-related Behavioral Phenotypes in an Inbred Rat Model, *Behav Brain Res* 2014