THE UNIVERSITY OF CHICAGO

CHARACTERIZING GENE-ENVIRONMENT RELATIONSHIPS IN DIVERSE POPULATIONS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY

DAYANA DELGADO

CHICAGO, ILLINOIS

JUNE 2021

To my husband, for his infinite support and for always believing in me,
To my daughter, for motivating me each day and being the light of my entire world,

and

To my parents, for teaching me that a strong work ethic pays off and character matters, and for working relentlessly to give me access to opportunities we could only dream of,

and

To my little brother and sister, for unknowingly shaping me and always cheering me on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Several individuals had a profound impact on the work presented in this dissertation, and my development through graduate school. First, I want to thank my mentor, Dr. Brandon Pierce, for his unconditionally support and encouragement both professionally and personally. I feel immensely blessed to have had the honor to work with such a talented and kind individual.

I would also like to thank the members of my committee: Drs. Lin Chen (Chair), Habibul Ahsan, and Donald VanderGriend, for their immeasurable contributions, thoughtful feedback, and guidance. A special thanks to Dr. Chen, for her significant input and statistical guidance. Thank you to multiple co-authors and collaborators, especially Drs. Ana Navas-Acien, Margaret Karagas, and Marc Gillard. I want to thank the study participants from the cohorts and consortia analyzed, for making this dissertation possible. I also want to thank the research staff whose efforts contributed vastly to the work presented in this dissertation.

I want to thank the Department of Public Health Sciences students, faculty, staff, and members of the University of Chicago community. A special thanks to Michele Thompson, Emma Collier, and Ryan Carter, for their logistic support and patience in navigating this PhD training program. Another special thanks to the current and previous members of the Pierce/Chen Lab, for lending an ear and always providing invaluable feedback.

This research was supported by funding from the Initiative for Maximizing Student Development Fellowship, Susan G. Komen Graduate Training in Disparities Research program, and NIH Research Supplement to Promote Diversity in Health-Related Research (R35ES02837).

Finally, I thank my family and friends: my husband and daughter, my parents, siblings, the Faulkner family, my Abuelita, my wonderful cousin Sandra Silva, and my dear friends, Melissa Villalba and Nadia Alvarez, for the love, laughs, and unconditional support.

ABSTRACT

This dissertation examines gene-environment relationships in diverse populations around the globe and provides novel insight regarding genetic and environmental susceptibility to human disease. Using three arsenic-exposed cohorts from diverse ancestral and environmental backgrounds, we estimate the impact of rare, protein-coding variation in the arsenic methyltransferase (*AS3MT*) gene on arsenic metabolism efficiency (AME) and identify population-specific and shared causal rare variants. Because genetic variants affecting AME are expected to influence internal dose of arsenic, they can be used as instrumental variables (i.e. proxies) in order to assess the effect of arsenic dose/exposure on disease risk, an approach known as "Mendelian randomization" (MR). Here, we utilize genetic determinants of AME in a Bangladeshi cohort to obtain a MR-based estimate the effect of AME on hypertension risk, a relationship that remains an area of debate due to inconsistent findings in prior studies. Finally, to elucidate whether the effect of inherited prostate cancer (PCa) risk loci on biological mechanisms underlying PCa disparities differ between African American (AA) and European American (EA) men, we examine the effect of PCa risk loci on DNA methylation features of AA and EA benign and tumor PCa tissue.

CHAPTER 1

INTRODUCTION

1.1 Gene-environment relationships

Over the past decade, there has been a rapid advancement in our understanding of genetic influences on human disease. We have also become aware of the complexity in the factors contributing to individual's susceptibility to disease. Both genetic and environmental factors play key roles in the contribution to susceptibility, and in many instances, these two factors act together to produce a measurable biological effect (i.e. urine metabolites, blood biomarkers, DNA methylation, gene expression). That is to say, for most complex diseases genetic differences can cause an individual to respond differently to the same environmental exposure as another person. In this dissertation, we explore two important gene-environment relationships and provide strong evidence of the potential mechanisms underlying susceptibility.

1.2.1 Overview of arsenic exposure as global health hazard

Naturally occurring inorganic arsenic (iAs) in drinking water and foods (primarily through rice and grain products) affect >230 million individuals globally [1-3]. Chronic exposure to levels of iAs above the World Health Organization (WHO) safety standard for drinking water (>10µg/L) has been recognized to pose a significant risk of adverse outcomes across multiple organ systems [4]. While, dietary exposure to iAs is an emerging concern that has received less regulatory focus [5]. Arsenic in water is typically found in the inorganic forms, As^{III} (arsenite) or As^{V} (arsenate), with normal levels ranging around a few micrograms per liter [1]. In affected areas, levels of arsenic in water can be as high as thousands of micrograms per liter (IARC 2004).

Ongoing epidemiological studies show chronic exposure to inorganic arsenic (iAs) is associated with adverse health effects increased risk for skin lesions [6, 7], cardiovascular disease [8], diabetes [9], cognitive dysfunction [10, 11], adverse birth outcomes [12, 13], cancer [9], and overall mortality [14]. The most common manifestation of arsenic toxicity is skin lesions, which appear as hyperpigmentation [7]. Skin lesions have been reported in individuals exposed to >100µg/L arsenic concentration in drinking water, although some studies also report skin lesions in individuals exposed to <50 µg/L arsenic concentrations [15, 16].

1.2.2 Arsenic Metabolism

Inorganic arsenic enters the body in the form of iAs^{III} (trivalent arsenite) or iAs^{V} (pentavalent arsenate) and is eliminated in urine primarily as a dimethylated metabolite (DMA^{III} and DMA^{V}), though a smaller percentage is eliminated in the monomethylated form (MMA^{III} and MMA^{V}) and smaller quantity remains as iAs. Sequential reduction and methylation reactions are catalyzed by glutathione and arsenic (+3 oxidation state) methyltransferase (*AS3MT*), respectively. The AS3MT gene spans 11 exons across 23 kilobases in the 10q24.32 region, and its protein product is responsible for the methylation of iAs to MMA and MMA to DMA. Methylation of iAs is a detoxification mechanism that facilitates urinary excretion of arsenic DMA is considered the least harmful metabolite as it is more rapidly expelled from the body compared to MMA [6, 17, 18]. Arsenic metabolism efficiency (AME) is often represented by the percentage of arsenic species in urine that are DMA (DMA%).

1.2.3 Inherited genetic variation influences arsenic metabolism efficiency

Over the last decade, a number of studies have shown considerable inter-individual variation in AME [19-21] (due in part to underlying genetic differences), which in turn affects susceptibility to arsenic-related disease. Our group [22] and others [23] have identified common

variation in the 10q24.32 region (containing the *AS3MT* gene) that is associated with AME as well as arsenic-related outcomes, such as skin lesion risk. Pierce et al. [18, 24], described two independent association signals for DMA% in the 10q24.32 region (represented by lead SNPs rs9527 and rs11191527) using data from a Bangladeshi cohort. These SNPs were also associated with MMA% and skin lesion risk, indicating the 10q24.32 region as a key source of individual variation in AME. These genetic variants have been validated in other populations and have also shown an association with arsenic-induced cytogenetic damage (i.e. chromosomal aberrations and micronucleus formation) [25].

Other than *AS3MT*, only one other gene, Formiminotransferase Cyclodeaminase (*FTCD)*, has been shown to contain inherited genetic variation that affects AME. Through a exome-wide study assessing the association between non-synonymous, protein coding variation and AME, Pierce et al. reported a missense variant in exon 3 of the FTCD gene, rs61735836, whose minor allele (A) was associated with decreased DMA% and increased risk for skin lesions in Bangladeshi individuals [26]. FTCD is primarily expressed in liver, where most of arsenic metabolism takes place, and is responsible for catalyzing two consecutive final reactions that feed into one-carbon metabolism. The one-carbon cycle is a critical component of arsenic metabolism because is produces the methyl groups *AS3MT* uses to methylate arsenic from iAs to MMA and finally to DMA. Together, common SNPs in the *AS3MT* and *FTCD* genes account for ~10% of the variation in DMA% [26]. Outside of these regions, there is very little evidence of additional variation that influences AME.

Rare inherited genetic variation has received little attention as a potential contributor to AME. Recent familial aggregation studies report heritability estimates for DMA% as high as 63%, suggesting a strong familial component, which may be attributable to low-frequency

variants (in addition to the effects of common variation) [27, 28]. The established role of 10q24/*AS3MT* common variants in AME points to *AS3MT* as a potential source of such rare variants. Discovering such variants is critical for identifying individuals at a high-risk for arsenic toxicities and understanding the biological mechanisms underlying inter-individual differences in susceptibility to arsenic toxicity.

1.2.4 Arsenic exposure and hypertension risk

Hypertension (HTN) is a leading cause of morbidity and mortality worldwide, affecting 31.1% of the global population (~1.4 billion people) [29]. Risk factors for HTN are unhealthy diet (high salt intake) [30], physical inactivity [31], consumption of tobacco and alcohol [32, 33], genetic predisposition [34], psychosocial stress [35], and environmental exposures [36-39]. Among the environmental exposures, experimental studies suggest several mechanisms by which iAs or its metabolites could impact HTN risk [40-43]. These studies indicate that iAs can induce HTN by the promotion of endothelial dysfunction [41, 43], oxidative stress [40], and inflammation [42]. *In vivo* and *in vitro* studies have shown that arsenic induces systemic HTN by exacerbating the vasoconstrictor, angiotensin II (AngII) [44]. The induction of AngII promotes changes in endothelial function, vascular reactivity, tissue remodeling and oxidative stress [45-48]. Additional evidence from mice studies show that chronic exposure to iAs increases left ventricular mass with a relative wall thickness that is suggestive of concentric hypertrophy, which is often associated with chronic HTN [49].

However, epidemiologic studies conducted in arsenic-endemic areas report inconsistent findings [9, 37]. While, several studies provide strong evidence for the association between arsenic exposure and increased hypertension risk [37, 50-52], some studies have also reported a lack of association at moderate to high levels of exposure [53] and low levels of exposure [54].

Additionally, the association between AME and HTN risk has only been analyzed by a handful of studies and remains inconclusive [9, 55]. A systematic review of five observational studies that assessed the association between AME and HTN risk [9] found that four of the five reported no association, and one reported a positive association between AME (i.e., DMA%) and HTN risk [56], suggesting individuals with higher proportions of DMA% have increased risk for hypertension. While, a recent MR study among individuals with presumed exposure to arsenic based on high rice consumption found that genetically predicted efficient arsenic metabolizers showed decreased risk for HTN [55].

Prior studies of the relationship of arsenic exposure and AME with HTN risk were limited in a number of ways, including lack of individual-level arsenic metabolite data, lack of a cohort with a historical exposure to arsenic, low exposure to arsenic rendering decreased power to detect associations, small sample size, and lack of prospective blood pressure data. A comprehensive analysis in a large cohort with a historic exposure to arsenic is needed in order to establish the relationship of arsenic exposure a metabolism with HTN risk.

1.3.1 Overview of prostate cancer disparities in the U.S.

Prostate cancer (PCa) is the second most common cancer and cause of cancer death among men. African American (AA) men are disproportionately affected by PCa compared to any other race and ethnicity [57]. Among AA men, the incidence and mortality rates of PCa are 1.7 times higher [58-60] and 2-3 times higher, respectively, compared to European American (EA) men [61, 62]. This growing disparity is attributed to a complex combination of factors including social, environmental, and genetic [59].

There are fundamental etiological differences in PCa between AA and EA men. For instance, AA men have higher testosterone levels and present with higher levels of PSA at

diagnosis compared to EA men [63, 64]. AA are diagnosed at a younger age, with more advanced stage and larger tumors compared to EA men [65-67]. After prostatectomy, AA men are more likely to present with higher positive surgical margins, pathologic upgrading, and upstaging [68]. In randomized clinical trials of AA and EA men with the same tumor stage and identical treatment and follow-up regimens, AAs experienced worse survival outcomes than EAs [69]. Tumor biology differences are uncaptured by known prognostic features (i.e. Gleason Score and stage) and are suspected to contribute to the difference in clinical outcomes between these two populations.

1.3.2 Contribution of inherited genetic variation to prostate cancer risk

PCa is one of most heritable common cancers with heritability estimates from twin studies as high as 60% [70]. To date, genome-wide association studies (GWAS) have identified more than 260 common risk alleles associated with PCa [71]. These common SNPs account for about 42.6% and 43.2% of the familial relative risk in EA and AA men, respectively. However, a majority of PCa GWAS have been conducted among EA men, while AA men are profoundly under-represented.

Studies of AA men report that risk alleles identified in EA men have weaker effects or are absent in AA men [72-77]. Similarly, several studies also report risk alleles unique to chromosomes of African ancestry (e.g. 17q21 and 8q24) are absent among EA men [78, 79]. Few studies also report that among 140 PCa-risk SNPs, about half are replicated in AA men and >50% of these PCa-risk SNPs have a higher risk allele frequency in AA men compare to EA men [73, 77]. Additionally, using risk scores generated from SNPs identified in EA cohorts are much less predictive for AAs compared to other ancestral groups, likely because of the lower linkage disequilibrium (LD) in individuals of African ancestry [80].

1.3.3 Role of DNA methylation in prostate cancer risk

In addition to inherited genetic variation, substantial evidence suggests disruption in DNA methylation features promotes neoplastic and malignant phenotypes in prostate cancer [81]. DNA methylation is a type of epigenetic alternation at clusters of cytosine-guanine di-nucleotides (CpG islands), which are frequently found in promoter regions [82, 83]. Alterations in DNA methylation patterns have implications for gene expression, as variation in DNA methylation reflects variation in gene expression [84]. Additionally, DNA methylation is cell and tissue-type specific. Internal factors at the molecular level and external environmental factors, like smoking and diet, influence DNA methylation features [85], mediating the effects of environmental risk factors on disease risk, progression, and treatment outcomes.

1.3.4 Influence of genetic risk factors on DNA methylation features

Genetic variation (i.e SNPs) exerts a strong effect on DNA methylation [86, 87]. Several studies have shown that PCa-risk SNPs alter the expression of genes involved in PCa biology [88-90] and the effect of these SNPs are known to vary significantly across ancestries, potentially due in part to differences in LD patterns. Genomes of African ancestry exhibit higher genetic diversity and lower levels of LD compared to other ancestral populations [77, 91]. DNA methylation status at individual CpG sites can be leveraged to assess whether a differential effect of PCa-risk SNPs on DNA methylation features exists between AA and EA PCa tissue and whether these differences contribute to PCa disparities. The regions where SNPs significantly impact DNA methylation status are known as methylation-Quantitative Trait Loci (mQTLs). To date, no studies have attempted to identify differences in the mQTLs of prostate tissue in AA and EA men. Given the long-standing racial differences in PCa incidence and mortality, interrogating the genetically driven differences in the DNA methylation features of PCa-associated loci can

shed light on the biological pathways underlying the differences in PCa susceptibility between AA and EA men.

1.4 Summary

In this work, we applied diverse yet robust statistical methods integrating genetic/epigenetic data to understand the effects of environment on disease susceptibility in diverse populations. This work was motivated by notable gaps in the literature relating to: 1) the contribution of rare genetic variation on arsenic metabolism efficiency, 2) inconsistent findings regarding the effect of arsenic exposure and arsenic metabolism efficiency on hypertension risk, and 3) the unknown causal mechanisms ascribed to the complex relationship between variation in environmental factors and genetic susceptibility factors that vary by race, driving prostate cancer disparities in AA men.

In Chapter 2, we estimated the impact of rare, protein-coding variation in *AS3MT* on AME using a multi-population approach to facilitate the discovery of population-specific and shared causal rare variants (Chapter 2). This study is innovative in part because it was the first to examine the role of rare variants in AME. In addition, this was the first study to use next generation sequencing data to identify arsenic susceptibility variants and the first to do so using genotype data from multiple arsenic-exposed population groups of different ancestries. These findings have critical implications for how we identify individuals at a high-risk for arsenic toxicities. In addition, these findings provide additional information about the biological mechanisms involved in arsenic metabolism and the role of rare variants in disease development.

In Chapter 3, we utilized genetic determinants of AME to estimate its effect on HTN risk using “Mendelian randomization” methods and longitudinal measures of blood pressure phenotypes. The MR approaches implemented in this work enabled stronger causal inference

because MR estimates are not vulnerable to the effects of potential confounding or reverse causation in comparison to observational studies. Additionally, gene-environment interaction (GxE) analyses were employed to determine whether genetic determinants of AME modify the effect of arsenic exposure on HTN risk. Here, we improve upon prior work by assessing a large prospective cohort with a historically high level of arsenic exposure and existing arsenic metabolite data.

In Chapter 4, we characterized the effects of PCa-risk SNPs on the DNA methylation at nearby CpG sites in both benign and cancerous prostate tissue of AA and EA men and identified co-occuring meQTLs and GWAS risk-loci. The role of most PCa-risk SNPs identified through GWAS is unknown. Our findings demonstrate a potential regulatory mechanism by which some PCa-risk SNPs modify local DNA methylation and/or the regulation of local gene expression and thereby influence PCa-risk. This work was also the first to include ethnically diverse PCa patients.

CHAPTER 2

RARE, PROTEIN-ALTERING VARIANTS IN AS3MT AND ARSENIC METABOLISM EFFICIENCY: A MULTI-POPULATION ASSOCIATION STUDY

Note about publication: This manuscript has been published in Environmental Health Perspectives, the full citation is in the acknowledgements

2.1 INTRODUCTION

More than 200 million people are exposed to inorganic arsenic (iAs) through drinking water worldwide [1]. Dietary exposure to iAs (primarily through rice and grain products) is an emerging concern that has received less regulatory focus [5]. Chronic exposure to levels of iAs above the World Health Organization (WHO) safety standard for drinking water (>10µg/L) has been recognized to pose a significant risk of adverse outcomes across multiple organ systems [4]. Epidemiological studies in arsenic-affected areas of South America, Asia, and North America have demonstrated that chronic exposure to iAs is associated with adverse health effects, including increased risk for cardiovascular disease [92], diabetes [93], cognitive dysfunction [10, 94], adverse birth outcomes [95], and overall mortality [14]. Additionally, iAs exposure increases the risk for cancers of the skin [96], lung [9, 97], bladder [9, 98], and kidney [99].

The metabolism of iAs in humans involves a series of reduction and methylation reactions. iAs enters the body as iAs^{III} (trivalent arsenite) or iAs^{V} (pentavalent arsenate). Sequential reduction and methylation reactions are catalyzed by glutathione and arsenic (+3

oxidation state) methyltransferase (*AS3MT*), respectively, producing monomethylated (MMA^{III} and MMA^{V}) and dimethylated (DMA^{III} and DMA^{V}) forms of arsenic. Consumed arsenic is eliminated in urine, primarily as a DMA, although smaller percentages are eliminated as MMA and iAs. Arsenic methylation facilitates excretion of arsenic in urine, as DMA is more rapidly expelled from the body compared to MMA or iAs [100-102]. There are a number of factors believed to impact individuals' ability to metabolize arsenic, including genetic differences, age, sex, body mass index (BMI), smoking status, nutritional status, and arsenic exposure level [19-21]. Arsenic metabolism efficiency (AME) is often represented by the percentage of arsenic species in urine that are DMA (DMA%) [103, 104].

Common inherited genetic variation in the 10q24.32 region (containing *AS3MT*) is known to impact AME and risk for arsenic-induced skin lesions. Two independent association signals for DMA% have been reported in the 10q24.32 region in a Bangladeshi cohort [22, 24]. These DMA%-associated variants, best represented by single nucleotide polymorphisms (SNPs) rs9527 and rs11191527, were also associated with MMA%, iAs%, and skin lesion risk, highlighting the 10q24.32 region as a key source of individual variability in AME and susceptibility to arsenic toxicity. Studies in other arsenic-exposed populations have also reported associations between *AS3MT* SNPs (single nucleotide polymorphisms) and AME phenotypes [105-108], including a study of American Indian populations in the United States with low to moderate levels of iAs exposure [109].

Rare genetic variation (i.e., minor allele frequency (MAF) <1%) may also impact AME. Gao et al. estimated the heritability of DMA% due to common SNPs to be ~16% (95% CI [-7.5%, 39.5%] based on a standard error of 12%) in a sample of unrelated Bangladeshi individuals; however, when restricting to only close relatives, the heritability estimate increased

to 63% (95% CI [31.6%, 94.4%] based on a standard error of 16%), potentially reflecting the contributions of rare variants [27]. Family studies in American Indian communities also reported DMA% heritability estimates >50% [28], also suggesting that rare variants may contribute to the heritability of AME. However, the contribution of rare variants to AME has not been assessed for any specific genes, including *AS3MT*.

In this study, our primary goal was to characterize the impact of rare, protein-altering variants in *AS3MT* on AME in three different arsenic-exposed populations. This multi-cohort approach allows us to assess the generalizability of our findings across cohorts of different ancestral and environmental backgrounds. Understanding the role of rare variation is critical for identifying individuals at a high-risk for arsenic toxicities and understanding the biological mechanisms underlying inter-individual differences in susceptibility to toxicity.

2.2 METHODS

2.2.1 Study Participants

This study uses data from three different studies of arsenic-exposed populations: the Health Effects of Arsenic Longitudinal Study (HEALS), the Strong Heart Study (SHS), and the New Hampshire Skin Cancer Study (NHSCS). We selected these three cohorts because of their known arsenic exposure, existing data on arsenic species, and available DNA for sequencing.

HEALS is a prospective cohort study designed to investigate health outcomes associated with chronic arsenic exposure through drinking water in Araihazar, Bangladesh [110]. Arsenic concentrations have been measured in >5,000 wells in the study area. 11,746 participants (5,042 males and 6,704 females, 18 to 75 years of age) were enrolled between 1999 and 2001 after providing informed consent. Participants completed a written questionnaire and participated in

clinical exams at baseline. Blood and urine samples were also collected from participants at baseline [110]. Participants were simultaneously assessed for arsenical skin lesions at baseline and every two years thereafter by trained physicians using a structured protocol [15]. Arsenic species in baseline urine samples were measured previously for 4,794 HEALS participants [19]. For this study, we selected 2,719 of these participants who had available DNA for sequencing, of which 273 were skin lesion cases at baseline. While many of these participants were randomly selected for arsenic species measurement in baseline urine, a sizable fraction (~50%) were selected based on outcomes they experienced at subsequent follow-up visits (i.e., skin lesions, respiratory symptoms, and cardiovascular conditions), in a case-cohort fashion. So while this group was relatively healthy at baseline, they were not randomly selected.

The SHS is the largest population-based cohort study of cardiovascular disease in American Indian men and women. SHS includes 12 American Indian tribes and communities and was designed to estimate cardiometabolic disease morbidity, mortality, and the prevalence of its risk factors [8]. Between 1989 and 1991, SHS recruited 4,549 American Indian men and women between the ages of 45 to 74. These participants have been exposed to low to moderate levels of iAs primarily through drinking water [111] and to a lower extent, diet (i.e. rice) [112]. Arsenic species were measured in 3,973 participants [8]. These individuals are members of large families, so we restricted to a group of 997 unrelated individuals with existing measures of arsenic species and available DNA for sequencing.

The NHSCS is a population-based case-control study of basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) of the skin [113]. Invasive SCC incident cases (n=510) , 25 to 74 years of age, were recruited from >90% of practicing dermatologist and pathologist clinics in New Hampshire and bordering areas. Controls (n=483) frequency-matched on sex and age were

selected from the New Hampshire Center for Medicare and Medicaid services and driver's license records. The enrollment period for cases and controls was between 2003 and 2009. Participants interviewers masked to the study hypothesis ascertained sociodemographic, lifestyle, medical, and sun-exposure information. Additionally, home water and spot urine samples were collected and used to measure total urinary arsenic concentration and arsenic species (post-diagnosis measurements for SCC cases) [113]. Of the 993 cases and controls, 288 individuals did not have sufficient DNA for sequencing and were excluded from the current study. This resulted in 706 individuals with existing data on arsenic species selected for the current sequencing study (349 SCC cases and 357 controls).

2.2.2 Measurement of Arsenic Metabolites in Urine

All participants selected for this study had existing data on arsenic species in urine. In all three cohorts, speciation analysis of arsenic metabolites was performed using high-performance liquid chromatography (HPLC) [114] followed by detection using ICPMS [8, 110, 115]. Details regarding the LOD for each metabolite and percent of samples below the LOD have been described previously [110, 111, 115]. Briefly, in HEALS the LOD was 1μg/L for iAs^{III}, iAs^{V}, MMA, DMA. In SHS, iAs^{III} was first oxidized to iAs^{V} in the urine, following the methods described in [114], this method minimizes undetectable iAs^{III} or iAs^{V} in urine samples from populations with low levels of exposure. The LOD was 0.1μg/L for iAs^{V} and 0.5μg/L for MMA and DMA in SHS. In NHSCS, the LOD was 0.15μg/L for iAs^{III}, 0.1μg/L for iAs^{V}, 0.14μg/L for MMA, and 0.11μg/L for DMA.

For this analysis, iAs^{III} and iAs^{V} were summed to obtain total iAs, and each arsenic species (iAs, MMA and DMA) is expressed as a percentage of the sum of each of these three

species (iAs+MMA+DMA). Arsenocholine (AsC) and arsenobetaine (AsB) are non-toxic forms of (organic) arsenic and are excluded from our analyses. We used DMA% as our primary measure of AME; MMA% and iAs% were used as secondary measures of AME. In SHS and NHSCS, measures of AME were not computed if ≥2 species were undetectable for an individual. If only one metabolite was undetectable, this missing metabolite was estimated as the $\frac{LOD}{\sqrt{2}}$ and was used in downstream analyses. For the current study, of the 2,719 HEALS participants, 523, 155, and 4 participants had values below the LOD for iAs^{III}, iAs^{V}, and MMA, respectively. We kept these participants in our analyses and set arsenic species values <LOD to zero. In SHS, we excluded 6 (of 997) participants with ≥2 arsenic species with undetectable values and calculated imputed values for iAs^{V} and MMA in 49 and 2 participants, respectively. In NHSCS, we excluded 39 (of 706) participants with ≥2 arsenic species with undetectable values and calculate imputed values for iAs^{III}, iAs^{V}, and MMA in 225, 598, and 29 participants, respectively.

2.2.3 DNA extraction and targeted sequencing

The details of sample collection and DNA extraction for HEALS and SHS participants have been described previously [22, 116]. In brief, genomic DNA for HEALS samples were extracted from clotted blood using the Flexigene DNA Kit (Cat # 51204). For SHS participants, buffy coats were extracted from fasting blood samples using organic solvents and were stored at MedStar Health Research Institute [116]. For NHSCS participants, frozen buffy coats retrieved for this study were thawed at room temperature and mixed by tube inversion and 5-second vortex. DNA was extracted using the Agencourt Genfind v2 Kit.

For all DNA samples, we conducted targeted sequencing focusing on the exons of *AS3MT*. Sequencing was carried out on an Illumina MiSeq instrument using Illumina's TruSeq Custom Amplicon Assay Kit. The custom kit was designed to sequence 781 small regions of ~450 bp in length (which includes additional content that is not relevant to the aims of the present study, including common variation in the 10q24.32 region).

2.2.4 Read alignment, variant calling, and quality control

Targeted sequencing data was processed at the University of Chicago Bioinformatics Core using GATK (GenomeAnalysisToolkit) [117] and the GATK Best Practices Workflow (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-). For all samples, raw paired-end reads were aligned to the reference genome (hg19) using the Novoalign software. HaplotypeCaller (in GVCF mode) was used to call variants (bi-allelic and indels) per-sample, generating intermediate GVCF files. Next, we used the GenomicsDBImport tool to consolidate the contents of GVCF files across samples into a single GVCF file, which was then used for joint calling across all samples, producing a set of joint-called bi-allelic variants and indels in a single VCF. We applied two sets of filters across all cohorts to the 414 raw variants and indels detected, using more stringent thresholds for indels since they are more prone to heterozygous false positives than bi-allelic variants. Variants matching any one of the following quality metric conditions were removed: QualbyDepth (QD < 2.0), FisherStrand (FS > 60), RMSMappingQuality (MQ < 40), StrandOddsRatio (SOR > 3), MappingQualityRankSumTest (MQRankSum < -12.5), or ReadPosRankSum < -8. We removed indels that met at least one of the following conditions: QD < 2.0, FS > 200, ReadPosRankSum > -20, InbreedingCoeff < -0.8, or SOR > 10. The first filter excluded 141 bi-allelic variants and 17

indels, resulting in 256 variants passing all the quality metrics. Additionally, we excluded all variants with call rates <90% resulting 135 for HEALS, 92 for SHS, and 120 for NHSCS. Overall, we were able to sequence 9 of the 11 exons in *AS3MT* (Illumina was unable to design flanking probes for exons 5 and 10) at a depth of coverage ranging between 173 and 729 per exon, across all cohorts (**Figure 1.1**).

Among the 2,719 HEALS participants selected for this study we excluded 260 participants due to low depth of coverage (<30x) and 25 participants who did not have genome-wide SNP data available for the generation of a kinship matrix, resulting in 2,434 remaining HEALS participants. Among the 997 SHS participants, we excluded 123 participants whose samples had extremely low coverage due to a very low number of reads and 6 participants with missing metabolite data, leaving 868 SHS participants. Among the 705 NHSCS participants selected for this study, we removed 39 samples due to missing metabolite data and one sample due to low number of reads, resulting in 666 NHSCS participants.

In an effort to validate our sequenced variants we used existing HEALS SNP array and exome array data (Pierce, et al. 2019) for 2,363 of the HEALS participants sequenced for this study. We identified 721 overlapping variants in both datasets (sequencing data and imputed SNP array data). We compared genotypes for each variant and found that 645 of 721 variants had >90% consistent genotypes across 2,363 individuals. Only 3 variants had lower than 60% consistency. Additionally, we used the Genome Aggregation Database (gnomAD) v2.1.1 [118] to validate observed variants and to obtain an estimate of the rare variants missed in our study (specifically in exons 5 and 10). GnomAD provides exome sequencing data for >125,000 unrelated individuals, most of which are classified into 6 ancestral populations (African American, Latino, East Asian, Finnish, non-Finish European, and South Asian).

2.2.5 Variant Annotation

All variants were annotated using the Annotate Variation (ANNOVAR) software [119],which provided information on the impact of each exonic variant on the amino acid sequence (i.e. non-synonymous, synonymous, frameshift insertion-deletion (indel), etc.). A variant was annotated as a predicted loss of function variant (pLoF) if they were categorized by ANNOVAR as frameshift, premature stop codon, loss of start or stop codon, or splice acceptor or donor. For this analysis, we focused on variants with a minor allele frequency (MAF <0.01) that altered the protein sequence. Thus, we excluded all synonymous, intronic, and intergenic variants detected. These exclusions resulted in 5, 4, and 5 *AS3MT* rare, protein-altering variants in HEALS, SHS, and NHSCS, respectively.

We used SIFT [120] and PolyPhen [121] to predict how variants coding for amino acid changes impact protein function. SIFT uses information on sequence homology and physical properties of amino acids to predict the impact on protein function. While Polyphen uses sequence, phylogenetic and structural information to empirically predict the effect of the substitution.. We also report Combined Annotation Dependent Depletion (CADD) scores [122] for the variants we analyze, which are based on evolutionary information, conversion metrics, functional genomic data, and transcription information. CADD scores are PHRED-like C-scores that rank variants relative to all possible substitutions of the human genome ($8.6x10^9$). The CADD score represents the potential level of deleteriousness. For instance, variants with CADD scores of 0 to 10 are in the top $10^{\%}$ most deleterious, CADD scores 10 to 20 are in the top 1%, CADD scores 20 to 30 are in the top 0.1%, and so on.

2.2.6 Identification of common variants in 10q24.32 region association with DMA%

In light of the well-established associations between common variants in the 10q24.32 region and AME, we used our sequencing data to genotype common variants in the 10q24.32 region and identify variants showing independent associations with DMA% in each population. We performed linear conditional, forward stepwise regression tests, stratified by cohort. We restricted analyses to 4,990 variants in the 10q24.32 target region and excluded variants with Hardy-Weinberg p-value < $1x10^{-10}$ (n=122) and variants with a MAF < 0.005 (n=4,485). This QC resulted in 383 variants across all cohorts (HEALS n=2,436, SHS n=874, and NHSCC n=752).

Linear conditional, forward stepwise regression tests consisted of a series of association analyses. To identify the primary association signal, we individually tested each of the 383 common variants for an association with DMA%, controlling for age, sex, and population structure in SHS and HEALS. We used PLINK [123] for SHS and NHSCS and Genome-wide Complex Trait Analysis (GCTA) [124] for HEALS (for adjustment of kinship matrix). To identify secondary independent association signals, we included the primary signal (identified in the previous analyses) as a covariate and repeated our association analyses. We repeated these analyses, adjusting for both the primary and secondary signals, to identify any remaining independent signals in each population. These analyses resulted in the identification of two independent lead variants in HEALS, rs12573221 ($P=6.4x10^{-12}$) and rs14553735 ($P=7x10^{-11}$), three independent lead variants in SHS, rs10786722 ($P=1x10^{-20}$), rs4919687 ($P=7.9x10^{-9}$), and rs10883846 ($P=4.7x10^{-16}$) and a single lead variant in NHSCS, rs76255497 ($P=7x10^{-5}$). We conducted analyses of MMA% and iAs% and found that the DMA% associated lead variants

were consistently among the top 5 lead independent variants in analyses of MMA% and iAs%, across all cohorts. Given this agreement in results across metabolites, we adjusted for the DMA% associated individual SNP alleles (0, 1, or 2) in our linear regressions. This adjustment ensured association estimates were not biased due to linkage disequilibrium (LD) between common variants known to impact AME and the rare variants studied in this work.

2.2.7 Statistical Analysis

The classical single-variant-based association test is not typically extended to analyses of rare variants due to lack of statistical power [125]. Instead, investigators have developed methods that aggregate variants in a biologically relevant region (i.e., a gene) and evaluate their cumulative effects. This approach is now the standard for rare variant studies and can provide reasonable power to detect association between a gene-based set of rare variants and a human trait [126]. In our study, the primary hypothesis was that carrying a rare, protein-altering variant in *AS3MT* reduces AME (represented by DMA%). To test this hypothesis, we conducted a burden test, where we assigned each individual a binary carrier status and tested whether the mean of DMA% differs between carriers and non-carriers (using linear regression). Individuals carrying at least one rare, protein-altering variant in *AS3MT* were assigned a carrier status of 1, and 0 otherwise. We did not observe any individuals carrying more than 1 rare, protein-altering variant in *AS3MT*. The burden test makes the strong assumption that all variants analyzed in the gene are causal and affect the trait in the same direction and with the same magnitude. Violation of these assumptions may result in loss of power [127].

We fit linear regression models for SHS and NHSCS, adjusted for age and sex, with carrier status as the predictor and arsenic metabolites (DMA%, MMA% and iAs%) as outcomes .

For HEALS, we conducted the burden test using a mixed-linear model based association test ('--mlma') and incorporated a kinship matrix as implemented in the GCTA software [124] in order to control for cryptic relatedness among participants. For SHS, we accounted for population structure by adjusting for the first five principle components (PCs) derived from existing genome-wide SNP data described previously (PCs provided by SHS) [128]. We were not able to control for population structure in NHSCS due to lack of existing genome-wide SNP data, however 99.5% of NHSCS participants included in this study are self-reported non-Hispanic white (only 3 individuals are self-reported Hispanic or Latino). To ensure minimal bias due to population structure we conducted the same burden test excluding the 3 individuals who are self-reported Hispanic or Latino. Additionally, we performed the burden test with adjustment and without adjustment for common SNPs in 10q24.32 region.

To address the possibility of bias in the association between rare variant carrier status and AME due to prevalent skin lesion and SCC case status, we repeated the burden tests excluding 293 prevalent skin lesion cases in HEALS (1 skin lesion cases was a carrier of the rs3523887 rare variant) and 349 SCC cases in NHSCS. We analyzed 2,076 HEALS individuals without prevalent skin lesions and 357 NHSCS controls and adjusted for age, sex, relatedness (in HEALS only), and common SNPs in 10q24.32 region. To estimate the association between carrier status and DMA% across all cohorts, we meta-analyzed the association estimates for each cohort using the 'metafor' package in R. We chose the meta-analyses approach instead of a pooled analysis because each analysis required a unique adjustment for population structure (i.e., in HEALS we adjusted for a kinship matrix and in SHS we adjusted for genotype principle components). Tests of heterogeneity were used to determine the appropriate selection of a fixed-effects versus a random-effects models. We used a random-effects model for a Cochran's Q-statistic with a p-

value < 0.10 and a fixed-effects model otherwise. We conducted similar analyses using MMA% and iAs% as outcomes, to ensure consistency with analyses of DMA%.

In secondary analyses, we estimated the association between the collective effect of rare, protein-altering variants in *AS3MT* on DMA% using SKAT (sequence kernel association test) [129], within each cohort. SKAT is a score-based variance-component test that allows the variants within a gene to have different directions and magnitudes of effect, where the SKAT statistic is primarily a representation of the variance of effect sizes within the set of variants. The SKAT software does not generate an effect estimate for the individual variants. Under the null hypothesis none of the variants in the gene set have an effect on the arsenic metabolites. Conversely, under the alternative hypothesis, at least one variant in the gene set must have an effect $\neq 0$, and a larger variance in phenotypic variation explained by the set of genetic variants would result in a larger SKAT statistic and a smaller p-value. The SKAT software restricts all analyses to variants with <1% missingness.

In SKAT analyses we made adjustments for age, sex, population structure, and dosage scores of common SNPs associated with DMA% in the 10q24.32 region. HEALS was analyzed using family-based Sequence Kernel Association Test (famSKAT) [130], which allows adjustment for a kinship matrix. We used both burden and non-burden (SKAT) gene-based tests because it was unknown which method would provide superior power and if the assumptions of the burden test were reasonable. The meta-analysis version of SKAT, MetaSKAT [131], was used to derived summary SKAT statistics of rare protein-altering variants across the three cohorts. The MetaSKAT framework is currently unable to implement kinship matrices and thus HEALS estimates in meta-analyses were not adjusted for relatedness. To assess any bias in the SKAT meta-analyses due to population structure in HEALS, we conducted sensitivity analyses

of the burden and SKAT test where we excluded related individuals (kinship coefficient $r^2 > 0.05$) resulting in 1,285 unrelated individuals.

Lastly, we attempted to extrapolate our results to approximate the carriers' increased risk of developing arsenic-related toxicities. We used previously reported odds ratios (ORs) for the risk of lung and bladder cancer in Chile [132] for a one unit percentage increase in MMA% (for lung and bladder cancer) . We then converted the reported ORs to a logORs and multiplied them by the estimated meta-analysis effect of rare variant carrier status on MMA%. These values were then exponentiated to obtain ORs for the risk of developing lung and bladder cancer for rare variant carriers. In addition, we used a SNP-based Mendelian randomization (MR) approach to estimate carriers' increased risk of skin lesions, using data from Bangladesh. We first estimated the impact of DMA% on skin lesion risk, using previously-reported association estimates for DMA%-associated *AS3MT* SNPs (rs9527 and rs11191527) and *FTCD* SNP (rs61735836) [24, 26]. We used a likelihood-based MR method that combines the summary statistics of multiple genetic variants into a single causal estimate for DMA% [133]. We multiplied this estimate by the association of carrier status with DMA%. Finally we exponentiated this value to obtain the OR describing the risk of developing skin lesions given a difference in DMA% due to rare variant carrier status.

2.3 RESULTS

Characteristics of the 2,434 HEALS, 868 SHS, and 666 NHSCS participants included in our analyses are described in **Table 1.1**. Total urinary arsenic (μg/L), represented by the sum of urinary concentrations of iAs^{III}, iAs^{V}, MMA, and DMA, varied across cohorts (**Table 1.1**), with

substantially higher total urinary arsenic in HEALS (mean=137.4 μg/L) compared to SHS (mean= 14.8 μg/L) and NHSCS (mean= 7.3 μg/L). Urinary DMA%, on average, was the lowest in HEALS (71.2%), higher in SHS (76.9%), and highest in NHSCS (81.1%) (Table 1). The distribution of arsenic metabolites by cohort are shown in Figure S1. The difference in metabolite percentages across cohorts is likely driven by the difference in overall level of iAs exposure. HEALS participants exhibit the lowest percent of urinary DMA and experience the highest level of iAs exposure. This is consistent with the reported inverse association between DMA% and iAs exposure, suggesting that the methylation machinery can become saturated at high doses or arsenic [6].

We identified 13 rare, protein-altering variants in *AS3MT* across all cohorts (**Table 1.2**). All but one of these variants (rs35232887 in HEALS and NHSCS) was specific to a single cohort, with 5, 4, and 5 variant sites detected in HEALS, SHS, and NHSCS respectively. These variants were primarily observed once or twice, with 23 total individuals carrying a rare allele at one of these sites (13 carriers in HEALS, 5 in SHS, and 5 in NHSCS). Individual-level percent of arsenic metabolites for rare-variant carriers are shown in Table S1. All five variants detected in HEALS were also present in gnomAD, and four out of five variants detected in NHSCS were present in gnomAD (**Table S1.2)**. Only one of the four variants detected in SHS was present in gnomAD, which may be due to the fact that gnomAD does not include data from individuals of Native American ancestry. Overall, 0.5% of HEALS and SHS participants were carriers of an *AS3MT* rare, protein-altering allele, and 0.7% were carriers in NHSCS. This is similar to, but slightly below the 0.9% of carriers of rare, protein-altering *AS3MT* variants in gnomAD [134] (variants in exons 5 and 10 were excluded from the percent of gnomAD carriers for consistency purposes).

Most variants were missense changes, but each cohort had at least one pLoF variant (i.e., stop gain, frameshift, or splice donor). In HEALS and NHSCS, the two pLoF variants were each observed once, rs528680133 (splice donor) and 10:104634377 (frameshift), with high CADD scores of 26 and 24.7, respectively. In SHS, we identified two pLoF variants, 10:104660404 coding for a stop-gain with a CADD score of 35 (observed in two individuals) and 10:104638178, coding for a frameshift insertion (observed in one individual).

Carrier status for rare, protein-altering variants in *AS3MT* was associated with lower mean DMA% on average compared to non-carriers, across all cohorts in analyses unadjusted for common variants in the 10q24.32 region (HEALS: β = -9.7, 95% CI: -14.4, -5.1, P = 3.9×10^{-5}; SHS: β = -10.2, 95% CI: -17.6, -2.9, P = 0.006; NHSCS: β = -9.3, 95% CI: -16.3, -2.4, P = 0.009). These results were consistent in adjusted analyses (HEALS: β = -9.4, 95% CI: -13.9, -4.8, P = 5.2×10^{-5}; SHS: β = -6.9, 95% CI: -13.5, -0.2, P = 0.04; NHSCS: β = -8.9, 95% CI: -15.6, -2.2, P = 0.01) (**Table 1.3 and Figure 1.2**). In SHS, at least 1 rare variant carrier had missing genotypes for the common variants rs10786722 and rs4919687, so we instead adjusted for dosage of proxy SNPs, 10:104723620 and 10:104685493, respectively, with no missing genotypes among carriers. Additionally, sensitivity analyses in NHSCS where we excluded 3 individuals who are self-reported Hispanic or Latino showed the same results (β = -8.9, P = 0.01) and thus confirm the inclusion of these individuals do not bias our results. Carrier status showed evidence of a positive association with MMA% and iAs% (Table 3) across all cohorts but the association exceeded the p-value < 0.05 threshold for SHS in analyses unadjusted and adjusted for common variants.

We conducted meta-analyses across the three cohorts (3,900 participants) for unadjusted and adjusted models (for common variants in 10q24.32) across all metabolites. Tests of

heterogeneity indicated fixed-effects models (P > 0.10) were appropriate for all metabolites in adjusted and unadjusted models (unadjusted: $P_{DMA\%} = 0.99$, $P_{MMA\%} = 0.93$, $P_{iAs\%} = 0.67$, and adjusted: $P_{DMA\%} = 0.81$, $P_{MMA\%} = 0.46$, $P_{iAs\%} = 0.86$). These meta-analyses indicated that, relative to non-carriers, being a carrier of a rare protein-altering variant in *AS3MT* was associated with 9.8% lower mean DMA% ($P = 2.2x10^{-8}$, 95% CI: -13.2, -6.3), 4.9% higher mean MMA% ($P = 1.3x10^{-6}$, 95% CI: 2.9, 6.9), and 4.9% higher mean iAs% ($P = 3.6x10^{-5}$, 95% CI: 2.6, 7.2) with no adjustment for common variants in 10q24.32. Meta-analyses adjusted for common variants showed similar results, where carriers had 8.7% ($P = 1.9x10^{-7}$, 95% CI: -11.9, -5.4) lower mean DMA%, 4.5% ($P = 9.7x10^{-6}$, 95% CI: 2.5, 6.4) higher mean MMA%, and 4.3% (P = 0.0002, 95% CI: 2.0, 6.6) higher mean iAs%.

Sensitivity analyses in HEALS and NHSCS restricted to individuals without prevalent skin lesions or SCC showed a small attenuation in the association of rare variant carrier status with DMA% in 2,076 HEALS participants ($\beta = -8.8$, 95% CI: -13.5, -4.1) and an increased strength of association in 357 NHSCS controls ($\beta = -9.3$, 95% CI: -16.1, -2.4) (**Table S1.3**) compared to analyses of all individuals (**Table 1.3**). These results remain consistent with the results prior to case status exclusion.

Using SKAT, we observed P-values <0.05 for the association between the collective effects of rare, protein-altering variants in *AS3MT* and DMA% in SHS (P = 0.03) and NHSCS (P = 0.005), but not in HEALS (P = 0.11) (**Table 1.4**). Meta-analyses across cohorts using SKAT provided evidence of association across all cohorts (P = 0.002). Analyses using MMA% and iAs% as outcomes in SKAT produced results similar to the burden test (**Table 1.4**).

It is important to note that the meta-analyses performed in the MetaSKAT software requires individual level data and do not allow adjustment for a kinship matrix in the HEALS

cohort. Thus, we performed burden and SKAT analyses of 1,285 unrelated individuals in the HEALS cohort across all metabolites without kinship adjustment. These analyses also excluded two carriers of one rare, protein altering variant in *AS3MT* (rs563943815). Overall, we found that excluding related individuals in HEALS slightly increases the effect of carrier status on arsenic metabolites in the burden test and decreases the P-value in SKAT (Table S4). These results also highlight the possibility that not all non-synonymous variants included in our rare variant test will impact protein function. These results suggest that accounting for kinship matrix in the meta-analyses has little impact on the estimated effect sizes from the burden test and the p-values from SKAT.

Using exome sequencing data from gnomAD, we assessed the extent of missed rare, protein-altering variants in exons 5 and 10 of *AS3MT*, which were not captured in the current study. In exon 5, among all participants, there were 19 rare (MAF<0.01), protein-altering variants (15 missense and 3 pLoF) (Table S5). In exon 10, there were 14 rare, protein-altering variants (12 missense and 2 pLoF). The carrier frequency in gnomAD was 0.09% (52 of 60,706) for exon 5 and 0.4% (242 of 60,706) for exon 10. Across all populations and all *AS3MT* exons in gnomAD, the percent of carriers of rare, protein-altering variants in AS3MT was 1.4% (859 of 60,706) (Table S6), approximately double what we observe in our three cohorts (0.5-0.7%).

We used results from prior studies to estimate the association of carrier status with disease outcomes (lung cancer, bladder cancer, and skin lesions) in exposed populations. A Chilean study [132] of lung and bladder cancer report logORs of 0.10 and 0.04 for a 1-unit increase in MMA%, respectively. These logORs and a 4.5% increase in MMA% among carriers (from meta-analysis, Table 3) correspond to a 59% (OR = 1.59) and 19% (OR = 1.19) estimated increase risk of developing lung and bladder cancer compared to non-carriers of *AS3MT* rare

variants, respectively (**Table S1.7**). For skin lesions, we used an MR approach to estimate the association of DMA% with skin lesion outcomes. We used previously reported association estimates and standard errors (from the HEALS cohort) for the: 1) of *AS3MT* SNPs and *FTCD* SNP with DMA% and 2) associations of *AS3MT* SNPs and *FTCD* SNP with skin lesions (**Table S1.8**) [24, 26]. Given an estimated logOR of 0.069 for a 1-unit decrease in DMA% and an 8.7% reduction in DMA% among carriers (from meta-analysis, Table 3), the estimated the risk of developing skin lesions for carriers in the HEALS cohort was 82% higher (OR = 1.82) compared to non-carriers of *AS3MT* rare variants (**Table S1.9**).

2.4 DISCUSSION

In this study, we conducted targeted sequencing of the coding regions of the *AS3MT* gene across three arsenic-exposed cohorts of different genetic ancestry. Our results suggest that on average, rare, protein-altering variants in *AS3MT* are associated with decreased AME (lower urinary DMA%), and these associations were independent of the effects of known common variants in this region. Our burden test results provide evidence that rare variants in *AS3MT*, which are predicted to affect protein structure and function, may reduce the efficiency of arsenic metabolism. These variants likely result in increased retention of arsenic in the body (as seen in *AS3MT* KO mice [135]) and increased risk of arsenic-related toxicities and health effects (as seen in humans carrying common alleles associated with AME). This is the same pattern of association observed among the HEALS study participants in Pierce et al. 2012; Pierce et al. 2013; Pierce et al. 2019 for common variants that are associated with AME, i.e, the risk allele decreases DMA% and increases MMA% and iAs%. This study represents one of the first

attempts to assess the effects of rare inherited variation on AME in humans, and we address this question in multiple population groups of different ancestries and arsenic exposure levels.

The impact of common variation in the *AS3MT* (10q24.32) region on AME is well established. Epidemiologic studies report consistent and reproducible associations between common SNPs in and around *AS3MT* and arsenic metabolites in urine across multiple populations [22, 105-109]. The association of these variants with risk for arsenic-induced skin lesions [24] highlights the relevance of these metabolism-related variants to arsenic toxicity risks. At least one of the causal variants in this region appears to have effects that are regulatory in nature, impacting expression of local genes including *AS3MT* [136, 137]. Other than *AS3MT*, only one other gene, *FTCD*, has been shown to contain inherited genetic variation that affects AME. In a Bangladeshi population, the minor allele (A) of a missense variant (rs61735836) in exon 3 of *FTCD* was shown to be associated with decreased DMA% and increased skin lesion risk in Bangladeshi individuals [26]. Common SNPs in *AS3MT* and *FTCD* account for ~10% of the variation in DMA% [26]. The *FTCD* association had not been identified at the time targeted sequencing was conducted, so we could not examine rare variants in *FTCD* in the current study.

Prior heritability studies suggest a role for rare variation in AME. For example, Gao et *al.* estimated the full narrow-sense heritability (h^2) of AME to be 41% (95% CI [7.7%, 74.3%] calculated based on reported standard error of 17%), which includes the additive effects of both common and rare variants, but excludes (adjusts for) the effects of common SNPs in the *AS3MT* region with known associations with AME [27]. They also estimated the heritability due to common SNPs to be only ~5% after similar adjustments. The difference in these two estimates suggests that rare variants (and/or unmeasured shared-environmental factors) may make an important contribution to inter-individual variation in AME. Similarly, in a family study

conducted within SHS, the heritability for DMA% was as high as 63% [28]. Together, these findings suggest that rare variation contributes to familial similarity in AME phenotypes.

AS3MT knock out (KO) mice can be used to understand the potential effects of *AS3MT* pLoF variants in humans. *AS3MT* KO studies have demonstrated the critical role of *AS3MT* in the elimination of arsenic from tissues, clearance of arsenic in urine, and prevention of arsenic toxicities. KO mice exhibit dramatically higher arsenic concentrations and higher proportions of iAs in numerous tissue types, including liver, bladder, and kidney, several hours post arsenic dosing compared to wild type (WT) mice with similar dose [135, 138-140]. Whole body clearance of iAs is substantially slower in KO compared to WT mice. For instance, Drobna et al. showed that at 24 hours post dosing, KO mice retained 50% of the dose while WT only retained 6% [135]. Additionally, KO mice are at an increased risk for arsenic toxicities [141-143]. The *AS3MT* KO mice models developed to date are homozygous for the deleted gene (as3mt $^{-/-}$). In contrast, the mutation carriers in this study are heterozygous for a *AS3MT* rare variant likely to be damaging. Thus, we do not expect the impaired arsenic metabolism in our participants to be as extreme as that observed in *AS3MT* KO mice. However, the pLoF variants we observe (i.e., rs528680133 splice donor in HEALS, 10:104660404 stop gain variant in SHS, and 10:104634377 frameshift in NHSCS) are likely to produce human phenotypes that most closely resemble those of the KO mice. In human studies, we are unable to obtain tissue specific measures of arsenic, and therefore cannot determine how the elimination of arsenic varies among tissue types. However, we do observe carriers of pLOF variants exhibit strikingly low metabolism efficiency, as these individuals are consistently in the bottom 1st decile of the DMA% distribution (Table 2). The effect of missense variants on AME is not as striking, and

this is expected because some of these amino acid substitutions may have small or negligible effects on protein function.

In the current study, we performed two types of gene-based tests to assess the association between rare *AS3MT* variants and AME phenotypes. The first was a burden test where we sum each individual's rare allele count (equivalent to carrier status in our case where no individual carries >1 rare protein-coding allele in *AS3MT*), and the second was a non-burden test (SKAT) where we test the variance component explained by the genetic variants among the total phenotypic variations. The burden test showed that on average, rare variants in *AS3MT* were inversely associated with DMA% across all three populations. Analyses using SKAT also showed clear association between AS3MT variants and DMA%. Both the burden test and SKAT methods also showed clear associations for AS3MT variants with both iAs% and MMA%. Meta-analyses across all 3 cohorts suggested that on average, carriers of rare, protein-altering variants in *AS3MT* have ~9% lower DMA% compared to non-carriers, a substantially larger effect compared to any single common variant associated with DMA% (i.e., *AS3MT* or *FTCD* SNPs). Furthermore, we expect that this estimated association of rare, protein-altering *AS3MT* variants with AME may be attenuated as compared to the association of known deleterious variants with AME, because some individuals may be carrying missense variants that have a negligible impact on *AS3MT* function.

Our study was limited by our lack of data on variants in exons 5 and 10 of *AS3MT*, due to limitations of the Illumina TruSeq custom amplicon kit. We assessed the extent of missingness in our study using exome sequencing data from gnomAD, and we estimated that due to our lack of data in exons 5 and 10 we were unable to capture approximately 0.5% of rare variant carriers. It is worth noting that exon 10 contains a rare variant with 205 carriers in gnomAD (frequency of

0.02%) (**Table S1.5**). Thus, it is likely that sequencing of exons 5 and 10 would identify a substantial number of additional variants not captured in the present study. Consequently, we expect the *AS3MT* protein-altering variant carrier proportion to be higher than what we report in this study.

Additional research is needed to further our understanding of the role of rare, protein-altering *AS3MT* variants in AME and in the etiology of arsenic-related health conditions. A broader representation of arsenic-exposed populations is necessary to replicate these findings, identify additional rare *AS3MT* variants, and assess their effects on the risk for arsenic toxicities. We are unable to examine the association of rare *AS3MT* variants with skin lesion status (in HEALS) and SCC (in NHSCS) directly due to the small number of cases who are carriers of rare variants (3 cases of prevalent and incident skin lesions and 0 cases of SCC were carriers). However, using previously reported risk estimates for skin lesions [24, 26] and lung and bladder cancer [132], we estimated the risk of developing skin lesions was 82% higher given an 8.7% reduction in DMA% and the risks of developing lung and bladder cancer were 59% and 19% higher, respectively, given a 4.5% increase in MMA%. A larger cohort study is needed to assess the impact of rare *AS3MT* variants on risk of arsenic toxicities and explore the interactions of rare-variant carrier status with sex, lifestyle factors, nutritional status and overall iAs exposure. Additional experimental research is also needed to fully understand the effects of these variants on *AS3MT* protein structure and function.

In summary, we used data from multiple arsenic-exposed cohorts to assess the association between rare, protein-altering variants in *AS3MT* and arsenic metabolism efficiency. We provide evidence that rare variants in *AS3MT* decrease arsenic metabolism efficiency, a finding consistent across multiple populations with distinct ancestral backgrounds. Although we

are unable to assess the effect of these rare, protein altering variants on arsenic-related disease outcomes (due to the small number of carriers), our findings suggest that at least some of the variants may affect internal dose of arsenic (through effects on arsenic metabolism and clearance) and, in turn, influence the risk of arsenic-related toxicities. Our results, together with our knowledge of common *AS3MT* variants, highlight *AS3MT* as a major susceptibility locus for arsenic metabolism efficiency with implications for arsenic toxicities.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (https://www.nih.gov/) grants R01 ES023834 (to B.L.P.), R35 ES028379 (to B.L.P.), R21 ES024834 (to B.L.P. and Maria Argos), P42ES010349 (to J.H.G.), R01 CA107431 (to H.A.), P30 ES027792 (to H.A. and Gail Prins), R24 ES028532 (to H.A.) and R24 TW009555 (to H.A.). The authors thanks all the men and women who participated in HEALS, SHS, and NHSCS and all the research staff who contributed to data collection.

Delgado DA, Chernoff M, Huang L, Tong L, Chen L, Jasmine F, Shinkle J, Cole SA, Haack K, Kent J, Umans J, Best LG, Nelson H, Griend DV, Graziano J, Kibriya MG, Navas-Acien A, Karagas MR, Ahsan H, Pierce BL. Rare, Protein-Altering Variants in AS3MT and Arsenic Metabolism Efficiency: A Multi-Population Association Study. Environ Health Perspect. 2021 Apr;129(4):47007. doi: 10.1289/EHP8152. Epub 2021 Apr 7. PMID: 33826413; PMCID: PMC8041273.

2.5 APPENDIX

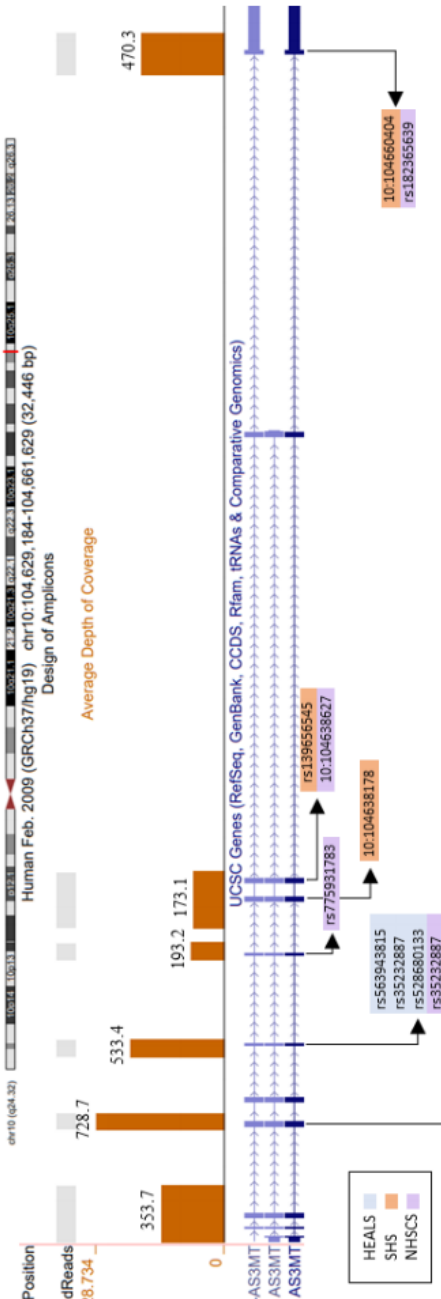


Figure 1. 1. Average aligned read depth (depth of coverage) of AS3MT exons

Table 1. 1. Participant characteristics stratified by cohort

Characteristics		**HEALS** (n=2,434)	**SHS** (n=868)	**NHSCS** (n=666)
Sex	Male (%)	1433 (58.9%)	348 (40%)	395 (59.3%)
	Female (%)	1001 (41.1%)	520 (60%)	271 (40.7%)
Age (years)	Mean ± SD	41.7 ± 10.5	56 ± 8.2	64.4 ± 7.6
BMI (kg/m2)[a]	Underweight (%)	1043 (43.3%)	7 (0.8%)	7 (1%)
	Normal (%)	1212 (50.3%)	143 (16.6%)	194 (29.6%)
	Overweight (%)	137 (5.7%)	300 (34.7%)	246 (37.5%)
	Obese (%)	17 (0.7%)	414 (47.9%)	209 (31.9%)
Total urinary arsenic (µg/L)	Mean ± SD	137.4 ± 162.8	14.8 ± 16.9	7.3 ± 8.9
25th Percentile		38.2	5.6	3.3
50th Percentile		82.6	9.8	5.0
75th Percentile		175.6	16.9	8.4
Min and Max		3.7, 1528.9	0.4, 161.9	0.7, 111.6
Urinary arsenic metabolites				
iAs%	Mean ± SD	15.2 ± 6.5	8.5 ± 5.2	7.8 ± 5.2
25th Percentile		11.0	5.2	4.3
50th Percentile		14.1	7.5	6.8
75th Percentile		18.3	10.7	9.8
Min and Max		0, 70.3	0.6, 59.7	0.5, 35.7
MMA%	Mean ± SD	13.7 ± 5.2	14.5 ± 5.5	10.4 ± 4.5
25th Percentile		9.8	10.7	7.3
50th Percentile		13.1	13.9	10.1
75th Percentile		16.7	17.7	13.2
Min, Max		0, 34.7	0.5, 45.9	0.6, 36.7
DMA%	Mean ± SD	71.2 ± 8.7	76.9 ± 8.8	81.8 ± 8.1
25th Percentile		66.1	72.1	77.0
50th Percentile		71.8	77.8	82.5
75th Percentile		77.2	83.3	87.2
Min, Max		27.4, 92.9	32.4, 94.5	35.1, 97.9

Abbreviations: BMI, body mass index; DMA%, percentage of dimethylarsinic acid (DMA%); MMA%, percentage of monomethylarsonic acid; iAs%, percentage of inorganic arsenic

[a] BMI categories were defined as underweight (<18.5), normal (18.5 – 24.9), overweight (25 – 29.9), and obese (≥30)

Table 1. 2. Rare variants observed in *AS3MT* across cohorts

rs ID	Position (BP)	Major Allele/minor allele	Amino acid change	Number of carriers	Exon	Mutation	Polyphen[a]	SIFT[b]	CADD score[c]	Carrier(s) DMA% Distribution[d]
HEALS[e]										
rs144713814	104632301	C/A	Ser/Arg	1	4	Missense	Probably damaging	Deleterious	18.9	2nd decile
rs759484385	104632326	G/A	Val/Met	2	4	Missense	Probably damaging	Deleterious	22	1st decile
rs563943815	104634388	A/C	Gln/His	7	6	Missense	Possibly damaging	Deleterious	14.4	5th decile
rs35232887[h]	104634407	C/T	Arg/Trp	2	6	Missense	Probably damaging	Deleterious	21.8	1st decile
rs528680133	104634419	G/A	NA	1	6	Splice donor	--	--	26	1st decile
SHS[f]										
NA	104632249	G/A	Cys/Tyr	1	4	Missense	Possibly damaging	Deleterious	27.6	2nd decile
NA	104638178	T/TC	Leu/--	1	8	Frameshift insertion	--	--	--	2nd decile
rs139656545	104638615	G/A	Arg/His	1	9	Missense	Benign	Deleterious	23.8	4th decile
NA	104660404	A/T	Arg/--	2	11	Stop gained	--	--	35	1st decile
NHSCS[g]										
rs770395949	104634377	G/GAT	Lys/--	1	6	Frameshift insertion	Probably damaging	Deleterious	24.7	1st decile
rs35232887[h]	104634407	C/T	Arg/Trp	1	6	Missense	Probably damaging	Deleterious	21.8	6th decile
rs775931783	104636754	A/T	Glu/Val	1	7	Missense	Benign	Tolerated	22.8	5th decile
NA	104638627	C/T	Ala/Val	1	9	Missense	Possibly damaging	Deleterious	27.0	2nd decile
rs182365639	104660417	A/T	Asp/Val	Singleton	11	Missense	Benign	Tolerated	11.5	2nd decile

Base pair (BP), Sorting Intolerant from Tolerant (SIFT), Combined Annotation Dependent Depletion (CADD), percentage of dimethylarsinic acid (DMA%), Not available (NA). a Polyphen and b SIFT predict the impact of amino acid changes. c CADD scores of 0 to 10 are in the top 10% most deleterious, scores 10 to 20 are in the top 1%, CADD scores 20 to 30 are in the top 0.1%, and so on. d This value is based on the distribution of DMA% across carriers and non-carriers and corresponds to the average decile when a variant has >1 carrier. e The proportion of rare variant carriers in HEALS was 13 of 2,434 (0.5%). f The proportion of rare variant carriers in SHS was 5 of 865 (0.5%). g The proportion of rare variant carriers in NHSCS was 5 of 666 (0.7%). h Only one variant (rs35232887) was observed in multiple cohorts

Table 1. 3. Association between carrier status of *AS3MT* rare, protein-altering variants and arsenic metabolism phenotypes

Cohort	Sample size	Number of carriers	DMA%		MMA%		iAs%	
			β (95% CI)	P-value	β (95%CI)	P-value	β (95%CI)	P-value
Unadjusted[a]								
HEALS[c]	2,369	13	-9.7 (-14.4, -5.11)	$3.9x10^{-5}$	4.9 (2.2, 7.5)	0.0003	4.8 (1.3, 8.3)	0.007
SHS[d]	865	5	-10.2 (-17.6, -2.9)	0.006	3.9 (-0.8, 8.5)	0.11	6.4 (2.0, 10.7)	0.004
NHSCS	666	5	-9.3 (-16.3, -2.4)	0.009	5.8 (1.9, 9.8)	0.004	3.5 (-0.9, 8.0)	0.12
Meta-analysis[e]	3,900	23	-9.8 (-13.2, -6.3)	$2.2x10^{-8}$	4.9 (2.9, 6.9)	$1.3x10^{-6}$	4.9 (2.6, 7.2)	$3.6x10^{-5}$
Adjusted[b]								
HEALS[c]	2,369	13	-9.4 (-13.9, -4.8)	$5.2x10^{-5}$	4.7 (2.1, 7.4)	0.0005	4.6 (1.2, 8.0)	0.008
SHS[d]	865	5	-6.87 (-13.5, -0.21)	0.04	2.0 (-2.39, 6.45)	0.37	4.8 (0.63, 9.05)	0.02
NHSCS	666	5	-8.9 (-15.6, -2.16)	0.01	5.6 (1.76, 9.44)	0.004	3.3 (-1.09, 7.69)	0.15
Meta-analysis[e]	3,900	23	-8.7 (-11.9, -5.4)	$1.9x10^{-7}$	4.5 (2.5, 6.4)	$9.7x10^{-6}$	4.3 (2.0, 6.6)	0.0002

Percentage of dimethylarsinic acid (DMA%), percentage of monomethylarsonic acid (MMA%), percentage of inorganic arsenic (iAs%). a Models adjusted for sex and age (continuous). b Models adjusted for sex, age (continuous), and common variants in the 10q24.32 region (variant dosage score, e.g., 0,1, or 2) (HEALS: rs12573221 and 145537350, SHS: 10:104723620, rs10883846, and 10:104685493, and NHSCS: rs76255497). c Burden test was conducted on a subset of 2,369 individuals with genome-wide SNP data available to adjust for kinship using the mixed-effects model in GCTA. d Included additional adjustment for first 5 principal components derived from genome-wide SNP data. e Performed using fixed-effect models (test of heterogeneity p-values across metabolites > 0.10)

Figure 1. 2. Percent of arsenic metabolism phenotypes by carrier status of rare, protein-altering *AS3MT* variants across cohorts

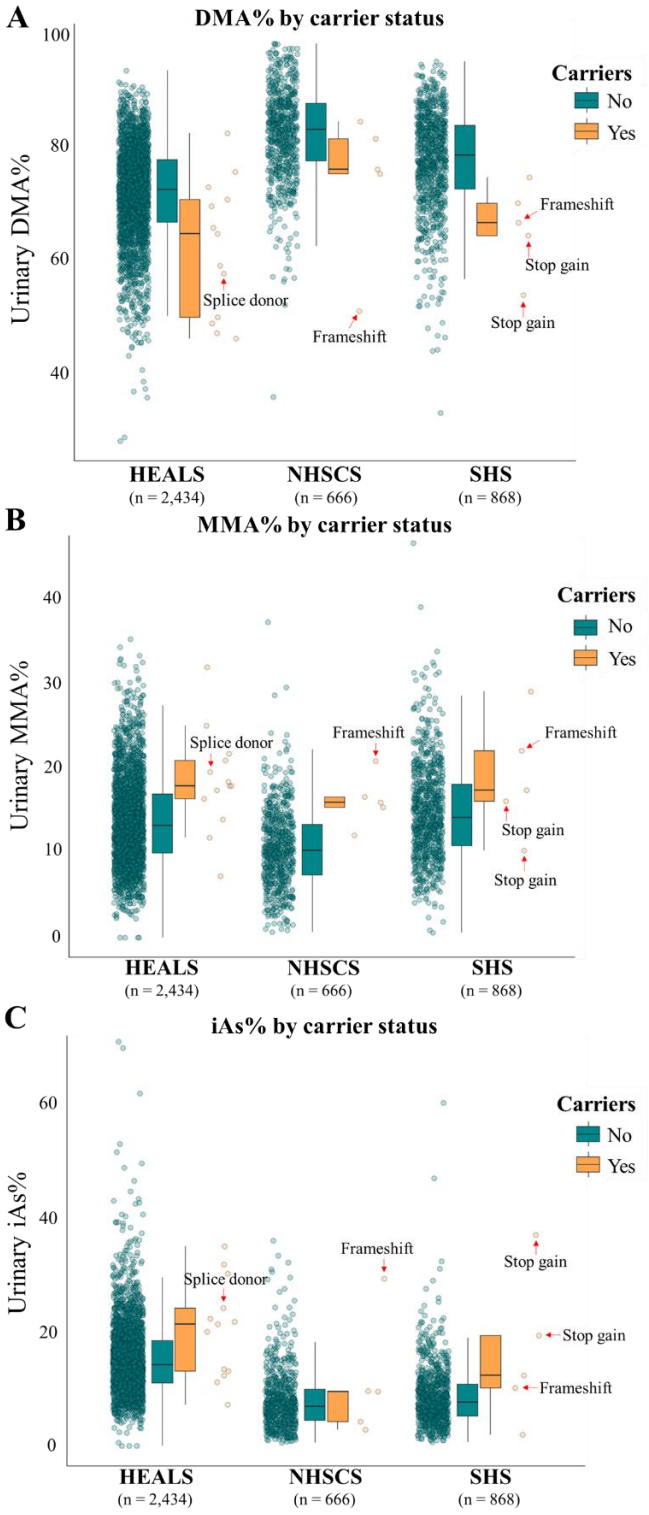


Table 1. 4. Association between rare, protein-altering variants in *AS3MT* and arsenic metabolism phenotypes across cohorts using a non-burden (SKAT[a]) testing method

Cohort	Sample size	Number of carriers	DMA%	MMA%	iAs%
			P-value	P-value	P-value
HEALS[b]	2,369	13	0.11	0.09	0.09
SHS	865	5	0.03	0.27	$4.3x10^{-6}$
NHSCS	666	5	0.005	0.08	0.003
Meta-analysis[c]	**3,900**	**23**	**0.002**	**0.1**	**$1.4x10^{-7}$**

Abbreviations: SKAT, Sequence kernel association test; DMA%, percentage of dimethylarsinic acid (DMA%); MMA%, percentage of monomethylarsonic acid; iAs%, percentage of inorganic arsenic

[a] SKAT is a score-based variance component test, under the null hypothesis all rare variants in the gene set have an effect = 0.

[b] SKAT analyses were implemented using the linear mixed model (EMMAX) built-in to the software to adjust for relatedness in HEALS.

[c] MetaSKAT software does not allow for adjust of a kinship matrix, thus SKAT meta-analyses were not adjusted for relatedness in HEALS cohort.

Table S1. 1. Individual level percentages of arsenic metabolism phenotypes among carriers of rare, protein-altering variants

Rare-variant Carrier	**DMA%**	**MMA%**	**iAs%**
HEALS (number of carriers = 13)			
rs144713814	64.1%	16.1%	19.8%
rs759484385			
Carrier 1	48.2%	17.1%	34.7%
Carrier 2	49.2%	19.2%	315.0%
rs563943815			
Carrier 1	72.3%	20.6%	7.1%
Carrier 2	70.1%	17.6%	12.3%
Carrier 3	65.1%	13.7%	21.2%
Carrier 4	75.0%	11.6%	13.4%
Carrier 5	68.9%	18.2%	13.0%
Carrier 6	81.8%	7.2%	11.0%
Carrier 7	57.0%	21.4%	21.5%
rs35232887			
Carrier 1	46.4%	31.5%	22.1%
Carrier 2	45.4%	24.7%	29.9%
rs528680133	58.4%	17.6%	23.9%
SHS (number of carriers = 5)			
10:104632249	69.5%	28.6%	1.9%
10:104638178	66.0%	21.8%	12.2%
rs139656545	74.0%	15.9%	10.1%
10:104660404			
Carrier 1	53.2%	10.1%	36.6%
Carrier 2	63.7%	17.1%	19.1%
NHSCS (number of carriers = 5)			
rs770395949	50.4%	20.5%	29.1%
rs35232887	84.0%	11.9%	4.1%
rs775931783	80.9%	16.3%	2.8%
10:104638627	74.7%	15.8%	9.5%
rs182365639	75.5%	15.2%	9.3%

Abbreviations: DMA%, percentage of dimethylarsinic acid (DMA%); MMA%, percentage of monomethylarsonic acid; iAs%, percentage of inorganic arsenic

Table S1. 2. Frequency of *AS3MT* variants in gnomAD

Variant	MAF	MAF in gnomAD	Population specific MAF in gnomAD[a]	Mutation type
HEALS				
rs144713814	0.0004	0.0004	0.0002	Missense
rs759484385	0.0008	0.0004	0.000006	Missense
rs563943815	0.002	0.00003	0.006	Missense
rs35232887	0.0008	0.001	0	Missense
rs528680133	0.0004	0.00004	0.0003	Splice donor
SHS				
10:104632249	0.001	NA	NA	Missense
rs139656545	0.001	0.001	NA	Missense
10:104638178	0.001	NA	NA	Missense
10:104660404	0.002	NA	NA	Stop gain
NHSCS				
rs35232887	0.002	0.001	0.0001	Missense
rs775931783	0.002	0.00002	0.00003	Missense
10:104638627	0.002	NA	NA	Missense
rs182365639	0.002	0.0001	0	Missense
10:104634377	0.002	0.00004	0.00006	Frameshift

Abbreviations: MAF, minor allele frequency; gnomAD, Genome Aggregation Database

[a] gnomAD South Asians and non-Finnish Europeans were used as reference for HEALS and NHSCS

Table S1. 3. Sensitivity analysis excluding Skin lesion cases and SCC cases

Cohort	**Sample size**	**Number of carriers**	**βcarrier**	**P-value**
HEALS	2,076	12	-8.8 (-13.5, -4.1)	0.0002
NHSCS	357	5	-9.3 (-16.1, -2.4)	0.009

Table S1. 4. Association between rare, protein-altering variants in *AS3MT* and arsenic metabolites across cohorts using burden, non-burden, and hybrid testing methods restricting to 1,285 unrelated individuals (r2<0.05) in HEALS[a]

Burden test[b]			**SKAT[b]**
β	**95% CI**	**P-value**	**P-value**
DMA%			
-12.0	-17.4, -6.6	$1.6x10^{-5}$	0.001
MMA%			
5.6	2.4, 10.6	0.0006	0.01
iAs%			
6.4	2.3, 10.4	0.002	0.006

Abbreviations: SKAT, Sequence kernel association test; CI, confidence intervals; DMA%, percentage of dimethylarsinic acid, MMA%, percentage of monomethylarsonic acid, iAs%, percentage of inorganic arsenic (iAs%).

[a] Two related carriers of rs563943815 were excluded, resulting in 11 rare variant carriers left in the analysis.

[b] Models include adjustment for sex, age (continuous), and dosage scores for common variants in 10q24.32 region (rs12573221 and rs145537350).

Table S1. 5. Rare, protein-altering variants in exons 5 and 10 of *AS3MT* across all population groups in the Genome Aggregation Database (gnomAD)

Chr	Position	Mutation type	Allele Count South Asian	Allele Count European	Allele Frequency	# Carriers
Exon 5						
10	104632861	missense	0	1	0.000008391	1
10	104632866	missense	0	0	0.000008352	1
10	104632873	missense	0	0	0.00004163	5
10	104632881	missense	0	23	0.000191	23
10	104632882	missense	0	1	0.000008301	1
10	104632886	missense	0	0	0.000008295	1
10	104632887	missense	0	3	0.00002488	3
10	104632888	frameshift	2	0	0.00001659	2
10	104632894	missense	1	0	0.000008288	1
10	104632894	frameshift	0	1	0.000008288	1
10	104632905	missense	0	1	0.000008285	1
10	104632933	missense	1	0	0.000008285	1
10	104632956	missense	0	0	0.00002486	3
10	104632959	stop gained	0	0	0.000008289	1
10	104632963	missense	1	0	0.000008291	1
10	104632978	missense	3	0	0.00002495	3
10	104632986	missense	0	0	0.00004166	5
10	104632986	missense	0	0	0.000008333	1
10	104632992	splice donor	1	0	0.000008342	1
Exon 10						
10	104650304	missense	0	0	0.000008576	1
10	104650305	missense	0	187	0.001756	205
10	104650313	missense	0	1	0.000017	2
10	104650318	frameshift	0	1	0.00000846	1
10	104650320	missense	2	0	0.00001689	2
10	104650332	missense	0	21	0.0001841	22
10	104650338	missense	1	0	0.000008334	1
10	104650348	missense	1	0	0.000008301	1
10	104650365	missense	0	1	0.000008288	1
10	104650374	missense	0	1	0.000008286	1
10	104650389	missense	0	0	0.00001657	2
10	104650393	missense	0	0	0.000008285	1
10	104650398	missense	0	1	0.000008285	1
10	104650435	splice donor	0	0	0.000008309	1

Table S1. 6. Rare, protein-altering variants in *AS3MT* across all population groups in the Genome Aggregation Database (gnomAD)

Chr	Position	Carriers	Allele Frequecy
10	104629577	1	0.000008282
10	104629588	1	0.000008283
10	104629588	1	0.000008283
10	104629592	1	0.000008284
10	104629602	1	0.000008285
10	104629603	7	0.000058
10	104629854	2	0.00001665
10	104629865	1	0.00000832
10	104629865	3	0.00002496
10	104629872	1	0.000008318
10	104629920	48	0.0003985
10	104629931	1	0.000008301
10	104629935	3	0.0000249
10	104629946	3	0.0000249
10	104629949	1	0.0000083
10	104629950	3	0.0000249
10	104629959	1	0.000008303
10	104629968	1	0.000008306
10	104632245	1	0.00000828
10	104632276	2	0.00001656
10	104632288	1	0.000008281
10	104632301	45	0.0003727
10	104632311	1	0.000008282
10	104632326	2	0.00001657
10	104632329	1	0.000008283
10	104632341	1	0.000008285
10	104632345	1	0.000008291
10	104632350	1	0.000008295
10	104632356	2	0.00001659
10	104634353	1	0.0000155
10	104634356	1	0.00001479
10	104634377	3	0.00003649
10	104634388	65	0.0007642
10	104634389	1	0.00001175
10	104634398	1	0.00001188
10	104634407	93	0.001127
10	104634419	3	0.00003853
10	104636712	1	0.000008281

Table S1. 6. , continued: Rare, protein-altering variants in *AS3MT* across all population groups in the Genome Aggregation Database (gnomAD)

10	104636720	1	0.000008281
10	104636726	1	0.000008281
10	104636738	4	0.00003312
10	104636745	2	0.00001656
10	104636754	2	0.00001656
10	104636773	1	0.00000828
10	104636775	1	0.00000828
10	104636780	1	0.00000828
10	104636791	5	0.00004141
10	104636791	1	0.000008281
10	104636794	20	0.0001656
10	104638136	5	0.00004143
10	104638141	2	0.00001657
10	104638148	2	0.00001656
10	104638163	1	0.000008281
10	104638168	1	0.000008281
10	104638174	1	0.000008281
10	104638186	3	0.00002484
10	104638201	1	0.000008281
10	104638207	1	0.000008281
10	104638210	1	0.000008282
10	104638223	2	0.00001656
10	104638225	1	0.000008282
10	104638229	13	0.0001077
10	104638231	2	0.00001657
10	104638264	1	0.000008292
10	104638267	7	0.00005805
10	104638611	1	0.000008288
10	104638614	3	0.00002486
10	104638615	116	0.0009612
10	104638646	2	0.00001657
10	104638650	1	0.000008283
10	104638651	1	0.000008283
10	104638683	2	0.00001657
10	104638704	1	0.000008285
10	104638732	1	0.0000083
10	104638735	6	0.00004984
10	104638746	1	0.000008338
10	104660351	1	0.000008284

Table S1. 6., continued: Rare, protein-altering variants in *AS3MT* across all population groups in the Genome Aggregation Database (gnomAD)

10	104660357	1	0.000008283
10	104660363	2	0.00001656
10	104660374	5	0.00004141
10	104660392	1	0.000008281
10	104660393	1	0.000008281
10	104660397	1	0.000008281
10	104660415	4	0.00003312
10	104660417	17	0.0001408
10	104660422	1	0.000008281
10	104660434	2	0.00001656
Number of carriers and proportion		**561**	**561 of 60,706 = 0.9%**

Table S1. 7. Estimated risk of developing lung and bladder cancer for a 4.5% increase in MMA%[a]

Outcome	**Predictor[b]**	**Original study OR**	**Original study logOR**	**logOR x $\beta_{carrier}$**	**Rare variant carrier OR**
Lung Cancer	MMA%	1.11	0.10	0.47	1.50
Bladder Cancer	MMA%	1.04	0.04	0.18	1.17

Abbreviations: OR, Odds ratio; MMA%, percentage of monomethylarsinic acid
[a] Percentage increase in MMA% (4.5%) is derived from the adjusted meta-analyses effect estimates reported in Table 3. [b] MMA% was treated as a continuous variables in original study.

Table S1. 8. Estimated pooled OR for skin lesions using IVW method

Summary statistics	***AS3MT* SNPS[a]**		***FTCD* SNP[b]**
	rs9527	**rs11191527**	**rs61735836**
$\beta_{DMA\%}$	-3.60	-2.22	-5.10
SE	0.47	0.34	0.51
$OR_{skin\ lesion\ \sim\ DMA\%}$	1.47	1.15	1.25
logOR	0.39	0.14	0.22
SE	0.07	0.05	0.09
Inverse-variance weighted Mendelian Randomization[c]			
IVW pooled OR	**95% CI**	**logOR**	**95% CI**
1.07	(1.04, 1.10)	0.07	(0.05, 0.09)

Abbreviations: OR, Odds ratio; SE, standard error, CI, confidence interval, DMA%, percentage of dimethylarsinic acid, IVW, inverse-variance weighted

[a]Summary statistics were derived from (Pierce et al. 2013, http://dx.doi.org/10.1093/ije/dyt182).

[b]Summary statistics were derived from (Pierce et al. 2019, https://doi.org/10.1371/journal.pgen.1007984).

[c]Estimates were calculated using the Likelihood-based method for combining summarized data on multiple genetic variants into a single causal estimate (Burgess et al. 2013, https://doi.org/10.1002/gepi.21758)

Table S1. 9. Estimated risk of developing skin lesions for an 8.7% decrease in DMA%[a]

Outcome	**Predictor[b]**	**Original study OR**	**Original study logOR**	**logOR x $\beta_{carrier}$**	**Rare variant carrier OR**
Skin Lesions	DMA%	1.07[c]	0.07	0.60	1.82

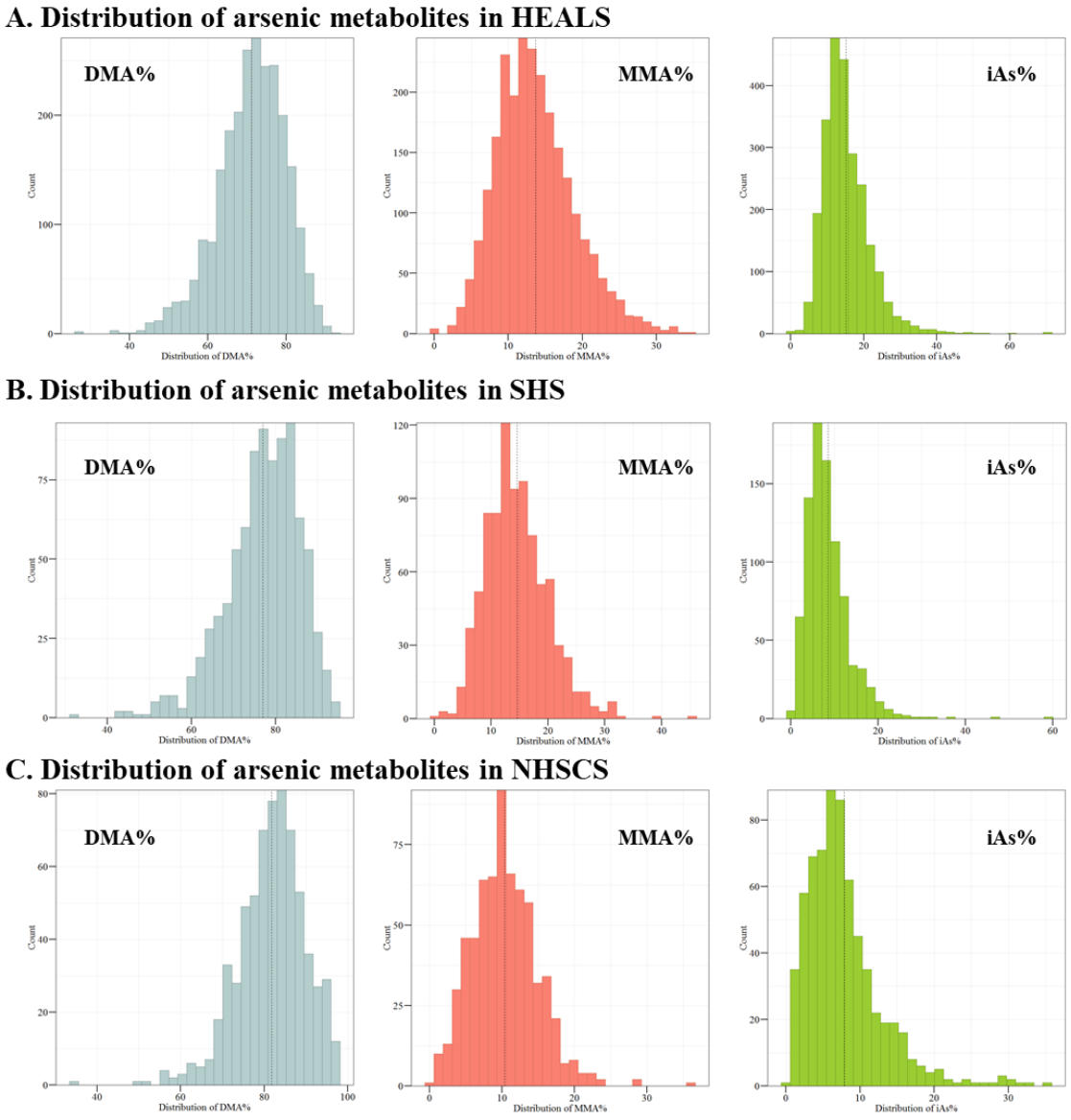
Abbreviations: OR, Odds ratio; DMA%, percentage of dimethylarsinic acid

[a] Percentage decrease in DMA% (8.7%) is derived from the adjusted meta-analyses effect estimates reported in Table 3.

[b] DMA% was treated as a continuous variables in the original studies original study.

[c] Details describing how this OR was estimated are reported in Table S8.

Figure S1. 1. Distribution of the percentage of arsenic metabolites across cohorts



CHAPTER 3

THE ASSOCIATION BETWEEN GENETIC DETERMINANTS OF ARSENIC METABOLISM EFFICIENCY AND HYPERTENSION RISK

3.1 INTRODUCTION

Inorganic arsenic (iAs) is a human toxicant and carcinogen, with more than 230 million individuals exposed worldwide primarily through drinking water and less often through dietary sources (e.g., rice and grain products) [2, 3]. Chronic exposure to levels of iAs above the World Health Organization (WHO) safety standard (>10μg/L) in drinking water has been recognized to pose a significant risk of adverse outcomes, including increased risk for cardiovascular disease [37, 144-146]. Several epidemiologic studies of arsenic exposed populations from different ancestral backgrounds also reported associations between arsenic exposure and elevated hypertension risk [37, 147-150]. However, some studies have also reported a lack of association at moderate to high levels of exposure [53] and low levels of exposure [54]. Additionally, the association between arsenic metabolism efficiency and hypertension remains inconclusive [9, 55].

Metabolism of iAs involves sequential reduction and methylation reactions, occurring primarily in the liver, and is a detoxification mechanism that facilitates urinary excretion of arsenic. Overall, conversion of iAs to dimethylated arsenic (DMA), the end product of iAs metabolism is considered protective, as DMA is more efficiently cleared from their body than iAs or MMA. Individuals who can efficiently produce DMA will have lower biologically effective dose.

Arsenic metabolism efficiency (AME) can be represented by the percentages of the arsenic species present in urine as DMA (i.e. DMA%). Genetic variation in the 10q24.32 region and the FTCD gene are known to influence arsenic metabolism and arsenic-related outcomes [22, 24, 26].

Genetic determinants of AME can be used to estimate its impact of AME on hypertension risk using Mendelian Randomization (MR) methods. MR is an approach used to infer causal relationships between a risk factor observed to be associated with a health outcome. Genetic variants are used as instrumental variables (i.e., predictors) of the risk factor [151]. Because genetic variants are 1) randomly and independently passed from parents to offspring and 2) and time-invariant, MR estimates are not vulnerable to bias due to unobserved cofounders or reverse causation. Thus, this approach enables stronger causal inference in determining whether an individual's ability to metabolize arsenic impacts hypertension risk. In the current study, we leveraged established genetic determinants of AME to estimate their effect on hypertension risk using a prospective cohort with exposure to arsenic and repeated measures of blood pressure.

3.2 METHODS

3.2.1 Study participants

This study uses data from two studies of arsenic-exposed individuals from Bangladesh: the Health Effects of Arsenic Longitudinal Study (HEALS) and the Bangladesh Vitamin E and Selenium Trial (BEST). HEALS is a prospective study designed to investigate health outcomes associated with chronic arsenic exposure through well drinking water in Araihazar, Bangladesh [152]. Briefly, 11,746 participants of ages 18 to 75 years old were recruited between October 2000 and May 2002. An expansion of the HEALS cohort (ACE) added 8,287 participants to

HEALS between 2006 and 2008. Blood and urine samples were collected from participants at baseline [110]. BEST is a 2x2 factorial, double-blind, randomized chemoprevention trial designed to assess the effect of vitamin E and selenium supplementation on non-melanoma skin cancer risk among 7000 adults. BEST participants were recruited from Araihazar, Matlab, and other surrounding districts of Bangladesh. Details of the BEST study methods have been described previously [153]. For current study, we selected 7,895 participants (4,824 from HEALS, 1,081 from ACE, and 1,990 from BEST) with complete genome-wide SNP data and baseline measures of blood pressure and hypertension case status.

3.2.2 Measurement of total urinary arsenic concentration

Baseline total urinary arsenic concentration was measured by graphite furnace atomic absorption using the Analyst 600 graphite furnace system with a detection limit of 1μg/L for 11,224 HEALS participants [110]. Urinary creatinine was measured by a colorimetric Sigma Diagnostics Kit (Sigma, St. Louis, MO, USA) for adjustment of total urinary arsenic concentration.

3.2.3 Measurement of arsenic metabolites

Arsenic species (iAs^{III}, iAs^{V}, MMA, and DMA) from baseline urine samples were measured for 4,814 HEALS participants [19]. We performed speciation analyses of arsenic metabolites using high-performance liquid chromatography (HPLC) [114] followed by detection using ICPMS [110]. The LOD was 1μg/L for iAs^{III} , iAs^{V}, MMA, DMA.

We summed iAs^{III} and iAs^{V} to obtain total iAs. Each arsenic species (iAs, MMA and DMA) is expressed as the percentage sum of each of these species (iAs+MMA+DMA). We excluded arsenocholine (AsC) and arsenobetaine (AsB), non-toxic forms of organic arsenic that

can be detected via ICPMS, from our analyses. We used DMA% as our primary measure of AME.

3.2.4 Blood pressure measurements and ascertainment of hypertension case status

In both HEALS and BEST, two systolic (SBP) and diastolic (DBP) blood pressure measures were taken by trained clinicians at baseline and at each follow-up using an automatic sphygmomanometer (HEM 712-C; Omron Healthcare GmbH, Hamburg, Germany) [147]. Details regarding the number of individuals with BP measurement at each time point are shown in **Supplementary Table 2.1**. HEALS participants had BP follow-up data for up to 4 time periods in addition to baseline and participation rate from baseline to their last follow-up was 83.4%. For ACE participants, we had follow-up data for up to two time periods in addition to baseline measurements and participation rate from baseline to follow-up 2 was 87.7%. BEST participants had BP data for up to 3 time periods in addition to baseline and participation rate from baseline to follow-up 3 was 90.3%. Measurements were taken while participants were in a seated position at rest for 5 minutes and using the upper left arm. For the current study, we took the arithmetic mean of the two measures taken at each time point. Information on medication use was also collected at baseline and each follow-up and then sorted into 44 medication categories. Anti-hypertensive (or blood pressure lowering) medications were used to adjust the observed values for SBP and DBP using the following formulas: SBP + 15 mmHg and DBP + 10 mmHg [154, 155]. Hypertension cases were defined as SBP $\geq$ 130 mmHg or DBP $\geq$ 85 mmHg, and among individuals taking anti-hypertensive medication we used adjusted SBP and DBP.

3.2.5 DNA extraction and genotyping

Using the HumanCytoSNP-12 v2.1 chips 5,499 HEALS and BEST DNA samples were genotyped and 257,747 SNPs were available post quality control. An additional 2,612 HEALS

participants were genotyped using Illumina Infinium Multi-ethnic EUR/EAS/SAS SNP arrays. Post QC, we retained high-quality data for 2,542 of the 2,612 samples and 855,984 SNPs remained. Genomic DNA for all HEALS samples were extracted from clot blood using the Flexigene DNA kit from Qiagen. For all BEST samples, we extracted DNA from whole blood using the QIAamp 96 DNA Blood Kit from Qiagen. Details of the blood collection, DNA extraction, genotyping, and quality control for genome-wide SNP data have been described previously [22, 24].We imputed these two batches (1) HEALS and BEST samples genotyped using HumanCytoSNP-12 v2.1 chips, 2) HEALS samples genotyped using the Illumina Infinium Multi-ethnic SNP array separately using the HRC reference panel in the Michigan Imputation Server. Post imputation, the number of high-quality (imputation score $r^2 > 0.3$) overlapping SNP across batches was 7,981,471.

3.2.6 Statistical Analyses

Unlike traditional observational studies where the association estimates can be biased due to unmeasured confounders and reverse causation, the MR approach is robust to confounders of the exposure-outcome relationship [156, 157]. This approach uses inherited genetic variants as instrumental variables (IV) (i.e. proxies) that predict a modifiable exposure [158]. The IV represents a component of variation in the exposure that is unaffected by the outcome or confounders. Additionally, MR studies leverage the random assortment of alleles from parent to offspring; thereby avoiding potential biases (i.e. environmental confounders) and robustly identifying causal associations [156].

In the current study, we used MR to estimate the causal effect of AME on HTN risk using two MR methods. We utilized genetic data from the HEALS/ACE and BEST cohorts with existing arsenic metabolite measurements (for a subset of participants) and HTN status, totaling

7,895 participants. Prior studies of these Bangladeshi cohorts have reported two independent SNPs near *AS3MT* [22] and one exonic variant in *FTCD* [26] that are robustly associated with AME. In the current study, we updated these association analyses, and identified three independent SNPs near *AS3MT* (rs4919690, rs117978529, and rs191177668) and confirmed the exonic variant in *FTCD* (rs61735836). A causal diagram depicting the MR approach to assess the effect of AME on hypertension risk is shown in **Figure 2.1**. *AS3MT* SNPs were identified through genome-wide association analyses of DMA% (i.e. AME) adjusting for age, sex, genotyping batch, and relatedness (through mixed-effects modeling using the GCTA software) among 3,968 HEALS participants with measured arsenic metabolites. To identify the secondary independent association signal, we included the primary signal as a covariate and repeated the association analysis. We performed a similar analysis, adjusting for the primary and secondary independent signals, to identify any remaining independent signals. The results of these analyses are shown in **Supplementary Table 2.2**. To confirm the association signal in the *FTCD* gene, we conducted exome-wide association analyses of DMA% among 1,660 HEALS participants with measured DMA%, adjusting for age, sex, and relatedness using the GCTA software (**Table S2.2**). MR methods were applied using these four independent genetic determinants of AME as IVs, where each IV is coded as 0, 1, or 2 high-efficiency alleles.

The first MR method we applied was the two-stage sequential regression. In the first-stage linear regression, urinary DMA% was regressed on four AME-related SNPs (as explained above). These analyses were conducted on a subset of individuals with complete arsenic metabolite and genotype data (for *AS3MT* SNPs we used subset of 3,698 individuals and for the *FTCD* SNP we used a subset of 1,660 individuals). Through this stage, we obtained effect size estimates for each IV-SNP (**Table S2.2**). These effect estimates were then used to generate

genetically predicted values of DMA% for 7,895 individuals (similar to a polygenic risk score). In the second stage, we assessed the association between genetically predicted DMA% and hypertension status using logistic regression adjusted for age and sex. We conducted linear regression analyses adjusted for age and sex using blood pressure measures as secondary outcomes.

The second MR method we applied was the inverse-variance weighted (IVW) meta-analysis as implemented in the "MendelianRandomization" package in R [133]. The IVW method uses summary statistics obtained from GWAS. Two sets of estimates were generated, one set represents the association between each genetic variant and urinary DMA%, and the other set represents the association between each genetic variant and hypertension status (separate analyses were conducted with BP measures). The IVW combination of the ratio of these two sets of estimates in a meta-analysis represents the causal effect of genetic variants on hypertension risk.

We tested the association between predicted AME and longitudinal measures of blood pressure using a linear mixed models. In this approach, we modeled predicted DMA% as a fixed-effect and repeated measures of BP as the outcome. We adjusted for sex and age (at each time point) as fixed-effects. Individuals were modeled as random-effects (each person has a person-specific intercept), in order to allow individuals to have different means for the BP phenotypes. In this analysis, we do not include time in the model, but we do account for age, which captures age-related changes in BP. We conducted similar analysis stratified by observed median age (across all follow-ups), and compared the association between predicted AME and longitudinal measures of BP in individuals <44 years of age and ≥44 years of age (adjusted for age and sex).

Additionally, we conducted mixed-effects analysis with measured DMA% (measured at baseline), which represents the observational association estimate complimentary to the MR analyses described above, among 4,814 individuals with arsenic metabolite data. We modeled measured DMA% as a fixed-effect and adjusted for the following fixed-effects: age, sex, BMI, ever smoker, ever betel use, land ownership, TV ownership, and educational length. Similarly, individuals were modeled as random-effects. We also conducted median age stratified analyses (<44 years of age and ≥44 years of age).

While many of the 4,814 participants were randomly selected for arsenic species measurement in baseline urine, a sizable fraction (~50%) were selected based on outcomes they experienced at subsequent follow-up visits (i.e., skin lesions, respiratory symptoms, and cardiovascular conditions), in a case-cohort fashion. So while this group was relatively healthy at baseline, they were not randomly selected. For this reason, we conducted sensitivity analyses restricting the mixed-effects approach to a subset of randomly sampled individuals (n = 1,606), which includes 1,226 non-cases, 41 CVD cases, 135 respiratory outcomes, 56 diseased, and 148 skin lesions (62 prevalent and 86 incident skin lesions).

Lastly, we performed exploratory gene-environment (GxE) analyses where we tested the interaction between predicted DMA% and urinary arsenic concentration (adjusted for creatinine) on hypertension status (and blood pressure measures as secondary analyses). Baseline urinary arsenic concentration was log-transformed to follow a normal distribution. We conducted GxE analysis treating urinary arsenic concentration and predicted DMA% as continuous variables and tertiles treated as ordinal variables among 4,775 HEALS participants. We restricted this analysis to HEALS participants because their baseline arsenic exposure represents historical arsenic exposure (prior to any arsenic mitigation efforts). Most BEST participants were likely aware of

their exposure to arsenic prior to baseline and may have modified their exposure levels prior to baseline. Because our sample includes some related individuals, we conducted sensitivity analyses restricted to a subset of unrelated individuals (relatedness coefficient $r^2 < 0.05$). Although, theoretically, predicted DMA% was derived from estimates adjusted for relatedness.

3.3 RESULTS

Baseline characteristics of the 7,895 participants included in our analyses are described in **Table 2.1**. On average ACE participants were younger at baseline (37 years) than HEALS (39 years) and BEST participants (43 years). We observed various male to female ratios across each cohort, in HEALS and BEST we found slightly more males than females, while in ACE we found 24% more females than males. The majority of Bangladeshi individuals across these cohorts fall within the underweight and normal ranges of BMI, with only a minority of individuals classified as overweight or obese. Mean SBP and DBP at baseline were lowest in HEALS compared to ACE and BEST (**Figure S2.1-S2.2**). At baseline, we observed a hypertension prevalence of 26% and identified a total of 2,053 hypertension cases (1,167 in HEALS, 275 in ACE, and 611 in BEST). Follow-up measures of SBP and DBP did not differ meaningfully (**Figure S2.3-S2.4**). In the current study, we computed genetically predicted DMA% (**Figure S2.5**) for 7,895 individuals using individual-level genotype high-efficiency dosage counts for 3 independent *AS3MT* SNPs and one *FTCD* SNP (**Table S2.2**). Among the 7,895 individuals included in this study, the frequency of the high-efficiency alleles for rs4919690, rs191177668, and rs61735836 were above >87% and one SNP (rs117978529) had a high efficiency allele frequency of 14%.

Using the two-stage sequential regression MR approach, we did not find an association between predicted DMA% and hypertension risk (OR = 1.01, P = 0.51) (**Table 2.2**). When analyzing SBP and DBP as outcomes, we found a one percentage unit increase in predicted DMA% was associated with 0.09 mmHg lower SBP (P = 0.21) and 0.05 mmHg lower DBP (P = 0.34), though these associations did not meet the $P < 0.05$ threshold. The inverse-variance weighted meta-analysis MR method showed evidence of an association between predicted DMA% and hypertension risk (OR = 1.03, P = 0.02) (**Table 2.2** and **Figure 2.2**), suggesting a one percentage unit increase in predicted DMA% is associated with a 3% increase in hypertension risk. These results were not directionally consistent when analyzing SBP (β = -0.05, P = 0.53) and DBP (β = -0.02, P = 0.66) as outcomes.

We incorporated longitudinal BP data from HEALS, ACE, and BEST (**Table S2.1**) using mixed-effects models to estimate the association between genetically predicted DMA% and repeated measures of SBP and DBP, adjusted for age and sex. We found predicted DMA% was not strongly associated with repeated measures of SBP (β = 0.03, 95% CI: -0.10, 0.15) or DBP (β = 0.01, 95% CI: -0.06, 0.08) (**Table 2.3**). Additionally, we conducted analyses stratified by observed median age (> 44 years of age and ≤ 44 years of age) and found no clear association between predicted DMA% and repeated measures of BP with very similar results between the two age groups (**Table 2.3**). Mixed-effects analyses using measured DMA% (n = 4,814) as the predictor (rather than genetically-predicted DMA%) also showed lack of association with repeated measures of SBP (β = 0.04, 95% CI: -0.02, 0.10) and DBP (β = -0.02, 95% CI: -0.05, 0.02) after adjusting for age, sex, BMI, ever smoker status, ever betel use, land ownership, TV ownership, and education (**Table S2.3**). These results were similar in models stratified by median age. We also conducted a sensitivity analysis restricted to a random sub-cohort of 1,555

individuals (these are randomly samples individuals) because a large portion of the individuals with measured DMA% were selected in case-cohort fashion. The results of this analysis confirm a lack of association between measured DMA% and repeated measures of SBP (β = 0.07, 95% CI: -0.02, 0.15) and DBP (β = -0.006, 95% CI: -0.06, 0.04), after adjusting for all measured confounders.

Lastly, we conducted an exploratory GxE analyses to measure the effect of the interaction between arsenic exposure (i.e., measure urinary arsenic) and predicted DMA% on hypertension risk in 4,775 HEALS participants. We did not observe strong evidence of interaction between continuous urinary arsenic and predicted DMA% (interaction OR = 1.01, P = 0.68) (**Table S2.4**). Interaction analyses between continuous urinary arsenic and predicted DMA% with SBP and DBP as outcomes also showed an inconsistent direction of effect and lack of interaction (β = 0.12, P = 0.20 and β = 0.92, P = 0.69, respectively). We conducted additional GxE analyses where we interacted tertiles of urinary arsenic (representing low, medium, and high arsenic exposure) and tertiles of predicted DMA% (representing low, medium, and high metabolism efficiency) on hypertension status, using low arsenic exposure – high metabolism efficiency as the reference category. These analyses showed a lack of interaction and inconsistencies in the direction of association (**Table S2.4** and **Figure S2.6**). Sensitivity analyses restricted to 1,496 unrelated HEALS participants ($r^2 < 0.05$) testing the interaction between continuous and tertiles of urinary arsenic and predicted DMA% also showed a lack of interaction (P > 0.05) (**Table S2.5**).

3.4 DISCUSSION

In this study, we used Mendelian randomization and mixed-effects effects models to assess the association between genetically predicted DMA% (i.e., a proxy for arsenic metabolism efficiency) and blood pressure phenotypes among individuals living in regions of Bangladesh affected by arsenic exposure in drinking water The IVW MR method suggest a unit increase in predicted DMA% is associated with a 3% increase in hypertension risk (**Table 2.2 and Figure 2.2**). MR analyses with SBP and DBP as outcomes were not consistent and did not reach $P < 0.05$ threshold. Similarly, we did not observe an association between genetically predicted DMA% and repeated measures of SBP and DBP (**Table 2.3**). Observational analyses using measured DMA% showed also showed no clear association with repeated measures of SBP and DBP. Exploratory GxE analyses consistently showed a lack of interaction between predicted DMA% and chronic arsenic exposure (using measured urinary arsenic concentration) on hypertension risk (**Table S2.4-S2.5**).

An association between arsenic exposure and increased hypertension risk has been reported in some prior publications [37, 147-150], though some have reported a lack of association at high and low levels of exposure [53, 54]. Additionally, only a handful of studies have attempted to study the impact of AME on hypertension risk, and the results are inconsistent. A systematic review of the association between AME and cardiovascular disease risk [9] examined 5 hypertension studies and only one found a positive association with DMA% [56], while the other four reported no association. The association Mendez et al. reports and the one we observed suggests that individuals with higher proportions of DMA% have increased risk for hypertension, which is inconsistent with our hypothesis that higher DMA% reflects enhanced clearance of arsenic from the body. This association has been observed before with other

cardiometabolic outcomes, and one explanation is that the toxic trivalent form of DMA, DMA^{III}, may be driving this association [159]. Due to the highly unstable nature of the pentavalent and trivalent forms of methylated arsenic metabolites, most epidemiological studies are not able to get individual measurements for each and rather obtain a sum of the two. This limits our ability to understand the individual effect of pentavalent and trivalent methylated arsenic metabolites on health risk associated with iAs exposure.

A recent MR study (not included in the Kuo et al. meta-analyses) that used previously reported association estimates for genetic variants in *AS3MT* (two independent SNPs) and *FTCD* (one SNP) genes found an inverse association between AME and hypertension risk among adults in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort [55], suggesting efficient arsenic metabolizers have decreased risk for hypertension. However, arsenic exposure in HCHS/SOL participants was not measured directly, but was inferred based on high rice consumption. In the current study, we used a similar MR approach and implemented two MR methods to assess the effect of genetically predicted DMA% on hypertension risk and BP measures.

Our study has several strengths including a sample size of close to 8,000 individuals. We were also able to leverage individual urinary arsenic metabolite measurements for 3,968 individuals with genome-wide SNP data to conduct updated association analysis of DMA% where identified three independent variants near the *AS3MT* gene, which we used as IVs in addition to the *FTCD* SNP. Additionally, the availability of urinary arsenic metabolites allowed us to conduct complimentary analyses and compare to results using predicted DMA%. Lastly, the Bangladeshi cohorts in our study represent a population with a historical exposure to high levels of iAs, which may have powered us to detect suggestive evidence that increased

genetically predicted DMA% is associated with increased hypertension risk through the IVW MR method. We also leveraged longitudinal measures of BP phenotypes to assess an association between predicted DMA% and over 10-years of follow-up BP data using a mixed-effects approach.

In summary, we leveraged genotype data for four arsenic metabolism efficiency (AME) -related SNPs in 7,895 individuals with a historically high level of arsenic exposure to evaluate the effect of genetically predicted AME on hypertension risk. Using the IVW MR method, we provide suggestive evidence that high AME is associated with increased risk of hypertension. However, this association was null in analyses of genetically predicted DMA% (using two-stage sequential regression MR) and repeated measures of SBP and DBP. We also provide evidence that predicted AME does not modify the effect of arsenic exposure on hypertension risk. Our results, along with prior analyses, provide evidence of a suggestive weak association between AME and hypertension risk (or lack thereof).

3.5 APPENDIX

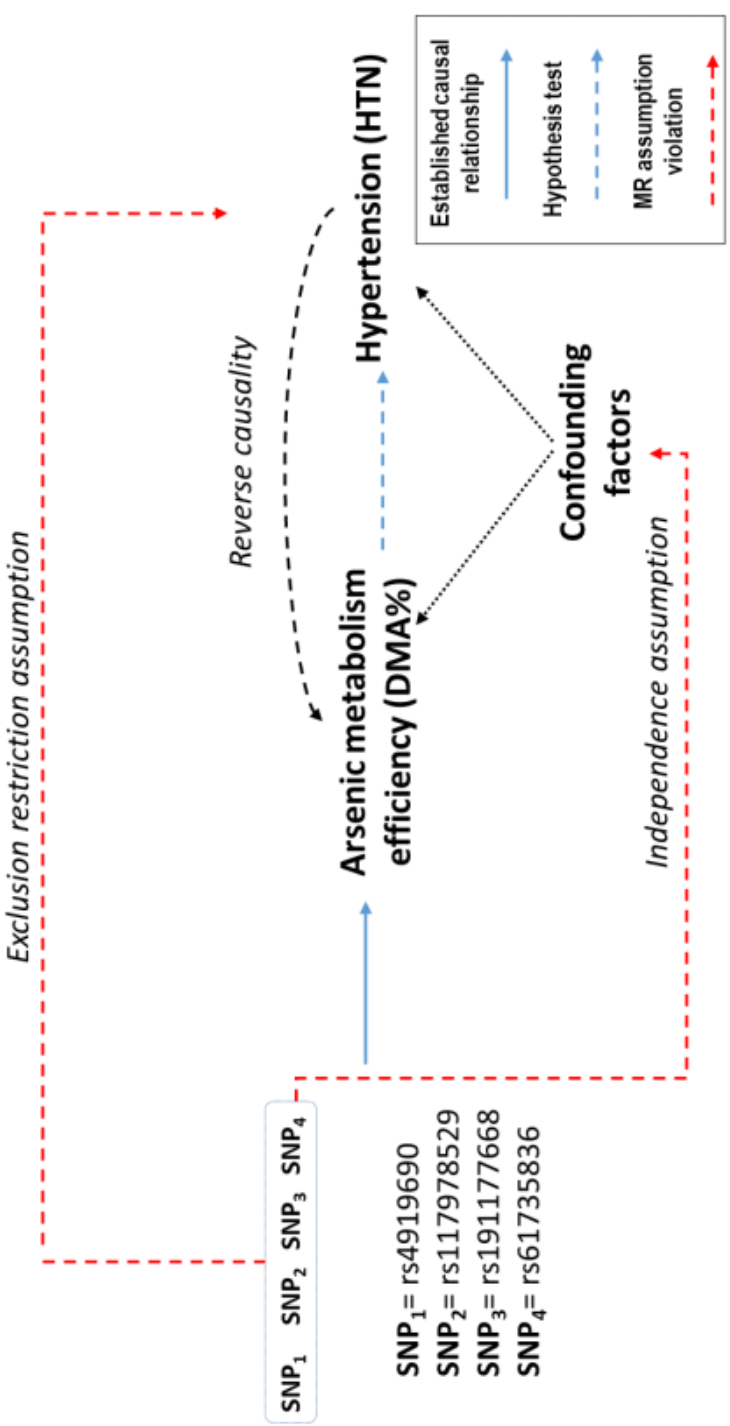


Figure 2. 1. Causal diagram depicting the Mendelian Randomization approach to assess the causal effect of arsenic metabolism efficiency on hypertension risk

Table 2. 1. Baseline characteristics of 7,895 participants stratified by cohort

Characteristics		**HEALS (n = 4,824)**	**ACE (n = 1,081)**	**BEST (n = 1,990)**
Age	mean (sd)	39 (11)	37 (11)	43 (11)
Sex	Male	2,517 (52%)	413 (38%)	1,075 (54%)
	Female	2,307 (48%)	668 (62%)	915 (46%)
BMI[a]	Underweight	1,959 (41%)	443 (41%)	754 (38%)
	Normal	2,508 (52%)	564 (52%)	1,058 (53%)
	Overweight	290 (6%)	65 (6%)	166 (8%)
	Obese	29 (1%)	6 (1%)	6 (1%)
Systolic blood pressure (mmHg)[b]	mean (sd)	116 (19)	120 (16)	118 (17)
Diastolic blood pressure (mmHg)	mean (sd)	74 (12)	76 (10)	77 (11)
Hypertension cases[c]		1,167	275	611

Abbreviations: BMI, body mass index

[a] BMI categories were defined as underweight (<18.5), normal (18.5 – 24.9), overweight (25 – 29.9), and obese (≥30).

[b] Systolic and diastolic blood pressure(SBP and DBP) measures were adjusted for use of antihypertensive medication using the following formulas: adjusted SBP = average SBP + 15 and adjusted DBP = average DBP + 10

[b] Hypertension cases were defined as SBP ≥ 130 mmHg or DBP ≥ 85 mmHg

Table 2. 2. Association between genetically predicted arsenic metabolism efficiency and hypertension phenotypes using two Mendelian Randomization methods

Outcome	**β or OR[a]**	**SE**	**P-value**
Two-stage sequential regression[b]			
Hypertension status	1.01	1.01	0.51
SBP	-0.09	0.07	0.21
DBP	-0.05	0.05	0.34
Inverse-variance weighted meta-analysis[c]			
Hypertension status	1.03	1.01	0.02
SBP	-0.05	0.07	0.53
DBP	-0.02	0.05	0.66

Abbreviations: OR, odds ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure

[a] An OR was reported for the hypertension status (binary outcome) and a β was reported for SBP and DBP (continuous outcomes).

[b] First-stage regression included 3,968 individuals for association between DMA% and AS3MT SNPs and 1,660 individuals for the association between DMA% and FTCD SNP. Second-stage regression included 7,895 individuals.

[c]Association between IV-SNPs (rs4919690, rs117978529, rs191177668, and rs61735836) and hypertension phenotypes was estimated among 7,895 individuals. Association between IV-SNPs and measure DMA% was estimated among 3,968 individuals.

Figure 2. 2. Association between IV-SNPs (predicted DMA%) and hypertension risk using the inverse-variance weighted (IVW) meta-analysis Mendelian Randomization method

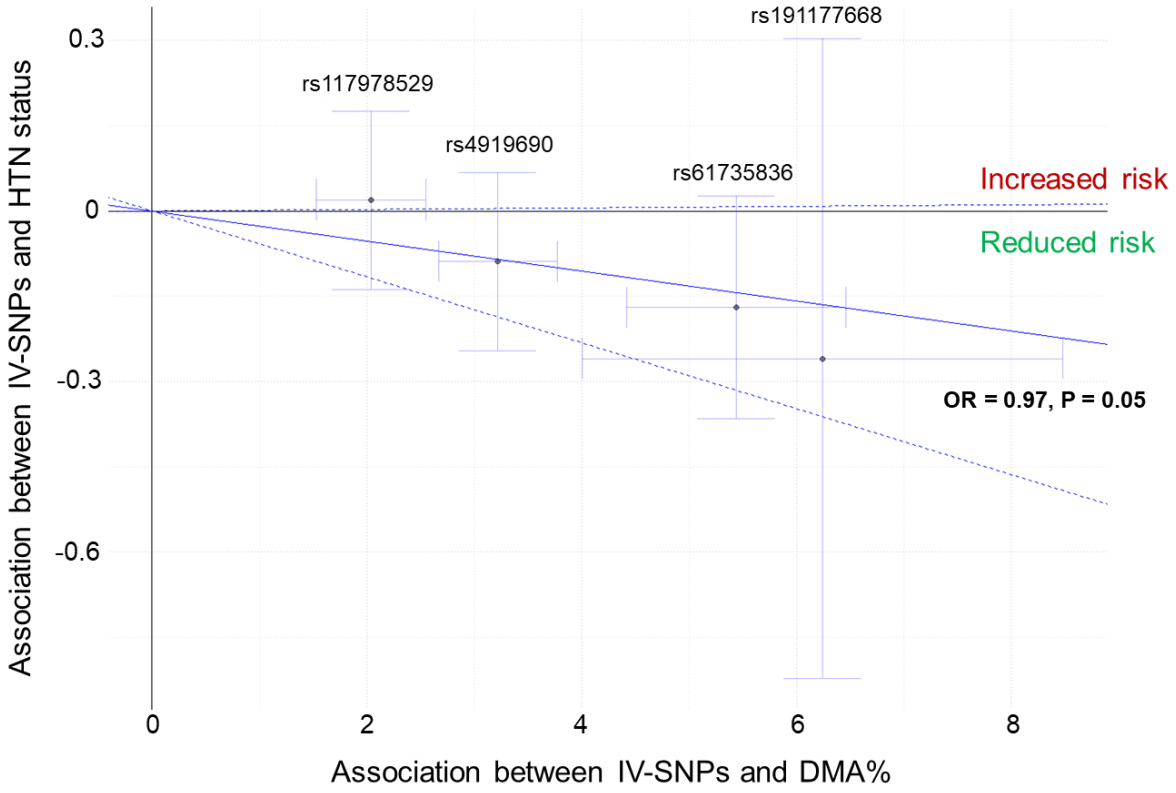


Table 2. 3. Association between genetically predicted arsenic metabolism efficiency and longitudinal measures of blood pressure

Outcome	Sample Size	Predictor	β	95% CI
Analysis based on genetic determinants of AME				
SBP	7,895	Predicted DMA%	0.03	(-0.10, 0.15)
DBP			0.01	(-0.06, 0.08)
Below median age (<44yrs)[a]				
SBP	4,704	Predicted DMA%	0.05	(-0.08, 0.19)
DBP			0.03	(-0.06, 0.12)
Above median age (≥44yrs)[a]				
SBP	4,574	Predicted DMA%	0.04	(-0.14, 0.23)
DBP			0.003	(-0.10, 0.11)

Abbreviations: CI, confidence intervals; AME, arsenic metabolism efficiency; SBP, systolic blood pressure; DBP, diastolic blood pressure

[a]Median age was calculated based on observed age across baseline and follow-ups.

Table S2. 1. Individuals with systolic and diastolic blood pressure measures at baseline and each follow-up by cohort

Cohort	Baseline	Follow-up 1	Follow-up 2	Follow-up 3	Follow-up 4
HEALS[a] (2000 – 2014)	4,824	4,574	4,378	4,272	4,022
ACE[b] (2007 – 2014)	1,081	955	948	--	--
BEST[c] (2006 – 2018)	1,990	1,889	1,811	1,796	--
Total	7,895	7,418	7,137	6,068	4,022

[a] Participation rate from baseline to last follow-up in HEALS was 83%.
[b] Participation rate from baseline to last follow-up in ACE was 88%.
[c] Participation rate from baseline to last follow-up in BEST was 90%.

Table S2. 2. Conditional genome-wide association analyses of measured DMA%

Independent hits	**SNPs identified**	**Nearby Gene**	**High efficiency allele**	**High efficiency allele frequency**	**β**	**SE**	**P-value**
1st	rs4919690[a]	*AS3MT*	T	0.87	3.22	0.28	$1.4x10^{-29}$
2nd	rs117978529[a]	*AS3MT*	A	0.14	2.04	0.26	$6.1x10^{-15}$
3rd	rs191177668[a]	*AS3MT*	A	0.99	6.24	1.14	$3.7x10^{-08}$
1st	rs61735836[b]	*FTCD*	G	0.99	5.29	0.53	$1.8x10^{-23}$

[a] Linear mixed model adjusted for age, sex, genotyping batch, and relatedness among 3,968 individuals with genome-wide SNP data and measured arsenic metabolites

[b] Linear mixed model adjusted for age, sex, and relatedness among 1,660 individuals with exome-wide SNP data and measured arsenic metabolites.

Table S2. 3. Association between measured DMA% and longitudinal measures of blood pressure

Outcome	Sample Size	Predictor	β	95% CI
Observational analysis based on measured DMA%[a]				
SBP	4,814	Measured DMA%	0.04	(-0.02, 0.10)
DBP			-0.02	(-0.05, 0.02)
Restricted to a random sub-cohort [a,b]				
SBP	1,555	Measured DMA%	0.07	(-0.02, 0.15)
DBP			-0.006	(-0.06, 0.04)
Below median age (<44yrs)[c]				
SBP	2,347	Measured DMA%	0.006	(-0.05, 0.07)
DBP			-0.004	(-0.04, 0.04)
Above median age (≥44yrs)[c]				
SBP	2,559	Measured DMA%	0.06	(-0.02, 0.15)
DBP			-0.02	(-0.07, 0.02)

Abbreviations: CI, confidence interval; SBP, systolic blood pressure; DBP, diastolic blood pressure

[a] Adjusted for age, sex, BMI, smoking, betel use, land ownership, TV ownership, and educational length

[b] Randomly sampled individuals, includes non-cases and cases of cardiovascular disease, respiratory outcomes, diseased, and skin lesions

[c] Median age was calculated based on observed age across baseline and follow-ups.

Table S2. 4. GxE analysis of the interaction between genetically predicted DMA% and measured urinary arsenic concentration among 4,775 HEALS participants

Variables	**β**	**OR**	**S.E**	**P**
GxE analysis with HTN as outcome				
Log UrAs	-1.38	0.25	1.23	0.26
pDMA%	-0.08	0.93	0.09	0.32
Log UrAs*pDMA%	0.02	1.02	0.02	0.31
GxE analysis with SBP as outcome				
Log UrAs	-13.2	--	9.53	0.16
pDMA%	-1.00	--	0.67	0.14
Log UrAs*pDMA%	0.16	--	0.12	0.20
GxE analysis with DBP as outcome				
Log UrAs	-4.83	--	4.91	0.33
pDMA%	2.41	--	2.36	0.31
Log UrAs*pDMA%	0.92	--	2.3	0.69
GxE analysis using tertiles of urinary arsenic and predicted DMA% with HTN as outcome				
Low UrAs*High pDMA	Reference			
Medium UrAs*Low pDMA	0.26	1.30	0.20	0.19
Medium UrAS*Medium pDMA	0.35	1.42	0.21	0.09
High UrAs*Low pDMA	-0.22	0.80	0.21	0.30
High UrAs*Medium pDMA	0.23	1.26	0.21	0.27

Abbreviations: OR; odds ratio, SE, standard error; HTN, hypertension; UrAs, urinary arsenic; pDMA%, predicted DMA%; SBP, systolic blood pressure; DBP, diastolic blood pressure

Table S2. 5. GxE analysis of the interaction between genetically predicted DMA% and measured urinary arsenic concentration among 1,496 unrelated HEALS participants ($r^2 < 0.05$)

Variables	**β**	**OR**	**S.E**	**P**
GxE analysis with HTN as outcome				
Log UrAs	-0.07	0.94	2.14	0.98
pDMA%	-0.007	0.99	0.15	0.96
Log UrAs*pDMA%	0.001	1.00	0.03	0.97
GxE analysis with SBP as outcome				
Log UrAs	12.4	--	17.2	0.47
pDMA%	0.99	--	1.19	0.41
Log UrAs*pDMA%	-0.17	--	0.22	0.44
GxE analysis with DBP as outcome				
Log UrAs	4.7	--	10.9	0.66
pDMA%	0.46	--	0.75	0.54
Log UrAs*pDMA%	-0.07	--	0.14	0.61
GxE analysis using tertiles of urinary arsenic and predicted DMA% with HTN as outcome				
Low UrAs*High pDMA	ref			
Medium UrAs*Low pDMA	0.13	1.14	0.34	0.71
Medium UrAS*Medium pDMA	0.14	1.15	0.37	0.70
High UrAs*Low pDMA	-0.20	0.82	0.37	0.59
High UrAs*Medium pDMA	0.53	1.70	0.38	0.16

Abbreviations: OR; odds ratio, SE, standard error; HTN, hypertension; UrAs, urinary arsenic; pDMA%, predicted DMA%; SBP, systolic blood pressure; DBP, diastolic blood pressure

Figure S2. 1. Distribution of baseline systolic blood pressure stratified by cohort

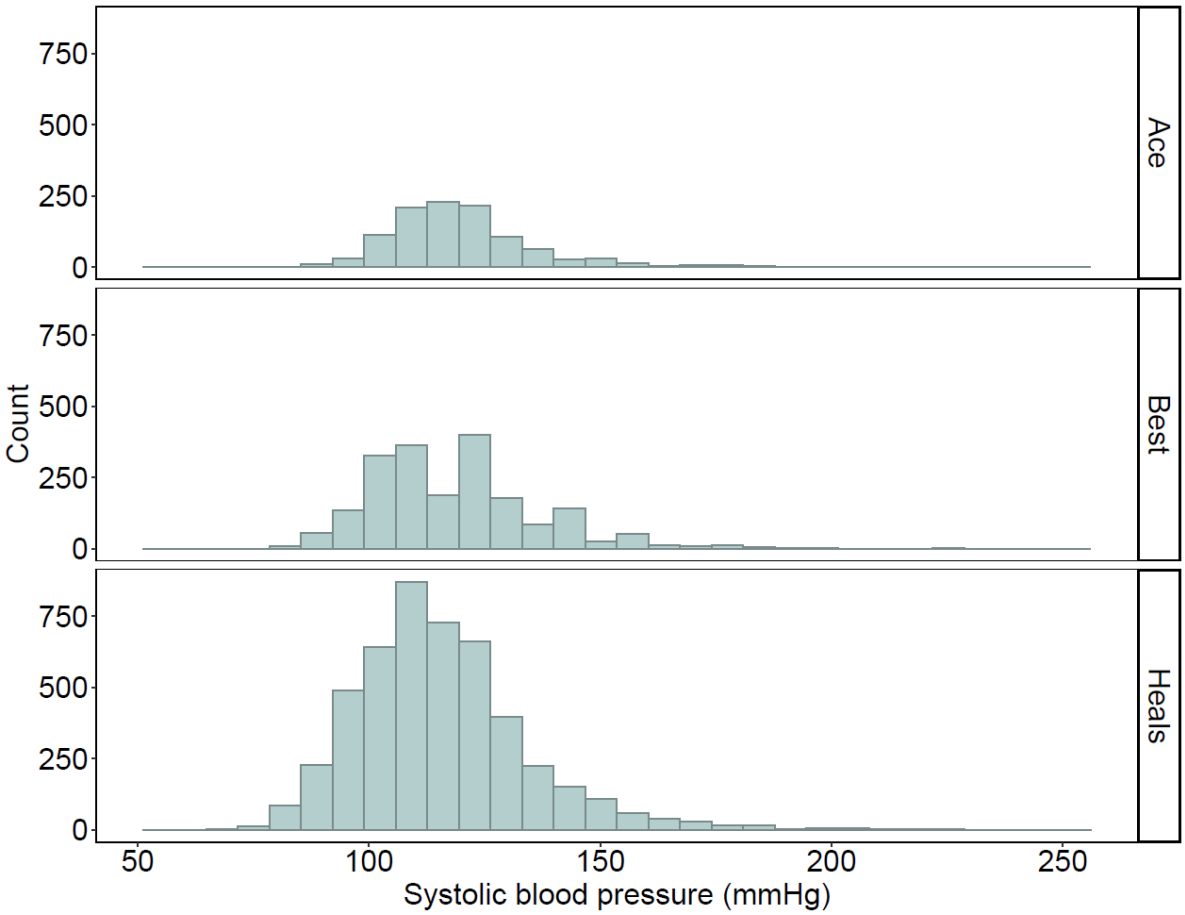


Figure S2. 2. Distribution of baseline diastolic blood pressure stratified by cohort

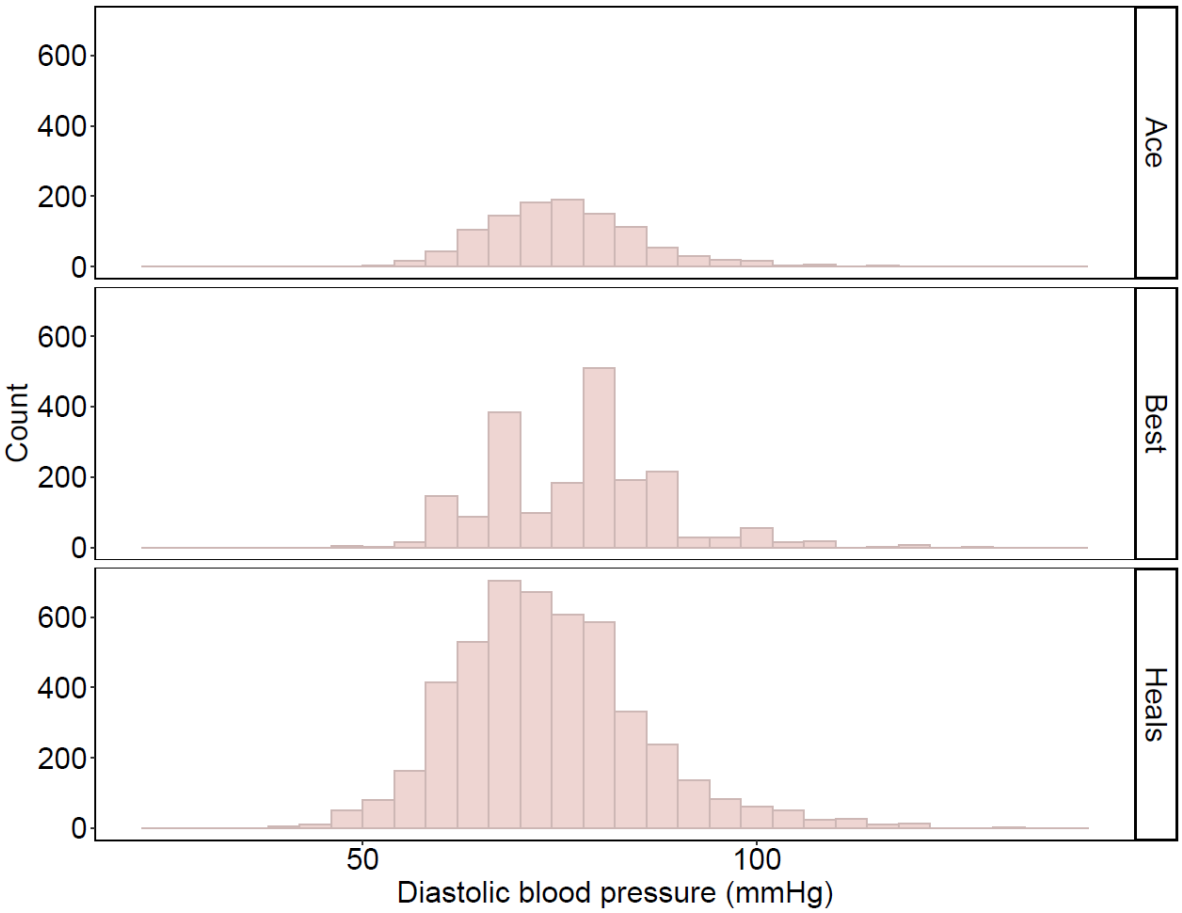


Figure S2. 3. Box plots of systolic blood pressure across follow-up

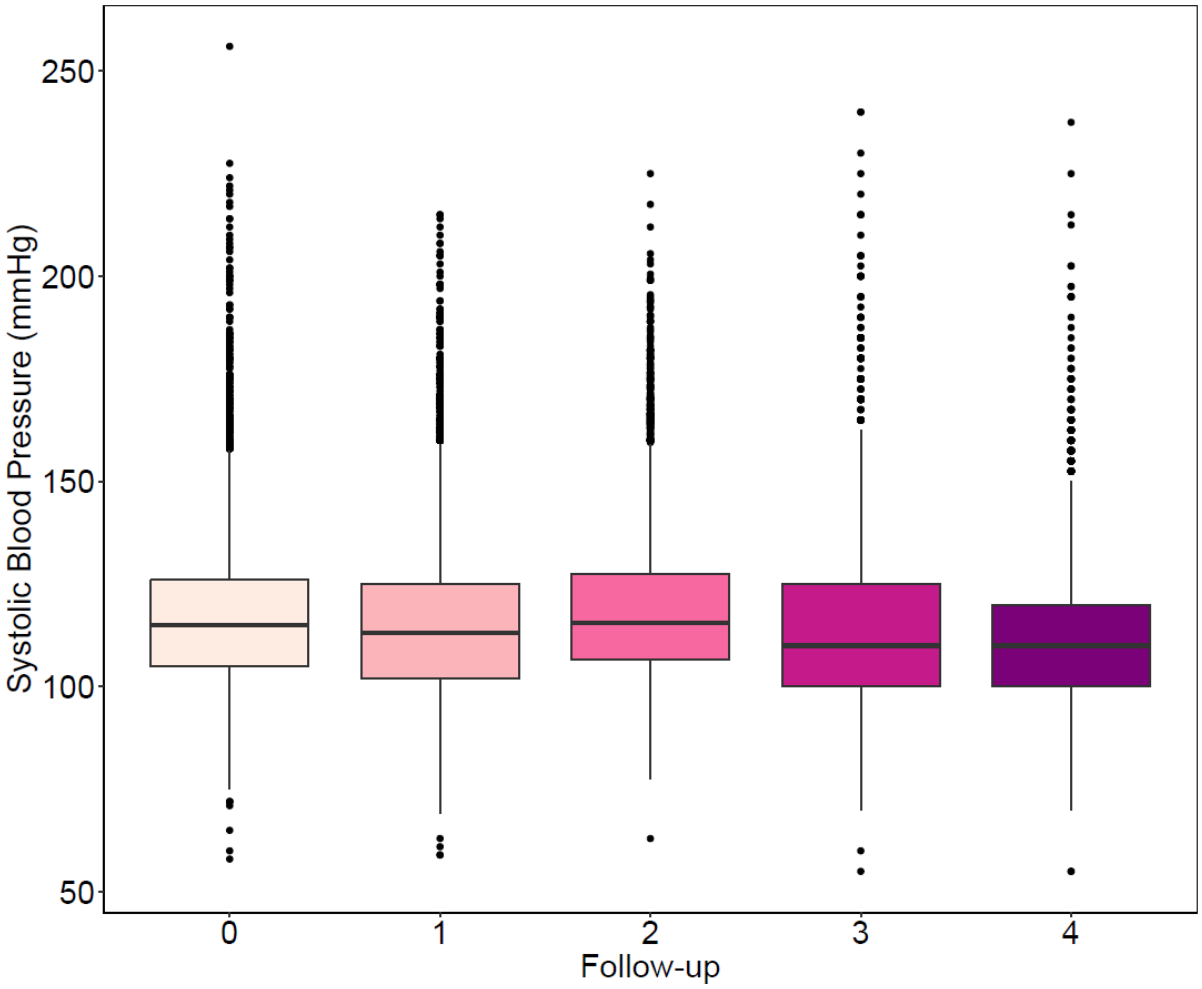


Figure S2. 4. Box plots of diastolic blood pressure across follow-up

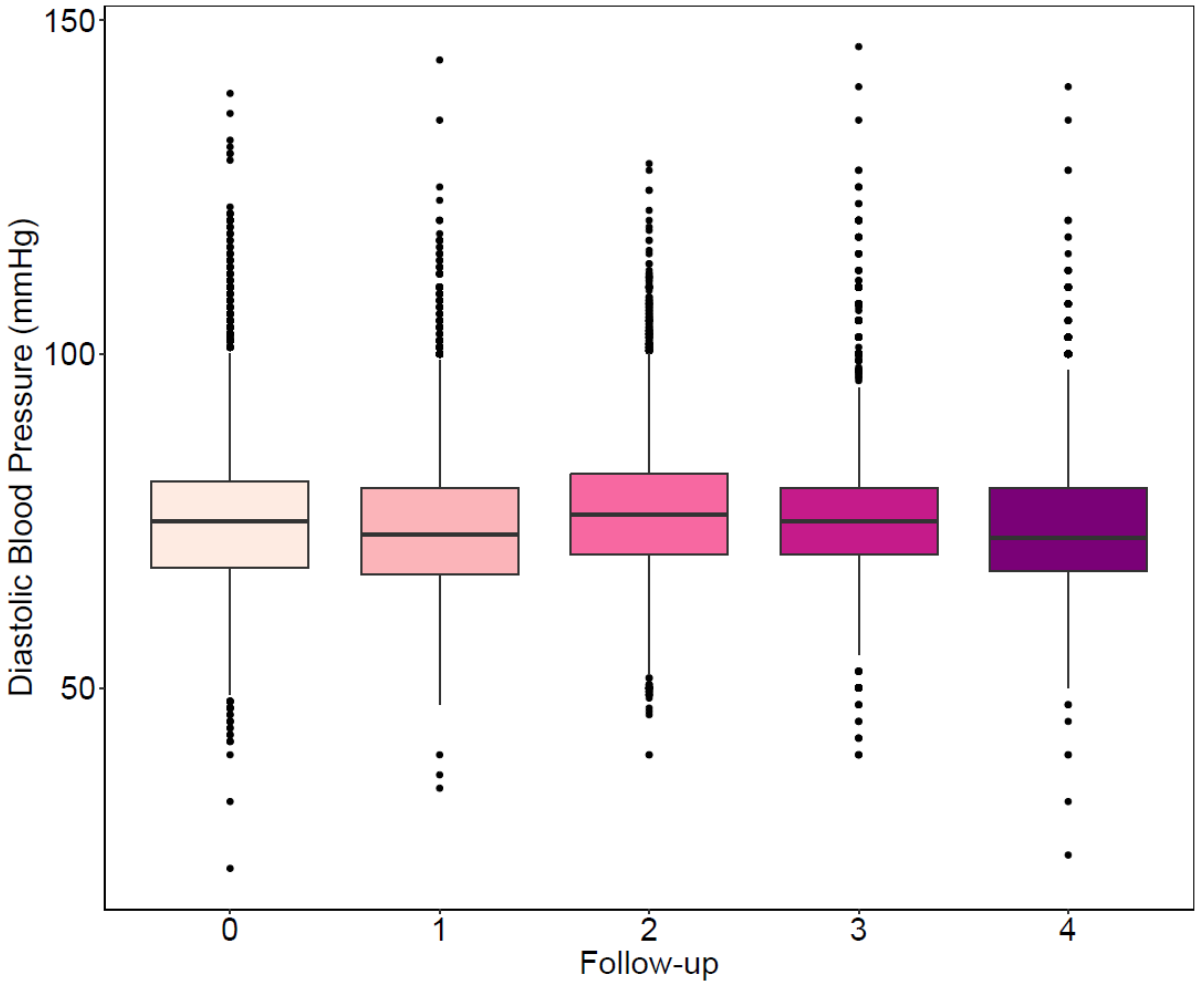


Figure S2. 5. Distribution of genetically predicted DMA% by cohort

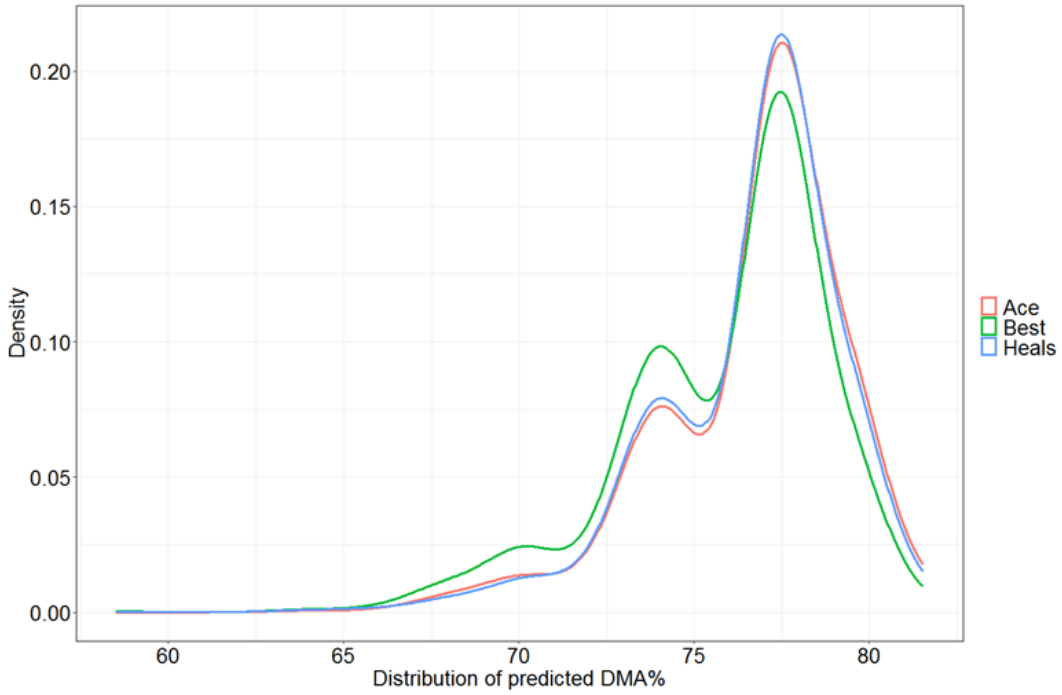
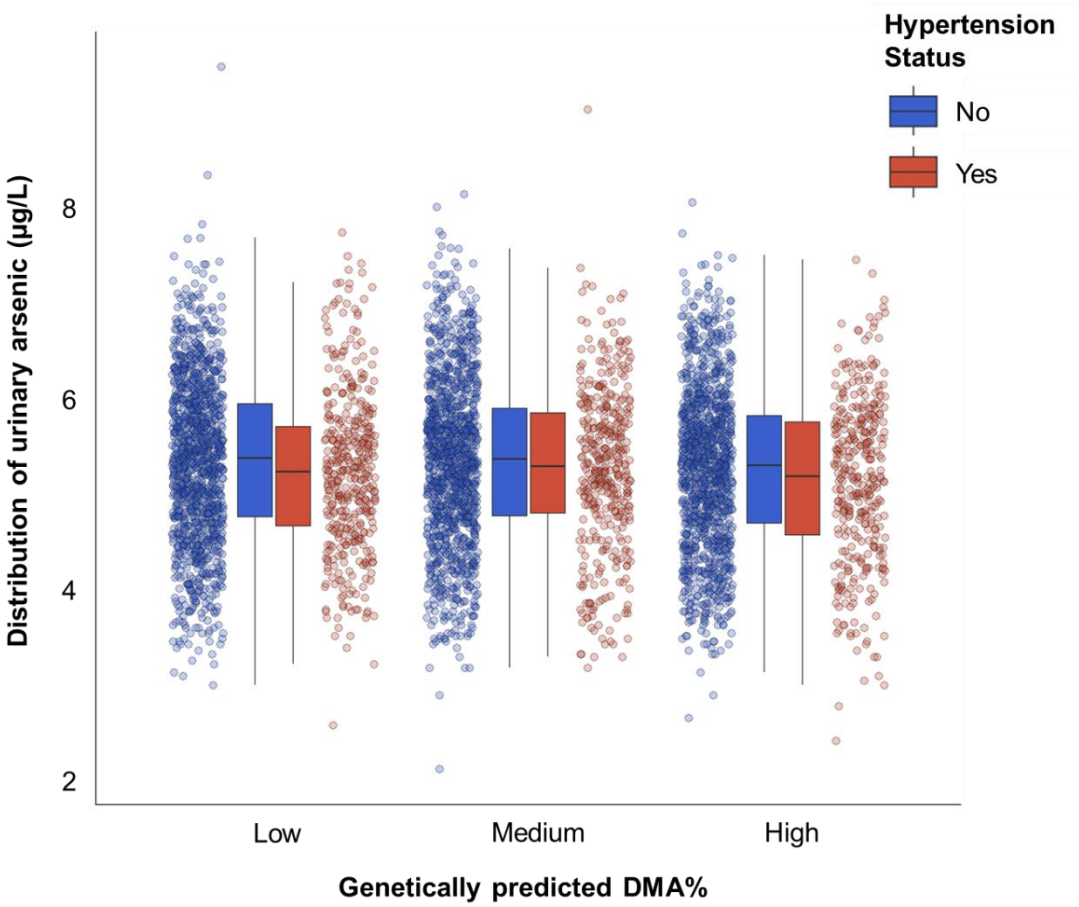


Figure S2. 6. Distribution of urinary arsenic by hypertension case status across tertiles of genetically predicted DMA%



CHAPTER 4

THE IMPACT OF INHERITED GENETIC VARIATION ON DNA METHYLATION IN PROSTATE TUMOR AND BENIGN TISSUES OF AFRICAN AMERICAN AND EUROPEAN AMERICAN MEN

4.1 INTRODUCTION

Prostate cancer (PCa) is the second most common cancer and cause of cancer death among men in the U.S [160]. African American (AA) men are disproportionately affected by PCa, with an incidence rate that is 1.7 times higher compared to men of European American (EA) [58-60]. Similarly, the risk of mortality from PCa is 2 to 3 times higher in AA men and as high as 4.2 fold among young-adults (45-49 years old) compared to EA men [61, 62]. The causes of these disparities are likely complex, with social, environmental, and genetic factors playing contributing roles [59].

Genome-wide association studies (GWAS) have identified 269 common risk alleles (single nucleotide polymorphisms, SNPs) associated with PCa susceptibility [71, 161]. More than 90% of PCa-risk SNPs are located in non-coding regions of the human genome, suggesting the biological mechanisms of the causal variants underlying these associations involve gene regulation. The largest GWAS to date [71] is ~80% European, but included 10,368 cases and 10,986 controls of African ancestry and reported that PCa-risk alleles with ORs>1.10 (71 SNPs) were more common in men of African ancestry (average risk allele frequency = 0.51) compare to men of European ancestry (average risk allele frequency = 0.48) [71]. Overall, these 269 PCa-

risk loci account for an estimated 42.6% and 43.2% of the familial relative risk in EA and AA men, respectively.

Despite the progress in identifying PCa-risk loci, the biological mechanisms by which these SNPs impact PCa risk are largely unknown. A common approach to understanding the mechanisms and underlying biology of GWAS findings is to assess the association of germline genetic variation on local gene expression and/or epigenetic features [162, 163]. The regions where SNPs affect gene expression and DNA methylation status are known as expression-Quantitative Trait Loci (eQTLs) and methylation-Quantitative Trait Loci (meQTLs), respectively. Only two studies to date attempted to identify meQTLs in prostate tissue, and both lack AA representation. One of those studies used tumor methylomes of 589 PCa patients (using the Illumina Infinium HumanMethylation450 BeadChip array) and reported 7,590 genome-wide cis-meQTLs, including 1,178 tumor-specific meQTLs (defined as loci associated with altered methylation in tumor but not in nonmalignant tissue) [164]. The other study focused on the 147 established PCa-risk SNPs and identified 93 PCa-risk SNPs that were associated with DNA methylation at nearby CpG sites in the tumor tissue samples of 355 PCa patients [165].

The goal of this study was to improve our understanding of the biological mechanisms by which PCa susceptibility SNPs influence PCa biology, by examining these SNPs' impact on local DNA methylation. To accomplish this goal, we characterized the effects of genetic variation on the DNA methylation at nearby CpG sites across the entire genome, in both benign and cancerous prostate tissue. In an attempt to identify biological mechanisms that may be relevant to PCa disparities, we analyzed tissue samples from both AA and EA PCa patients and conducted all analyses stratified by ancestry. We examined the 147 PCa-risk SNPs (reported by Schumacher, et al. 2018) to identify PCa-relevant cis-meQTLs in both tumor and benign tissue.

We leveraged existing GWAS summary statistics to determine whether GWAS and cis-meQTL association signals (identified in ancestry-specific analyses) are likely to share a common causal variant (using co-localization methods). The identification of co-occurring methylation QTLs and GWAS risk-loci can help prioritize GWAS findings and provide insight on the epigenetic mechanisms by which SNPs influence PCa risk.

4.2 METHODS

4.2.1 Study population

Subjects included in this work are PCa patients who underwent robotic-assisted laparoscopic prostatectomy at University of Chicago Medical Center (UCMC) Urology Clinical between 2011 and 2017. Trained interviewers from the Epidemiology Research Recruitment Core (ERRC) consented 75 AA and 75 EA eligible men for the collection of questionnaire data, prostate cancer tissue samples, and access to medical records. All eligible participants were diagnosed with Gleason scores of at least seven. The Institutional Review Board of The University of Chicago approved this study.

4.2.2 Bio-specimen collection

After surgery, prostate specimens were sent to the Human Tissue Resource Center (HTRC) at the University of Chicago. Prostatic tissue cores were extracted by HTCR staff and bisected for storage as FFPE (formalin-fixed paraffin-embedded) tissue and frozen tissue (OCT). Each prostate specimen underwent histological examination and Gleason grading by pathologists at the University of Chicago. Gleason grade was confirmed through hematoxylin and eosin (H&E) stains. In addition, the presence of adenocarcinoma was confirmed by the alpha-

methylacyl-coenzyme-A racemase (AMACR), a sensitivity marker for prostate cancer. For each patient, H&E stained tissue slides from the diagnostic blocks were reviewed by a genitourinary pathologist (Dr. Paner) to select areas to sample for DNA extraction. Benign epithelial tissue (prostatic glands), benign stromal tissue, and tumor tissue were collected.

The following criteria were used for the selection of benign tissue: 1) benign epithelium and stroma were selected from blocks that were free of tumor tissue; 2) benign tissue from the peripheral zone of the prostate was prioritized; 3) if no tumor free areas were available in the peripheral zone, benign tissue was collected from the central zone of the prostate. In 10 cases, neither the peripheral, nor the central zone of the prostate were suitable for sampling, so BPH tissue was collected. Areas of tumor tissue to sample were selected in the index focus of the tumor. Collection of prostatic tissue was performed either by using a 1mm biopunch or by laser capture microdissection of 100 μm2 of tissue. In cases where two consecutive diagnostic blocks showed acceptable areas to sample (continuous benign or tumor tissue running through two following blocks of prostate from base to apex), tissue was collected by punching through the base-most diagnostic block. For all other cases, tissue was collected by laser capture microdissection of 8μm thick tissue sections using a Leica LMD 6500 system.

4.2.3 DNA extraction

The Gentra Puregene Tissue Kit (QIAgen) was used to extract 1-2μg of DNA from benign prostate tissue. We assessed DNA concentration and quality with the NanoDrop and Agilent BioAnalyser. We excluded DNA samples with a concentration <40ng/μL and/or 260/280 ratio outside the range of <1.6 to ≥2.1 and/or fragmented DNA <2Kb. The Illumina Infinium HD FFPE restoration kit was used to restore FFPE DNA (according the manufacturer's protocol, including qPCR-based QC of the DNA samples prior to restoration).

4.2.4 SNP genotyping and imputation

Genome-wide SNP data was generated using the Illumina Infinium Multi-Ethnic Global-8 v1.0 array at the University of Chicago Genomics Core Facility. Individual genotypes were called using a GenCall Threshold 0.15. Prior to quality control, our genotype data consisted of 150 individuals (75 AA and 75 EA) with 1,707,345 autosomal SNPs measured. We excluded 224,774 SNPs with low call rates (< 99%). Minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) thresholds were applied separately for AA and EA samples. For AA samples, we excluded 855,033 SNPs with a MAF < 0.05 and 53 SNPs with HWE p-values < 10-5, resulting in 627,485 high-quality SNPs. For EA samples, we excluded 932,143 SNPs with a MAF < 0.05 and 23 SNPs with HWE p-value < 10-5, resulting in 550,405 high-quality SNPs.

We performed imputation using the Haplotype Reference Consortium (HRC, Version r1.1 2016) reference panel through the Michigan Imputation Server. Of the 627,485 and 550,405 post-QC SNPs in AA and EA, respectively, 563,473 SNP in AA and 498,015 SNPs in EA matched the HRC reference panel and met the QC thresholds set by the Michigan imputation server. A total of 39 million SNPs were imputed in both EA and AA cohorts. In AA, we excluded 23M SNPs with and imputation accuracy (r2) ≤ 0.3 and 9.5M SNPs with MAF ≤ 0.07, resulting in 6,463,658 SNPs. Similarly, in EA, we removed 29.3M SNPs with r2 ≤ 0.3 and 4.9M SNPs with MAF ≤ 0.07, resulting in 4,900,500 SNPs. The correlation between MAF and r2 was 0.15 in AA and 0.11 in EA. We conducted all downstream analyses on the resulting 6.4M SNPs for AAs and 4.9M SNPs for EAs.

4.2.5 DNA methylation

The Illumina Infinium MethylationEPIC array kit was used to interrogate >850,000 CpG cites from benign and tumor tissue (at the University of Chicago Genomics Core Facility). This

assay uses ~250ng of bisulfite-converted DNA and employs Infinium I and II assay chemistry technologies. Infinium I assay design uses 2 bead types (for methylated and unmethylated states) per CpG locus, while the infinium II design uses 1 bead type for the methylated state, which is determined at the single base extensions step after hybridization. The methylation state of each CpG site is the ratio of methylated fluorescent signal and unmethylated signal using the BeadArray Readers. Signal intensity from raw data was extracted using the GenomeStudio software. This array provides dense coverage across CpG islands (>95%), shores (>80%), and shelves (>90%).

Methylation data was normalized using the BMIQ function in the ChAMP software and were expressed as β values ranging from 0 (completed unmethylated) to 1 (completed methylated). We processed the methylation data combining across ancestry but stratified by tissue type. We removed probes with a detection p-value > 0.01 (95,481), beadcount < 3 in at least 5% of samples (106), non-CpG probes (2,228), non-specific probes that align to multiple locations (47), and underperforming probes (69,244) [166]. This QC resulted in 698,812 benign tissue probes and 682,694 tumor tissue probes that were included in downstream analyses.

4.2.6 Cis-meQTL Analyses

We used the FastQTL software to conduct genome-wide *cis*-meQTL analyses, where we tested the local (*cis*) association of SNPs and CpG sites <500 kb apart. Methylation β values were rank normalized to satisfy linear model assumptions. To identify CpGs affeted by a meQTL, we computed CpG-level empirical p-values obtained from the approximation of the beta-distribution using adaptive permutations as implemented in the FastQTL software (--

permute 1000 10000). A false discovery rate (FDR) ≤ 5% for significant *cis*-meQTLs was applied at the level of the CpG.

All meQTL analyses were conducted separately for AA and EA men, and were adjusted for age, 5 genotyping principal components (PCs), and 10 methylation surrogate variables (SVs). Methylation SVs were estimated for AA and EA men separately, using the SVA software package in Bioconductor. To determine the number of SVs that rendered optimal power to detect meQTLs, we conducted cis-meQTL analysis of chromosome 1 for each ancestry and tissue type using 5, 10, 15, and 20 SVs. Principal component analyses were conducted using 214,706 independent SNPs for AA and 109,927 independent SNPs for EA in Plink (--indep-pairwise 50 5 0.2).

4.2.7 Identifying GWAS and meQTL association signals likely to share a causal variant

Our workflow to prioritize regions for co-localization is shown in **Figure 3.1**. For our meQTL results, we restricted to the lead SNP for each CpG site having a meQTL (FDR < 0.05). Next, we restricted to the 20,646 GWAS SNPs that met the genome-wide significance threshold ($p < 5x10^{-8}$ based on summary results from Schumacher et al.) and identified SNPs that were also a lead meQTL SNP. We conducted co-localization tests (described below) at loci where the lead GWAS SNP was in high linkage disequilibrium (LD, $r^2 > 0.5$) with the lead meQTL SNP. LD estimates were obtained from LDlink (https://ldlink.nci.nih.gov/) (for AA men we used the Americans of African Ancestry in SW USA (ASW) reference population and for EA men we used the European reference populations (EUR)). We conducted all GWAS-meQTL co-localization analyses stratified by ancestry and tissue type.

4.2.8 GWAS-meQTL co-localization analyses

Publically available SNP summary statistics were not available for the most recent PCa GWAS to date [71] at the time of the current study. Thus, we obtained SNP summary statistics from [161] to identify GWAS and *cis*-meQTL association signals likely to share a common causal variant using Bayesian co-localization analysis (*coloc* software) [167]. This software uses the summary statistics of the cis-meQTL and GWAS analyses and restricts the analysis to SNPs that are present in both sets of summary statistics along with the start and end genomic positions (basepairs) defined by the meQTL analysis. Coloc requires three prior probabilities to be specified: p_1, the probability a specific variant is causal for trait 1 (GWAS), p_2, the probability a specific variant is causal for trait two (*cis*-meQTL or *cis*-eQTL), and p_{12} , the probability a variant is causal for both traits. The coloc software uses these priors to calculate the posterior probability of a common causal variant (P(CCV)). To determine a prior for the GWAS trait (PCa), we used a recent estimate of the number of independent common PCa susceptibility variants (4,530) [168]. Given there were ~20M genome-wide SNPs tested in recent GWAS, the probability of a SNP being a causal GWAS SNP is 4,530/20M, very close to $1x10^{-4}$. This probability is equal to the sum of $p_1 + p_{12}$. For meQTL prior determination, we used the previously detected 7,590 *cis*-meQTLs among 4,894,225 SNPs [164], indicating the probability of a SNP being a causal meSNP is close to $5x10^{-3}$. This probability corresponds to the sum $p_2 + p_{12}$ (for GWAS-meQTL co-localization). These priors were chosen informed by the literature but are not intended to be exact. We chose to vary the value of p_{12} to correspond to probabilities of a causal GWAS SNP being a causal meSNP of 10%, 25%, 50%, and 75%. We chose to vary p_{12} as has been done in previous meQTL studies [169, 170], because the true value of p_{12} is unknown. The resulting values for p_1, p_2, and p_{12} are shown in **Supplementary Table 3.1**.

4.2.9 Identification of eQTLs among co-localized meSNPs

We searched for eQTLs among the GWAS-meQTLs (meSNPs) that reached the co-localization threshold of P(CCV) > 80%. We used the Genotype-Tissue Expression Project (GTEx v8) prostate tissue gene expression data (n=221) and search for eQTLs associated with the PCa-risk SNP. To prepare the eQTL and GWAS summary statistics for coloc, we restricted to the common SNPs and the start and end genomic positions (basepairs) defined by the meQTL analysis. Genomic coordinates for GTEx SNP data were converted from GRCh38/hg38 to GRCh37/h19 using the NCBI Genome Remapping Service.

4.2.10 GWAS-eQTL co-localization analyses

GWAS-eQTL co-localization analyses were restricted to loci where we estimated a probability of P(CCV) of >80% for the GWAS and meQTL association signals. We conducted similar analyses for GWAS-eQTL pairs using *coloc*. GTEx identified 7,356 eQTLs in prostate tissue among 11.5M SNPs, suggesting the probability that a SNP is a causal eSNP in prostate tissue is ~$6x10^{-4}$. This probability corresponds to the sum $p_2 + p_{12}$ (for GWAS-eQTL co-localization). We also varied the value of p_{12} to correspond to probabilities of a causal GWAS SNP being a causal eSNP of 10%, 25%, 50%, and 75% (**Supplementary Table 3.1)**.

4.3 RESULTS

4.3.1 Overview of samples

Characteristics of the 75 AA and 75 EA PCa patients included in our analyses are described in **Table 3.1**. AA and EA patients were on average 65 years-old at time of diagnosis. Gleason scores were slightly higher in EAs compared to AAs, for example, 14 EA patients had a

Gleason Score >8 compared to 6 AA patients. Similarly, PSA was similar in EAs (58.2ng/mL) compared to AAs (57.5ng/mL). Tumor volume, defined as the percent of prostate that is tumor, was also higher in EAs (22.5%) compared to AAs (18.2%). Overall, EAs had slightly less favorable clinical characteristics compared to AAs. This is likely because EA men with more advanced PCa disease residing outside the UCMC catchment area were seeking care at the University of Chicago, with our AA patients being more likely to reside in the catchment area.

4.3.2 Cis-meQTLs in benign tissue

We identified 6,298 and 6,960 genome-wide *cis*-meQTLs (FDR < 0.01) in benign prostate tissue of AA and EA men, respectively (**Table 3.2**). Target CpGs affected by meQTLs (mCpGs) were on average 23,741 bp away from the lead SNP in AA men and 30,487 bp away in EA men. We observed more mCpGs in non-CpG islands in both AAs and EAs (65% and 63%, respectively) compared to all other locations (islands, shores, and shelves) on the EPIC array (**Figure S3.1a**). Approximately 64% (4,060) and 63% (4,409) of mCpGs in AA and EA were assigned to a gene (based on Illumina's annotations) as compared to 73% in non-CpG targets (**Figure S3.2a**). We observed enrichment of mCpGs in enhancer regions ($p < 0.001$) for both AAs (4%) and EAs (4%) compare to the all other CpGs (3%) (**Figure S3.2a**). In AA men, we also observed enrichment of mCpGs in DNase hypersensitivity sites (DHS) regions (7%) compared to EA (6%) and all other CpGs in EPIC array (6%) ($p < 0.001$) (**Figure S3.2a**).

Of the 6,298 cis-meQTLs (FDR < 0.05) detected in AA benign tissue, we replicated 4,269 (68%) at a $p < 0.01$ and 3,865 (61%) at $p < 0.001$ in EA benign tissue. The correlation (R) between the beta coefficients observed in AA and EA tissue for the 4,269 cis-meQTLs was 0.97 and 4,258 (99%) were consistent in the directionality of the effect (**Figure S3.3a**). Similarly, among the 6,960 cis-meQTLs (FDR< 0.05) detected in EA benign tissue, we replicated 4,372

(63%) at $p < 0.01$ and 3,576 (51%) at $p < 0.001$ in AA benign tissue. The correlation between the beta coefficients for the 4,372 cis-meQTLs was 0.95 and 4,343 (99%) were consistent in the directionality of the effect (**Figure S3.3b**).

4.3.3 Cis-meQTLs in tumor tissue

In tumor prostate tissue of AA and EA we identified 2,641 and 1,700 genome-wide *cis*-meQTLs (FDR < 0.05), respectively (**Table 3.2**). The average distance between the mCpG and the meQTL lead SNP was 25,647 bp in AAs and 28,085 bp in EAs. The locational distribution of tumor cis-meQTLs CpG targets was similar to benign tissue, with more mCpGs in non-CpG islands for both AA and EAs (60% and 57%, respectively) than all other locations (**Figure S3.1b**). We observed a higher percent of tumor mCpG targets for both AA and EA (69% and 67%, respectively) assigned to a gene (**Figure S3.2b**) as compared 73% to non-CpG targets ($p < 0.001$). Tumor CpG targets were depleted in promoters for both AA and EA (13% for both) compared to 20% non-CpG targets ($p < 0.001$) and enriched in DHS regions for both AA and EA (7% for both) compare to 7% non-CpG targets ($p < 0.001$).

Among the 2,641 cis-meQTLs (FDR < 0.05) detected in AA tumor tissue, we replicated 1,667 (63%) at $p < 0.01$ and 1,457 (55%) at $p < 0.001$ in EA tumor tissue. The correlation among the beta coefficients from AA and EA corresponding to the 1,667 replicated cis-meQTLs was 0.96, and 1,659 (99%) were consistent in their direction of effect (**Figure S3.3c**). Additionally, of the 1,700 cis-meQTLs (FDR < 0.05) detected in EA tumor tissue, we replicated 1,131 (67%) at $p < 0.01$ and 996 (59%) at $p < 0.001$ in AA tumor tissue. The correlation between the beta coefficients in EA vs AA was 0.96 among the 1,131 cis-meQTLs and 1,126 (99%) were consistent in their direction of effect (**Figure S3.3d**).

4.3.4 Tissue specificity of cis-meQTLs in AA and EA men

Among the 6,298 benign-tissue cis-meQTLs observed in AAs, we replicated 62% at a $p < 0.01$ and 51% at a $p < 0.001$ in AA tumor tissue (**Table S3.2**). Similarly, in EA, we replicated 55% benign-tissue cis-meQTLs in tumor tissue at $p < 0.01$ and 42% at a $p < 0.001$. Thus, our results suggest that less than half of benign-tissue cis-meQTLs in AAs and EAs are specific to benign tissue, 38% and 45%, respectively. We observed a larger percent of replication of tumor-tissue cis-meQTLs in benign tissue results of AAs and EAs, 79% at $p < 0.01$ for both and 74% and 75% at $p < 0.001$ for AA and EA, respectively (**Table S3.2**). Only 21% of tumor-tissue cis-meQTLs are specific to tumor tissue in both AA ad EA. Overall, we conclude that the majority of meQTLs in tumor and benign are shared across tissue types.

4.3.5 Replication of meQTLs from prior studies

We attempted replication of the 7,590 cis-meQTLs identified in 589 localized prostate tumors of EA men [164] using various p-value thresholds (**Table S3.3**). In AA benign and tumor tissue, we had data available for 5,231 of the 7,590 reported cis-meQTLs. Using a threshold of $P < 0.001$, we found that ~80% of the 5,231 target CpGs were also the target of a meQTL in both benign and tumor AA tissues. We also searched for replication of the specific CpG-SNP pairs reported using a $P < 0.01$ threshold and observed replication of <50% of the 5,231 cis-meQTLs in AA men (47% in benign tissue and 44% in tumor tissue). In EA benign and tumor tissue, we examined 5,590 of the 7,590 cis-meQTLs with available data (**Table S3.3**). We replicated 68% and 60% of the target CpGs at a $P < 0.001$ in benign and tumor tissue, respectively. Additionally, we replicated 59% of the Houlahan et al. CpG-SNP pairs in benign tissue and 49% in tumor tissue. For both AA and EA benign and tumor tissue, >99% of meQTLs replicated ($p < 0.01$) were also consistent in the direction of effect. Overall, we observed more replication of the target

CpGs in AA than EAs. However, replication of CpG-SNP pairs was higher in EAs compared to AAs. We also observed more replication in benign tissue than tumor tissue.

In addition to genome-wide meQTLs, Houlahan et al. identified 75 PCa-risk cis-meQTLs, which describe the association between 27 unique PCa-risk loci and 73 CpG sites. Houlahan et al. validated 52 of these cis-meQTLs in an independent cohort. We attempted replication for 41 and 45 of the 52 cis-meQTLs in AA and EA tissues, respectively (**Table S3.4**). In AA benign and tumor tissue samples, we replicated 12% and 31% of the 41 target CpGs at a $P < 0.001$, respectively. Using $P<0.01$ threshold, we replicated 29% and 31% of the 42 CpG-SNP pairs in benign and tumor tissue, respectively. In EA benign and tumor tissue samples, we replicated 20% and 47% of the 45 target CpGs at a $P < 0.001$, respectively. Replication of the 45 CpG-SNP pairs at a $P < 0.01$ was 31% and 60% for EA benign and tumor tissues, respectively.

We also sought to replicate 110 autosomal *cis*-meQTLs, which passed a threshold $p < 1x10^{-9}$ (same as Houlahan et al. threshold), identified in the prostate tumor tissues of 355 EA men [165] (**Table S3.5**). Of the 110 *cis*-meQTLs, we had summary results for 60 *cis*-meQTLs in AAs and 67 *cis*-meQTLs in EAs. We attempted replication using various thresholds in both benign and tumor tissues. Overall, we replicated <30% of the lead CpGs ($P < 0.001$) in AA benign and tumor tissues, and <40% in EA benign and tumor tissues. Similar to the Houlahan et al. replication, replication of meQTLs from Dia et al. ($P < 0.01$) was highest in EA tumor tissue (49%) and AA tumor tissue (42%) and lowest in AA benign tissue (15%) and EA benign tissue (28%), suggesting differences in meQTL profiles in cancer vs. benign tissue.

4.3.6 Co-localized GWAS-meQTL pairs in benign tissue

Among our 6,298 and 6,960 meQTLs identified in AA and EA benign tissue, respectively, we attempted to identify *cis*-meQTL residing in the same location as previously

reported PCa-risk loci (**Figure 3.1** and **Table S3.6-S3.7**). Restricting to lead SNPs with p $<5x10^{-8}$ in the Schumacher et al. GWAS of PCa, we searched for meQTL lead SNPs with an LD (r^2) > 0.5 with a GWAS lead SNP. For AA men we used ASW reference population and for EA men we used the European reference populations (EUR), https://ldlink.nci.nih.gov/). Pairs identified were formally tested for co-localization, resulting in 8 tested loci for AA men and 17 for EA men.

We found evidence of co-localization (P(CCV) >80%) (based on 75% prior probability that a GWAS SNP is an meSNP, **Table S3.1**) for 1 GWAS-meQTL pair in AA men and 7 GWAS-meQTL pairs in EA men (**Table 3.3** and **Figure 3.2**). The probabilities that a GWAS SNP is a causal meSNP, indicated by $P(CCV)_1$, $P(CCV)_2$, and $P(CCV)_3$, correspond to 25%, 50%, and 75% (**Table 3.3**). The co-localization signal unique to AA men was near a promoter of *HAUS6*. In EA men, the 7 likely co-localized GWAS-meQTL pairs were located near the intergenic region of *MLPH*, gene body of *EEFSEC*, gene body of *TNS3*, promoter near *MSMB*, intergenic region of *LARP4B*, promoter near *MRPL52/MMP14*, and intergenic region of *COPRS/UTP6* . Among the seven GWAS-meQTL pairs, 4 (*MLPH, EEFSEC, MSMB*, and *MRPL52/MMP14*) were specific to benign tissue and the other three were replicated in EA tumor tissue (**Table 3.4**).

4.3.7 Co-localized GWAS-meQTL pairs in tumor tissue

In tumor tissue, we identified *cis*-meQTLs in the same location as 38 and 33 previously reported PCa-risk loci (p $<5x10^{-8}$) in AA and EA men, respectively (**Figure 3.1** and **Table S3.8-S3.9**). Restricting to loci with LD (r^2) > 0.5 between the lead GWAS and lead meQTL SNP, we identified 3 and 14 loci in AA and EA men, respectively, to test for co-localization. There was evidence of co-localization (based on 75% prior probability that a GWAS SNP is a meSNP,

Table S3.1) for 3 GWAS-meQTL pairs in AAs and 5 GWAS-meQTL pairs in EAs (**Table 3.4, Figure 3.3**). Two co-localization signals were shared between AAs and EAs; one near the gene body of *IRX4* and the other in the 5'UTR of *MMP7*. In AA men, one additional region showing evidence of co-localization was located near the gene body of *MYO9B*. The 3 additional co-localization signals identified in EA men were located near a promoter near *GGCX*, gene body of *TNS3*, and intergenic region of *LARP4B*. In AA tumor tissue, only the signal near IRX4 was unique to tumor tissue, all other signals in AA and EA tumor tissue were also replicated in benign tissue.

4.3.8 Co-localized GWAS-eQTL pairs

Among 13 unique co-localized GWAS-meQTL pairs across both tissue types and ancestries, we identified eQTLs for six PCa-risk SNPs (rs12653946, rs11568818, rs11666569, rs10993994, rs10187424, and rs7767188) in GTEx normal prostate tissue. We conducted co-localization analyses for the six PCa-risk SNPs and the corresponding eGenes (*IRX4*, *MMP7*, *MYO9B*, *MSMB*, *GGCX*, and *TRIM26*, respectively). Using the four sets of priors listed in **Supplementary Table 3.1,** we found strong evidence of shared common causal variants affecting both GWAS and eQTL traits for five of the six SNP-eGenes tested (**Table 3.5 and Figure 3.4**). These results suggest rs12653946, rs11568818, rs11666569, rs10993994, and rs10187424 may affect PCa risk by regulating both local methylation and gene expression.

4.4 DISCUSSION

In this study, we performed a genome-wide search for *cis*-meQTLs and identified 6,298 and 6,960 *cis*-meQTLs in the benign tissue of AA and EA men, respectively, and 2,641 and 1,700 *cis*-meQTLs in the tumor tissue of AA and EA men, respectively. Our focus was on

characterizing the biological effect of 147 previously reported PCa susceptibility loci. Because the mechanisms of most PCa-risk SNPs are unknown, we hypothesized some of these SNPs may affect local methylation, as well as expression of nearby genes. To test this hypothesis, we integrated summary statistics from the largest PCa GWAS available at the time of this work [161] and searched for GWAS-meQTL pairs that share a common causal variant (CCV). Using a Bayesian test for co-localization (*coloc*), we identified one GWAS-meQTL pair in benign tissue and three GWAS-meQTL pairs in tumor tissue of AA men. In EA men, we found evidence of co-localization for seven and five GWAS-meQTL pairs in benign and tumor tissue, respectively. Among these co-localized GWAS-meQTL pairs, we found co-occurring GWAS-eQTL pairs in six eGenes (*IRX4, MMP7, MYO9B, MSMB,* and *GGCX*), all of which have known roles in tumorigenesis. These findings demonstrate a potential regulatory mechanism by which some PCa-risk SNPs can modify local DNA methylation and/or the regulation of local gene expression.

In this genome-wide search for SNPs that modify DNA methylation at nearby CpGs, we found nearly triple the number of cis-meQTLs in benign tissue compared to tumor tissue in both AA and EA samples. One potential explanation for this difference is the relative cellular homogeneity of benign tissue, as compared to tumor tissue, which may increase power for meQTL detection (as described further below). We attempted to replicate meQTLs in the opposite tissue of which they were discovered and found that 79% of tumor-tissue meQTLs in both AA and EA were also likely present in benign tissue at a $p < 0.01$. We observed less replication of benign-tissue meQTLs in tumor tissue (62% in AAs and 55% in EAs). We suspect these differences in number of mQTLs detected are largely driven by differences in cell type composition. Benign tissue samples are likely more homogeneous, with patient samples

consisting largely of prostate epithelial and stromal cells. In contrast, tumor tissue samples will contain cancer cells, which can differ in various cellular phenotypes across individuals (and within an individual), as well as adjacent normal (epithelial and stromal cells). PCa is one of the most heterogeneous cancers with respect to morphology and genetic/molecular characteristics [171]. Assuming the meQTL associations differ to some extent across cell types, power to detect cell type-specific QTLs will be higher in more homogeneous cell types mixtures and reduced in more heterogeneous cell type mixtures (due to smaller populations of individual cell types).

To gauge how our results compare to other cis-meQTL analyses of prostate tissue, we incorporated summary statistics from two recent studies that were conducted among EA men [164, 165]. Based on Houlahan et al. results, we observed a higher level of replication of cis-meQTLs (SNP-CpG pair) in EA men than AA, which is expected, as the ancestry of our EA samples should be similar to that of the Houlahan et al., 2019 study. We also observed more replication in benign tissue compared to tumor tissue across both ancestries, and this is likely due to the increased level of cellular heterogeneity and person-to-person differences in tumor tissue. As stated previously, benign tissue is more homogenous and lends for more power to detect cis-meQTLs compared to tumor tissue. However, we also observed more replication of the target CpGs from Houlahan et al. ($p < 0.001$) in AAs (in both benign and tumor tissue) than EAs. One potential explanation for higher lead CpG replication in AA men is because we tested about 1.5M more SNPs in AAs than EAs, which increases the likelihood of a CpG reaching the $p < 0.001$ replication threshold. Replication of PCa-risk cis-meQTLs reported in Houlahan et al., 2019 and Dai et al., 2020 ($p < 0.01$) was highest among EA men and particularly tumor tissue. Similarly, replication of the lead CpGs reported in both studies ($p < 0.001$) was higher in EA

tumor tissue. These results are expected since both Houlahan et al., 2019 and Dai et al., 2020 are studies of EA men, and their analyses were conducted in tumor tissue.

Our study has several strengths compared to the prior work in this area. First, our study employed co-localization tests, which are critical for determining SNP associations with PCa and DNAm are due to common causal variants. Our work suggests that the number of PCa risk loci potentially explained by meQTLs is likely smaller that reported by prior studies. Second, both of these studies only conducted meQTL analyses in tumor tissue and did not have paired-benign tissue. Third, both of these studies lacked AA representation. In the current study, we specifically address these limitations by analyzing equal numbers of AA and EA benign and tumor PCa tissue samples and conducting co-localization analyses of GWAS-meQTL pairs.

In this study, we leveraged PCa GWAS summary data and found 8 PCa-risk common causal variants driving both GWAS and meQTL signals in benign (1 in AA and 7 in EA men) and tumor tissue (3 in AA and 5 in EA men). To assess whether these meQTLs are unique to prostate tissue, we used the QTLbase database (QTLbase Home (mulinlab.org)) and searched for the CpG-SNP pair (PCa-risk SNP). QTLbase provides genome-wide QTL summary statistics from 257 QTL studies for greater than 70 tissues/cell types. We also used GTEx whole blood eQTL summary statistics for 755 samples to assess whether the PCa-risk meSNP is an eQTL in whole blood. Among the 8 co-localized meQTLs in benign tissue, only 2 are meQTLs and eQTLs in whole blood (both in EA men, rs2292884-cg14458575: eGene = *MLPH* and rs10993994-cg17030820: eGene = *AGAP7P*) (**Table S3.10**). For tumor tissue, 4 of the 8 co-localized meQTLs were also meQTLs in whole blood, one of which was shared among AA and EA men (rs11568818- cg25511807, near MMP7), one that was unique to AA men (rs11666569-cg19418318, near MYO9B), and two unique to EA men (rs10187424- cg02493740, near GGCX

and rs12768349- cg20503657, near LARP4B) (**Table S3.10**). Among these, only two were also eQTLs in whole blood (rs11666569- cg19418318: eGene = *MYO9B* and rs10187424- cg02493740: eGene = *GGCX*). Overall, the majority of PCa-risk meSNPs identified in benign tissue are PCa tissue specific in their effect on DNAm, while 50% of the meSNPs identified in tumor tissue are not. Additionally, the majority of PCa-risk eSNPs identified in both tumor and benign tissue appear to uniquely impact gene expression in PCa tissue.

We found evidence that five PCa-risk loci were common causal variants for GWAS, meQTL, and eQTL traits. These loci are best represented by rs12653946 near *IRX4*, rs11568818 near *MMP7*, rs11666569 near *MYO9B*, rs10993994 near the *MSMB*, and rs10187424 near *GGCX*. For the loci near MYO9B and GGCX, there is little evidence of the tumor biology or direct association with PCa. However, the GWAS risk signal (rs11666569) and *MYO9B* eQTL were co-localized based on TCGA data [161]. *IRX4* is a transcription factor that mediates ventricular differentiation during cardiac development and suppresses expression of MyHC3 by forming an inhibitory complex with vitamin D receptor (VDR) and retinoic X receptor (RXR) that binds the response element of the MyHC3 promoter [172]. By interacting with the VDR pathway, IRX4 could suppress PCa growth. Associations between expression of *IRX4* and rs12653946 have been previously identified in PCa tumor tissue and replicated in adjacent benign tissue [165, 173, 174]. Similarly, an meQTL for rs12653946 in prostate tumor tissue has also been previously reported [165]. In vitro studies using PCa cell lines, suggest that rs12653946 may function as an enhancer element that substantially decreases *IRX4* expression, specifically in prostate tissue [173, 175].

We also identified a co-occurring meQTL (in AA and EA men) and eQTL for PCa-risk SNP rs11568818, eGene *MMP7*, and CpG cg25511807. *MMP7* is a known cancer prognostic

marker and is over-expressed in epithelial cells in human PCa [176, 177]. Mouse studies show that *MMP7* induced by pro-inflammatory cytokine, IL-17, enhances epithelial-to-mesenchymal transition (EMT) by releasing β-catenin and increasing expression of mesenchymal markers [178, 179]. Mice with double *MMP7* knockout ($Mmp7^{-/-}$) exhibit decreased cellular proliferation, increased apoptosis in prostate lesions, decreased angiogenesis, and overall decreased formation of invasive prostate adenocarcinoma compared to $Mmp7^{+/+}$ (wild type) and $Mmp7^{+/-}$ mice. Additionally, in human normal and tumor prostates from the Cancer Genome Atlas (TCGA), IL-17 mRNA levels were positively correlated with MMP7 mRNA levels [178].

Our meQTL analysis of EA benign tissue suggest the T allele of rs10993994 is associated with increased methylation at the cg17030820 CpG site that lies near a promoter of *MSMB,* and this association was unique to benign tissue. *MSMB* encodes one the most abundant proteins secreted in the prostate, β-microseminoprotein (β-MSP), and is typically under-expressed or entirely absent in prostate tumor tissue [180]. β-MSP is thought to manifest fungicidal activity [181] and induce tumor suppressive properties [182]. An meQTL for rs10993994 has been previously reported in prostate tumor tissue [165]. Evidence from a recent study of cis- and trans- eQTLs, highlights rs10993994 as a trans-eQTL for *SNHG11*, *NDRG1*, and *SPON2* whose effect was mediated through a cis-eQTL for *MSMB* [174]. These results were confirmed through co-localization analysis of the cis- and trans-eQTLs and genome editing of rs10993994 in LNCaP cell lines using CRISPR/Cas9. Interestingly, this study also found that the SNPs driving *MSMB* expression differ in benign vs. tumor tissue, which gives rise to the theory that regulatory network rewiring during oncogenic transformation might occur at this locus. A second signal (rs7098889) for MSMB expression, independent of rs10993994, was reported at this locus [183]. Through CRISPR/Cas9 genome editing experiments, researchers found that rs7098889 is

positioned in a regulatory region that specifically regulates *MSMB* expression in prostate cells and its deletion results in a 9.5 fold increase in *MSMB* expression in LNCaP cell lines.

One of the primary challenges of conducting co-localization analyses using meQTL results from AA men is the LD mismatch with the largely European ancestry GWAS of PCa risk. LD patterns are known to vary across ancestral populations, and admixed AA populations exhibit higher genetic diversity and lower levels of LD compared to populations of European ancestry [77, 91]. Overall, LD patterns in African populations show low concordance with those of European ancestry, which likely hindered our ability to detect colocalization in AA individuals. This LD discordance has been shown to affect the generalizability of PCa-risk alleles identified in studies of European ancestry men (Cook et al. 2014; Haiman et al. 2011; Han et al. 2015; Hoffmann et al. 2015; Virlogeux et al. 2015). Less than half of the 147 established PCa-risk SNPs have been replicated in AA men and more than 50% of these PCa-risk SNPs have higher risk allele frequencies in individuals of African ancestry [73]. Additionally, in the current study, we observed several examples of LD differences between AA and EA men, which has implications for the GWAS-meQTL pairs selected for co-localization testing (GWAS-meQTL pairs with LD (r^2) > 0.5 between the GWAS lead SNP and meQTL lead SNP to test for co-localization). Through this process, in AA men, we found 79% (29 of 37) benign and 92% (35 of 38) tumor tissue GWAS-meQTL pairs (FDR < 0.05) did not meet the LD threshold (r^2 > 0.5) between lead GWAS SNP and lead meQTL SNP. For example, in AA benign tissue, we identified a meQTL (cg23694490- rs834603) near *TNS3* but the lead meQTL SNP was in low LD (r^2 = 0.15) with the lead GWAS SNP (rs56232506). However, in EA men these same SNPs were in strong LD (r^2 = 0.86) and were co-localized. Similarly, in tumor tissue, the meQTL (cg20503657-rs72776479) near LARP4B were not tested for co-localization because the lead

meQTL SNP rs72776479 and lead GWAS SNP rs141536087 were in low LD ($r^2 = 0.11$) in AA men but were in strong LD in EA men ($r^2 = 0.79$). We performed co-localization analyses of these two examples and found evidence of likely co-localization for the GWAS-meQTL pair near LARP4B in AA tumor tissue (**Table S3.10** and **Figure S3.4**). These examples highlight the importance of LD structure and the need for large GWAS of AAs, particularly, if we are interested in exploring the functional role of GWAS hits in altering the epigenetic landscape.

Our study had limited small sample size, and we were likely underpowered to detect weak meQTLs. Additionally, our sample size limited our ability to test the 141 previously reported autosomal PCa-risk SNPs, as our meQTL analyses were restricted to SNPs with a MAF > 0.07, which excluded 21 PCa-risk SNPs in AA men and 16 in EA men. Overall, we had data available for 113 and 119 of the 141 autosomal PCa-risk SNPs in AA and EA men, respectively. While one strength of this study is the use of histologically benign paired samples, there is also the possibility that these samples may have some cancer-related molecular characteristics (as they were taken from prostates with detectable cancer) and are not representative of normal prostate tissue from PCa-free men of comparable age. Regarding generalizability, it is important to note our AA and EA cohorts are likely not representative of the PCa disparities observed nationwide, as in our study, EA patients have less favorable clinical features than the AA patients do. In the current study, we did not generate gene expression data, so we leveraged eQTL results from GTEx normal prostate tissue. Ideally, having paired gene expression data for both ancestries and tissue types would have allowed us to assess the direct impact of co-localized PCa-risk SNPs on gene expression in an ancestry and tissue specific manner. To better understand the biological mechanisms by which PCa susceptibility SNPs influence PCa biology, larger and more diverse methylation studies are needed.

In conclusion, we conducted a comprehensive search for meQTLs in tumor and paired benign prostate tissues of AA and EA men. We incorporated summary statistics from prior GWAS of PCa risk and conducted co-localization analyses to identify PCa susceptibility loci whose biological mechanism involves alteration of local DNA methylation. Complimentary GWAS-eQTL co-localization analyses provide additional insight into the functional role of these PCa-risk SNPs and DNA methylation features in gene regulation. Overall, these results provide a platform to explore the differences in the meQTL profiles of AA and EA men and highlights a few susceptibility regions providing a functional pathway for these SNPs in PCa.

4.5 APPENDIX

Figure 3. 1. Co-localization workflow of GWAS-meQTL pairs

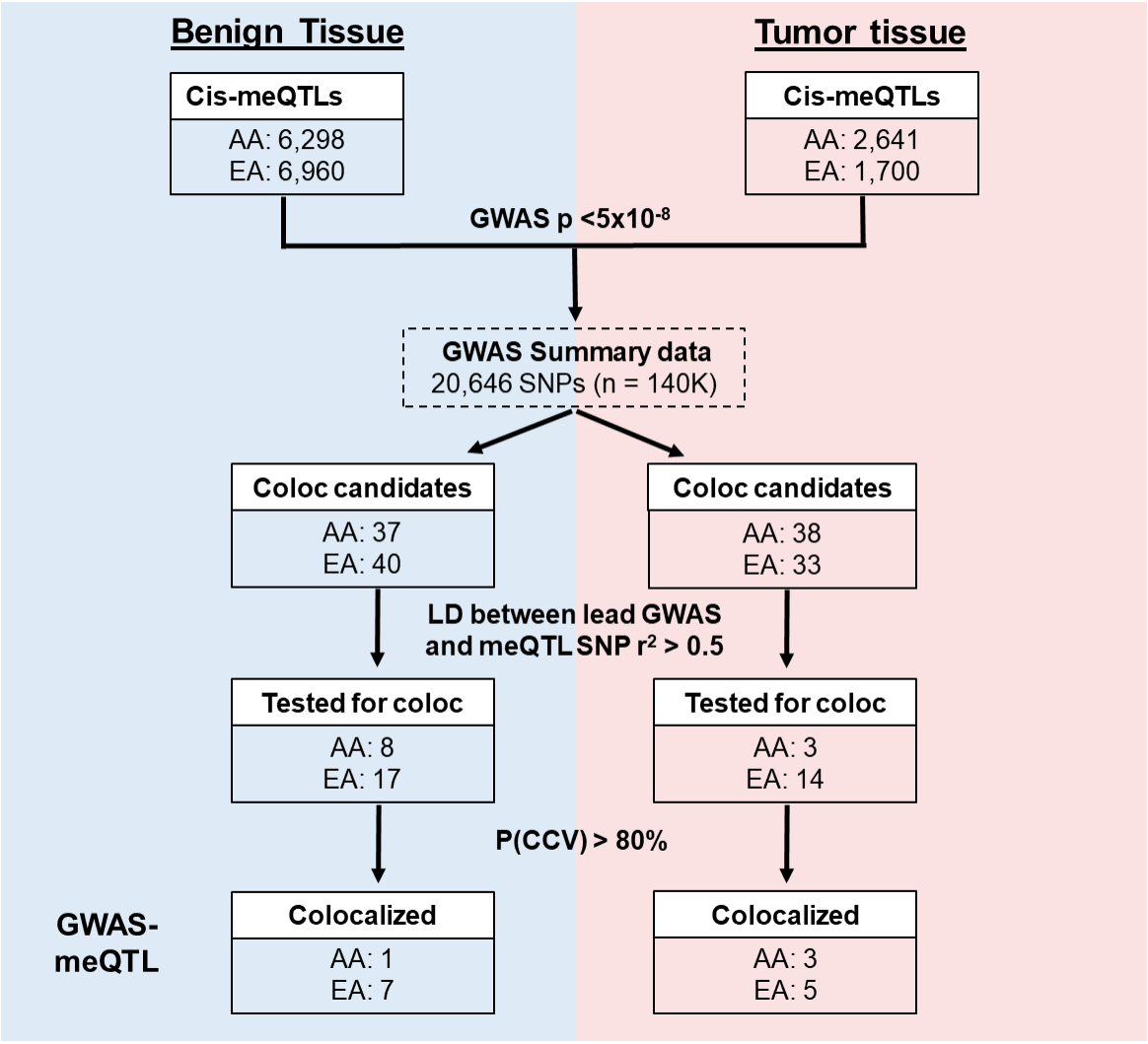


Table 3. 1. Characteristics of 150 prostatectomy patients recruited at the University of Chicago Medical Center

Patient Characteristics		**African American (N= 75)**	**European American (N= 75)**	**P**
Age	Years	65 (8.5)	65 (7.2)	0.93
Prior Cancer History	Yes	21 (28%)	16 (21.3%)	0.13
	No	30 (40%)	46 (61.3%)	
	Missing	24 (32%)	13 (17.4%)	
Family History of PCa	Yes	14 (18.7%)	13 (17.3%)	0.56
	No	37 (49.3%)	49 (65.3%)	
	Missing	24 (32%)	13 (17.4%)	
Gleason Score	<=6	4 (5.3%)	0	0.03
	7(3+4)	50 (66.7%)	41 (54.6%)	
	7(4+3)	15 (20%)	20 (26.7%)	
	>=8	6 (8%)	14 (18.7%)	
Prostate-specific Antigen (PSA)	ng/mL	57.5 (31.8)	58.2 (29.8)	0.9
pTNM	T2	51 (68%)	30 (40%)	0.001
	T3	24 (32%)	45 (60%)	
Percent of tumor volume	mean (SD)	18.3 (11.2)	22.5 (11.5)	0.33

Abbreviations: PCa, prostate cancer; pTNM, pathological tumor node metastasis staging

Table 3. 2. Summary of genome-wide cis-meQTLs identified in analyses stratified by ancestry and tissue type

	AA *cis*-meQTL analysis (n = 75)		**EA *cis*-meQTL analysis (n = 75)**	
	Benign	**Tumor**	**Benign**	**Tumor**
Genome-wide SNPs	n = 6,463,658		n = 4,900,500	
Genome-wide CpG sites	698,812	682,694	698,812	682,694
Significant SNP-CpG pairs[a]	6,298	2,641	6,960	1,700
Unique lead SNPs	5,855	2,434	6,496	1,586
Average distance between CpG and lead SNP (bp)	23,741	25,647	30,487	28,085

Abbreviations: bp, nucleotide basepairs

[a] SNP-CpG pairs meet false discovery rate (FDR) < 0.05 threshold

Table 3. 3. GWAS-meQTL pairs with a likely common causal variant in benign tissue

Chr	PCa-risk SNP	Direction of effect	Effect Allele	CpG	CpG Location	Feature	Nearest gene	$P(CCV)_1$	$P(CCV)_2$	$P(CCV)_3$	Tissue Specific
AA benign prostate tissue (n = 75)											
9	rs1048169	↓	C	cg10236024	Open Sea	Promoter	*HAUS6*	60.1%	82.0%	93.2%	No
EA benign prostate tissue (n = 75)											
2	rs2292884	↑	G	cg14458575	Shelf	Intergenic	*MLPH*	42.4%	69.0%	87.0%	Yes
3	rs10934853	↑	A	cg09718996	Open Sea	Body	*EEFSEC*	46.9%	72.7%	88.9%	Yes
6	rs7767188	↓	A	cg25540824	Shore	Body	*TRIM31*	25.4%	51.3%	76.1%	No
7	rs56232506	↑	A	cg23694490	Open Sea	Body	*TNS3*	58.7%	81.1%	92.8%	No
10	rs10993994	↓	C	cg17030820	Open Sea	Promoter	*MSMB*	81.2%	92.9%	97.5%	Yes
10	rs12768349*	--	G	cg20503657	Open Sea	Intergenic	*LARP4B*	35.0%	61.9%	83.1%	No
14	rs1004030	↓	C	cg18366651	Shore	Promoter	MRPL52	98.1%	99.4%	99.8%	Yes
17	rs142444269	↓	C	cg11677712	Open Sea	Intergenic	*COPRS*	77.9%	91.4%	97.0%	No

Figure 3. 2. Examples of co-localized GWAS-meQTL pairs in AA and EA benign tissue

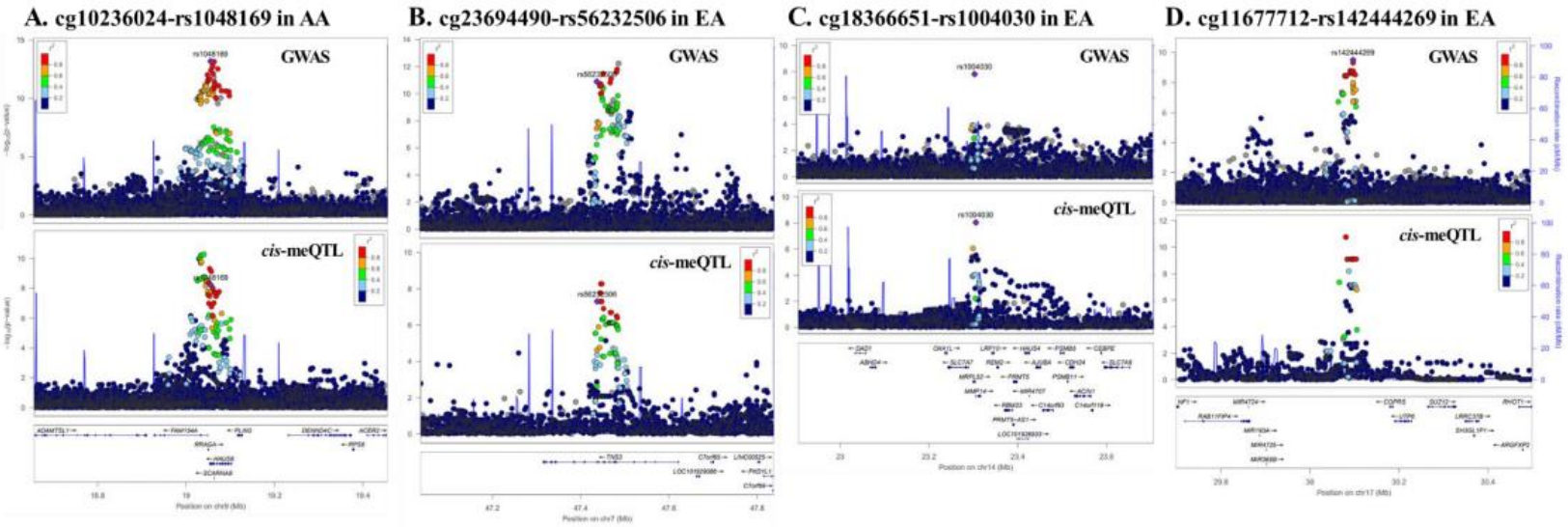


Table 3. 4. GWAS-meQTL pairs with a likely common causal variant in tumor tissue

Chr	PCa-risk SNP	Direction of effect	Effect allele	CpG	CpG Location	Feature	Nearest gene	$P(CCV)_1$	$P(CCV)_2$	$P(CCV)_3$	Tissue Specific
AA tumor prostate tissue (n = 75)											
5	rs12653946	↓	T	cg01859299	Open Sea	Body	*IRX4*	93.1%	97.6%	99.2%	Yes
11	rs11568818	↑	C	cg25511807	Open Sea	1st Exon/5'UTR	*MMP7*	96.0%	98.6%	99.5%	No
19	rs11666569	↑	T	cg19418318	Open Sea	Body	*MYO9B*	70.5%	87.8%	95.6%	No
EA tumor prostate tissue (n = 75)											
2	rs10187424	↓	C	cg02493740	Shore	Promoter	*GGCX*	53.2%	77.4%	91.2%	No
5	rs12653946	↑	T	cg09672187	Island	Intergenic	*IRX4*	93.2%	97.7%	99.2%	No
7	rs56232506	↑	A	cg04778099	Open Sea	Body	*TNS3*	56.6%	79.8%	92.2%	No
10	rs12768349*	↑	G	cg20503657	Open Sea	Intergenic	*LARP4B*	39.6%	66.4%	85.6%	No
11	rs11568818	↑	C	cg25511807	Open Sea	1st Exon/5'UTR	*MMP7*	98.3%	99.4%	99.8%	No

Figure 3. 3. Examples of co-localized GWAS-meQTL pairs in AA and EA tumor tissue

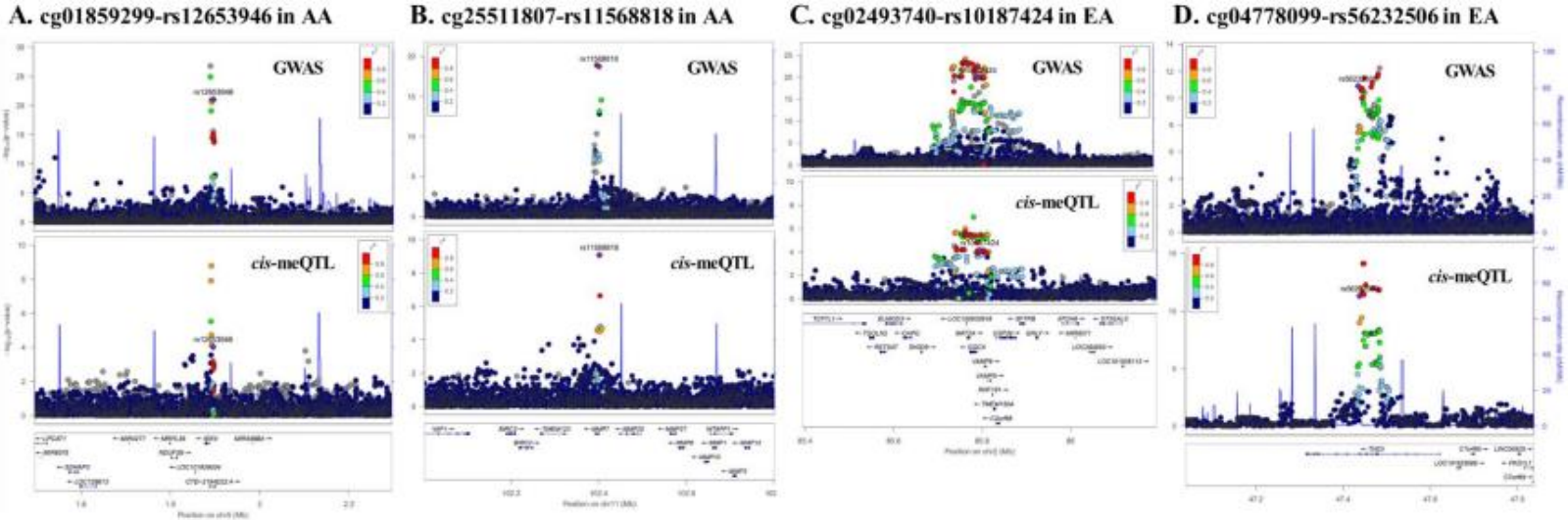


Table 3. 5. GWAS-eQTLpairs that are likely to share common causal variants based on GTEx normal prostate eQTL summary statistics

PCa-risk SNP	eGene	eQTL[a]		Co-localization P(CCV)[b]
		β	P-value	
Signals present in AA and EA tumor tissue				
rs12653946	*IRX4*	-0.4	1x10-14	100%
rs11568818	*MMP7*	-0.3	1x10-7	100%
Signal present AA tumor tissue				
rs11666569	*MYO9B*	-0.22	1x10-6	99.80%
Signal present in EA benign tissue				
rs10993994	*MSMB*	0.4	2x10-16	100%
Signal present in EA tumor tissue				
rs10187424	*GGCX*	0.48	1x10-10	99.90%

Abbreviations: eQTL, expression-Quantitative Trait Loci; P(CCV), probability of common causal variant

[a] eQTL summary statistics were obtained from the Genotype-Tissue Expression project (GTEx) normal prostate tissue

[b] The reported probability of common causal variant is based on a 75% prior probability that a GWAS SNP is an eSNP (see Table S3.1).

Figure 3. 4. Co-localized GWAS-eQTL pairs

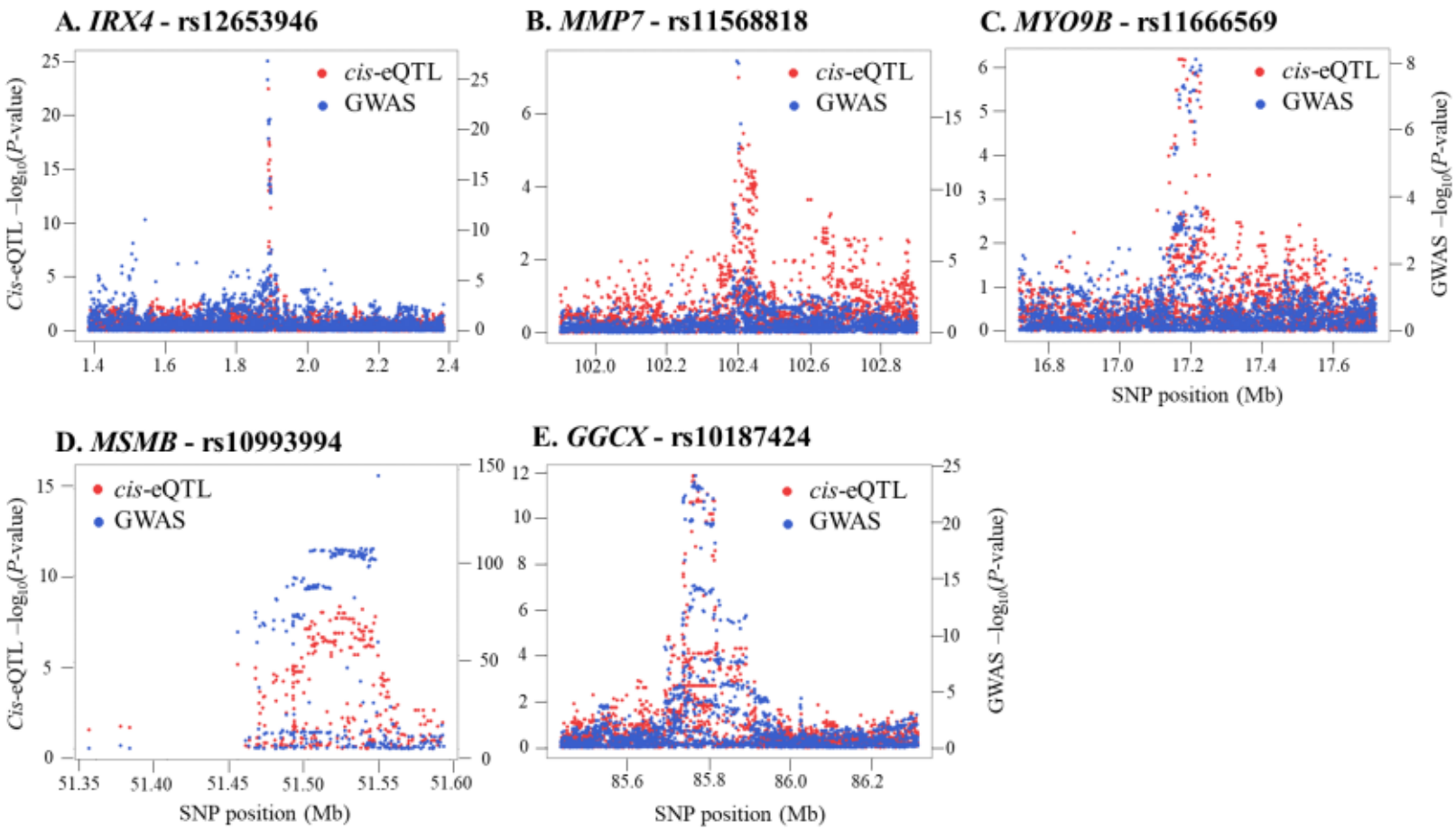


Table S3. 1. List of priors for GWAS-QTL co-localization analyses

Prior probability that a GWAS SNP is an meSNP/eSNP	**PP a SNP affects QTL and PCa (p_{12})[a]**	**PP a SNP affects PCa only (p_1)[b]**	**PP a SNP is meQTL only (p_2)[c]**	**PP a SNP is eQTL only (p_2)[c]**
10%	$1x10^{-5}$	$9x10^{-5}$	0.00499	0.00054
25%	$2.5x10^{-5}$	$7.5x10^{-5}$	0.00498	0.00045
50%	$5x10^{-5}$	$5x10^{-5}$	0.00495	0.0003
75%	$7.5x10^{-5}$	$2.5x10^{-5}$	0.00492	0.00015

Abbreviations: PP, prior probability, PCa, prostate cancer

[a] Probability a specific variant is causal for two traits (PCa GWAS and QTL)

[b] Probability a specific variant is causal for trait 1 (PCa GWAS)

[c] Probability a specific variant is causal for trait 2 (cis-meQTL or cis-eQTL)

Table S3. 2. Tissue specificity of cis-meQTLs in AA and EA men

Cohort	Benign tissue meQTLs	Replicated in tumor tissue (P < 0.01)[a]	Not replicated in tumor	Tumor tissue meQTLs	Replicated in benign tissue (P < 0.01)[a]	Not replicated in benign tissue
AA	6,298	3,945 (62%)	2,353 (38%)	2,461	2,089 (79%)	552 (21%)
EA	6,960	3,795 (55%)	3,165 (45%)	1,700	1,349 (79%)	351 (21%)

[a] Replication of the cis-meQTL in the opposite tissue does not imply consistent direction of effect across tissues.

Table S3. 3. Replication of 7,590 genome-wide tumor cis-meQTLs reported in Houlahan et al., 2019

Cohort	meQTLs tested	Lead CpG P<0.001	SNP-CpG pair P<0.01[a]
AA Benign	5,231 (69%)	4,179 (80%)	2,490 (47%)
AA Tumor		4,121 (79%)	2,294 (44%)
EA Benign	5,590 (74%)	3,814 (68%)	3,287 (59%)
EA Tumor		3,354 (60%)	2,761 (49%)

[a] Consistency in direction of effect for SNP-CpG pairs (meQTLs) replicated at p<0.01 was >99% across all cohorts and tissue types.

Table S3. 4. Replication of 52 PCa-risk cis-meQTLs reported in Houlahan et al., 2020

Cohort	meQTLs tested	Lead CpG P<0.001	SNP-CpG pair P<0.01[a]
AA Benign	41 (54%)	5 (12%)	12 (29%)
AA Tumor		13 (31%)	13 (31%)
EA Benign	45 (59%)	9 (20%)	14 (31%)
EA Tumor		21 (47%)	27 (60%)

[a] Consistency in direction of effect for SNP-CpG pairs (meQTLs) replicated at p<0.01 was >99% across all cohorts and tissue types

Table S3. 5. Replication of 110 autosomal PCa-risk cis-meQTLs ($p < 1x10^{-9}$) reported in Dai et al., 2020

Cohort	meQTLs tested	Lead CpG P<0.001	SNP-CpG pair P<0.01[a]
AA Benign	60 (55%)	14 (23%)	9 (15%)
AA Tumor		17 (28%)	25 (42%)
EA Benign	67 (61%)	19 (28%)	19 (28%)
EA Tumor		25 (37%)	33 (49%)

[a] Consistency in direction of effect for SNP-CpG pairs (meQTLs) replicated at p<0.01 was >99% across all cohorts and tissue types

Table S3. 6. Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in AA benign tissue

Chr	Pca-risk SNP	CpG	meQTL-lead SNP	fdr	Nearest Gene	r^2 (ASW)[a]
1	rs1218582	cg06221963	1:154845484	1.35E-06	*KCNN3*	0.49
1	rs1218582	cg09359103	1:154845484	0.026660272	*KCNN3*	0.49
2	rs721048	cg21115199	2:63106784	1.41E-07	*EHBP1*	0.03
2	rs12621278	cg07833011	2:173419805	0.02738296		0.07
2	rs12621278	cg05186902	2:173419805	6.23E-05		0.07
2	rs2292884	cg14458575	2:238379582	0.000417387	*MLPH*	0.03
3	rs2660753	cg04477660	3:87156130	0.004005837	*LINC00506*	0.58
3	rs10934853	cg02080175	3:127812349	0.022012275	*EEFSEC*	0.12
3	rs10934853	cg15808905	3:127991149	0.0041958	*EEFSEC*	0.93
3	rs10934853	cg21710443	3:128032991	1.39E-07	*EEFSEC*	0.2
6	rs9296068	cg06460587	6:31601022	0.048483276		0.08
6	rs9296068	cg25769566	6:31601022	1.07E-05		0.08
6	rs9296068	cg11991824	6:31601022	0.002376144		0.08
6	rs9296068	cg25417675	6:31601513	0.008264934		0.08
6	rs9296068	cg01297670	6:31601513	0.008379021		0.08
6	rs9296068	cg07180897	6:32628407	0.00541341		0
6	rs9296068	cg16345566	6:32628420	0.001741196		0
6	rs9296068	cg19887824	6:32797537	0.041108523		0.01
6	rs9296068	cg25744682	6:32797620	2.49E-07		0.01
6	rs9364554	cg12753515	6:160700769	0.020731466		0.02
7	rs56232506	cg23694490	7:47447921	0.005933071	*TNS3*	0.15
7	rs56232506	cg04778099	7:47456846	0.003529794	*TNS3*	0.14
7	rs6465657	cg14507403	7:97746387	2.01E-07	*LMTK2*	0.02
9	rs1048169	cg10236024	9:19040237	0.000185396	*HAUS6*	0.72
10	rs141536087	cg20503657	10:835542	0.000230498	*LARP4B*	0.11
11	rs7127900	cg03628333	11:2214804	0.042616238	*INS-IGF2*	0.005
11	rs7127900	cg25635251	11:2229361	0.003407111	*INS-IGF2*	0.70
11	rs7931342	cg16166568	11:68986495	0.001318723		1
12	rs902774	cg07143715	12:53354788	0.028558121	*KRT8*	0.02
12	rs902774	cg05393297	12:53357713	0.003616986	*KRT8*	0
12	rs902774	cg14581129	12:53358352	3.46E-05	*KRT8*	0
12	rs902774	cg16329197	12:53359396	0.000722543	*KRT8*	0
19	rs11672691	cg03275582	19:42093112	0.035110309	*PCAT19*	0.57
20	rs6062509	cg00331541	20:62221249	5.96E-06	*GMEB2*	0
20	rs6062509	cg13591364	20:62221249	0.005727494	*GMEB3*	0

Table S3. 6., continued: Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in AA benign tissue

20	rs6062509	cg04747225	20:62221249	3.59E-12	*GMEB4*	0
20	rs6062509	cg03425785	20:62223238	1.64E-11	*GMEB4*	0

Abbreviations: FDR, false discovery rate

[a] Linkage disequilibrium between PCa-risk SNP and meQTL lead SNP using the Americans of African Ancestry in SW USA reference panel obtained from LDlink (https://ldlink.nci.nih.gov/)

Table S3. 7. Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in EA benign tissue

Chr	Pca-risk SNP	CpG	meQTL-lead SNP	fdr	Nearest Gene	r^2 (EUR)[a]
1	rs17599629	cg18288576	1:150554876	0.045298131	ENSA	0.02
1	rs1218582	cg06221963	1:154916106	0.000575044	KCNN3	0.68
1	rs1218582	cg09359103	1:154917314	0.000664881	KCNN3	0.89
2	rs721048	cg21115199	2:63208460	1.15E-05		0.12
2	rs721048	cg27648677	2:63208460	0.010804854		0.12
2	rs2292884	cg14458575	2:238354982	2.88E-07	MLPH	0.62
2	rs2292884	cg14070755	2:238419959	0.002013617	MLPH	0.47
2	rs2292884	cg06484157	2:238446678	0.036931166	MLPH	0.62
3	rs10934853	cg09718996	3:127898501	0.004530767	EEFSEC	0.88
3	rs10934853	cg21710443	3:128029008	3.05E-06	EEFSEC	0.33
4	rs12500426	cg12486630	4:95486094	0.001493265		0.78
6	rs7767188	cg25540824	6:30139699	7.29E-06	TRIM31	0.59
6	rs7767188	cg27136023	6:30163955	0.002124842	TRIM31	0.04
6	rs7767188	cg03084824	6:30163955	0.001732241	TRIM31	0.04
6	rs3129859	cg07180897	6:32634952	0.000830442		0.03
6	rs9364554	cg13221458	6:160123425	0.021517143		0.001
6	rs9364554	cg20372956	6:160148348	5.80E-05		0.001
7	rs10486567	cg14849571	7:27482574	0.014816555		0.12
7	rs56232506	cg23694490	7:47447921	0.001581175	TNS3	0.86
7	rs6465657	cg14507403	7:97715892	1.74E-06		0.3
7	rs6465657	cg03626541	7:98039459	0.000190608		0.4
10	rs141536087	cg20503657	10:835015	0.000148491	LARP4B	0.84
10	rs10993994	cg17030820	10:51549496	0.046778372	MSMB	1

Table S3. 7., continued: Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in EA benign tissue

10	rs3850699	cg11453585	10:104227661	0.008658471		0.24
10	rs3850699	cg13349042	10:104444615	0.011893743	TRIM8	0.76
11	rs7127900	cg25635251	11:2229645	6.25E-06	INS-IGF2	0.52
11	rs7127900	cg13130588	11:2284590	0.010421232		0.11
11	rs7931342	cg16166568	11:68986495	5.32E-05		0.99
12	rs902774	cg06553033	12:53269393	0.010033005		0.02
12	rs902774	cg16329197	12:53355557	4.54E-05		0.29
12	rs902774	cg14581129	12:53359229	2.15E-09		0.22
12	rs902774	cg05393297	12:53359416	0.000146349		0.22
14	rs1004030	cg18366651	14:23305649	0.002816148	MRPL52/MMP14	1
17	rs684232	cg15660573	17:507120	0.000133497	VPS53	0.49
17	rs142444269	cg11677712	17:30080257	3.78E-05		0.91
20	rs6091758	cg24591615	20:52417890	8.60E-06		0.34
20	rs6062509	cg00331541	20:62221249	2.06E-07		0.16
20	rs6062509	cg04747225	20:62221249	4.05E-05		0.16
20	rs6062509	cg03425785	20:62221249	0.013696971		0.16
22	rs58133635	cg05033341	22:40428433	0.048557405	Nearest Gene	0.5

Abbreviations: FDR, false discovery rate

[a] Linkage disequilibrium between PCa-risk SNP and meQTL lead SNP using European reference populations obtained from LDlink (https://ldlink.nci.nih.gov/)

Table S3. 8. Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in AA tumor tissue

Chr	Pca-risk SNP	CpG	meQTL-lead SNP	fdr	Nearest Gene	r^2 (ASW)[a]
2	rs58235267	cg13883027	2:62895132	2.07E-08	*OTX1*	0.11
2	rs58235267	cg21115199	2:63081759	0.019459356	*OTX1*	0.16
2	rs2292884	cg21367213	2:238398268	0.003539668	*MIR6811*	0.004
3	rs10934853	cg07228336	3:127812349	0.001223566	*EEFSEC*	0.12
3	rs10934853	cg03115093	3:127868321	0.03911163	*EEFSEC*	0.1
3	rs10934853	cg21710443	3:128018447	0.009313477	*EEFSEC*	0.2
5	rs12653946	cg01859299	5:1891800	0.003479609	*IRX4*	0.43
6	rs7767188	cg27136023	6:30167835	2.21E-06	*TRIM31*	0.04
6	rs7767188	cg03084824	6:30167835	9.92E-05	*TRIM26*	0.04
6	rs7767188	cg14799927	6:30169327	0.042546379	*TRIM40*	0.04
6		cg02304584	6:31601012	0.034150752		
6		cg02389040	6:31601022	0.034264083		
6	rs3096702	cg24147543	6:32602396	0.002715729	*C6orf10*	0.04
6	rs3096702	cg07180897	6:32629091	8.66E-06	*C6orf10*	0.004
6	rs3096702	cg12296550	6:32629146	0.008164084		0
6	rs3129859	cg25744682	6:32797620	5.88E-06	*HLA*	0.06
6	rs1933488	cg15031989	6:153381725	4.72E-07	*RGS17*	0.32
6	rs1933488	cg16924337	6:153445965	0.048043418	*RGS17*	0.38
7	rs10486567	cg22168087	7:27642318	0.000760697	*JAZF1*	0.006
7	rs56232506	cg04778099	7:47463842	0.003099609	*TNS3*	0.15
7	rs56232506	cg23694490	7:47463937	0.001623792	*TNS3*	0.15
7	rs6465657	cg14507403	7:97798117	1.59E-07	*LMTK2*	0.02
10	rs141536087	cg20503657	10:835542	0.000357923	*LARP4B*	0.11

Table S3. 8., continued: Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in AA tumor tissue

10	rs3850699	cg11453585	10:104290462	0.002715729	*TRIM8*	0.14
11		cg25635251	11:2229361	0.000114256		
11	rs11568818	cg25511807	11:102401661	0.001420392	*MMP7*	1
12	rs902774	cg06553033	12:53268217	0.033504607		0.14
12	rs902774	cg08553726	12:53270845	0.000387728		0.19
12	rs902774	cg05393297	12:53355955	2.43E-06	*KRT18*	0
12	rs902774	cg14581129	12:53356411	0.041747845	*KRT18*	0
12	rs902774	cg16329197	12:53357710	0.001210907	*KRT18*	0
12	rs902774	cg25025879	12:53359416	0.021842567	*KRT18*	0
19	rs11666569	cg19418318	19:17227335	0.005588234	*MYO9B*	0.5
20	rs6062509	cg00331541	20:62222392	1.23E-05	*RTEL1*	0.002
20	rs6062509	cg04747225	20:62230706	1.42E-07	*RTEL1*	0.005
20	rs6062509	cg03425785	20:62230706	8.36E-05	*RTEL1*	0.005
20	rs6062509	cg06363034	20:62251525	0.023888337	*RTEL1*	0.003
21	rs1041449	cg04615233	21:42893908	0.004572751		0.1

Abbreviations: FDR, false discovery rate

[a] Linkage disequilibrium between PCa-risk SNP and meQTL lead SNP using the Americans of African Ancestry in SW USA reference panel obtained from LDlink (https://ldlink.nci.nih.gov/)

Table S3. 9. Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in EA tumor tissue

Chr	Pca-risk SNPs	CpG	meQTL-lead SNP	fdr	Nearest Gene	r^2 (EUR)[a]
1		cg12898220	1:205759195	8.78E-07		
1		cg24503407	1:205759195	7.03E-07		
2	rs721048	cg13883027	2:62895132	2.04E-07	*EHBP1*	0.1
2	rs721048	cg21115199	2:63204611	0.00047636	*EHBP1*	0.14
2	rs10187424	cg02493740	2:85780536	0.04406587	*GGCX*	0.42
5	rs12653946	cg14823763	5:1890877	0.00214902	*IRX4*	0.58
5	rs12653946	cg09672187	5:1890877	0.00064843	*IRX4*	0.58
5	rs12653946	cg06874119	5:1890877	0.01241014	*IRX4*	0.58
5	rs12653946	cg03587843	5:1891174	0.00472073	*IRX4*	0.78
5	rs12653946	cg26195178	5:1891174	0.00083819	*IRX4*	0.78
5	rs12653946	cg11279838	5:1891174	3.72E-05	*IRX4*	0.78
5	rs12653946	cg00626856	5:1891174	0.00045354	*IRX4*	0.78
5	rs12653946	cg14051264	5:1891174	3.82E-05	*IRX4*	0.78
6	rs3129859	cg18067840	6:32399187	0.03124031	*HLA*	0.14
6	rs3129859	cg13081526	6:32409656	0.00346305	*HLA*	0.15
6	rs3129859	cg07180897	6:32628698	0.00032266	*HLA*	0.04
6	rs9296068	cg25744682	6:32797620	0.00539261	*HLA*	0.05
7	rs10486567	cg22168087	7:27564744	0.00897933	*JAZF1*	0.21
7	rs56232506	cg04778099	7:47445812	6.27E-07	*TNS3*	0.86
7	rs56232506	cg23694490	7:47445812	0.00042648	*TNS3*	0.86
7	rs6465657	cg14507403	7:97739317	0.00108286	*LMTK2*	0.3
10	rs141536087	cg20503657	10:836211	0.0003236	*LARP4B*	0.79
11	rs7127900	cg03628333	11:2201059	0.04617367		0.22
11	rs7127900	cg25635251	11:2233951	8.16E-06		0.63
11	rs11568818	cg25511807	11:102396607	0.00051913	*MMP7*	0.99
12	rs902774	cg06553033	12:53268734	0.00059873	*KRT8*	0.02

Table S3. 9., continued: Cis-meQTLs residing in same location as PCa-risk SNPs with $p < 5x10^{-8}$ in the Schumacher et al., 2018 GWAS of PCa in EA tumor tissue

12	rs902774	cg08553726	12:53269393	0.00027723	*KRT8*	0.02
12	rs902774	cg14581129	12:53359416	0.04650477	*KRT8*	0.22
12	rs902774	cg05393297	12:53359416	6.56E-05	*KRT8*	0.22
12	rs902774	cg25025879	12:53359416	2.34E-05	*KRT8*	0.22
12	rs902774	cg16329197	12:53359416	3.48E-05	*KRT8*	0.22
20	rs6062509	cg19962434	20:62238083	0.0174872	*ZGPAT*	0.18
20	rs6062509	cg00331541	20:62268149	0.00666881	*ZGPAT*	0.14

Abbreviations: FDR, false discovery rate

[a] Linkage disequilibrium between PCa-risk SNP and meQTL lead SNP using European reference populations obtained from LDlink (https://ldlink.nci.nih.gov/)

Table S3. 10. Co-localization analyses of AA GWAS-meQTL pairs where lead GWAS and lead meQTL SNPs do not meet LD < 0.5 threshold

Chr	PCa-risk SNP	CpG	meQTL SNP	LD (r^2)	Nearest gene	$P(CCV)_1$	$P(CCV)_2$	$P(CCV)_3$
7	rs56232506	cg23694490	rs834603	0.15	*TNS3*	27.5%	53.4%	77.6%
10	rs12768349*	cg20503657	rs72776479	0.11	*LARP4B*	43.2%	69.6%	87.4%

*Proxy for rs141536087 ($r^2 > 0.9$)

Figure S3. 1. Locational (with relation to island) distribution of benign and tumor tissue CpG targets

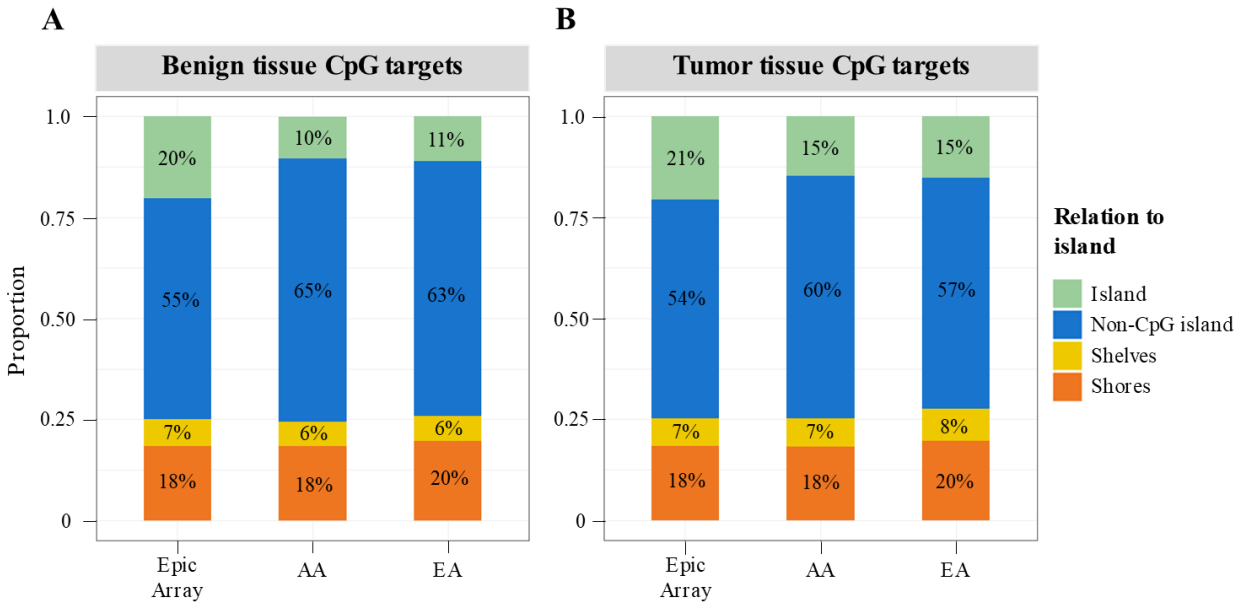


Figure S3. 2. Genomic features of benign and tumor tissue CpG targets

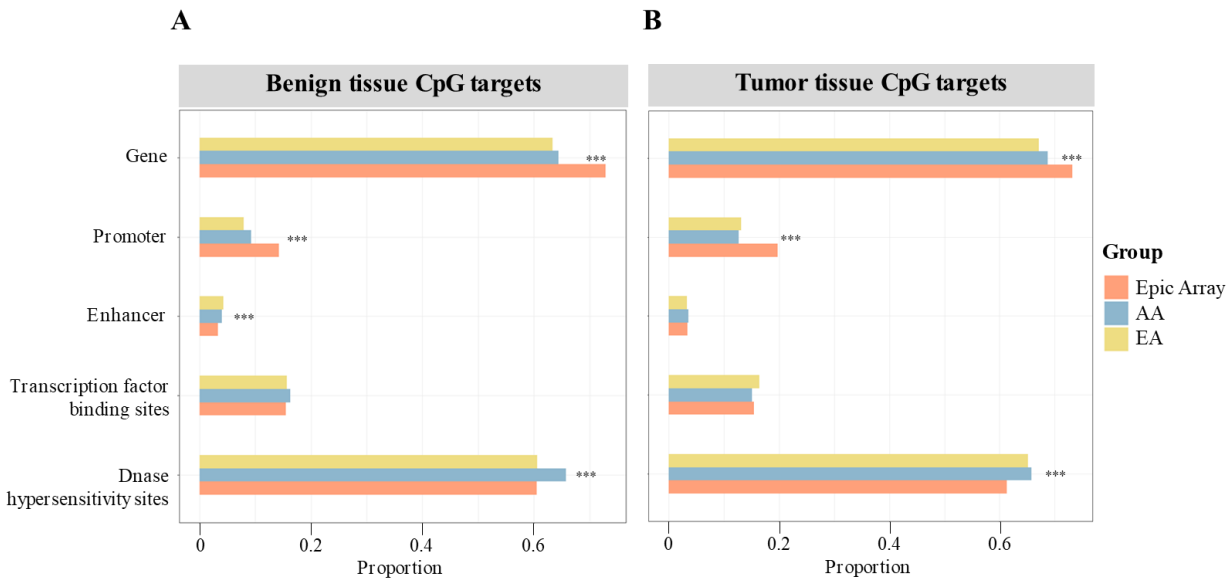


Figure S3. 3. Replication of *cis*-meQTLs in the opposite ancestry

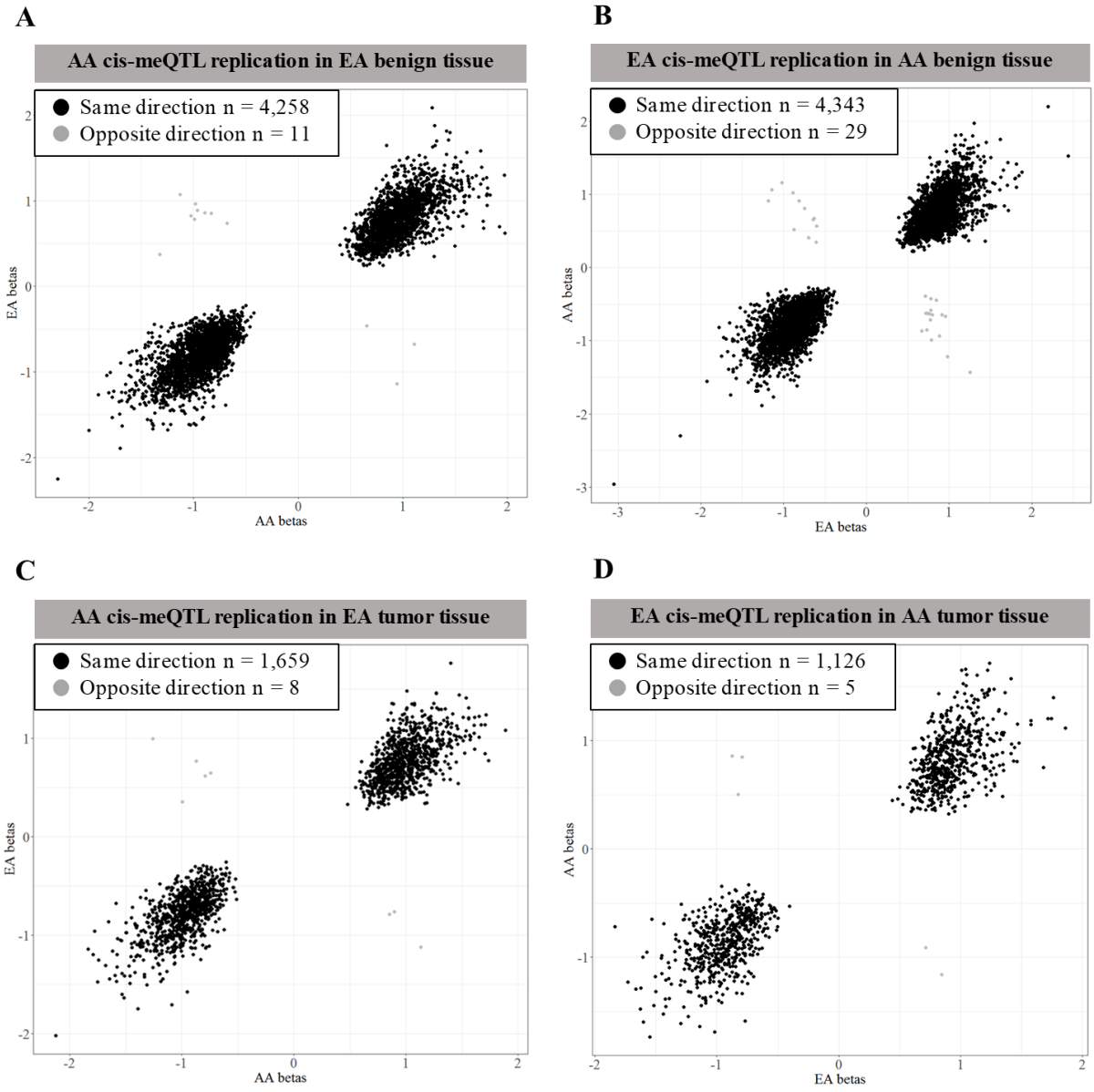
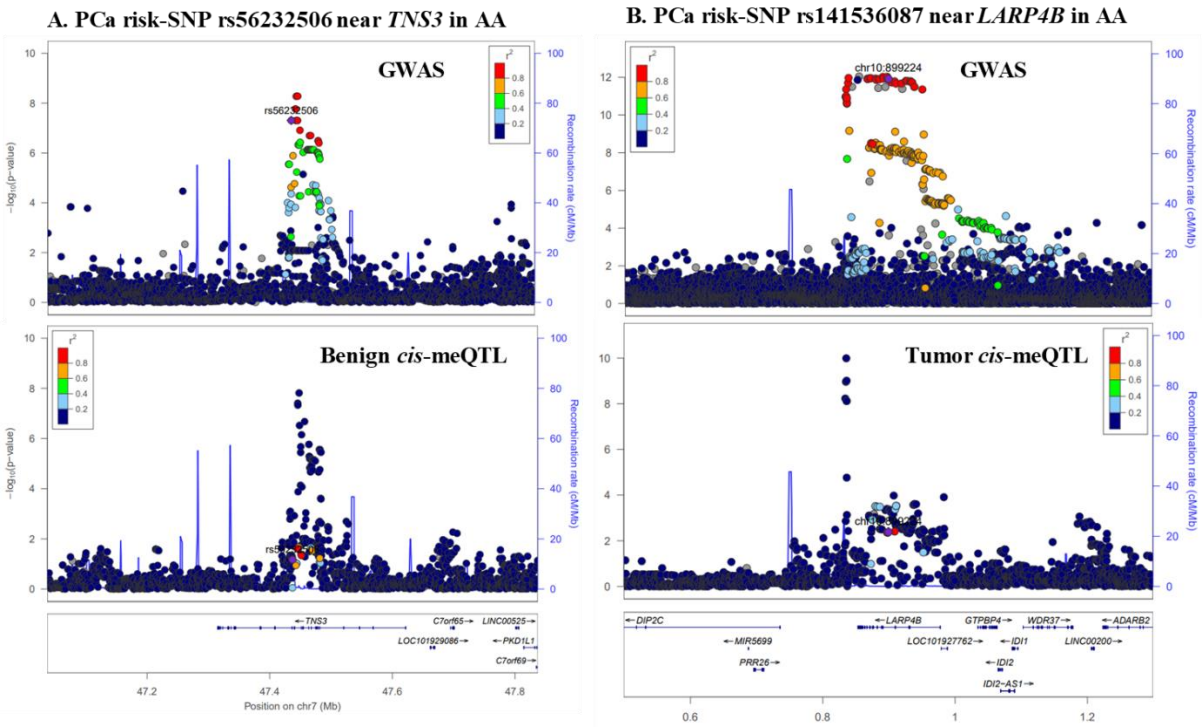


Figure S3. 4. Regional association plots of GWAS-meQTL pairs in AA where lead GWAS and lead meQTL SNPs do not meet LD < 0.5 threshold



CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

5.1 SUMMARY

In this work, we leveraged genetic, epigenetic, and biomarkers of environmental exposures in multiple ancestral populations, including Bangladeshi, Native American, African American, and European American, to investigate gene-environment relationships. We provide evidence that rare variants in *AS3MT* decrease arsenic metabolism efficiency (AME), a finding that was consistent across multiple populations with distinct ancestral backgrounds. We also report suggestive evidence that high AME is associated with a decreased risk of hypertension (HTN). Lastly, we demonstrate a potential regulatory mechanism by which some prostate cancer (PCa) risk SNPs modify local DNA methylation and local gene-expression, while also highlighting specific limitations to understanding the biological mechanisms underlying PCa disparities attributed to the lack of AA representation in large genetic and epigenetic studies.

In Chapter 2, we report for the first time a targeted sequencing study of rare variants in the coding regions of the *AS3MT* gene across three arsenic-exposed cohorts (HEALS, SHS, and NHSCC) of different genetic ancestry. We hypothesized that rare variants in *AS3MT* contribute to inter-individual variation in AME. To test our hypothesis, we performed burden tests stratified by cohort and conducted complementary analyses using the SKAT framework. We used both burden and non-burden (SKAT) gene-based tests because it was unknown which method would provide superior power and if the assumptions of the burden test were reasonable. Meta-analyses

estimates for the effect of rare variants in *AS3MT* on AME across cohorts under the burden test and SKAT framework were also estimated. Our analyses revealed 13 rare, protein-altering variants in *AS3MT*, and carrying on of these variants was associated with a 9% decrease in AME, an association that was independent of the effects of known common variants in this region. These results were also consistent in analyses of MMA% and iAs% as secondary outcomes. This work was limited in that we lacked data for exons 5 and 10. Due to our lack of data in exons 5 and 10, we estimate that approximately 0.5% of our participants are rare variant carriers we were unable to classify as carriers. Additionally, we were unable to examine the association between rare variants in *AS3MT* and arsenic related outcomes due to the small number of rare variant carriers. However, we attempted to address this limitation by incorporating previously reported risk estimates for lung and bladder cancer and skin lesions. We estimate that being a carrier of a rare variant in *AS3MT* increases the risk of skin lesions by 82%, lung cancer by 59%, and bladder cancer by 19%.

In Chapter 3, we use Mendelian Randomization (MR) methods (two-stage sequential regression and inverse-variance weighted meta-analysis) to demonstrate a small inverse association between genetically predicted DMA% (measure of AME) and HTN risk among 7,895 Bangladeshi individuals. MR methods suggest a unit increase in DMA% is associated with a 3% decrease in hypertension risk. We leverage longitudinal blood pressure data (systolic and diastolic blood pressure, SBP and DBP, respectively) in a mixed-effects approach and report a lack of a clear association between predicted DMA% and repeated measures of SBP and DBP. Complimentary analyses using measured DMA% (on a subset of 4,814 individuals with metabolite data) support a lack of association with repeated measures of SBP and DBP. Overall, these comprehensive analyses along with prior analyses, suggest AME has a weak effect, if any,

on HTN risk. Prior studies may have been unable to detect such a weak association due to lack of statistical power. Our work was limited in that a subset of our measures of arsenic metabolism phenotypes were generated among individuals from case-groups (i.e., cardiovascular disease, respiratory outcomes, diseased, and skin lesions), which is not ideal for looking at environmental effects. For this reason, we do not thoroughly assess the association between arsenic exposure and hypertension phenotypes in this work, but focus on genetic effects. Although, we did conduct exploratory analyses where we assessed whether predicted AME modified the association between arsenic exposure and hypertension status, and found no clear evidence of interaction.

In Chapter 4, we implemented a genome-wide analysis to characterize the effects of genetic variation on the DNA methylation at nearby CpG sites, known as cis-methylation-quantitative trait loci (cis-meQTLs), in the tumor and benign tissues of AA and EA men. This characterization of meQTLs in PCa tissue was the first such study to focus on AA men. Additionally, we used our meQTL results to identify PCa GWAS and cis-meQTL association signals that share a common causal variant, using co-localization methods. We report one GWAS-meQTL pair in benign tissue and three GWAS-meQTL pairs in tumor tissue of AA men. In EA men, we found evidence of co-localization for seven and five GWAS-meQTL pairs in benign and tumor tissue, respectively. By integrating GTEx normal prostate tissue eQTL summary data, we provide evidence suggesting at least five PCa-risk SNPs increase risk for PCa by modifying local DNA methylation leading to measurable changes in gene expression at nearby genes (*IRX4, MMP7, MYO9B, MSMB,* and *GGCX*), all of which have known roles in tumorigenesis. One of the primary limitations and challenges in conducting co-localization analyses using meQTL results from AA was the LD mismatch with the largely European

ancestry GWAS of PCa risk. Because of this discordance in LD patterns, we detected fewer GWAS-meQTL co-localized pairs in AA men. However, through this analysis we were able to make a stronger argument for future genetic studies to have adequate representation of AA men (and generally all under-represented ancestries) in genetic studies. Another limitation was the small sample size, which contributed to limited power to detect meQTLs and also limited the number of PCa-risk SNPs we were able to test (because of MAF restrictions). Additionally, lack of gene expression data prevented us from assessing the impact of PCa-risk meSNPs on the transcription of nearby genes in a tissue and ancestry specific manner. We attempted to assess this relationship using GTEx eQTL summary results, however, only normal prostate tissue from primarily EA men was available.

5.2 FUTURE DIRECTIONS

The presented research addresses a number of gaps in the literature in terms of individual genetic variation contributing to susceptibility to arsenic toxicity and the underling biology contributing to PCa risk in AA and EA men. Our findings highlight a few promising area for further development.

In this dissertation, we present evidence supporting the role of rare, protein altering *AS3MT* variants in AME across multiple populations and environmental backgrounds (Chapter 2). A natural next step would be to validate these findings in other arsenic-exposed populations (i.e., Mexico, Chile, Argentina, India, China, Cambodia, Vietnam, and Pakistan) and identify additional shared and population-specific rare variants. This type of work would be very informative not only for understanding disease etiology and risk, but also for understanding evolutionary pressures at play in arsenic endemic regions. Another area for future research is to

assess the impact of rare *AS3MT* variants on risk for arsenic toxicities in a larger cohort, to confirm that inefficient metabolism results in increased risk for arsenic-related disease. Interactions of AME with sex, lifestyle factors, nutritional status and overall iAs exposure, are particularly useful to understand genetic risk in different contexts. Groundwater (the primary source of naturally occurring arsenic) is becoming increasingly used worldwide to compensate for growing populations and water scarcity. Thus, elucidating the rare and common genetic components that influence an individual's arsenic metabolism efficiency will become increasingly important in order to identify individuals at high-risk for arsenic toxicities. This research also has important implications for policymakers and health organizations worldwide, including the World Health Organization (WHO).

Another future direction for additional research involves a large collaboration across multiple cohorts (in addition to HEALS and BEST) with different levels of exposure (low, moderate, and high) to establish the effect of AME on HTN risk. Our study suggests a weak association between AME and HTN. Thus, a large meta-analysis could provide stronger evidence of this association (or lack thereof). Ideally, genetic data on the determinants of AME and metabolite data would be available across cohorts. This type of analysis would require validation of AME-related SNPs across ancestries.

Finally, in a post-GWAS era, the availability of GWAS summary statistics are critical in order to elucidate the biological mechanisms through which GWAS hits affect PCa susceptibility. Because LD patterns are known to vary substantially across ancestral populations, the generalizability of PCa-risk alleles discovered in largely European ancestry GWAS, show low concordance in individuals of African ancestry. Our results from Chapter 4 highlight the importance of LD structure and the need for larger GWAS of AAs (and other underrepresented

populations). The largest GWAS to date [1] is ~80% European, but included 10,368 cases and 10,986 controls of African ancestry, which is the largest representation of AA men in PCa GWAS. The summary statistics from this recent GWAS were not available at the time of our analysis. These data represent a unique opportunity to understand PCa disparities. However, larger and more diverse methylation and gene expression studies are also needed in order to execute integrative multi-omics analyses that will shed light on the functional role of GWAS hits.

In conclusion, the work presented in this dissertation addressed notable gaps in our understanding of genetic susceptibility in terms of arsenic exposure, metabolism, and toxicity for arsenic-related outcomes, and the functional mechanisms underlying PCa-risk SNPs in AA and EA men. We demonstrated the contribution of *AS3MT* rare variants to AME, provided suggestive evidence for the association between AME and hypertension risk, and highlighted a few susceptibility regions providing a functional pathway for PCa-risk SNPs in PCa while also providing a platform to continue to explore the differences in the meQTL profiles of AA and EA men.

REFERENCES

1. Naujokas, M.F., et al., *The broad scope of health effects from chronic arsenic exposure: update on a worldwide public health problem.* Environ Health Perspect, 2013. **121**(3): p. 295-302.
2. Podgorski, J. and M. Berg, *Global threat of arsenic in groundwater.* Science, 2020. **368**(6493): p. 845-850.
3. Shaji, E., et al., *Arsenic contamination of groundwater: A global synopsis with focus on the Indian Peninsula.* Geoscience Frontiers, 2021. **12**(3): p. 101079.
4. Mohammed Abdul, K.S., et al., *Arsenic and human health effects: A review.* Environmental Toxicology and Pharmacology, 2015. **40**(3): p. 828-846.
5. Nachman, K.E., et al., *Mitigating dietary arsenic exposure: Current status in the United States and recommendations for an improved path forward.* The Science of the total environment, 2017. **581-582**: p. 221-236.
6. Ahsan, H., et al., *Arsenic metabolism, genetic susceptibility, and risk of premalignant skin lesions in Bangladesh.* Cancer Epidemiol Biomarkers Prev, 2007. **16**(6): p. 1270-8.
7. Sengupta, S.R., N.K. Das, and P.K. Datta, *Pathogenesis, clinical features and pathology of chronic arsenicosis.* Indian J Dermatol Venereol Leprol, 2008. **74**(6): p. 559-70.
8. Moon, K.A., et al., *Association between exposure to low to moderate arsenic levels and incident cardiovascular disease. A prospective cohort study.* Ann Intern Med, 2013. **159**(10): p. 649-59.
9. Kuo, C.C., et al., *The Association of Arsenic Metabolism with Cancer, Cardiovascular Disease, and Diabetes: A Systematic Review of the Epidemiological Evidence.* Environ Health Perspect, 2017. **125**(8): p. 087001.
10. Tyler, C.R. and A.M. Allan, *The Effects of Arsenic Exposure on Neurological and Cognitive Dysfunction in Human and Rodent Studies: A Review.* Current Environmental Health Reports, 2014. **1**(2): p. 132-147.
11. Tolins, M., M. Ruchirawat, and P. Landrigan, *The Developmental Neurotoxicity of Arsenic: Cognitive and Behavioral Consequences of Early Life Exposure.* Annals of Global Health, 2014. **80**(4): p. 303-314.
12. Almberg, K.S., et al., *Arsenic in drinking water and adverse birth outcomes in Ohio.* Environ Res, 2017. **157**: p. 52-59.
13. Ettinger, A.S., et al., *Arsenic levels among pregnant women and newborns in Canada: Results from the Maternal-Infant Research on Environmental Chemicals (MIREC) cohort.* Environ Res, 2017. **153**: p. 8-16.
14. Argos, M., et al., *Arsenic exposure from drinking water, and all-cause and chronic-disease mortalities in Bangladesh (HEALS): a prospective cohort study.* Lancet, 2010. **376**(9737): p. 252-258.
15. Argos, M., et al., *A Prospective Study of Arsenic Exposure From Drinking Water and Incidence of Skin Lesions in Bangladesh.* American Journal of Epidemiology, 2011. **174**(2): p. 185-194.
16. Smith, A.H. and C.M. Steinmaus, *Health Effects of Arsenic and Chromium in Drinking Water: Recent Human Findings.* Annual review of public health, 2009. **30**: p. 107-122.
17. Kile, M.L., et al., *A pathway-based analysis of urinary arsenic metabolites and skin lesions.* Am J Epidemiol, 2011. **173**(7): p. 778-86.

18. Pierce, B., et al., *Genome-wide association study identifies chromosome 10q24.32 variants associated with arsenic metabolism and toxicity phenotypes in Bangladesh.* PLoS Genet, 2012. **8**(2): p. e1002522.
19. Jansen, R.J., et al., *Determinants and Consequences of Arsenic Metabolism Efficiency among 4,794 Individuals: Demographics, Lifestyle, Genetics, and Toxicity.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(2): p. 381-90.
20. Kordas, K., et al., *Low-level arsenic exposure: Nutritional and dietary predictors in first-grade Uruguayan children.* Environmental Research, 2016. **147**: p. 16-23.
21. Shen, H., et al., *Factors Affecting Arsenic Methylation in Arsenic-Exposed Humans: A Systematic Review and Meta-Analysis.* International Journal of Environmental Research and Public Health, 2016. **13**(2): p. 205.
22. Pierce, et al., *Genome-Wide Association Study Identifies Chromosome 10q24.32 Variants Associated with Arsenic Metabolism and Toxicity Phenotypes in Bangladesh.* PLoS Genetics, 2012. **8**(2): p. e1002522.
23. Agusa, T., et al., *Individual Variations in Inorganic Arsenic Metabolism Associated with AS3MT Genetic Polymorphisms.* International Journal of Molecular Sciences, 2011. **12**(4): p. 2351-2382.
24. Pierce, B.L., et al., *Arsenic metabolism efficiency has a causal role in arsenic toxicity: Mendelian randomization and gene-environment interaction.* International Journal of Epidemiology, 2013. **42**(6): p. 1862-1872.
25. Das, N., et al., *Association of single nucleotide polymorphism with arsenic-induced skin lesions and genetic damage in exposed population of West Bengal, India.* Mutat Res, 2016. **809**: p. 50-56.
26. Pierce, B.L., et al., *A missense variant in FTCD is associated with arsenic metabolism and toxicity phenotypes in Bangladesh.* PLOS Genetics, 2019. **15**(3): p. e1007984.
27. Gao, J., et al., *The Genetic Architecture of Arsenic Metabolism Efficiency:A SNP-Based Heritability Study of Bangladeshi Adults.* Environ Health Perspect, 2015. **123**(10): p. 985-92.
28. Tellez-Plaza, M., et al., *Heritability and preliminary genome-wide linkage analysis of arsenic metabolites in urine.* Environ Health Perspect, 2013. **121**(3): p. 345-51.
29. Mills, K.T., et al., *Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries.* Circulation, 2016. **134**(6): p. 441-50.
30. Gay, H.C., et al., *Effects of Different Dietary Interventions on Blood Pressure: Systematic Review and Meta-Analysis of Randomized Controlled Trials.* Hypertension, 2016. **67**(4): p. 733-9.
31. Bakker, E.A., et al., *Physical activity and fitness for the prevention of hypertension.* Curr Opin Cardiol, 2018. **33**(4): p. 394-401.
32. Fernández-Solà, J., *Cardiovascular risks and benefits of moderate and heavy alcohol consumption.* Nat Rev Cardiol, 2015. **12**(10): p. 576-87.
33. Saladini, F., et al., *Effects of smoking on central blood pressure and pressure amplification in hypertension of the young.* Vasc Med, 2016. **21**(5): p. 422-428.
34. Gouveia, M.H., et al., *Trans-ethnic meta-analysis identifies new loci associated with longitudinal blood pressure traits.* Sci Rep, 2021. **11**(1): p. 4075.
35. Liu, M.Y., et al., *Association between psychosocial stress and hypertension: a systematic review and meta-analysis.* Neurol Res, 2017. **39**(6): p. 573-580.

36. Tellez-Plaza, M., et al., *Cadmium exposure and hypertension in the 1999-2004 National Health and Nutrition Examination Survey (NHANES).* Environ Health Perspect, 2008. **116**(1): p. 51-6.
37. Abhyankar, L.N., et al., *Arsenic exposure and hypertension: a systematic review.* Environ Health Perspect, 2012. **120**(4): p. 494-500.
38. Laclaustra, M., et al., *Serum Selenium Concentrations and Hypertension in the US Population.* Circulation: Cardiovascular Quality and Outcomes, 2009. **2**(4): p. 369-376.
39. Vaziri, N.D., *Mechanisms of lead-induced hypertension and cardiovascular disease.* Am J Physiol Heart Circ Physiol, 2008. **295**(2): p. H454-65.
40. Aposhian, H.V., et al., *Oxidation and detoxification of trivalent arsenic species.* Toxicol Appl Pharmacol, 2003. **193**(1): p. 1-8.
41. Balakumar, P., T. Kaur, and M. Singh, *Potential target sites to modulate vascular endothelial dysfunction: current perspectives and future directions.* Toxicology, 2008. **245**(1-2): p. 49-64.
42. Duan, X., et al., *Acute arsenic exposure induces inflammatory responses and CD4+ T cell subpopulations differentiation in spleen and thymus with the involvement of MAPK, NF-kB, and Nrf2.* Molecular Immunology, 2017. **81**: p. 160-172.
43. Lee, M.Y., et al., *Arsenic-induced dysfunction in relaxation of blood vessels.* Environ Health Perspect, 2003. **111**(4): p. 513-7.
44. Waghe, P., et al., *Arsenic causes aortic dysfunction and systemic hypertension in rats: Augmentation of angiotensin II signaling.* Chemico-Biological Interactions, 2015. **237**: p. 104-114.
45. Liu, Y., et al., *Differential activation of ERK, JNK/SAPK and P38/CSBP/RK map kinase family members during the cellular response to arsenite.* Free Radic Biol Med, 1996. **21**(6): p. 771-81.
46. Ludwig, S., et al., *The stress inducer arsenite activates mitogen-activated protein kinases extracellular signal-regulated kinases 1 and 2 via a MAPK kinase 6/p38-dependent pathway.* J Biol Chem, 1998. **273**(4): p. 1917-22.
47. Simeonova, P.P., et al., *c-Src-dependent activation of the epidermal growth factor receptor and mitogen-activated protein kinase pathway by arsenic. Role in carcinogenesis.* J Biol Chem, 2002. **277**(4): p. 2945-50.
48. Wang, F., et al., *Arsenic induces the expressions of angiogenesis-related factors through PI3K and MAPK pathways in SV-HUC-1 human uroepithelial cells.* Toxicology Letters, 2013. **222**(3): p. 303-311.
49. Sanchez-Soria, P., et al., *Chronic low-level arsenite exposure through drinking water increases blood pressure and promotes concentric left ventricular hypertrophy in female mice.* Toxicol Pathol, 2012. **40**(3): p. 504-12.
50. Huda, N., et al., *Elevated levels of plasma uric acid and its relation to hypertension in arsenic-endemic human individuals in Bangladesh.* Toxicol Appl Pharmacol, 2014. **281**(1): p. 11-8.
51. Farzan, S.F., et al., *Blood Pressure Changes in Relation to Arsenic Exposure in a U.S. Pregnancy Cohort.* Environmental Health Perspectives, 2015. **123**(10): p. 999-1006.
52. Rahman, M., et al., *Hypertension and arsenic exposure in Bangladesh.* Hypertension, 1999. **33**(1): p. 74-8.

53. Chen, Y., et al., *Arsenic Exposure from Drinking Water, Dietary Intakes of B Vitamins and Folate, and Risk of High Blood Pressure in Bangladesh: A Population-based, Cross-sectional Study.* American Journal of Epidemiology, 2007. **165**(5): p. 541-552.
54. Jones, M.R., et al., *Urine arsenic and hypertension in US adults: the 2003-2008 National Health and Nutrition Examination Survey.* Epidemiology, 2011. **22**(2): p. 153-61.
55. Scannell Bryan, M., et al., *Mendelian randomization of inorganic arsenic metabolism as a risk factor for hypertension- and diabetes-related traits among adults in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort.* International Journal of Epidemiology, 2019. **48**(3): p. 876-886.
56. Mendez, M.A., et al., *Chronic exposure to arsenic and markers of cardiometabolic risk: A cross-sectional study in Chihuahua, Mexico.* Environmental Health Perspectives, 2016. **124**(1): p. 104-111.
57. Kheirandish, P. and F. Chinegwundoh, *Ethnic differences in prostate cancer.* British Journal of Cancer, 2011. **105**(4): p. 481-485.
58. Torre, L.A., et al., *Global cancer statistics, 2012.* CA Cancer J Clin, 2015. **65**(2): p. 87-108.
59. Rawla, P., *Epidemiology of Prostate Cancer.* World J Oncol, 2019. **10**(2): p. 63-89.
60. Pietro, G.D., et al., *Racial Differences in the Diagnosis and Treatment of Prostate Cancer.* Int Neurourol J, 2016. **20**(Suppl 2): p. S112-119.
61. DeSantis, C.E., et al., *Cancer statistics for African Americans, 2016: Progress and opportunities in reducing racial disparities.* CA Cancer J Clin, 2016. **66**(4): p. 290-308.
62. Kelly, S.P., et al., *Trends in the Incidence of Fatal Prostate Cancer in the United States by Race.* Eur Urol, 2017. **71**(2): p. 195-201.
63. Ross, R., et al., *Serum Testosterone Levels in Healthy Young Black and White Men2.* JNCI: Journal of the National Cancer Institute, 1986. **76**(1): p. 45-48.
64. Gapstur, S.M., et al., *Serum Androgen Concentrations in Young Men: A Longitudinal Analysis of Associations with Age, Obesity, and Race.* The CARDIA Male Hormone Study, 2002. **11**(10): p. 1041-1047.
65. Cotter, M.P., et al., *Role of family history and ethnicity on the mode and age of prostate cancer presentation.* Prostate, 2002. **50**(4): p. 216-21.
66. Hoffman, R.M., et al., *Racial and Ethnic Differences in Advanced-Stage Prostate Cancer: the Prostate Cancer Outcomes Study.* JNCI: Journal of the National Cancer Institute, 2001. **93**(5): p. 388-395.
67. Thompson, I.M., et al., *Association of African-American Ethnic Background With Survival in Men With Metastatic Prostate Cancer.* JNCI: Journal of the National Cancer Institute, 2001. **93**(3): p. 219-225.
68. Sanchez-Ortiz, R.F., et al., *African-American men with nonpalpable prostate cancer exhibit greater tumor volume than matched white men.* Cancer, 2006. **107**(1): p. 75-82.
69. Albain, K.S., et al., *Racial disparities in cancer survival among randomized clinical trials patients of the Southwest Oncology Group.* J Natl Cancer Inst, 2009. **101**(14): p. 984-92.
70. Hjelmborg, J.B., et al., *The Heritability of Prostate Cancer in the Nordic Twin Study of Cancer.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2014. **23**(11): p. 2303-2310.

71. Conti, D.V., et al., *Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction.* Nature Genetics, 2021. **53**(1): p. 65-75.
72. Haiman, C.A., et al., *Characterizing Genetic Risk at Known Prostate Cancer Susceptibility Loci in African Americans.* Plos Genetics, 2011. **7**(5).
73. Lachance, J., et al., *Genetic Hitchhiking and Population Bottlenecks Contribute to Prostate Cancer Disparities in Men of African Descent.* Cancer Research, 2018. **78**(9): p. 2432-2443.
74. Han, Y., et al., *Generalizability of established prostate cancer risk variants in men of African ancestry.* Int J Cancer, 2015. **136**(5): p. 1210-7.
75. Hoffmann, T.J., et al., *A Large Multiethnic Genome-Wide Association Study of Prostate Cancer Identifies Novel Risk Variants and Substantial Ethnic Differences.* Cancer Discovery, 2015. **5**(8): p. 878-891.
76. Cook, M.B., et al., *A genome-wide association study of prostate cancer in West African men.* Hum Genet, 2014. **133**(5): p. 509-21.
77. Virlogeux, V., et al., *Replication and heritability of prostate cancer risk variants: impact of population-specific factors.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(6): p. 938-43.
78. Haiman, C.A., et al., *Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21.* Nature Genetics, 2011. **43**(6): p. 570-U103.
79. Han, Y., et al., *Prostate Cancer Susceptibility in Men of African Ancestry at 8q24.* J Natl Cancer Inst, 2016. **108**(7).
80. Hoffmann, T.J., et al., *A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences.* Cancer Discov, 2015. **5**(8): p. 878-91.
81. Sharma, S., T.K. Kelly, and P.A. Jones, *Epigenetics in cancer.* Carcinogenesis, 2010. **31**(1): p. 27-36.
82. Baylin, S.B., *DNA methylation and gene silencing in cancer.* Nature Clinical Practice Oncology, 2005. **2**: p. S4.
83. Herman, J.G. and S.B. Baylin, *Gene silencing in cancer in association with promoter hypermethylation.* N Engl J Med, 2003. **349**(21): p. 2042-54.
84. Rodriguez-Paredes, M. and M. Esteller, *Cancer epigenetics reaches mainstream oncology.* Nat Med, 2011. **17**(3): p. 330-9.
85. Teschendorff, A.E., et al., *Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.* Genome Res, 2010. **20**(4): p. 440-6.
86. Gibbs, J.R., et al., *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.* PLoS Genet, 2010. **6**(5): p. e1000952.
87. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.
88. Larson, N.B., et al., *Comprehensively evaluating cis-regulatory variation in the human prostate transcriptome by using gene-level allele-specific expression.* Am J Hum Genet, 2015. **96**(6): p. 869-82.
89. Penney, K.L., et al., *Association of prostate cancer risk variants with gene expression in normal and tumor tissue.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(1): p. 255-60.

90. Thibodeau, S.N., et al., *Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set.* Nat Commun, 2015. **6**: p. 8653.
91. Vergara-Lope, A., et al., *Linkage disequilibrium maps for European and African populations constructed from whole genome sequence data.* Scientific Data, 2019. **6**(1): p. 208.
92. Moon, K.A., et al., *A dose-response meta-analysis of chronic arsenic exposure and incident cardiovascular disease.* International Journal of Epidemiology, 2017. **46**(6): p. 1924-1939.
93. Sung, T.-C., J.-W. Huang, and H.-R. Guo, *Association between Arsenic Exposure and Diabetes: A Meta-Analysis.* BioMed Research International, 2015. **2015**: p. 368087.
94. Karim, Y., et al., *Dose-dependent relationships between chronic arsenic exposure and cognitive impairment and serum brain-derived neurotrophic factor.* Environ Int, 2019. **131**: p. 105029.
95. Milton, A.H., et al., *A Review of the Effects of Chronic Arsenic Exposure on Adverse Pregnancy Outcomes.* Int J Environ Res Public Health, 2017. **14**(6).
96. Karagas, M.R., et al., *Drinking Water Arsenic Contamination, Skin Lesions, and Malignancies: A Systematic Review of the Global Evidence.* Curr Environ Health Rep, 2015. **2**(1): p. 52-68.
97. Lamm, S.H., et al., *A Systematic Review and Meta-Regression Analysis of Lung Cancer Risk and Inorganic Arsenic in Drinking Water.* International Journal of Environmental Research and Public Health, 2015. **12**(12): p. 15498-15515.
98. Gamboa-Loira, B., et al., *Arsenic metabolism and cancer risk: A meta-analysis.* Environmental Research, 2017. **156**: p. 551-558.
99. Ferreccio, C., et al., *Case-control study of arsenic in drinking water and kidney cancer in uniquely exposed Northern Chile.* Am J Epidemiol, 2013. **178**(5): p. 813-8.
100. Gamble, M.V., et al., *Folate and arsenic metabolism: a double-blind, placebo-controlled folic acid-supplementation trial in Bangladesh.* The American journal of clinical nutrition, 2006. **84**(5): p. 1093-1101.
101. Gamble, M.V., et al., *Folic acid supplementation lowers blood arsenic.* The American journal of clinical nutrition, 2007. **86**(4): p. 1202-1209.
102. Peters, B.A., et al., *Renal function is associated with indicators of arsenic methylation capacity in Bangladeshi adults.* Environmental research, 2015. **143**(Pt A): p. 123-130.
103. Vahter, M., *Methylation of inorganic arsenic in different mammalian species and population groups.* Sci Prog, 1999. **82 (Pt 1)**: p. 69-88.
104. Hopenhayn-Rich, C., et al., *Methylation study of a population environmentally exposed to arsenic in drinking water.* Environ Health Perspect, 1996. **104**(6): p. 620-8.
105. Engstrom, K.S., et al., *Genetic variation in arsenic (+3 oxidation state) methyltransferase (AS3MT), arsenic metabolism and risk of basal cell carcinoma in a European population.* Environ Mol Mutagen, 2015. **56**(1): p. 60-9.
106. Garcia-Alvarado, F.J., et al., *[Polymorphisms of the Arsenite Methyltransferase (As3MT) gene and urinary efficiency of arsenic metabolism in a population in northern Mexico]*. Rev Peru Med Exp Salud Publica, 2018. **35**(1): p. 72-76.
107. Agusa, T., et al., *Genetic polymorphisms in AS3MT and arsenic metabolism in residents of the Red River Delta, Vietnam.* Toxicol Appl Pharmacol, 2009. **236**(2): p. 131-41.

108. Engstrom, K., et al., *Polymorphisms in arsenic(+III oxidation state) methyltransferase (AS3MT) predict gene expression of AS3MT as well as arsenic metabolism.* Environ Health Perspect, 2011. **119**(2): p. 182-8.
109. Balakrishnan, P., et al., *Association of Cardiometabolic Genes with Arsenic Metabolism Biomarkers in American Indian Communities: The Strong Heart Family Study (SHFS).* Environ Health Perspect, 2017. **125**(1): p. 15-22.
110. Ahsan, et al., *Health Effects of Arsenic Longitudinal Study (HEALS): description of a multidisciplinary epidemiologic investigation.* J Expo Sci Environ Epidemiol, 2006. **16**(2): p. 191-205.
111. Navas-Acien, A., et al., *Urine arsenic concentrations and species excretion patterns in American Indian communities over a 10-year period: the Strong Heart Study.* Environ Health Perspect, 2009. **117**(9): p. 1428-33.
112. Nigra, A.E., et al., *Dietary determinants of inorganic arsenic exposure in the Strong Heart Family Study.* Environ Res, 2019. **177**: p. 108616.
113. Gilbert-Diamond, D., et al., *A Population-based Case–Control Study of Urinary Arsenic Species and Squamous Cell Carcinoma in New Hampshire, USA.* Environmental Health Perspectives, 2013. **121**(10): p. 1154-1160.
114. Scheer, J., et al., *Arsenic species and selected metals in human urine: validation of HPLC/ICPMS and ICPMS procedures for a long-term population-based epidemiological study.* Analytical methods : advancing methods and applications, 2012. **4**(2): p. 406-413.
115. Gilbert-Diamond, D., et al., *Rice consumption contributes to arsenic exposure in US women.* Proceedings of the National Academy of Sciences, 2011. **108**(51): p. 20656-20660.
116. Lee, E.T., et al., *THE STRONG HEART STUDY A STUDY OF CARDIOVASCULAR DISEASE IN AMERICAN INDIANS: DESIGN AND METHODS.* American Journal of Epidemiology, 1990. **132**(6): p. 1141-1155.
117. Poplin, R., et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2018: p. 201178.
118. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
119. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Research, 2010. **38**(16): p. e164-e164.
120. Ng, P.C. and S. Henikoff, *SIFT: predicting amino acid changes that affect protein function.* Nucleic Acids Research, 2003. **31**(13): p. 3812-3814.
121. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
122. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic Acids Research, 2018. **47**(D1): p. D886-D894.
123. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* American journal of human genetics, 2007. **81**(3): p. 559-575.
124. Yang, J., et al., *GCTA: A Tool for Genome-wide Complex Trait Analysis.* American Journal of Human Genetics, 2011. **88**(1): p. 76-82.
125. Asimit, J. and E. Zeggini, *Rare variant association analysis methods for complex traits.* Annu Rev Genet, 2010. **44**: p. 293-308.

126. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* American journal of human genetics, 2014. **95**(1): p. 5-23.
127. Neale, B.M., et al., *Testing for an unusual distribution of rare variants.* PLoS Genet, 2011. **7**(3): p. e1001322.
128. Matise, T.C., et al., *The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study*. Am J Epidemiol, 2011. **174**(7): p. 849-59.
129. Wu, Michael C., et al., *Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.* American Journal of Human Genetics, 2011. **89**(1): p. 82-93.
130. Wang, X., et al., *Rare variant association test in family-based sequencing studies.* Brief Bioinform, 2017. **18**(6): p. 954-961.
131. Lee, S., et al., *General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies.* The American Journal of Human Genetics, 2013. **93**(1): p. 42-53.
132. Melak, D., et al., *Arsenic methylation and lung and bladder cancer in a case-control study in northern Chile.* Toxicology and applied pharmacology, 2014. **274**(2): p. 225-231.
133. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data.* Genetic Epidemiology, 2013. **37**(7): p. 658-665.
134. Karczewski, K.J., et al., *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes.* bioRxiv, 2019: p. 531210.
135. Drobna, Z., et al., *Disruption of the arsenic (+3 oxidation state) methyltransferase gene in the mouse alters the phenotype for methylation of arsenic and affects distribution and retention of orally administered arsenate.* Chem Res Toxicol, 2009. **22**(10): p. 1713-20.
136. Engström, K.S., et al., *Efficient Arsenic Metabolism — The AS3MT Haplotype Is Associated with DNA Methylation and Expression of Multiple Genes Around AS3MT.* PLOS ONE, 2013. **8**(1): p. e53732.
137. Chernoff, M., et al., *Genetic determinants of reduced arsenic metabolism efficiency in the 10q24.32 region are associated with reduced AS3MT expression in multiple human tissue types.* Toxicological Sciences, 2020.
138. Chen, B., et al., *Mouse arsenic (+3 oxidation state) methyltransferase genotype affects metabolism and tissue dosimetry of arsenicals after arsenite administration in drinking water.* Toxicol Sci, 2011. **124**(2): p. 320-6.
139. Currier, J.M., et al., *Oxidation state specific analysis of arsenic species in tissues of wild-type and arsenic (+3 oxidation state) methyltransferase-knockout mice.* J Environ Sci (China), 2016. **49**: p. 104-112.
140. Hughes, M.F., et al., *Arsenic (+3 oxidation state) methyltransferase genotype affects steady-state distribution and clearance of arsenic in arsenate-treated mice.* Toxicol Appl Pharmacol, 2010. **249**(3): p. 217-23.
141. Yokohira, M., et al., *Severe systemic toxicity and urinary bladder cytotoxicity and regenerative hyperplasia induced by arsenite in arsenic (+3 oxidation state) methyltransferase knockout mice. A preliminary report.* Toxicol Appl Pharmacol, 2010. **246**(1-2): p. 1-7.

142. Douillet, C., et al., *Knockout of arsenic (+3 oxidation state) methyltransferase is associated with adverse metabolic phenotype in mice: the role of sex and arsenic exposure.* Arch Toxicol, 2017. **91**(7): p. 2617-2627.
143. Negro Silva, L.F., et al., *Effects of Inorganic Arsenic, Methylated Arsenicals, and Arsenobetaine on Atherosclerosis in the Mouse Model and the Role of As3mt-Mediated Methylation.* Environ Health Perspect, 2017. **125**(7): p. 077001.
144. Mateen, F.J., et al., *Chronic arsenic exposure and risk of carotid artery disease: The Strong Heart Study.* Environ Res, 2017. **157**: p. 127-134.
145. Wu, F., et al., *Interaction between arsenic exposure from drinking water and genetic polymorphisms on cardiovascular disease in Bangladesh: a prospective case-cohort study.* Environ Health Perspect, 2015. **123**(5): p. 451-7.
146. Hall, E.M., et al., *Hypertension among adults exposed to drinking water arsenic in Northern Chile.* Environ Res, 2017. **153**: p. 99-105.
147. Jiang, J., et al., *Association between Arsenic Exposure from Drinking Water and Longitudinal Change in Blood Pressure among HEALS Cohort Participants.* Environmental Health Perspectives, 2015. **123**(8): p. 806-812.
148. Islam, M.R., et al., *Association between Hypertension and Chronic Arsenic Exposure in Drinking Water: A Cross-Sectional Study in Bangladesh.* International Journal of Environmental Research and Public Health, 2012. **9**(12): p. 4522-4536.
149. Yu, Y., et al., *A perspective of chronic low exposure of arsenic on non-working women: Risk of hypertension.* Science of The Total Environment, 2017. **580**: p. 69-73.
150. Zhong, Q., et al., *Arsenic Exposure and Incident Hypertension of Adult Residents Living in Rural Areas Along the Yangtze River, Anhui, China.* Journal of Occupational and Environmental Medicine, 2019. **61**(4).
151. Zheng, J., et al., *Recent Developments in Mendelian Randomization Studies.* Current Epidemiology Reports, 2017. **4**(4): p. 330-345.
152. Ahsan, H., et al., *Arsenic Exposure from Drinking Water and Risk of Premalignant Skin Lesions in Bangladesh: Baseline Results from the Health Effects of Arsenic Longitudinal Study.* American Journal of Epidemiology, 2006. **163**(12): p. 1138-1148.
153. Argos, M., et al., *Baseline Comorbidities in a Skin Cancer Prevention Trial in Bangladesh.* European journal of clinical investigation, 2013. **43**(6): p. 579-588.
154. Shih, Y.-H., et al., *Gravidity, parity, blood pressure and mortality among women in Bangladesh from the HEALS cohort.* BMJ Open, 2020. **10**(8): p. e037244.
155. Tobin, M.D., et al., *Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure.* Stat Med, 2005. **24**(19): p. 2911-35.
156. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.
157. Didelez, V. and N. Sheehan, *Mendelian randomization as an instrumental variable approach to causal inference.* Stat Methods Med Res, 2007. **16**(4): p. 309-30.
158. Boef, A.G.C., O.M. Dekkers, and S. le Cessie, *Mendelian randomization studies: a review of the approaches used and the quality of reporting.* International Journal of Epidemiology, 2015. **44**(2): p. 496-511.
159. Del Razo, L.M., et al., *Exposure to arsenic in drinking water is associated with increased prevalence of diabetes: a cross-sectional study in the Zimapán and Lagunera regions in Mexico.* Environmental Health, 2011. **10**(1): p. 73.

160. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA: A Cancer Journal for Clinicians, 2020. **70**(1): p. 7-30.
161. Schumacher, F.R., et al., *Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci.* Nature Genetics, 2018. **50**(7): p. 928-936.
162. Do, C., et al., *Genetic–epigenetic interactions in cis: a major focus in the post-GWAS era.* Genome Biology, 2017. **18**(1): p. 120.
163. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nature Reviews Genetics, 2015. **16**(4): p. 197-212.
164. Houlahan, K.E., et al., *Genome-wide germline correlates of the epigenetic landscape of prostate cancer.* Nature Medicine, 2019. **25**(10): p. 1615-1626.
165. Dai, J.Y., et al., *DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs.* PLOS Genetics, 2020. **16**(3): p. e1008667.
166. Zhou, W., P.W. Laird, and H. Shen, *Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes.* Nucleic Acids Research, 2016. **45**(4): p. e22-e22.
167. Giambartolomei, C., et al., *Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics.* PLOS Genetics, 2014. **10**(5): p. e1004383.
168. Zhang, Y.D., et al., *Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers.* Nature Communications, 2020. **11**(1): p. 3353.
169. Pierce, B.L., et al., *Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms.* Nature Communications, 2018. **9**(1): p. 804.
170. Wallace, C., *Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses.* PLOS Genetics, 2020. **16**(4): p. e1008720.
171. Tolkach, Y. and G. Kristiansen, *The Heterogeneity of Prostate Cancer: A Practical Approach.* Pathobiology, 2018. **85**(1-2): p. 108-116.
172. Wang, G.F., et al., *Irx4 forms an inhibitory complex with the vitamin D and retinoic X receptors to regulate cardiac chamber-specific slow MyHC3 expression.* J Biol Chem, 2001. **276**(31): p. 28835-41.
173. Xu, X., et al., *Variants at IRX4 as prostate cancer expression quantitative trait loci.* Eur J Hum Genet, 2014. **22**(4): p. 558-63.
174. Bicak, M., et al., *Prostate cancer risk SNP rs10993994 is a trans-eQTL for SNHG11 mediated through MSMB.* Human Molecular Genetics, 2020. **29**(10): p. 1581-1591.
175. Ha Nguyen, H., et al., *IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility.* Human Molecular Genetics, 2012. **21**(9): p. 2076-2085.
176. Pajouh, M.S., et al., *Expression of metalloproteinase genes in human prostate cancer.* J Cancer Res Clin Oncol, 1991. **117**(2): p. 144-50.
177. Knox, J.D., et al., *Matrilysin expression in human prostate carcinoma.* Mol Carcinog, 1996. **15**(1): p. 57-63.
178. Zhang, Q., et al., *Interleukin-17 promotes prostate cancer via MMP7-induced epithelial-to-mesenchymal transition.* Oncogene, 2017. **36**(5): p. 687-699.
179. Zhang, Q., et al., *Interleukin-17 promotes formation and growth of prostate adenocarcinoma in mouse models.* Cancer Res, 2012. **72**(10): p. 2589-99.
180. Whitaker, H.C., et al., *The potential value of microseminoprotein-β as a prostate cancer biomarker and therapeutic target.* The Prostate, 2010. **70**(3): p. 333-340.

181. Edström Hägerwall, A.M.L., et al., *β-Microseminoprotein Endows Post Coital Seminal Plasma with Potent Candidacidal Activity by a Calcium- and pH-Dependent Mechanism.* PLOS Pathogens, 2012. **8**(4): p. e1002625.
182. Pomerantz, M.M., et al., *Analysis of the 10q11 Cancer Risk Locus Implicates MSMB and NCOA4 in Human Prostate Tumorigenesis.* PLOS Genetics, 2010. **6**(11): p. e1001204.
183. Wang, X., et al., *Validation of prostate cancer risk variants rs10993994 and rs7098889 by CRISPR/Cas9 mediated genome editing.* Gene, 2021. **768**: p. 145265.