Identifying Genetic Risk Factors for Coronary Artery Angiographic Stenosis in a Genetically Diverse Population

Zhi Liu
University of Miami, zhiliu.um@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/1722

This Embargoed is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
IDENTIFYING GENETIC RISK FACTORS FOR CORONARY ARTERY ANGIOGRAPHIC STENOSIS IN A GENETICALLY DIVERSE POPULATION

By

Zhi Liu

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

August 2016
UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

IDENTIFYING GENETIC RISK FACTORS FOR CORONARY ARTERY
ANGIOGRAPHIC STENOSIS IN A GENETICALLY DIVERSE POPULATION

Zhi Liu

Approved:

Gary W. Beecham, Ph.D.
Assistant Professor of Human
Genetics

Liyong Wang, Ph.D.
Associate Professor of Human
Genetics

Eden R. Martin, Ph.D.
Professor of Human Genetics

Guillermo Prado, Ph.D.
Dean of the Graduate School

Tatjana Rundek, M.D., Ph.D.
Professor of Neurology
Coronary artery disease (CAD) is the leading cause of death world-wide. It is a multifactorial genetic disease with an estimated heritability of 40%. Many genetic studies had been performed on CAD, but they explained only 10% of the total heritability, partially due to most of these studies were performed on Europeans only. Admixed populations, such as Hispanics, were understudied, despite their higher risks of CAD than Europeans. To explain additional heritability of CAD, 2023 individuals were ascertained from the Miami Cardiovascular Registry catheterization labs, which were representative of the South Florida population – a mixture of Hispanics, Europeans, and African Americans.

The prevalence of CAD varies across different race-ethnicity populations. A causal risk allele with large distinct ancestral allele frequency differences can lead to local ancestry deviations at these causal loci. Therefore, I performed local ancestry inference using LAMP-LD/LAMP-ANC software, a Hidden Markov Model based method on our population. Summing local ancestry across the entire population, the study population was composed of 75% European ancestry, 20% African ancestry and 5% Native American ancestry. These results were corresponding to principal component method identified ancestry, which did not utilize any ancestry informative markers.
Collateralization, a natural bypass process that protects against harmful CAD events, showed sex and ethnic differences in the presence of obstructive artery disease. In the study dataset, men had higher rate in collateralization than women (p= 0.000175). Interestingly, collateralization had distinct prevalence among race-ethnicity groups, with Hispanics being the highest (58%), followed by African Americans (50%) and Caucasians (47%). Due to different prevalence among race-ethnicity groups, I performed admixture mapping with inferred local and global ancestry and identified two genetic regions where local ancestry was associated with collateralization: Native American ancestry was associated with the presence of collaterals at a region on chromosome 17 (chr17:36120051–40782083, β = 0.55, p-value = 0.0001); African ancestry associated with collaterals on chromosome 17 but at a different location (chr17:32796666–33440166, β = 0.38, p-value = 0.0007). Gene-based test results showed that 8 genes within the Native American region and 5 genes within the African region were significantly associated with collateralization.

Angiographically defined and clinically diagnosed CAD shared common and unique genetic variants. Previous studies on CAD mostly focused on clinically defined CAD, which included various sub-phenotypes such as chronical angina, myocardial infarction, etc. Instead, I used angiographic stenosis through catheterization imaging to define CAD and performed genome-wide scanning of genetic variants associated with it. I identified 6 SNPs and 7 genes that were significantly associated with both clinically and angiographically defined CAD.

In my research, I studied genetic components for angiographic stenosis of CAD in a genetically diverse population. This phenotype allowed us to identify genetic components specific to the coronary angiographic stenosis rather than the complex disease. I also
identified genetic variants and contributing factors for collateralization, a protective mechanism towards CAD. This may help us explain why some atherosclerosis patients had worse clinical outcomes than others, and whether the race-ethnicity prevalence differences in CAD could be partially attributable to collateralization. Clinically, genetic factors influencing CAD may lead to novel treatment for CAD and promote new opportunities for prevention of patients suffer from severe obstructive atherosclerosis. Last, studying this genetically diverse population allowed us to eliminate health disparities and therefore benefited all populations as a whole.
To my family.
Acknowledgements

I would like to bring my gratitude to my doctorate mentor Dr. Gary W. Beecham for his excellent mentorship and endless support throughout years of my Ph.D. study. His guidance, patience and professional knowledge had motivated me to grow into a well-rounded statistical geneticist.

I would like to thank my committee members Drs. Liyong Wang, Eden R. Martin and Tatjana Rundek for their sound guidance and generous help on my research projects.

I want to thank the other team members in the Miami Cardiovascular Registry project for devoting their time and efforts into this research, and without them my research project would not be possible.

I express my appreciation to Dr. Bill Scott for his scientific inputs and support throughout my study. I thank instructors of classes I had taken for their generousness of sharing their knowledge with me, and thanks for their kind help for me to get the most from their lectures. Special thanks to Ms. Dori McLean who took care of every administrative process I needed so I could focus my energy on my research works. I was extremely thankful to have you as our program coordinator.

Last but not least, I want to thank my family members, for their unconditional love and care; thanks to my boyfriend and all my friends for their company, support and positive energies that helped me go through graduate school.
# Table of Contents

List of Figures .......................................................................................... vii

List of Tables .......................................................................................... xi

CHAPTER 1. INTRODUCTION ............................................................................. 1
  1.1. Historical studies of CAD ...................................................................... 1
  1.2. Clinical heterogeneity of current genetic association studies on CAD .......... 3
  1.3. The importance of multi race-ethnicity studies of CAD ......................... 6
  1.4. Studying angiographically defined coronary artery stenosis as the phenotype .... 8
  1.5. Collateralization as a protective mechanism against harmful cardiovascular events .......................................................... 8
  1.6. Identify genetic susceptibility loci for coronary artery disease related phenotypes ................. 10
  1.7. Summary ........................................................................................... 13

CHAPTER 2. GENERAL METHODOLOGY ............................................................ 14
  2.1. Data quality control (QC) ................................................................. 14
  2.2. Power calculation ........................................................................... 19

CHAPTER 3. LOCAL ANCESTRY INference IN A GENETICALLY DIVERSE POPULATION ................................................................. 22
  3.1. Overview ........................................................................................ 22
  3.2. Experimental approach .................................................................... 24
    3.2.1. Local ancestry inference ............................................................ 23
    3.2.2. Global ancestry inference ......................................................... 27
  3.3. Results and summary ...................................................................... 27
    3.3.1. Sample description ................................................................. 27
    3.3.2. Local ancestry inference results ................................................. 28
    3.3.3. Global ancestry inference results ................................................ 31
    3.3.4. Self-identified vs. principal component (PC) based ancestry ............... 32

CHAPTER 4. CORONARY COLLATERALIZATION SHOWS SEX AND ETHNIC DIFFERENCES IN THE PRESENCE OF OBSTRUCTIVE ARTERY DISEASE ................................................................................. 34
  4.1. Overview ........................................................................................ 34
  4.2. Experimental approach .................................................................... 36
    4.2.1. Sample assessment ................................................................. 36
    4.2.2. Phenotyping ........................................................................... 37
    4.2.3. Statistical analysis ................................................................... 38
    4.2.4. Admixture mapping ............................................................... 38
    4.2.5. Permutation correction ............................................................ 39
    4.2.6. Likelihood ratio test ................................................................ 39
List of Figures

Figure 1.1. Normal vs. narrowing coronary artery. A) Heart location. B) Normal coronary artery. C) Narrowing of coronary artery vessel, which limits the blood flow. The yellow buildup is a plaque, which is developed throughout many years. (Figure adapted from www.nhlbi.nih.gov) .......................................................... 5

Figure 1.2. American heart association 16 segment coronary artery tree. Figure retrieved from book “CT Evolution of Coronary Artery Disease”. (Pavone et al., 2009) .... 10

Figure 2.1. A flowchart showing the Affymetrix 6.0 genotyping platform SNP quality control steps that resulted in how the final SNP-set was created (804,235 SNPs). The number of SNPs shown is the number that remains after all markers that do not satisfy each quality control criterion are dropped. HWE: Hardy-Weinberg Equilibrium ............... 16

Figure 2.2. A flowchart showing the Affymetrix 6.0 genotyping platform sample quality control steps that resulted in how the final sample-set was created (1978 samples). The number of samples shown is the number that remains after all samples that do not satisfy each quality control criterion are dropped. GRR: graphical representation of relationship (Abecasis et al., 2001) .......................... 16

Figure 2.3. A flowchart showing the Illumina Exome Chip genotyping platform SNP quality control steps that resulted in how the final SNP-set was created (161,658 SNPs). The number of SNPs shown is the number that remains after all markers that do not satisfy each quality control criterion are dropped. HWE: Hardy-Weinberg Equilibrium ...... 17

Figure 2.4. A flowchart showing the Illumina Exome Chip genotyping platform sample quality control steps that resulted in how the final sample-set was created (2,086 samples). The number of samples shown is the number that remains after all samples that do not satisfy each quality control criterion are dropped. GRR: graphical representation of relationship (Abecasis et al., 2001) ......................... 18

Figure 2.5. Power calculation for collateral phenotype where 473 cases and 406 controls were included in the study. A) Estimated population prevalence = 0.004; B) estimated population prevalence = 0.01; C) estimated population prevalence = 0.1. In all three figures, x- axis represented allele frequency and y axis represented the statistical power. Blue indicated GRR =1.6; orange indicated GRR= 1.7 and gray indicated GRR =1.8 .................................................................................. 20

Figure 2.6. Power calculation for angiographic CAD phenotype where 435 cases and 484 controls were in the study. A) Estimated population prevalence = 0.06; B) estimated population prevalence = 0.1; C) estimated population prevalence = 0.2 in people over 65 years old. In all three figures, x- axis represented allele frequency and y axis represented the statistical power. Blue indicated GRR =1.6; orange indicated GRR= 1.7 and gray indicated GRR =1.8 .................................................................................. 21
Figure 3.1. Pipeline for local ancestry inference and calculating average local ancestry percentage across 100 pairs of inferred haplotypes.

Figure 3.2. Principal component analysis shows the genetic diversity of MCR samples. Eigenvector 1 separated European and African descent. Eigenvector 2 separated Native American and all others descent.

Figure 3.3. Comparing local ancestry inference based on averaging 100 pairs of haplotypes and the best-guess haplotype in self-identified white non-Hispanic, African American and white Hispanic. A) Averaging 100 pairs of haplotypes for a self-identified Caucasian. B) Best-guess haplotype for the same Caucasian. C) Averaging 100 pairs of haplotypes for a self-identified African American. D) Best-guess haplotype for the same African American. E) Averaging 100 pairs of haplotypes for a self-identified white Hispanic sample. F) Best-guess haplotype for the same white Hispanic sample. In all three figures, blue: European ancestry; Red: African ancestry; Yellow: Native American ancestry. Each line from top to bottom represented chromosome 1-22.

Figure 3.4. Scatter plots showed the correlation between eigenvectors and ancestry component averaged from 100 pairs. A) Correlation between eigenvector 1 and African component. B) Correlation between eigenvector 2 and Native American component.

Figure 3.5. Principal components determined race-ethnicity for subjects participated in our study. X-axis represented eigenvector 2 and y-axis represented eigenvector 5. Red: Hispanics; green: Non-Hispanics whites; blue: African Americans.

Figure 4.1. Manhattan plots showed associations between collateralization and A) delta Native American ancestry; B) delta African ancestry (differences between global and local Ancestry) in admixture mapping. The peak regions of delta Native American ancestry and delta African ancestry were both located on chromosome 17. Native American ancestry was associated with the presence of collaterals at a region on chromosome 17 (chr17:36120051-40782083, beta = 0.55, minimal p-value = 1.27E-04). Effective number of testing: 1543; P-value cut off: 3.2E-05; -log10 p-value: 4.5. African ancestry showed association with collaterals at a different region on chromosome 17 (chr17:32796666-33440166, beta = 0.38, minimal p-value = 7.2E-04). Effective number of testing: 1931; P-value cut off: 2.6E-05; -log10 p-value: 4.6. X-axis indicated chromosomes 1 to 22. Y-axis indicated –log10 of delta Native American/African ancestry p-values.

Figure 4.2. Chromosome 17 regional plot showed peak association between delta African/Native American ancestries with collateralization in admixture mapping. Local Native American ancestry was highly associated with the presence of collaterals at a region on chromosome 17 (hg19:chr17:36120051–40782083, min p-value = 0.000127, -log10 (p-value) = 3.90). Local African ancestry also showed association with collaterals at a different region on chromosome 17 (hg19:chr17:32796666–33440166, min p-value = 0.00072, -log10 (p-value) = 3.14). X-axis represented the base pair location on chromosome 17; Y-axis represented –log10 p-value.
Figure 4.3. Regional plot showed admixture mapping in only Hispanic population. Top results of A) Native American ancestry associated with collateralization in only Hispanics was located on chromosome 17, black line illustrated the associated of all individuals, while grey line represented Hispanics only population. Figure B) showed the top results of African ancestry associated with only Hispanics was located on chromosome 5. Red represented all populations, while pink represented only Hispanics population. In both figures, X axis represented base pair position in MB, and y axis represented $-\log_{10}(p$-value). …………………………………57

Figure 4.4. Manhattan Plot showed admixture mapping in Hispanic population: A) local Native American ancestry was associated with collateralization on chromosome 17: 36074979–41009445 (hg19) as the top signal. Effective number of testing: 1911; P-value cut off: 2.6E-05; $-\log_{10}$ p-value: 4.6. B) Local African ancestry was associated with collateralization on chromosome 5: 9195233–9509811 as the top signal. Effective number of testing: 2283; P-value cut off: 2.2E-05; $-\log_{10}$ p-value: 4.7. In both figures, x axis represented chromosome 1 to 22. Y axis represented $-\log_{10}$ p-value………………………………58

Figure 4.5. Single genetic variants associated with collateralization when adjusting for local or global ancestry. A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry. In all figures, x-axis represented $-\log_{10}$ of expected P-values; y-axis represented $-\log_{10}$ of observed P-values………………………………………61

Figure 4.6. Manhattan plots showed the association results of collateralization with single genetic variants. A) Common genetic variants from Affymetrix 6 chip when adjusting for global ancestry; B) common genetic variants from Affymetrix 6 chip when adjusting for local ancestry; C) rare genetic variants from Illumina Exome chip when adjusting for global ancestry; D) rare genetic variants from Illumina Exome chip when adjusting for local ancestry. In all figures, x-axis represented chromosome 1 to 22; y-axis represented $-\log_{10}$ of P-values……………………………………………………..62

Figure 4.7. QQ plots showed gene-based test p-values distribution. A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry. In all figures, x-axis represented $-\log_{10}$ of expected p-values; y-axis represented $-\log_{10}$ of the observed p-values………………..65

Figure 4.8. Pathway analyses showed top enriched gene-sets that were associated with collateralization. A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry…………………………………………………………66

Figure 5.1. Coronary catheterization is a process that evaluates inside a coronary artery. Figure A. A catheter is inserted into an artery in the groin area and is guided through an X-ray into the artery. Figure B. Catheterization image shows the narrowing of a coronary artery through an X-ray examination. Figures retrieved from: heart.org……..71
Figure 5.2. Left main and 3 major branches of coronary arteries. Figure retrieved from http://my.clevelandclinic.org..................................................72

Figure 5.3. QQ Plots for single genetic variants test of Angiographically-defined CAD phenotype. A) Affymetrix 6 genotyping platform when adjusting for global ancestry. B) Affymetrix 6 genotyping platform when adjusting for local ancestry. C) Illumina exome chip genotyping platform when adjusting for global ancestry. D) Illumina exome chip genotyping platform when adjusting for local ancestry. In all figures, x-axis indicated – log10 of expected P-values; y-axis represented –log10 of observed P-values.............80

Figure 5.4. Manhattan plots for single genetic variants test of angiographically-defined CAD. A) Affymetrix 6 genotyping platform adjusting for global ancestry. B) Affymetrix 6 genotyping platform adjusting for local ancestry. C) Illumina exome chip genotyping platform adjusting for global ancestry. D) Illumina exome chip genotyping platform adjusting for local ancestry. In all figures, x-axis represented chromosome 1 to 22; y-axis represented –log10 of observed P-values.............81

Figure 5.5. QQ plots showed the P-value distribution for gene-based test of angiographically-defined CAD phenotype. A) Affymetrix 6 genotyping chip results when adjusting for global ancestry; B) Affymetrix 6 genotyping chip results when adjusting for local ancestry; C) Illumina exome chip results when adjusting for global ancestry; D) Illumina exome chip results when adjusting for local ancestry. In all figures, x-axis represented –log10 expected P-values, and y-axis represented –log10 of observed P-values.................................................................................83

Figure 5.6. Pathway analyses showed the top associated pathways with angiographic CAD. Figures showed the top three enriched pathways when we used A) Affymetrix chip adjusting for global ancestry; B) Affymetrix chip adjusting for local ancestry; C) Illumina exome chip adjusting for global ancestry; D) Illumina exome chip adjusting for local ancestry........................................................................................................86
List of Tables

Table 2.1. Description of overall Miami Cardiovascular Registry (MCR) dataset for catheterization patients who has been clinically diagnosed with CAD .............................. 15

Table 2.2. Summary of MCR patient samples for each race-ethnicity ............................ 15

Table 4.1. Patients clinical characteristics under each collateral grades ............................ 42

Table 4.2. EIGENSTRAT defined race-ethnicity differences in collateral grades within obstructive CAD patients ................................................................................................. 44

Table 4.3.1. Affy6 Chip: adjusting for global ancestry (African) ........................................ 48

Table 4.3.2. Affy6 Chip: adjusting for local ancestry (African) ........................................ 48

Table 4.3.3. Exome Chip: adjusting for global ancestry (African) ........................................ 49

Table 4.3.4. Exome Chip: adjusting for local ancestry (African) ........................................ 49

Table 4.4.1. Affy6 Chip: adjusting for global ancestry (Native American) ......................... 52

Table 4.4.2. Affy6 Chip: adjusting for local ancestry (Native American) ......................... 52

Table 4.4.3. Exome Chip: adjusting for global ancestry (Native American) ......................... 53

Table 4.4.4. Exome Chip: adjusting for local ancestry (Native American) ......................... 53

Table 5.1. Study sample race-ethnicities identified by principal components ..................... 75

Table 5.2. Classify cases and controls in angiographically-defined CAD study .................. 76

Table 5.3. Comparing previously published 65 SNPs that showed genome-wide significant association with clinically defined CAD to angiographically defined CAD in MCR dataset. Table shows SNPs that were significant in both clinically and angiographically defined CAD ........................................................................................................ 88

Table 5.4. Comparing 58 previously published genome-wide significant loci from meta-analysis additive association with clinically defined CAD results with angiographically defined CAD gene-based test on MCR dataset Affymetrix 6 chip/Exome Chip results. Tables show genetic loci that were significant in both clinical and angiographic CAD. A) Affymetrix 6 chip data. B) Exome chip data .................................................................................................................. 89
CHAPTER 1. INTRODUCTION

Coronary artery disease (CAD) is the leading cause of death in the United States (Heron et al. 2009) with an estimate of 500,000 deaths per year (Kochanek et al. 2011, 1-51). Atherosclerosis is the major etiology of CAD. It is a condition of cholesterol plaques buildup in the arteries to the heart, narrowing the small blood vessels that supply blood and oxygen to the heart (Gaziano, Ridker, and Libby 2011). Atherosclerosis often has mild to no symptoms until the plaque ruptures or causes complete occlusion of the coronary blood vessels.

1.1. Historical studies of CAD

To identify the common factors that contribute to cardiovascular disease, the Framingham Heart Study was started in 1948. It is a long-term, ongoing prospective cohort study (Dawber, Meadors, and Moore 1951, 279-281). By following a large group of participants of their development over a long period of time, researchers in Framingham study have identified major cardiovascular disease risk factors: high blood pressure, high blood cholesterol, smoking, obesity (Hubert et al. 1983, 968-977), diabetes, and physical inactivity, as well as a great amount of information on the effects of related factors such as blood triglyceride and high-density lipoprotein levels, age (Franklin et al. 2001, 1245-1249), gender, and psychosocial issues. Although the Framingham cohort is primarily Caucasian, the importance of the major risk factors identified in this group have been shown to apply almost universally across racial and ethnic groups, though the trait distributions may vary from group to group.

Another well-known research is the Atherosclerosis Risk in Communities (ARIC) study, which began in 1985 and is currently ongoing. It is a prospective cohort study
designed to ascertain the etiology and predictors of cardiovascular disease, which enrolled over 15,000 middle-aged adults from four U.S. communities (Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD) between 1987 and 1989 (Atherosclerosis Risk in Communities (ARIC) Study 1989, 687-702). These studies have greatly contributed to our current knowledge in cardiovascular disease.

CAD is a complex genetic disease with an estimated heritability of 40% (Peden and Farrall 2011, R198-205). Before the genome-wide association study (GWAS) era, efforts to identify genes contributing cardiovascular disease were largely restricted to studies of candidate genes or to analyze variations in animal models. With these methods, people have identified a handful of genes, such as connexin 37 (CX37), plasminogen-activator inhibitor type 1 (PAI1), thrombospondin (THBS), but the findings were limited and the disease was still less understood (Wenzel et al. 1996, 15-20; Elghannam et al. 2000, 562-568; Lusis 2000, 233-241). As the development of 1000 Genome Project and HapMap Project, GWAS have fruitfully identified many common single-nucleotide polymorphisms (SNPs) associated with CAD (Lusis 2012, 267-275). Researchers have mapped over 65 common variants for association with CAD, most notably 9p21 locus, which was repeatedly reported associated with CAD in several studies (McPherson et al. 2007, 1488-1491; Helgadottir et al. 2007, 1491-1493; Samani et al. 2007, 443-453; Wellcome Trust Case Control Consortium 2007, 661-678; Nikpay et al. 2015, 1121-1130). The region is adjacent to the cyclin-dependent kinase inhibitor 2A (CDKN2A) and 2B (CDKN2B) genes. CARDIoGRAM (Schunkert et al. 2011, 333-338) and C4D studies (Coronary Artery Disease (C4D) Genetics Consortium 2011, 339-344) together scanned nearly 40K coronary artery disease cases for susceptibility gene signal and discovered common variants
associated with CAD. Most of these studies were case-control and were performed on a large scale of sample size. However, typical effect sizes for individual single nucleotide polymorphisms (SNPs) were relatively small, which explained only 8 to 13% of the total 40% heritability of CAD and up to 4% of inter-individual variation in disease risk (Peden and Farrall 2011, R198-205).

1.2. Clinical heterogeneity of current genetic association studies on CAD

So far, large collaborative genetic association studies of case-control CAD tend to focus on phenotypes that are clinically heterogeneous (Peden and Farrall 2011, R198-205; Schunkert et al. 2011, 333-338; Coronary Artery Disease (C4D) Genetics Consortium 2011, 339-344). That is, they often include patients with different diagnosis subgroups, such as including patients presenting with chronic stable angina alongside patients presenting with myocardial infarction (Peden and Farrall 2011, R198-205) or other diagnoses. The less clinically heterogeneous studies tend to focus on clinical endpoints such as hypertension (International Consortium for Blood Pressure Genome-Wide Association Studies et al. 2011, 103-109; Ehret 2010, 17-25; Newton-Cheh et al. 2009, 666-676) or MI (Schunkert, Erdmann, and Samani 2010, 918-925; Myocardial Infarction Genetics Consortium et al. 2009, 334-341) that may still have heterogeneous etiologies. People have identified at least eight loci associated with blood pressure (BP) (Newton-Cheh et al. 2009, 666-676). All variants found associated with continuous BP were also associated with dichotomous hypertension. SNPs at over nine loci were reproducibly associated with myocardial infarction, including 9p21, which also associated with the risk of CAD (Myocardial Infarction Genetics Consortium et al. 2009, 334-341).
Atherosclerosis is a major etiology of cardiovascular disease (Lusis, Mar, and Pajukanta 2004, 189-218) (Figure 1.1), which has both genetic components and environmental risk factors such as diet and smoking, as well as extensive comorbidities such as diabetes, obesity, and dyslipidemia that may exacerbate the severity of other cardiovascular phenotypes (Fox et al. 2003, 397-401). Atherosclerosis, particularly coronary atherosclerosis, is a critical phenotype to understand, as it is highly associated with cardiovascular disease and thrombosis (Faxon et al. 2004, 2617-2625). If a coronary plaque ruptures, the unstable materials inside the plaque are exposed to the bloodstream, potentially causing the vessel to clot, resulting in unstable angina or myocardial infarction (Rauch et al. 2001, 224-238; Hansson 2005, 1685-1695). The more severe and vulnerable the atherosclerotic plaque buildup, the more likely the patient will have a plaque rupture, increasing the risk for MI (Finn et al. 2010, 1282-1292). One primary measure of atherosclerosis is the level of stenosis in the coronary bed. Proximal coronary stenosis has a strong genetic component, with an estimated heritability (h²) 50%, (Fischer et al. 2005, 855-862) and is a strong predictor of future cardiac events. Despite this high heritability and the clinical importance of coronary stenosis, genetic studies directly on coronary artery stenosis are limited (Bjornsson et al. 2015, 1526-1531; Pleva et al. 2015, 135-015-0128-8). Studies on atherosclerosis have primarily focused on carotid atherosclerosis (intima-media thickness or carotid plaque) as an intermediate, preclinical phenotype of vascular disease (Bartels, Franco, and Rundek 2012, 139-145; Rundek et al. 2008, 1200-1207; Zivotic et al. 2016, 1210-1216). Large population studies identified BCAR1-CFDP1-TMEM170A locus as a determinant of carotid intima-media thickness (Gertow et al. 2012, 656-665). In the Northern Manhattan Study (NOMAS), presence of plaque was associated
with a 2.8-fold [HR 2.76, 95% CI, 2.1–3.63] increased risk of stroke, myocardial infarction and vascular death during a mean follow up of 6.9 years (Rundek et al. 2008, 1200-1207).

Together, these studies offer important contributions to our understanding of the genetics of broadly defined “heart disease” (Schunkert et al. 2011, 333-338; Coronary Artery Disease (C4D) Genetics Consortium 2011, 339-344; Schunkert et al. 2008, 827-835; Ripatti et al. 2010, 1393-1400; Harismendy et al. 2011, 264-268), MI (Schunkert, Erdmann, and Samani 2010, 918-925; Myocardial Infarction Genetics Consortium et al. 2009, 334-341), and carotid atherosclerosis (Rundek et al. 2008, 1200-1207; Gertow et al. 2012, 656-665; Bis et al. 2011, 940-947). However, their implications on the genetics of coronary disease remain unclear, due to the vast clinical heterogeneity involved.

**Figure 1.1. Normal vs. narrowing coronary artery.** A) Heart location. B) Normal coronary artery. C) Narrowing of coronary artery vessel, which limits the blood flow. The yellow buildup is a plaque, which is developed throughout many years. (Figure adapted from www.nhlbi.nih.gov)
1.3. The importance of multi race-ethnicity studies of CAD

The other major limitation in our current understanding of the genetics of coronary atherosclerosis is with regards to the role of race and ethnicity in genetic risk of disease. There are clear differences in the prevalence of CAD among racial/ethnic populations, especially among women: with Native American women being the highest (8.4%, [6.8-10.4%], 95% confidence interval), followed by African Americans (5.9%, [5.4-6.3%]), Hispanics (5.3%, [4.7-5.9%]), European Ancestry (4.2%, [4.1-4.3%]) and Asian-Pacific (2.3%, [1.7-3.2%]) (Fang, Shaw, and Keenan 2011, 1377-1381). Multi-Ethnic Study of Atherosclerosis (MESA) found that African Americans had the highest blood pressure, but the lowest total cholesterol. Hispanics had the highest LDL-cholesterol but the lowest HDL-cholesterol. Caucasians had the lowest rate of diabetes. (Bild et al. 2005, 1313-1320) American Indian or Alaska Native men showed highest rate of Smoking (one of the top three leading risk factors for CAD) (26%), followed by Caucasian men (22%), and lowest among Hispanic women (7%) and Asian women (5%) (Mozaffarian et al. 2015, e29-322). The Atherosclerosis Risk in Communities (ARIC) study discovered that K55R variant allele was significantly associated with higher risk of incident CAD in Caucasian but the significant association was not observed in African-Americans (Lee et al. 2006, 1640-1649). Despite these differences, a majority of genetic studies of CAD have been with European (McPherson et al. 2007, 1488-1491; Samani et al. 2007, 443-453; Wellcome Trust Case Control Consortium 2007, 661-678; Erdmann et al. 2009, 280-282; Soranzo et al. 2009, 1182-1190) and Asian populations (Coronary Artery Disease (C4D) Genetics Consortium 2011, 339-344; Saleheen et al. 2009, 329-338; Mente et al. 2010, 2390-2398) with underrepresented populations remaining understudied.
This is especially true of Hispanics. As researches moving forward, studies on less
representative populations began: the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild
et al. 2005, 1313-1320) studied the characteristics of subclinical cardiovascular disease and
the risk factors that predicted progression to clinically overt cardiovascular disease.
Atherosclerosis Risk in Communities (ARIC) studies (Atherosclerosis Risk in
Communities (ARIC) Study, 1989, 687-702) were designed to investigate the causes of
atherosclerosis and its clinical outcomes, variation in cardiovascular risk factors, disease
by race, gender, and locations. Studying these less-representative populations, especially
recently admixed populations such as Hispanics, provided us a great opportunity to
determine disease variants with substantial allele frequency differences in ancestral
populations (Zhu et al. 2011, 2285-2295).

It is likely that racial differences in cardiovascular disease may be partly attributable to
genetic differences between populations. For example, there are significant racial
differences in coronary angiographic stenosis (Budoff et al. 2002, 408-412) and the
characteristics of CAD lesions (Bild et al. 2005, 1313-1320). In Caucasians, significant
obstruction on angiogram was 71%. Compared with Caucasians, African Americans have
an angiographic disease 49% (p<0.01), while Hispanics have an angiographic obstruction
58% (p<0.01) and Asians are not significantly different in angiographic stenosis (64%,
p=0.30) (Budoff et al. 2002, 408-412). The increased risk of MI in African Americans
(Nguyen and Stack 2006, 1716-1723) may be related to genes that are associated with
African ancestry (Topol et al. 2006, R117-23). Admixed populations offer a unique
opportunity for mapping disease susceptibility loci that have large allele frequency
differences between ancestral populations (Zhu et al. 2011, 2285-2295), therefore studying
them allows us to understand whether differences of risk allele frequency are responsible for the diverse racial prevalence of cardiovascular disease.

Although there are some studies focusing on CAD or MI, very few have extensive clinical phenotype information from direct imaging of the coronary anatomy, and none of them so far have been performed in a predominantly Hispanic population. In addition, current knowledge of genetic risk factors across racial/ethnic groups is quite limited.

1.4. Studying angiographically defined coronary artery stenosis as the phenotype

Coronary artery stenosis can predict the risk of thrombosis, which is a condition when the materials within the coronary artery plaque release into the blood stream and cause the blood to clot. A patient who undergoes thrombosis may suffer from angina or myocardial infarction when heart muscle cells die from lack of nutrients or oxygen. So far, very few studies have been performed on angiographically defined coronary artery stenosis, not only because it is difficult and expensive to acquire the phenotype, it is also difficult to perform catheterization on a large scale of population. To evaluate angiographically defined coronary artery stenosis, American Heart Association 16-segment model was utilized to compile a coronary artery tree (Figure 1.2). The presence and extent (number of segments) of stenosis were recorded for 16 branches. In addition, overall evaluation of the Cardiovasculature was recorded: presence, location and grade of coronary artery collaterals, calcification, small vessel disease, etc.

1.5. Collateralization as a protective mechanism against harmful cardiovascular events

Collateral circulation provides alternative blood flow when the original coronary arteries failed to provide sufficient blood (Koerselman et al. 2003, 2507-2511). It is an important
natural bypass process that benefits CAD patients – the newly formed blood vessels bypass the obstructed region of the coronary arteries to establish functional circulation, and limit hypoxia. Factors influencing coronary collateral circulation include myocardial ischemia, pressure gradient and shear stresses and growth factors, such as vascular endothelial growth factor (VEGF), transforming growth factor (TGF-α) and acidic fibroblast growth factor (a-FGF) (Koerselman et al. 2003, 2507-2511). However, it is still unclear why some patients develop collateralization while some do not, and the genetic components behind collateralization are understudied. Studying the contributing factors and genetic components of collateralization may help us explain why some atherosclerosis patients have worse clinical outcomes than others, and whether the race-ethnicity differences in CAD prevalence can be partially attributable to collateralization. Clinically, it can help identify targets for treatment (van Royen et al. 2001, 543-553) and promote new opportunities for prevention of patients suffer from severe obstructive atherosclerosis and its relevant clinical symptoms (Koerselman et al. 2003, 2507-2511).
1.6. Identify genetic susceptibility loci for coronary artery disease related phenotypes

1.6.1. Analyzing genome-wide genotyping data to identify common genetic variants associated with CAD related phenotypes

Figure 1.2. American heart association 16 segment coronary artery tree. Figure retrieved from book “CT Evolution of Coronary Artery Disease” (Pavone, Fioranelli, and Dowe 2009).

Genome-wide association studies (GWAS) have fruitfully determined common single genetic variants associated with CAD. For example, the Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) consortium identified genetic variants that were associated with aortic-valve calcification across multiple ethnic groups and with incident clinical aortic stenosis (Thanassoulis et al. 2013, 503-512). Since the phenotype used in this study is angiographically-defined coronary artery stenosis, a phenotype similar to aortic stenosis, I decided to search for common variants at the first step. Additionally, since no candidate genes have been determined to be associated with angiographically defined coronary artery stenosis, a non-hypothesis driven test is optimum. GWAS is a standard tool to search for candidate genetic variants and it also allows us to validate
previously findings in association with novel phenotypes of CAD, therefore a genome-wide search for common variants was performed. Lastly, densely mapped common variants cover ancestry informative markers, which can be utilized to identify ancestry in our genetically diverse population.

1.6.2. Analyzing exome chip data to identify rare genetic variants associated with CAD related phenotypes

Exons represent the regions in genes that are translated into proteins, and genetic variants in these regions alter protein functions. Using exome chip data allowed us to study the genetic variants in coding regions of genes. As a complementary tool to GWAS, which covers common variants, exome chip was designed to cover both common and rare variants (over 240,000 genetic variants). In the genome of an average individual, 8000 to 10,000 non-synonymous variants, 200 to 300 splice variants and 80 to 100 stop altering variants were expected to be observed. The chip used was designed to cover 97-98% of the non-synonymous variants, 94-95% of the splice altering variants and 94-95% of the stop altering variants detected in average genome through exome sequencing. In addition, it covers tags for all previously published GWAS hits by mid-2011 and ancestry informative markers which showed strong differentiation between African - European and Native American - European ancestry samples sequenced by the 1000 Genome Project (1000 Genomes Project Consortium et al. 2012, 56-65). It is a cost-effective alternative as compared to whole exome sequencing while still allows us to study the genetic effect of rare variants, which may have larger effect size as compared with common variants.
1.6.3. Identify genes associated with CAD related phenotypes

Traditional GWAS methods have been remarkably successful in identifying genetic associations of single SNP. However, in order to have adequate power to find disease associated SNPs, very large sample sizes are needed – especially for SNPs with lower minor allele frequency. Gene-based test is a promising approach to overcome limitations of statistical power: it aggregates true genotype-phenotype signal across multiple single variants within a gene and reduces the multiple-testing penalty by reducing the number of tests being conducted (Petersen et al. 2013, e62161). There are a few advantages to use gene-based test: first, a single gene may contain multiple variants that contribute independently to a trait, and these variants, which locate in protein-coding and adjacent regulatory region, are likely to have functional relevance. Second, gene-based tests allow us to directly compare different populations, despite the potential for racial differences in linkage disequilibrium (LD) patterns or functional alleles.

1.6.4. Gene-set enrichment analyses to identify pathways that were associated with CAD related phenotypes

Another effective way to overcome the requirement of large sample size to increase power is to test the association through pathway analysis, which contains multiple genes with small effect size, but the pathway in aggregate could have larger effect size. Currently, two pathways have been well established to be associated with coronary artery disease: Interleukin-6 receptor pathways (IL6R Genetics Consortium Emerging Risk Factors Collaboration et al. 2012, 1205-1213) and the IL-33-ST2L Pathway (Tu et al. 2013, 652-660).
1.7. Summary

In total, studying angiographically-defined CAD allowed us to identify genetic risk factors in the phenotype with minimized clinical heterogeneity from the traditional clinically defined CAD. Understanding collateralization helped us to identify the genetic components behind protective mechanisms of CAD. Using this understudied genetically diverse population, especially Hispanics, enabled us to explain missing heritability of coronary artery disease and therefore benefit the population as a whole.
CHAPTER 2. GENERAL METHODOLOGY

2.1. Data quality control (QC)

2023 patients were ascertained from those who underwent cardiac catheterization procedure. Patient DNA samples were extracted from blood and genotyped on Affymetrix 6.0 Genechip Human Mapping 1 Million SNP array (Santa Clara, CA) and Illumina Exome chip array (San Diego, CA) using standard protocols. Affymetrix chip targets mostly common genetic variants and Illumina chip targets mostly rare exonic variants. Patient samples are described in Table 2.1 and 2.2.

2.1.1. Affymetrix 6.0 genotype array

906,600 SNPs were obtained after genotyping and the following QCs were performed: call rate threshold >95% to ensure SNPs were genotyped in high quality; Hardy-Weinberg Equilibrium deviation (p > 1E-05) to avoid genotyping errors, batch effects, etc.; minor allele frequency threshold > 0.01 to ensure only common genetic variants are included in the analysis. After SNP quality control, 804,235 SNPs were included in the final dataset (Figure 2.1).

For sample QC, the following steps were performed on the initial 2023 patient samples: Graphical Representation of Relationships (GRR) (Abecasis et al. 2001, 742-743) to remove related individuals; genotype call rate threshold > 95% to guarantee high quality of genotyping; sex check (X-homo/heterozygosity) to make sure self-reported sex matches genotyped sex; EIGENSTRAT (Price et al. 2006, 904-909) was used to detect population substructure and remove population outliers. After sample QC, 1978 unrelated individuals were included in the final dataset for my analysis (Figure 2.2).
Table 2.1. Description of overall Miami Cardiovascular Registry (MCR) dataset for catheterization patients who has been clinically diagnosed with CAD.

<table>
<thead>
<tr>
<th>Description</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1978</td>
</tr>
<tr>
<td>Average age (±SD) at diagnosis</td>
<td>68 (±13)</td>
</tr>
<tr>
<td>Sex (Female %)</td>
<td>34%</td>
</tr>
<tr>
<td>Self-report race ethnicity</td>
<td>13% African American, 29% white non-Hispanics, 36% White Hispanics, 2% Black Hispanics and 20% Unknown</td>
</tr>
</tbody>
</table>

Table 2.2. Summary of MCR patient samples for each race-ethnicity.

<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>Hispanics (n=1007)</th>
<th>Caucasians (n=526)</th>
<th>African Americans (n=305)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (% Male)</td>
<td>68%</td>
<td>70%</td>
<td>54%</td>
</tr>
<tr>
<td>Age, years (±SD)</td>
<td>67(±13)</td>
<td>73(±13)</td>
<td>63(±11)</td>
</tr>
<tr>
<td>Hypertension (%)</td>
<td>80%</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>Diabetes (%)</td>
<td>31%</td>
<td>18%</td>
<td>35%</td>
</tr>
<tr>
<td>Anti-cholesterol Meds (%)</td>
<td>31%</td>
<td>29%</td>
<td>26%</td>
</tr>
<tr>
<td>Smoking (%)</td>
<td>50%</td>
<td>58%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Hispanics, Caucasians and African Americans were identified based on principal components.
Figure 2.1. A flowchart showing the Affymetrix 6.0 genotyping platform **SNP quality control** steps that resulted in how the final SNP-set was created (804,235 SNPs). The number of SNPs shown is the number that remains after all markers that did not satisfy each quality control criterion were dropped. HWE: Hardy-Weinberg Equilibrium.

- Affymetrix 6.0 genotyping array
  - 906,600 SNPs
- SNP call rate ≥ 0.95
  - 859,153 SNPs
- Remove minor allele frequency < 0.01
  - 840,459 SNPs
- Remove non-autosomal SNPs
  - 804,235 SNPs
- Final dataset
  - 804,235 SNPs

Figure 2.2. A flowchart showing the Affymetrix 6.0 genotyping platform **sample quality control** steps that resulted in how the final sample-set was created (1978 samples). The number of samples shown is the number that remains after all samples that did not satisfy each quality control criterion were dropped. GRR: graphical representation of relationship (Abecasis et al. 2001, 742-743).

- Genotyped
  - 2023 Samples
- Sample call rate ≥ 0.95
  - 2009 Samples
- Remove gender mismatch
  - 1986 Samples
- Final dataset
  - 1978 Samples
- Remove related samples based on GRR
  - 1978 Samples
- EIGENSTRAT race-ethnicity check
  - 1986 Samples
2.1.2. Illumina exome-chip array

SNPs QCs were performed on the initial 247,870 genotyped SNPs. After standard QC steps (SNP call rate ≥ 0.98, remove monomorphic and HWE failed SNP, and retained only autosomal SNPs), 161,658 SNPs were retained (Figure 2.3).

Sample QC: 2122 patient samples were initially genotyped. After retaining samples that had call rate ≥ 0.98, removing gender mismatch, related individuals and using EIGENSTRAT to check race/ethnicity, the final dataset contained 2086 unrelated individuals (Figure 2.4).

Figure 2.3. A flowchart showing the Illumina Exome Chip genotyping platform SNP quality control steps that resulted in how the final SNP-set was created (161,658 SNPs). The number of SNPs shown is the number that remains after all markers that did not satisfy each quality control criterion were dropped. HWE: Hardy-Weinberg Equilibrium.
Figure 2.4. A flowchart showing the Illumina Exome Chip genotyping platform sample quality control steps that resulted in how the final sample-set was created (2,086 samples). The number of samples shown is the number that remains after all samples that did not satisfy each quality control criterion were dropped. GRR: graphical representation of relationship (Abecasis et al. 2001, 742-743).

- Genotyped 2122 Samples
- Sample call rate ≥ 0.98 2118 Samples
- Remove gender mismatch 2099 Samples
- Remove related samples based on GRR 2086 Samples
- EIGENSTRAT race-ethnicity check 2099 Samples
- Final dataset 2086 Samples
2.2. Power calculation

Statistical power was calculated for logistic regression test using Quanto software (Gauderman 2006). Collateralization and angiographic CAD phenotype power were calculated respectively. Since both phenotypes in this study were dichotomous, the option of unmatched case-control power calculation was utilized (control:case = 406:473 = 0.86 for collateralization; control:case = 484:435 = 1.12 for angiographically defined CAD). The hypothesis was gene-only. Allele frequency was within the range for common variants, from 0.05 to 0.5, by 0.05 increment. Inheritance mode was log-additive. Disease risk parameters were estimated by population risk of 0.004, 0.01 and 0.1 for collateralization (estimated collateral population prevalence was 0.004 (Lee et al. 2015, 313-321)); 0.06, 0.1 and 0.2 for angiographically defined CAD (estimated CAD population prevalence was 0.06 in all age groups; for people over 65 years old, the estimated prevalence was 0.2). Type I error rate was 5E-08 and two-sided test was used. For a genetic variant with MAF = 0.3, the genetic relative risk (GRR) of above 1.9 would be well powered (> 80% power) to detect. As population prevalence increased, statistical power with fixed sample size decreased.
Figure 2.5. Power calculation for collateral phenotype where 473 cases and 406 controls were included in the study. A) Estimated population prevalence = 0.004; B) estimated population prevalence = 0.01; C) estimated population prevalence = 0.1. In all three figures, x axis represented allele frequency and y axis represented the statistical power. Blue indicated GRR =1.6; orange indicated GRR= 1.7; gray indicated GRR =1.8; yellow indicated GRR =1.9.
Figure 2.6. **Power calculation for angiographic CAD phenotype where 435 cases and 484 controls were in the study.** A) Estimated population prevalence = 0.06; B) estimated population prevalence = 0.1; C) estimated population prevalence = 0.2 in people over 65 years old. In all three figures, x axis represented allele frequency and y axis represented the statistical power. Blue indicated GRR = 1.6; orange indicated GRR = 1.7; gray indicated GRR = 1.8; yellow indicated GRR = 1.9.

**Power Calculation for Angiographic CAD**

(Case = 435; Control = 484)
CHAPTER 3. LOCAL ANCESTRY INFERENCE IN A GENETICALLY DIVERSE POPULATION

3.1. Overview

Admixed populations are populations mixed with recent ancestries from two or more distinct continents, due to historical reasons such as migration and transatlantic slave trade in the last few hundred years. Examples of admixed populations are African Americans and Hispanics. These populations are often understudied despite their high risks of genetic diseases. Due to their population complexity, these populations have both advantages and disadvantages for complex disease genetic studies (Seldin, Pasaniuc, and Price 2011, 523-528).

3.1.1. Hispanic population structure

Hispanics represent a large proportion of residents in South Florida region and one of the fastest growth populations (U. S. Census Bureau 2012). They make up 17% of the entire U.S. population but their genetic risks of CAD are understudied (U. S. Census Bureau 2012). Hispanic populations have a complex genetic structure that reflects recent admixture among Native American, European, and West African ancestral populations. Hispanic/Latino populations show very large variations in admixture population proportions, not only across geographic regions, but also within countries themselves (Bryc et al. 2010, 8954-8961). However, association analyses in Hispanics are complicated by population stratification, which refers to a systematic difference in allele frequencies between cases and controls rather than association of genes with disease (Freedman et al. 2004, 388-393). Population stratification may lead to association with loci unlinked to the disease locus, therefore causing false positive association signals (Freedman et al. 2004, 388-393; Wang et al. 2011, 670-677). Gene flow among European, African and Native
American groups resulted in admixed populations with variable distributions of chromosomes comprised of non-uniform segments from different ancestral (i.e., geographic) populations (Halder et al. 2012, 146-155). Allele frequency differences among ethnic groups can result in frequent false-positive results or reduced power in genetic studies (Tian, Gregersen, and Seldin 2008, R143-50). Thus, I inferred local ancestry - the ancestry of each chromosome segments, in order to adjust population substructure and capture the ancestry specific genetic effect. On the other hand, I also calculated global ancestry by genome-wide averaging of the local ancestry estimates (Liu et al. 2013, 1-7364-7-1).

By now, to adjust for local or global ancestry in association studies is still debatable. Ongoing studies showed that in simulation studies, statistical power was generally higher in admixed population than non-admixed population through adjusting global ancestry for variants with similar effect size in all populations; global ancestry adjustment was best when used for discovery purpose while local ancestry was best for localization purpose (Zhang and Stram 2014, 502-515). Another study had shown that when population stratification was due to local ancestry, adjusting for global ancestry might inflate Type I error (Wang et al. 2011, 670-677). Since part of the study was to discover new genetic variants, considering the complexity of the study population, I adjusted both global and local ancestry in association studies and compared results.
3.2. Experimental Approach

3.2.1. Local ancestry inference

3.2.1.1. Haplotype phasing

To determine the possible haplotypes from genotyping data of each individual, I phased our haplotype with SHAPEIT (Delaneau et al. 2014, 3934) based on the reference populations of European, African and Asian from 1000 Genome Project (1000 Genomes Project Consortium et al. 2012, 56-65). Among the reference populations, I selected unrelated samples of Africans (total 185 individuals): Luhya in Webuye, Kenya (LWK) (97) and Yoruba in Ibadan, Nigeria (YRI) (88); Europeans (183): Utah residents with Northern and Western European ancestry (CEU) (85) and Tuscany in Italia (TSI) (98); Asians (186): Han Chinese in Beijing, China (CHB) (97) and Japanese in Tokyo, Japan (JPT) (89).

SHAPEIT used Hidden Markov Model (HMM) based tree representation algorithm to infer unknown haplotype from known reference haplotypes. I first built up a phasing graph of our population which contained all possible haplotypes, and then I extracted 100 pairs of possible haplotypes from the built graph. The 100 pairs of haplotypes well represented the distribution of possible haplotypes and capture phasing uncertainty due to lack of family data.

3.2.1.2. Local ancestry inference procedure

Local ancestry of the 100 possible haplotypes pairs was inferred using LAMP-LD/LAMP-ANC (Baran et al. 2012, 1359-1367), which used an algorithms combine hidden Markov Models (HMMs) of haplotype diversity within a window-based framework (Figure 3.1). Since samples used in this study were mostly White Non-Hispanics, African Americans
and Hispanic populations, I used references of the same European and African ancestral population from 1000 Genome Project as in the phasing step, and Native American from Human Genome Diversity Project (HGDP) (Cann et al. 2002, 261-262); (Rosenberg et al. 2002, 2381-2385). Since Native American samples were very limited in quantity and EIGENSTRAT plot showed that these samples were relatively close in genetics, I combined all Native American samples from Central and South Americas: Pima (14 individuals) from Mexico; Colombian (7 individuals) from Colombia; Surui (8 individuals) and Karitiana (13 individuals) from Brazil; Maya (21 individuals) from Mexico.

SNPs from the phased 100 pairs of haplotypes were trimmed for those with LD (r²)>0.8. I then extracted common SNPs per window, number of state of 10, and with smoothing. After ancestry inference, I counted the ancestry states at each locus among European, African and Native American for the 100 pairs of haplotypes, and calculated the average of ancestry percentage for each haplotype. I then averaged across the two haplotypes of each individual and get the percentage of ancestry for each locus. The final ancestry percentage was used for ancestry mapping.
Figure 3.1. Pipeline for local ancestry inference and calculating average local ancestry percentage across 100 pairs of inferred haplotypes.

Averaging from 100 pairs of haplotypes:
1, 0 0.58, 0.42 0.52, 0.48 1, 0
1, 0 0.58, 0.42 0.58, 0.42 1, 0

Averaging haplotypes for genotype:
\[
\begin{align*}
&\text{Average ancestry percentage across 100 pairs} \\
&= (1, 0 0.58, 0.42 0.52, 0.48 1, 0 \\
&+ 1, 0 0.58, 0.42 0.58, 0.42 1, 0) / 2 \\
&= 1, 0 0.58, 0.42 0.55, 0.45 1, 0
\end{align*}
\]
3.2.2. Global ancestry inference

Local ancestry was averaged across the genome to calculate the global components of European, African and Native American ancestry. I also used EIGENSTRAT (Price et al. 2006, 904-909) - a principal component based method, to estimate the global ancestry across the entire genome without utilizing ancestry specific markers.

3.3. Results and summary

3.3.1. Sample description

As seen in Figure 3.2, our population was composed of a mixture of Hispanics, Caucasians, and African Americans, which well represented the demography of Miami area. The distance towards ancestral populations in eigenvalues indicated how close the population was to the ancestral population (Africans: YRI, LWK; Caucasians: CEU, TSI; Native Americans). Population in between ancestral populations indicated admixture.
3.3.2. Local ancestry inference results

Local ancestry was inferred by averaging the 100 pairs of haplotype ancestries. This approach showed some similarities with local ancestry inferred from the best haplotype pairs, but better captured phasing uncertainty (Figure 3.3).
Figure 3.3. Comparing local ancestry inference based on averaging 100 pairs of haplotypes and the best-guess haplotype in self-identified white non-Hispanic, African American and white Hispanic. A) Averaging 100 pairs of haplotypes for a self-identified Caucasian. B) Best-guess haplotype for the same Caucasian. C) Averaging 100 pairs of haplotypes for a self-identified African American. D) Best-guess haplotype for the same African American. E) Averaging 100 pairs of haplotypes for a self-identified white Hispanic sample. F) Best-guess haplotype for the same white Hispanic sample. In all three figures, blue: European ancestry; Red: African ancestry; Yellow: Native American ancestry. Each bar from top to bottom represented chromosome 1-22.
3.3.3. Global ancestry inference results

By averaging across all individuals and chromosomes, our samples were composed of approximately 75% European descent, 20% African descent and 5% Native American descent. Besides, global ancestry estimated by our approach and principal component analysis (PCA) were highly correlated (correlation coefficient between eigenvector 1 and African ancestry component $r^2=0.99$; for eigenvector 2 and Native American, $r^2=-0.87$) (Figure 3.4). Since there might be a few Asian samples presented in our original population, and Asian samples did not contain any of the ancestral populations that we used as references, a little swoop for Eigenvector 2 was seen in Figure 3.4 B.

Figure 3.4. Scatter plots showed the correlation between eigenvectors and ancestry component averaged from 100 haplotype pairs. A) Correlation between eigenvector 1 and African component. B) Correlation between eigenvector 2 and Native American component.
3.3.4. Self-identified vs. principal component (PC) based ancestry

The vast majority of the study population could precisely self-identify their race-ethnicity. However, for some individuals there were ambiguity. In order to utilize race-ethnicity information in genetic association analyses, I categorized our population based on eigenvalues from principal component analysis instead of self-report. There were several reasons: first, in Miami, population admixture was more complicated than elsewhere in the US, therefore it could be difficult for individuals to trace back admixture history of their ancestors and know exactly their ancestry category. Second, individuals might not understand differences between each ancestry categories so they self-identified their ancestry category based on vague knowledge. Therefore, to eliminate ambiguity, PC based
ancestry was considered more reliable than self-identified ancestry. In this case, eigenvector 1 and 5 illustrated the best separation of race-ethnicities estimation - PC based race-ethnicity agreed with the majority of self-report race-ethnicity, therefore PC based race-ethnicity was categorized based on eigenvector 1 and 5. (Figure 3.5).

**Figure 3.5. Principal components determined race-ethnicity for subjects participated in our study.** X-axis represented eigenvector 1 and y-axis represented eigenvector 5. Red: Hispanics; green: Caucasians; blue: African Americans.
CHAPTER 4. CORONARY COLLATERALIZATION SHOWS SEX AND ETHNIC DIFFERENCES IN THE PRESENCE OF OBSTRUCTIVE ARTERY DISEASE

4.1. Overview

4.1.1. Biological function of coronary collateral

Coronary artery disease (CAD) is the leading cause of death worldwide, with atherosclerosis being the major etiology. Currently, much is known about CAD pathogenesis and risk factors (Grobbee et al. 1995, 478-482), such as age, gender, smoking, serum cholesterol, metabolism rate and hypertension (Lim et al. 2012, 2224-2260). While this knowledge is useful in predicting CAD, it is often insufficient to adequately predict clinical outcomes after cardiac events, such as myocardial infarction (MI). As such, other factors may play a critical role in clinical outcomes. One such factor is the presence of a collateral circulation. Collateral circulation is a natural bypass process, whereby vessels are extended or expanded to supply blood flow to cardiac muscle that may otherwise be lacking oxygen (Eagle et al. 2004, e213-310; Seiler 2003, 1352-1357). Collateralization is protective and likely reduces mortality (Schaper 2012, 564-566; Meier et al. 2012, 614-621), and has been shown to have a protective effect among those with CAD (Koerselman et al. 2003, 2507-2511; Meier et al. 2012, 614-621). Indeed, the extent of collateralization significantly determines the severity of MI in acute coronary occlusion—the better developed collateral circulation, the less severe of the MI outcomes (Seiler et al. 2013, 2674-2682).

4.1.2. Genetic components of coronary collaterals

The extent of the collateral circulation varies significantly among healthy individuals as a function of both genetic background (Zhang et al. 2010, 923-934; Meier et al. 2007, 975-983) and vascular risk factors such as aging, diabetes and smoking (Kinnaird et al. 2008,
In animal studies, genetic heterogeneity leads to different levels of collateral formation, suggesting genetic predisposition (Zhang et al. 2010, 923-934; la Sala et al. 2012, 494-501; Wang et al. 2012, e31910). Array-based expression profiling in mice showed 783 out of 12000 genes were differently expressed between femoral artery ligation group and control group. Among those differentially expressed genes, a number of angiogenesis related genes were upregulated in the femoral artery ligation group as compared to the control group, such as MCP1, placental growth factor and cysteine-rich protein-61 (Lee et al. 2004, 474-482); (Teunissen, Horrevoets, and van Royen 2012, 897-904). More work need to be done for identifying the genetic loci underlying this wide collateral variation, and candidate genes regulating collateral formation.

**4.1.3. Sex differences and race-ethnicity differences in collateralization**

The rates of collateralization between men and women remains unclear. Some studies showed that there was no significant effect of sex on collateral vessel development (Tatli et al. 2007, 97-99). However, others showed in acute coronary syndrome, men tended to develop greater collateral circulation than women (Lewis 2014). The opposite was reported in another study, where collaterals were more frequent in women than in men with multivessel disease (Waldecker et al. 2002, 243-248). While race-ethnicity groups clearly show prevalence differences in CAD, it is unclear if rates of collateralization differ across race-ethnicity groups. Given differing genetic backgrounds (i.e., population substructure), and differing environmental exposure, rates of collateralization likely differ across racial groups.

To investigate factors influencing coronary collateralization, 868 obstructive CAD patient samples from cardiac catheterization lab were included in this study. These patients
were investigated for collateral circulation grades to assess the impact of sex, race-ethnicity, and other covariates on collateralization. Additionally, I performed admixture mapping to identify regions of the genome that may contribute to different collateralization across the race-ethnic groups.

4.2. Experimental approach

4.2.1. Sample assessment

The study was approved by the University of Miami IRB. All subjects initially presented to the cardiac catheterization laboratories at the University of Miami Medical Center and Jackson Memorial Hospital for coronary angiography. Since the prevalence of collateral was different among patients with or without obstructive coronary artery disease, I decided to analyze the genetic effect of collateralization in obstructive CAD patients only, as collateral circulation was functionally critical when obstructive CAD presented. 868 patients with obstructive CAD (over 50% stenosis presented in any of the four major coronary branches: left main (LM), left anterior descending (LAD), left circumflex (LCX), right coronary artery (RCA) or branch vessel > 2.0 mm in diameter (Eagle et al. 2004, e213-310)) were analyzed from our cardiac catheterization patient cohort. Each patient’s collateral level was assessed based on Rentrop collateral grading system (Rentrop et al. 1985, 587-592). According to the filling of the recipient artery, 4 degrees of collateralization were distinguished. As for grade 0, no filling was presented; grade 1 indicated filling of the side branch of the recipient artery without filling of the main epicardial artery; grade 2 indicated the main epicardial recipient artery was partially filled; and grade 3 indicated the main epicardial recipient artery was completely filled. Other patient characteristics were also assessed: overall presence/absence of collateralization, age, sex, race-ethnicity, blood pressure, current comorbidities such as diabetes, smoking
history, presence/absence of hypertension or cholesterol lowering medications and reason for procedures. Blood was drawn from the subjects prior to the start of the cardiac catheterization. All samples were collected after a written informed consent was obtained from the participant.

The cardiac catheterization procedures were performed in the course of clinical care. Because patients were undergoing a CT angiography for safety reasons, those were excluded if they have a contrast dye or iodine allergy, renal disease, active asthma or bronchospasm, or cardiac conduction abnormalities on a screening electrocardiogram. Pregnant women, children and elderly patients who could not provide informed consent were excluded from study subjects. Patients who had non-obstructive CAD determined by one of the study cardiologist were also excluded due to the different prevalence of collateralization between obstructive and non-obstructive CAD patients.

4.2.2. Phenotyping

The collateral phenotype was defined on the coronary CT angiographic images. These images were assessed by an experienced cardiologist/reader to determine the presence and location of collateralization, as well as other coronary traits (e.g., presence of stenosis, etc). The initial dataset includes 2023 samples from the Miami Cardiovascular Registry with completed coronary angiography. Subjects in this study were required to have obstructive coronary artery disease, defined as over 50% stenosis in any major epicardial vessel (Eagle et al. 2004, e213-310) (N=868). Coronary collateral was defined as an anastomotic connections without an intervening capillary bed between portions of the same coronary artery and between different coronary arteries (Koerselman et al. 2003, 2507-2511). Based
on the Rentrop collateral flow classification, the maximum collateral grade at any given location was used in our analyses.

**4.2.3. Statistical analysis**

Analysis of variance (ANOVA) test with two-tail hypothesis were performed on continuous traits (e.g., age) to compare the difference of means between four collateral classification groups. Chi-square tests were performed to compare categorical trait differences (such as sex, race-ethnicity, etc.) between classification groups.

Ordinal logistic regression was performed using the polr function in R. Among all the covariates recorded, I selected covariates that showed differences in collateralization and performed a generalized linear model with categorical trait to test their association with the phenotype. Variables of collateral classification as the endpoint, with sex, smoking status, diabetes and race-ethnicity as the independent variables \( \text{Collaterals} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Smoke} + \beta_3 \text{Diabetes} + \beta_4 \text{Race-ethnicity} + \varepsilon \). Covariates which did not show a difference (P-value > 0.05) among collateral classification were excluded from this model, such as cholesterol lowering medication and hypertension.

**4.2.4. Admixture mapping**

Admixture mapping uses a population of mixed ancestry to identify genetic loci that contribute to different disease prevalence found among ancestral populations. In our study, logistic regression model was used to estimate the effect of global ancestry and the difference between global and local ancestry on collaterals. After controlling for global ancestry represented by averaging local ancestry across the entire genome, I calculated the association between the phenotype and difference of global and local ancestry at each
locus. I extracted $\beta$ and p-value for each ancestry component to analyze the effect of ancestry on collateral phenotype.

$$\text{logit (Y)} = \beta_0' + \beta_1' G_{AA} + \beta_2' G_{NA} + \beta_3' (G_{AA} - L_{AA}) + \beta_4' (G_{NA} - L_{NA}) + \beta_5' \text{Age} + \beta_6' \text{Sex} + \varepsilon$$

Where, $Y$ was the collaterals phenotype coded as 0 no-collaterals and 1 collaterals, $G_{AA}$ was a measure of the global African ancestry, $G_{NA}$ was a measure of the global Native American ancestry, $L_{AA}$ was the locus-specific local African ancestry, and $L_{NA}$ was the locus-specific local Native American ancestry. Age represented the age at exam.

4.2.5. Permutation correction

To correct for multiple testing, I performed a permutation test. The phenotype was permuted 2000 times; for each permutation, p-values of correlation coefficients for the differences between global and local ancestry ($\beta_3'$ and $\beta_4'$ as described in 4.2.4) were calculated across the genome. Subsequently, I recorded the minimum p-value across the entire genome. I then estimated the effective number of tests by dividing 0.05 to the 5th percentile distribution of the minimum p-values of the 2000 iterations. To correct for multiple testing, the genome-wide significant P-value threshold was calculated by dividing 0.05 to the effective number of test.

4.2.6. Likelihood ratio test

To further investigate the effect of local ancestry in addition to global ancestry in collateralization, I also performed likelihood ratio test of full model vs. reduced model.

Full model: $\text{logit (Y)} = \beta_0 + \beta_1 G_{AA} + \beta_2 G_{NA} + \beta_3 (G_{AA} - L_{AA}) + \beta_4 (G_{NA} - L_{NA}) + \beta_5 \text{Age} + \beta_6 \text{Sex} + \varepsilon$

Reduced model: $\text{logit (Y)} = \beta_0' + \beta_1' G_{AA} + \beta_2' G_{NA} + \beta_3' \text{Age} + \beta_4' \text{Sex} + \varepsilon$
Since adjacent local ancestry is highly correlated with each other due to linkage disequilibrium, I also performed 2000 permutation test to calculate the effective number of test, and the method is similar with 4.2.5.

4.2.7. Genetic components for coronary collateralization

4.2.7.1. Power calculation

Statistical power was calculated for collateral phenotype (Figure 2.5). The estimated prevalence of collateral in the general population was 0.4% (Lee et al. 2015, 313-321). Two-sided test was performed.

4.2.7.2. Identify common genetic variants

Genotyping for common variants was performed on Affymetrix 6.0 genotyping chip (Chapter 2). PLINK was used (Purcell et al. 2007, 559-575) to analyze common genetic variant association with collateral phenotype. When adjusting for global ancestry, the statistical model was: \( \text{logit}(Y) = \beta_0 + \beta_1 \text{Genotype} + \beta_2 \text{Eigenvector1} + \beta_3 \text{Eigenvector2} + \beta_4 \text{Eigenvector5} + \beta_5 \text{Age} + \beta_6 \text{Sex} + \epsilon \)

I selected eigenvector1, eigenvector2, eigenvector5, age and sex as covariates since these covariates were significantly correlated with collateral phenotype (P-value < 0.05). I also used R package leaps (Lumley 2009) to check for selected covariates.

Despite adjusting for global ancestry, I also adjusted local ancestry and compared the results with when adjusting for global ancestry. R was used to perform the association test, and the statistical model was: \( \text{logit}(Y) = \beta_0 + \beta_1 \text{Genotype} + \beta_2 \text{Local}_\text{African}_\text{Ancestry} + \beta_3 \text{Local}_\text{Native}_\text{American}_\text{Ancestry} + \beta_4 \text{Age} + \beta_5 \text{Sex} + \epsilon \)
4.2.7.3. Rare genetic variants

Genotyping for the rare variates were performed on Illumina Exome Chip platform (chapter 2). PLINK (Purcell et al. 2007, 559-575) was used to analyze rare single genetic variant association test. Statistical models were consistent with common genetic variants analyses.

4.2.7.4. Gene-based test

SNPs that were within the longest transcript of a gene or its 40 kb flanking region (genome build 37.1) were assigned to that gene: if SNPs were mapped within more than one gene, all such genes were included. Using this approach, I was able to map as much single variants to genes as possible. The P-value of the SNP showing the strongest evidence for allelic association was then be taken as the gene-wide significance measurement.

To identify the association between genes that contain multiple single variants and the phenotype, I performed gene-based analysis using Versatile Gene-based Association Study (VEGAS) (Liu et al. 2010, 139-145). The VEGAS software utilized single-variant tests and knowledge of linkage disequilibrium to simulate a null distribution. Since local or global ancestry adjustment were performed in single genetic variant association level, there was no need to further adjust for ancestry in gene-based test. VEGAS was used for both common and rare variant gene-based test analyses.

4.2.7.5. Pathway analyses

Pathway analyses were performed using gene-set enrichment analyses (GSEA) (Wang, Li, and Bucan 2007). All genes were assigned to a series of functional categories defined by a web-based source Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/). GSEA then ranked all genes in order of a gene-wide
association statistic and tested whether the genes in a particular gene set had higher rank overall than would be expected by chance.

4.3. Results

4.3.1. Dataset description

Among the 868 subjects with obstructive CAD, 404 (47%) had no collateralization (grade 0); 142 (16%) had maximum grade 1 collateral at any location; 205 (24%) had maximum grade 2 collateralization at any locations and 117 (13%) had maximum grade 3 collateralization at any location. There was no difference in the average age among patients with different grade of collateralization (P-value > 0.068). Men were more likely to develop collateralization than women in all collateral groups, with chi-square test P-value = 0.000175 (Table 4.1). Smoking status was also significant contributor to collateralization, especially among higher grade collateral patients. Smokers tend to develop higher rate of collateralization than non-smokers (chi-square p-value = 0.00223). Diabetes, systolic and diastolic blood pressure (BP), cholesterol, and BP and lipid lowering medicine use, did not show statistical significant differences among patients with different grade of collateralization (Table 4.1).

| Table 4.1. Patients clinical characteristics under each collateral grades. |
|---------------------------------|---------|---------|---------|---------|----------|
| Age (sd)                        | Grade 0 | Grade 1 | Grade 2 | Grade 3 | P-Value  |
| Sex (Female)                    | 32%     | 27%     | 20%     | 14%     | 0.000175 |
| DBP (sd)                        | 76 (13) | 76 (13) | 77 (14) | 76 (13) | 0.889    |
| SBP (sd)                        | 141 (23)| 140 (24)| 139 (25)| 136 (25)| 0.291    |
| Diabetes                        | 43%     | 53%     | 41%     | 51%     | 0.068    |
| Smoking                         | 51%     | 50%     | 63%     | 66%     | 0.00223  |
| Anti-cholesterol meds           | 30%     | 33%     | 36%     | 38%     | 0.401    |
4.3.2. Sex was associated with collateralization, however smoking status might be a confounder

To better understand the relationship between sex, smoking, and collateralization, I also tested the association between sex and smoking status to see if any of them was a confounder. Indeed, men were more likely to smoke as compared to women (62% vs 38% smokers, respectively, P-value < 0.0001). Covariates that showed moderate to significant differences in collateralization were included into ordinal logistic regression model (sex, smoking status, diabetes and race-ethnicity), and the result indicated that sex was statistically associated with collateralization while including smoking, diabetes and race-ethnicity in the model (beta = 0.59, P-value = 0.000162). Smoking status also showed significant association with collateralization (beta = 0.38, P-value = 0.00523) while controlling for sex, diabetes and race-ethnicity. However, when an interaction term of sex and smoking were included into the regression model, neither smoking nor the interaction term was significant (P-value > 0.05); however, sex was still significant (P-value < 0.001).

4.3.3. Hispanics is statistically significantly different in collateralization as compared to Caucasians and African Americans

Based on principal components of the genotype, individuals were grouped into Caucasians (247 individuals), Hispanics (487 individuals) and African Americans (92 individuals) (Table 4.2). There were 42 individuals with unclear race/ethnicities based on principal components that were excluded from this analysis. From a total of 826 individuals, different race-ethnicity groups showed different rates of collateralization: overall, 59% of Hispanic patients had collaterals as compared to 50% in African Americans and 48% in Caucasians (chi-square P-value = 0.017; df = 2). Hispanic patients tended to have higher
rate of collateralization in grade 1 and grade 3 as compared to the other two populations
(Chi-square P-value = 0.0257) (Table 4.2).

Table 4.2. EIGENSTRAT defined race-ethnicity differences in collateral grades within obstructive CAD patients.

<table>
<thead>
<tr>
<th></th>
<th>Grade 0</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>47 (51%)</td>
<td>13 (14%)</td>
<td>23 (25%)</td>
<td>9 (10%)</td>
<td>92 (100%)</td>
</tr>
<tr>
<td>CAUC</td>
<td>129 (52%)</td>
<td>30 (12%)</td>
<td>61 (25%)</td>
<td>27 (11%)</td>
<td>247 (100%)</td>
</tr>
<tr>
<td>HISP</td>
<td>202 (41%)</td>
<td>94 (19%)</td>
<td>114 (23%)</td>
<td>77 (16%)</td>
<td>487 (100%)</td>
</tr>
<tr>
<td>Total</td>
<td>378 (47%)</td>
<td>137 (16%)</td>
<td>198 (24%)</td>
<td>113 (13%)</td>
<td>826 (100%)</td>
</tr>
</tbody>
</table>

Chi-square P-value = 0.0257, df = 6. AA: African Americans; CAUC: Caucasians; HISP: white Hispanics/Latinos.

In ordinal logistic regression model, after controlling for sex, diabetes and smoking status, race-ethnicity comparisons also showed statistically significant difference in collaterals between Hispanics and Caucasians (beta = -0.39, P-value = 0.00845). However, the comparison between Hispanics and African Americans was not statistically significant (beta = -0.23, P-value = 0.276). Collateralization was also classified by presence/absence, and results were similar as compared to classify collaterals by grade.

4.3.4. Admixture mapping identified two regions associated with collateral phenotype
4.3.4.1. Admixture mapping showed Native American and African ancestry are significantly associated with collateral phenotype at regions on chromosome 17

Admixture mapping were performed on 868 subjects of obstructive CAD patients, with presence/absence of collateralization as the endpoint phenotype. Four covariates were tested for their association with collateral phenotype: global African ancestry, global Native American ancestry; the difference between global and local ancestry for African ancestry and Native American ancestry (delta African and delta Native American). Global African and Native American ancestry did not associate with collateral phenotype (the top ranking P-value > 0.05). Delta Native American ancestry was highly associated with the
presence of collaterals at a region on chromosome 17 (hg19:chr17:36120051-40782083 (hg38:chr17:37760062-42630065), beta = 0.55, min P-value = 0.000127) (Figure 4.1.A & Figure 4.2). Delta African ancestry also showed association with collaterals at a different region on chromosome 17 (hg19:chr17:32796666-33440166 (hg38: chr17:34469647-35113147), beta = 0.38, min P-value = 0.00072) (Figure 4.1.B & Figure 4.2). These P-values did not survive a multiple-testing correction based on our permutation test (1931 effective tests in the local African Ancestry analysis, P-value threshold = 2.59E-05; 1543 effective tests in the Native American ancestry analysis, P-value threshold = 3.249E-05).

4.3.4.2. Likelihood ratio test showed regions that were associated with collateral phenotype

In the likelihood ratio test between full and reduced model, the top association result was also on chromosome 17 (hg19:chr17:32353020-39736083, beta=17.49, min P-value = 1.59E-04). Based on the permutation test, the effective number of test is 2372, therefore the significant P-value threshold was 2.11E-05. Our result did not reach the genome-wide significant threshold, however, the likelihood ratio test agreed with admixture mapping test in a region where local ancestry played an additional role in collateralization upon global ancestry.

4.3.4.3. Candidate genes that were within these two chromosome 17 regions identified by admixture mapping

Based on our gene-based tests, I extracted those that were within the two regions on chromosome 17 (±100KB) determined by admixture mapping.
Figure 4.1. Manhattan plots showed associations between collateralization and A) delta Native American ancestry; B) delta African ancestry (differences between global and local Ancestry) in admixture mapping. The peak regions of delta Native American ancestry and delta African ancestry were both located on chromosome 17. Native American ancestry was associated with the presence of collaterals at a region on chromosome 17 (chr17:36120051-40782083, beta = 0.55, minimal P-value = 1.27E-04). Effective number of testing: 1543; P-value cut off: 3.2E-05; -log10 P-value: 4.5. African ancestry showed association with collaterals at a different region on chromosome 17 (chr17:32796666-33440166, beta = 0.38, minimal P-value = 7.2E-04). Effective number of testing: 1931; P-value cut off: 2.6E-05; -log10 P-value: 4.6. X-axis indicated chromosomes 1 to 22. Y-axis indicated –log10 of delta Native American/African ancestry P-values.
Figure 4.2. Chromosome 17 regional plot showed peak association between delta African/Native American ancestries with collateralization in admixture mapping. Local Native American ancestry was highly associated with the presence of collaterals at a region on chromosome 17 (hg19:chr17:36120051–40782083, min P-value = 0.000127, -log10 (P-value) = 3.90). Local African ancestry also showed association with collaterals at a different region on chromosome 17 (hg19:chr17:32796666–33440166, min P-value = 0.00072, -log10 (P-value) = 3.14). X-axis represented the base pair location on chromosome 17; Y-axis represented –log10 p-value.
### Table 4.3.1. Affy6 Chip: adjusting for global ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>NLE1</td>
<td>22</td>
<td>33459395</td>
<td>33469302</td>
<td>71.00972</td>
<td><strong>0.0045</strong></td>
<td>rs8071313</td>
<td>0.000239</td>
</tr>
<tr>
<td>17</td>
<td>FNDC8</td>
<td>18</td>
<td>33448673</td>
<td>33457455</td>
<td>47.44485</td>
<td><strong>0.00999</strong></td>
<td>rs8071313</td>
<td>0.000239</td>
</tr>
<tr>
<td>17</td>
<td>RAD51D</td>
<td>18</td>
<td>33427956</td>
<td>33446670</td>
<td>37.94023</td>
<td><strong>0.02991</strong></td>
<td>rs8071313</td>
<td>0.000239</td>
</tr>
<tr>
<td>17</td>
<td>SLC35G3</td>
<td>28</td>
<td>33520263</td>
<td>33521319</td>
<td>67.45813</td>
<td><strong>0.03223</strong></td>
<td>rs8076291</td>
<td>0.007231</td>
</tr>
<tr>
<td>17</td>
<td>UNC45B</td>
<td>37</td>
<td>33475236</td>
<td>33513610</td>
<td>79.37885</td>
<td><strong>0.03295</strong></td>
<td>rs8076291</td>
<td>0.007231</td>
</tr>
</tbody>
</table>

### Table 4.3.2. Affy6 Chip: adjusting for local ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>NLE1</td>
<td>22</td>
<td>33459395</td>
<td>33469302</td>
<td>52.37766</td>
<td><strong>0.024</strong></td>
<td>rs8071313</td>
<td>0.007806</td>
</tr>
<tr>
<td>17</td>
<td>RAD51D</td>
<td>18</td>
<td>33427956</td>
<td>33446670</td>
<td>35.36699</td>
<td><strong>0.04183</strong></td>
<td>rs8071313</td>
<td>0.007806</td>
</tr>
</tbody>
</table>
### Table 4.3.3. Exome Chip: adjusting for global ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>C17orf102</td>
<td>4</td>
<td>32904511</td>
<td>32906321</td>
<td>2.083409</td>
<td>0.737</td>
<td>exm1311048</td>
<td>0.232</td>
</tr>
<tr>
<td>17</td>
<td>LIG3</td>
<td>9</td>
<td>33310002</td>
<td>33331575</td>
<td>5.277177</td>
<td>0.788</td>
<td>exm1311291</td>
<td>0.05518</td>
</tr>
<tr>
<td>17</td>
<td>CCT6B</td>
<td>12</td>
<td>33255025</td>
<td>33288412</td>
<td>6.055161</td>
<td>0.901</td>
<td>exm1311291</td>
<td>0.05518</td>
</tr>
</tbody>
</table>

### Table 4.3.4. Exome Chip: adjusting for local ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>RAD51D</td>
<td>28</td>
<td>33427956</td>
<td>33446670</td>
<td>27.54135</td>
<td>0.459</td>
<td>exm1311562</td>
<td>0.119128</td>
</tr>
<tr>
<td>17</td>
<td>FNDC8</td>
<td>31</td>
<td>33448673</td>
<td>33457455</td>
<td>29.28343</td>
<td>0.532</td>
<td>exm1311562</td>
<td>0.119128</td>
</tr>
<tr>
<td>17</td>
<td>RFFL</td>
<td>5</td>
<td>33338964</td>
<td>33353623</td>
<td>4.187229</td>
<td>0.538</td>
<td>exm1311491</td>
<td>0.196297</td>
</tr>
</tbody>
</table>
Within the regions of chr17:32796666–33440166 ± 100KB, where local African ancestry was highly associated with collateral phenotype, there were 11 genes in this regions. In Affymetrix 6.0 platform, when adjusting global ancestry as a covariate, 3 out of the 11 genes were significant in our gene-based test: NLE1 (P-value=4.50E-03); FNDC8 (P-value=9.99E-03); RAD51D (P-value=0.030); SLC35G3 (P-value=0.032); UNC45B (P-value=0.033) (Table 4.3.1). When adjusting local ancestry as a covariate, 2 out of the 11 genes were significant in our gene-based test: NLE1 (P-value=0.024); RAD51D (P-value=0.042) (Table 4.3.2).

In Illumina Exome Chip platform, when adjusting for either global or local ancestry as a covariate, there was no gene that reached statistically significant in our gene-based test. The top associated results were shown in Table 4.3.3 and Table 4.3.4.

Within the regions of chr17:36120051–40782083 ± 100KB, where local Native American ancestry was highly associated with collateral phenotype based on admixture mapping, there were 148 genes in this region.

In Affymetrix 6.0 platform, when adjusting for global ancestry as a covariate, 3 out of 148 genes in our gene-based test were significantly associated with collateralization: PLXDC1 (P-value = 0.019); MED1 (P-value = 0.026); KRT23 (P-value = 0.033) (Table 4.4.1).

When adjusting for local ancestry as a covariate, 6 out of 148 genes were significantly associated with collateralization in our gene-based test: PLXDC1 (P-value = 7.99E-03); ZPBP2 (P-value = 0.011); GSDMB (P-value = 0.023); ORMDL3 (P-value = 0.027); KRT23 (P-value = 0.038); KRT20 (P-value = 0.049). Among these 6 genes, 3 of them had the same best SNP (Table 4.4.2).
In Illumina Exome Chip platform, when adjusting for global ancestry, no gene reached statistically significant level (Table 4.4.3). However, when adjusting for local ancestry, 4 out of 148 genes were significantly associated with collaterals: ZPBP2 (P-value = 0.017); GSDMB (P-value = 0.023); ORMDL3 (P-value = 0.029); IKZF3 (P-value = 0.049). All of these 4 genes had the same best SNP (Table 4.4.4).

4.3.4.4. Functions of these candidate genes within two chromosome 17 regions identified by admixture mapping

For genes that were within the region where African ancestry associated with collaterals (chr17:32796666–33440166 ± 100KB), I looked up their functions.

Unc-45 Myosin Chaperone B (UNC45B), also known as Cardiomyopathy Associated 4, acts as a co-chaperone for HSP90 and is required for proper folding of the myosin motor domain. It plays a role in sarcomere formation during muscle cell development, and it is necessary for normal early lens development. mRNA of this gene is mainly found in blood and heart, plus epiblast (early embryonic tissues) and skeletal muscle (musculoskeletal system). mRNA level indicates that this gene is overexpressed in heart-left ventricle (x18.5) and heart–atrial appendage (x10.1) (source: GeneCards).

Notchless Homolog 1 (NLE1) regulates Notch activity. It plays a role in regulating the expression of CDKN1A and several members of the WNT signaling pathway, probably via its effects on Notch activity. It is required during embryogenesis for inner mass cell survival (source: GeneCards).
### Table 4.4.1. Affy6 Chip: adjusting for global ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>PLXDC1</td>
<td>38</td>
<td>37224094</td>
<td>37296114</td>
<td>89.24197</td>
<td>0.01863</td>
<td>rs8068922</td>
<td>0.000869</td>
</tr>
<tr>
<td>17</td>
<td>MED1</td>
<td>5</td>
<td>37563777</td>
<td>37607275</td>
<td>17.97599</td>
<td>0.02621</td>
<td>rs7226222</td>
<td>0.01781</td>
</tr>
<tr>
<td>17</td>
<td>KRT23</td>
<td>47</td>
<td>39079188</td>
<td>39092876</td>
<td>101.0352</td>
<td>0.03289</td>
<td>rs10491135</td>
<td>0.003159</td>
</tr>
</tbody>
</table>

### Table 4.4.2. Affy6 Chip: adjusting for local ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th>P-value</th>
<th>Best-SNP</th>
<th>SNP-p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>PLXDC1</td>
<td>38</td>
<td>37224094</td>
<td>37296114</td>
<td>104.3129</td>
<td>0.00799</td>
<td>rs6503745</td>
<td>0.001536</td>
</tr>
<tr>
<td>17</td>
<td>ZPBP2</td>
<td>13</td>
<td>38024570</td>
<td>38033085</td>
<td>47.77292</td>
<td>0.01148</td>
<td>rs11078927</td>
<td>0.003563</td>
</tr>
<tr>
<td>17</td>
<td>GSDMB</td>
<td>17</td>
<td>38061054</td>
<td>38073583</td>
<td>53.25027</td>
<td>0.02303</td>
<td>rs11078927</td>
<td>0.003563</td>
</tr>
<tr>
<td>17</td>
<td>ORMDL3</td>
<td>13</td>
<td>38078751</td>
<td>38080474</td>
<td>42.18244</td>
<td>0.02655</td>
<td>rs11078927</td>
<td>0.003563</td>
</tr>
<tr>
<td>17</td>
<td>KRT23</td>
<td>47</td>
<td>39079188</td>
<td>39092876</td>
<td>97.81497</td>
<td>0.03816</td>
<td>rs10491135</td>
<td>0.005532</td>
</tr>
<tr>
<td>17</td>
<td>KRT20</td>
<td>29</td>
<td>39032637</td>
<td>39041415</td>
<td>55.43304</td>
<td>0.04873</td>
<td>rs10221230</td>
<td>0.016474</td>
</tr>
</tbody>
</table>
### Table 4.4.3. Exome Chip: adjusting for global ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th><strong>P-value</strong></th>
<th>Best-SNP</th>
<th>SNP-pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>ORMDL3</td>
<td>17</td>
<td>38078751</td>
<td>38080474</td>
<td>29.83573</td>
<td>0.09856</td>
<td>rs2872507</td>
<td>0.02195</td>
</tr>
<tr>
<td>17</td>
<td>ZPB2</td>
<td>22</td>
<td>38024570</td>
<td>38033085</td>
<td>35.12355</td>
<td>0.11095</td>
<td>rs2872507</td>
<td>0.02195</td>
</tr>
<tr>
<td>17</td>
<td>GSDMB</td>
<td>24</td>
<td>38061054</td>
<td>38073583</td>
<td>36.10887</td>
<td>0.123</td>
<td>rs2872507</td>
<td>0.02195</td>
</tr>
</tbody>
</table>

### Table 4.4.4. Exome Chip: adjusting for local ancestry

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>nSNPs</th>
<th>Start</th>
<th>Stop</th>
<th>Test</th>
<th><strong>P-value</strong></th>
<th>Best-SNP</th>
<th>SNP-pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>ZPB2</td>
<td>22</td>
<td>38024570</td>
<td>38033085</td>
<td>57.96908</td>
<td><strong>0.0165</strong></td>
<td>rs2872507</td>
<td>0.006631</td>
</tr>
<tr>
<td>17</td>
<td>GSDMB</td>
<td>24</td>
<td>38061054</td>
<td>38073583</td>
<td>57.8318</td>
<td><strong>0.02257</strong></td>
<td>rs2872507</td>
<td>0.006631</td>
</tr>
<tr>
<td>17</td>
<td>ORMDL3</td>
<td>17</td>
<td>38078751</td>
<td>38080474</td>
<td>43.30217</td>
<td><strong>0.02904</strong></td>
<td>rs2872507</td>
<td>0.006631</td>
</tr>
<tr>
<td>17</td>
<td>IKZF3</td>
<td>28</td>
<td>37921991</td>
<td>38020421</td>
<td>45.66973</td>
<td><strong>0.04923</strong></td>
<td>rs2872507</td>
<td>0.006631</td>
</tr>
</tbody>
</table>
RAD51 Paralog D (RAD51D) has recombinase activity, which involves in DNA repair during DNA replication or induced DNA-damaging agents. Solute Carrier Family 35 Member G3 (SLC35G3) is a membrane protein, and it has unknown functions. Fibronectin Type III Domain Containing 8 (FNDC8) also has unknown function.

For genes that were within the region where Native American ancestry associated with collateral phenotype (chr17:36120051–40782083 ± 100KB), their individual functions are: Plexin Domain Containing 1 (PLXDC1) is a tumor endothelial marker. Disease associated with this gene include endocardium disease. It plays a critical role in endothelial cell capillary morphogenesis (source: GeneCards). It is detected in endothelial cells from cancer cells, such as colorectal cancer, lung, liver, pancreas, breast and brain. It is not detectable in endothelial cells from normal tissue. It is expressed in fibrovascular membrane, with increased expression in individuals with proliferative diabetic retinopathy. Tumor angiogenesis can be induced by hypoxia, and hypoxia can also induce collaterals.

Mediator Complex Subunit 1 (MED1), is a component of the Mediator complex, which is a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes (source: GeneCards).

Keratin 23, Type I (KRT23) is a member of the keratin family. The keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into cytokeratins and hair keratins. The type I cytokeratins consist of acidic proteins which are arranged in pairs of heterotypic keratin chains (source: GeneCards).

Gasdermin B (GSDMB) encodes a member of the Gasdermin-domain containing protein family. Gasdermin-family genes are implicated in the regulation of apoptosis in epithelial cells, and are linked to cancer. This gene may play a role as secretory or metabolic
product involved in secretory pathway. It may also play a role in achieving and maintaining the final differentiation state of epithelial cells (source: GeneCards).

Keratin 20 Type I (KRT20) plays a significant role in maintaining keratin filament organization in intestinal epithelia. When phosphorylated, it plays a role in the secretion of mucin in the small intestine (source: GeneCards).

IKAROS Family Zinc Finger 3 (IKZF3), encodes a member of the Ikaros family of zinc-finger proteins. Three members of this protein family (Ikaros, Aiolos and Helios) are hematopoietic-specific transcription factors involved in the regulation of lymphocyte development. This gene product is a transcription factor that is important in the regulation of B lymphocyte proliferation and differentiation. It plays an essential role in regulation of B-cell differentiation, proliferation and maturation to an effector state. It also involved in regulation BCL2 expression and controlling apoptosis in T-cells in an IL2-dependent manner (source: GeneCards).

ORMDL Sphingolipid Biosynthesis Regulator 3 (ORMDL3) negatively regulates sphingolipid synthesis, and it may indirectly regulate endoplasmic reticulum-mediated Ca\(^{2+}\) signaling. Zona Pellucida Binding Protein 2 (ZPBP2) has unknown function, and it may be implicated in gamete interaction during fertilization (source: GeneCards).

### 4.3.5. Admixture mapping in Hispanic population

Among 868 study samples, 493 individuals were Hispanics. Hispanic population was extracted and performed a separated admixture mapping analysis. Admixture mapping in Hispanic-only population showed that local Native American was highly associated with collaterals phenotype on chromosome 17 (Hispanic only: chr17:36074979–41009445, beta = 0.55, minimal P-value = 2.8E-04), same region but in a wider range as compared with
combining individuals from all race ethnicities (all individuals: chr17:36120051–40782083, beta = 0.55, minimal P-value = 1.27E-04) (Figure 4.3.A and Figure 4.4.A).

Hispanics-only data also showed local African ancestry was highly associated with collaterals on chromosome 5 (Hispanics-only: chr5: 9195233–9509811, beta = -0.46, minimal P-value = 1.10E-03). Previously, when combining all race-ethnicities, local African ancestry showed top association signal with collaterals on chromosome 17 (all individuals: chr17:32796666–33440166, beta = 0.38, minimal P-value = 7.2E-04). However, in Hispanics-only population, chromosome 17 region was still significant (Hispanics-only: chr17: 32971763–33892087, beta = 0.40, minimal P-value = 2.31E-03), but it was not where the top association signal located (Figure 4.3.B and Figure 4.4.B).

4.3.6. Genetic components for collateral phenotype

4.3.6.1 Common genetic variants

When adjusting for global ancestry, single genetic variant association test showed the top association results were: chr5: rs16890371, P-Value = 5.18E-06; chr13: rs17497524, P-Value = 7.86E-06; chr18: rs4892107, P-Value = 8.08E-06. None of our results reached genome-wide significance (Figure 4.5.A and 4.6.A).

Based on single genetic variants, gene-based test was performed. The top associated genes were: chr16: SNAI3, P-Value = 8.53E-05; chr16: RNF166, P-Value = 9.98E-05; chr16: PIEZO1, P-Value = 1.45E-04 (Figure 4.7.A).
Figure 4.3. Regional plot showed admixture mapping in Hispanics-only population. Top results of A) Native American ancestry associated with collateralization in Hispanics-only was located on chromosome 17, black line illustrated the associated of all individuals, while grey line represented Hispanics-only population. Figure B) showed the top results of African ancestry association in Hispanics-only was located on chromosome 5. Red represented all populations, while pink represented Hispanics-only population. In both figures, X axis represented base pair position in MB, and y axis represented $-\log_{10} P$-value.
Figure 4.4. Manhattan Plot showed admixture mapping in Hispanic population: A) local Native American ancestry was associated with collateralization on chromosome 17: 36074979–41009445 (hg19) as the top signal. Effective number of testing: 1911; P-value cut off: 2.6E-05; -log10 P-value: 4.6. B) Local African ancestry was associated with collateralization on chromosome 5: 9195233–9509811 as the top signal. Effective number of testing: 2283; P-value cut off: 2.2E-05; -log10 P-value: 4.7. In both figures, x axis represented chromosome 1 to 22. Y axis represented –log10 P-value.
In gene-set enrichment analyses, totally 178 pathways were tested. These pathways were chosen based on KEGG database – a commonly used database which includes pathways representing molecular interaction and reaction networks such as metabolism, human diseases, cellular processes, etc. Among tested pathways, no gene set was significant at FDR < 25%. 2 gene sets were significantly enriched at nominal P-value < 1% and 15 gene sets were significantly enriched at nominal P-value < 5%. Pathway analyses indicated that SNARE interactions in vesicular transport (size = 37, NOM P-value = 0.001, FDR q-value = 0.486), pentose phosphate pathway (size = 23, NOM P-value = 0.016, FDR q-value = 0.296), and RNA polymerase (size = 29, NOM P-value = 0.010, FDR q-Value = 0.361) were the top associated pathways (Figure 4.8.A).

When adjusting for local ancestry, single genetic variants association test showed the top association results were: chr5: rs16890371, P-Value = 1.21E-06, beta = 0.19; chr11: rs10894957, P-Value = 6.68E-06, beta = 0.12; chr13: rs17497524, P-Value = 7.19E-06, beta = 0.16. No single genetic variants passed genome-wide significant threshold (Figure 4.5.B and 4.6.B).

I also performed gene-based test on single genetic variant results. The top association results were the same as compared to when global ancestry was adjusted, with slightly different statistical values. chr16: SNAI3, P-Value = 9.98E-05; chr16: RNF166, P-Value = 1.05E-04; chr16: PIEZO1, P-Value = 1.51E-04 (Figure 4.7.B).

In gene-set enrichment analyses, there were also 178 pathways tested based on KEGG database. Among tested pathways, 2 gene set were significant at FDR <25%. 4 gene sets were significantly enriched at nominal P-value <1% and 8 gene sets were significantly enriched at nominal P-value <5%. Pathway analyses showed top association results were
also the same as when global ancestry was adjusted, while the ranking was different: RNA polymerase (size = 29, NOM P-value = 0.001, FDR q-Value = 0.152); SNARE interactions in vesicular transport (size = 37, NOM P-value = 0.001, FDR q-Value = 0.120); pentose phosphate pathway (size = 23, NOM p-value = 0.015, FDR q-value = 0.312) (Figure 4.8.B).

4.3.6.2. Rare genetic variants

For rare genetic variants genotyped on the Illumina exome chip platform, when I adjusted for global ancestry, the top single genetic variant association results were: chr13: 73916628, P-Value = 5.20E-05, beta = -4.05; chr2: 35190105, P-Value = 7.24E-05, beta = 3.97; chr13: 36663302, P-Value = 7.48E-05, beta = -3.96 (Figure 4.5.C and 4.6.C).

Based on single genetic variants, gene-based test showed that chr13: DCLK1 (P-Value = 6.68E-04); chr4: PIGG (P-Value = 8.40E-04); chr4: ZNF721 (P-Value = 2.42E-03) were the top associated genes with collateral phenotype (Figure 4.7.C).

I also performed pathway analyses on gene-based test results. In gene-set enrichment analyses, totally 178 pathways were tested. Among tested pathways, 1 gene set was significant at FDR < 25%. 4 gene sets were significantly enriched at nominal P-value < 1% and 18 gene sets were significantly enriched at nominal P-value < 5%. The top associated pathways were: glycosylphosphatidylinositol GPI anchor biosynthesis (size = 24, NOM P-value = 0.010, FDR q-value = 0.213); mismatch repair (size = 22, NOM P-value = 0.037, FDR q-value = 0.773); viral myocarditis (size = 66, NOM P-value = 0.003, FDR q-value = 0.768) (Figure 4.8.C).
Figure 4.5. Single genetic variants associated with collateralization when adjusting for local or global ancestry. A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry. In all figures, x-axis represented –log10 of expected P-values; y-axis represented –log10 of observed P-values.
Figure 4.6. Manhattan plots showed the association results of collateralization with single genetic variants. A) Common genetic variants from Affymetrix 6 chip when adjusting for global ancestry; B) common genetic variants from Affymetrix 6 chip when adjusting for local ancestry; C) rare genetic variants from Illumina Exome chip when adjusting for global ancestry; D) rare genetic variants from Illumina Exome chip when adjusting for local ancestry. In all figures, x-axis represented chromosome 1 to 22; y-axis represented $-\log_{10}$ of P-values.
When adjusting for local ancestry, the top associated single genetic variants were: chr4: 46039478, P-value = 6.33E-05, beta = -0.10; chr13: 36663302, P-value = 8.58E-05, beta = -0.12; chr19: 49376582, P-value = 1.45E-04, beta = 0.11 (Figure 4.5.D and 4.6.D).
Based on single genetic variants, gene-based test top association results were similar to when adjusting for global ancestry, however the ranking was slightly different. chr13: DCLK1 (P-Value = 2.91E-04), chr4: PIGG (P-value = 8.69E-04), and chr12: OR6C76 (P-value = 1.08E-03) were the top associated genes with collateral phenotype (Figure 4.7.D).

In gene-set enrichment analyses, totally 178 pathways were tested. Among tested pathways, no gene set was significant at FDR < 25%. 2 gene sets were significantly enriched at nominal P-value < 1% and 13 gene sets were significantly enriched at nominal P-value < 5%. Pathway analyses top results were: ascorbate and aldarate metabolism (size = 25, NOM P-value = 0.001, FDR q-value = 0.264); glycosylphosphatidylinositol GPI anchor biosynthesis (size = 24, NOM P-value = 0.010, FDR q-value = 0.339); propanoate metabolism (size = 32, NOM P-value = 0.014, FDR q-value = 0.518) (Figure 4.8.D).

4.4. Discussions

4.4.1. Sex differences in collateralization

Coronary collateralization is a natural neovascularization process that provides additional blood supply to bypass highly stenotic region. It is a protective mechanism against CAD. In our dataset, men have significantly higher rate of collaterals as compared to women. In the past two decades, more women than men die of CAD (Jacobs 2009, 69-78), which can be attributed to the fact that women undergo less aggressive treatment or are under-representative in clinical trial studies, but also can partially be due to have significantly lower rate of collaterals as compared to men. As collateralization is a protective mechanism after a cardiovascular event such as MI, men are more likely to have better outcomes following MI than women because of the additional collateral vessels circulate oxygen-enriched blood for heart muscle to recover and therefore prevent further damage.
Figure 4.7. **QQ plots showed gene-based test P-values distribution.** A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry. In all figures, x-axis represented –log10 of expected P-values; y-axis represented -log10 of the observed P-values.
Figure 4.8. Pathway analyses showed top enriched gene-sets that were associated with collateralization. A) Common genetic variants when adjusting for global ancestry; B) common genetic variants when adjusting for local ancestry; C) rare genetic variants when adjusting for global ancestry; D) rare genetic variants when adjusting for local ancestry.
4.4.2. Effect of smoking

In our data, smokers had higher rate of collaterals as compared to non-smokers, and smoking status showed significant association with collaterals after accounting for sex, diabetes and race-ethnicity differences. While smoking status was significantly associated when testing by itself, neither it, nor its interaction term was significant when including an interaction term between sex and smoking. Since sex and smoking were highly correlated, smoking could be confounded with sex and showed a difference in collateralization, due to more men smokers than women smokers in our data set. However, it is also possible that smoking status can truly affect collateralization through inducing hypoxia, which triggers collateralization, but the effect of smoking was not as strong as sex to show a significant association when including the interaction term. One study previously showed that smoking was positively associated with the presence of collateralization, while pack years of smoking was not related (Koerselman et al. 2007, 191-198). To test the true effect of smoking on collateralization, a study needed a large enough sample size if a weak effect was present.

4.4.3. Candidate genes within these two regions on chromosome 17

Within these two regions identified on chromosome 17, where African and Native American ancestries were associated with collateralization, I found several candidate genes from literatures. Region chr17:37760062-42630065(hg38) harbors several interesting genes where Native American ancestry was associated with collateralization in our data. For example, human junction plakoglobin (JUP) has a function of stimulating VE-cadherin in endothelial cells (Nottebaum et al. 2008, 2929-2945). Titin cap protein (TCAP) specifically expressed in heart and skeletal muscle. It is responsible for muscle assembly
regulation, and its mutation has been shown to affect cardiomyopathy (Hirtle-Lewis et al. 2013, 628-633). There are also genes that have been known to have significant roles in other biological functions, such as breast cancer 1 (BRCA1), a well-studied breast cancer gene (Rennert et al. 2007, 115-123) and signal transducer and activator of transcription 3 (STAT3), an intermediate components in signaling pathways for cytokines and growth factors such as IL-6 (Anglesio et al. 2011, 2538-2548) and VEGF (Niu et al. 2002, 2000-2008). In region chr17:34469647-35113147(hg38) where African ancestry was highly associated with collateralization, candidate gene is mitochondrial rRNA methyltransferase 1 (MRM1), which has been shown to be positively associated with blood pressure determination, cholesterol and LDL (Benjamin et al. 2007, S11).

### 4.4.4. Race-ethnicity differences in collateralization

Hispanics is an admixed population composed of different proportions of European, African and Native American ancestries. Our dataset showed that Hispanics had the highest rate of collaterals as compared to African Americans and Caucasians. Both Native American and African ancestry showed association with collateralization on chromosome 17, at two loci located close to each other. African ancestry showed association with collateralization with a slightly lower association signal than Native American ancestry. Since both Native American and African ancestry had a positive effect size on collateral, it is possible that these two ancestries contribute to a higher rate of collaterals in Hispanic populations. Hispanics had the highest rate of collateral may due to the highest proportion of Native American than Non-Hispanic Whites and African Americans. It was likely that due to the present of African Ancestry at this specific locus, African Americans had higher
rate of collateral than Non-Hispanic Whites, because Caucasians usually did not have either African or Native American at these loci.

4.4.5. Limitation of the study

Inflammation factors may also induce collateral growth, such as the pro-inflammatory agent lipopolysaccharide (LPS) increases the degree and speed of collateralization (Hoffman and others 2012). Therefore, some autoimmune diseases may potentially affect collateralization. Since the inflammation data were incomplete in our dataset, those patients who may have autoimmune and inflammation diseases could not be excluded. However, from those records that were available, the prevalence of inflammation/autoimmune disease in our dataset was very low (<2%), therefore, I assumed that without excluding autoimmune disease samples would not significantly impact the direction of our results.

Due to small sample size, the study had limited statistical power, which might explain that although our findings were interesting, they did not reach genome-wide significance after correcting for multiple testing.

4.5. Summary

In summary, I studied sex and race-ethnicity differences in collateralization. I showed that men had higher rate of collateralization than women, which may in part explained why women had poorer outcomes after MI than men. Different race-ethnicity groups had different rates of collateralization, which might contribute our knowledges to racial disparities in MI outcomes. Finally, I identified two regions on chromosome 17 that were likely to harbor genetic variations that influence collateralization. Clinically, it allowed us to use genetic tools to develop rejuvenated treatment for CAD and promote new opportunities for prevention of patients suffer from severe obstructive atherosclerosis.
5. ANGIOGRAPHICALLY DEFINED AND CLINICALLY DIAGNOSED CORONARY ARTERY DISEASE SHARE COMMON AND UNIQUE GENETIC VARIANTS.

5.1. Overview

5.1.1. Clinically diagnosed CAD

By definition, CAD is a complex disease with atherosclerosis being the major etiology, where cholesterol-containing plaques build up inside the coronary arteries, decrease blood flow to the heart, and eventually the decreased blood flow may cause clinical symptoms such as angina (chest pain) and shortness of breath. If a plaque ruptures, the content within the plaque was released to the blood stream, blocking the blood flow and causing a heart attack (Hansson 2005, 1685-1695). Atherosclerosis is a slow process that usually takes decades to develop, and that explains why most of the patients will not notice a symptom until late age, when they have a significant blockage or a heart attack.

Clinically, CAD is diagnosed based on angiography or functional testing (Felker, Shaw, and O'Connor 2002, 210-218), such as a patient’s medical and family histories, CAD risk factors and a series of physical exams such as electrocardiogram, stress testing, blood tests, coronary angiography and cardiac catheterization.

Previous genetic studies on CAD were mostly focused on clinically defined diagnosis that co-existed multiple subphenotypes (Samani et al. 2007, 443-453; Tregouet et al. 2009, 283-285), such as myocardial infarction; acute coronary syndrome, chronic stable angina or coronary stenosis (Nikpay et al. 2015, 1121-1130).
5.1.2. Angiographically defined CAD

Cardiac catheterization (angiogram) is an effective way to evaluate the inside of the coronary blood vessels using special X-rays angiograms. The doctor injects a special dye into the coronary arteries. Through X-ray images, the dye outlines narrow spots and blockages. If there is a blockage that requires treatment, the doctor can dilate the artery by pushing a balloon through the catheter and inflating it to improve the blood flow in the coronary arteries (Michaels and Chatterjee 2002, e187-90). When severe occlusion or a poor heart function present, patients usually need to undergo Coronary artery bypass graft (CABG) surgery and percutaneous coronary intervention (PCI) for treatment (Deb et al. 2013, 2086-2095) (Figure 5.1).

**Figure 5.1. Coronary catheterization is a process that evaluates inside a coronary artery.** Figure A: A catheter is inserted into an artery in the groin area and is guided through an X-ray into the artery. Figure B: Catheterization image shows the narrowing of a coronary artery through an X-ray examination. Figures retrieved from: heart.org
Base on angiography images, it can be observed that the heart has three major coronary arteries: left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA). The part between the aorta and the bifurcation is known as the left main (LM) coronary artery, which is a critical portion of the left coronary artery (Figure 5.2).

**Figure 5.2. Left main and 3 major branches of coronary arteries.** Figure retrieved from http://my.clevelandclinic.org

Left main disease is defined as over 50% stenosis in left coronary artery vessel. It is found in 4 to 6 percent of all patient who underwent coronary angiography (Ragosta et al. 2006, 357-362). Occlusion of this vessel will compromise blood flow to at least 75 percent of the left ventricle, therefore patients with left main disease are usually symptomatic and at high risk for cardiovascular events, unless the patient has collaterals or gone through coronary artery bypass graft to protect them from cardiovascular events (Eagle et al. 2004, e213-310; Conley et al. 1978, 947-952). It is uncommon for a left main stem lesion to be present by itself, and in over 70% of cases of left main stem disease additional coronary
artery disease is present, which increases the complexity of revascularization (Serruys and Garg 2009, 719-725). Depending on the number of vessels that were blocked, patients are diagnosed to have one, two or three-vessel disease, and three vessel disease is the most severe type among all three, which is defined as over 50% stenosis simultaneously present on all three major vessels.

Therefore, grouping patients with severe symptoms such as left main and three vessel disease seems a reasonable approach to define angiographic CAD cases.

### 5.1.3. Advantages of using angiographically defined CAD as compared to clinically defined CAD

Angiographically defined CAD is a clearer cut phenotype, it is more conservative and relatively less heterogeneous as compared to clinically defined CAD. Using this phenotype allowed us to identify genetic variants that targeted coronary artery stenosis, rather than a complex disease that contains multiple symptoms, such as angina, MI, which may or may not directly due to coronary artery stenosis. Even if it is due to coronary artery stenosis, the level of stenosis is hard to be quantified unless patients go thru catheterization procedures.

So far, very few genetics studies has been done on coronary artery stenosis rather than the conventional clinically defined CAD. A few studies focused on the candidate genes for clinical CAD rather than performing a genome-wide scanning coronary artery stenosis or restenosis (Bjornsson et al. 2015, 1526-1531; Pleva et al. 2015, 135-015-0128-8). In addition, most of these studies were performed on Caucasians (Bjornsson et al. 2015, 1526-1531; Pleva et al. 2015, 135-015-0128-8) rather than Hispanics - a rapidly growing and understudied population. To bridge the gap, a genome-wide scanning of genetic variants
was performed using angiographic stenosis as the phenotype to strictly define CAD, and compared our results of angiographically defined CAD with clinically defined CAD.

Since all of our patients were from catheterization lab, they more or less had some clinical symptoms of CAD. Therefore, patients were categorized as severe angiographically defined CAD vs disease free patients, in which severe angiographically defined CAD cases were patients with either left main or three vessel disease, and disease free patients had neither LM nor any major vessel disease, but may have mild stenosis.

5.2. Experimental approach

5.2.1. Phenotype definition

To evaluate coronary artery stenosis, the American Heart Association 16-segment model (Cerqueira et al. 2002, 539-542) was utilized to assess stenosis level for each patient (Figure 1.2).

5.2.1.1. Clinical phenotypes

The phenotype is angiographically-defined coronary artery disease (AD-CAD) stenosis, where the CT angiographic images were assessed by an experienced cardiologist. Initial phenotyping was utilizing ordinal measures of anatomic stenosis burden across the entire coronary artery vasculature. Points were assigned to represent the stenotic narrowing percentage of the most severe lesion within each of the 16 coronary artery segments designated from the American Heart Association (AHA) model. Zero to four points were given respectively for lesions in each segment as follows: 0% (no stenosis, 0), 1-25% (minimal, 1), 26-50% (mild, 2), 51-75% (moderate, 3) and 76-100% (severe, 4). An at least 50% luminal narrowing in any of the 16 segments of main coronary vessels (left main (LM), left anterior descending (LAD), left circumflex (LCX) and right coronary artery...
(RCA)) defined a significant angiographic obstruction (i.e., one vessel disease, 2 vessel disease, etc.).

5.2.1.2. Analysis phenotypes

Our primary analysis phenotype used a case/control analyses to summarize stenosis burden across the coronary vasculature. This indicated the severity of CAD, which was based on the stenosis level of all major branches (LM, LAD, LCX and RCA).

Cases were defined as patients with 50% stenosis in either left main (left main disease), or simultaneously presented in all three major vessels, LAD, LCX and RCA, which is also known as patients with three vessel disease (3VD).

Table 5.1. Study sample race-ethnicities identified by principal components.

<table>
<thead>
<tr>
<th>Race</th>
<th>Total Sample (N= 919)</th>
<th>Left main disease (LM) (N=162)</th>
<th>Three Vessel Disease (3VD) (N=356)</th>
<th>LM+3VD (N=435)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanics</td>
<td>54%</td>
<td>55%</td>
<td>57%</td>
<td>58%</td>
</tr>
<tr>
<td>Caucasians</td>
<td>26%</td>
<td>37%</td>
<td>31%</td>
<td>31%</td>
</tr>
<tr>
<td>African Americans</td>
<td>20%</td>
<td>8%</td>
<td>12%</td>
<td>11%</td>
</tr>
</tbody>
</table>
5.2.2. Identifying common and rare single genetic variants associated with angiographically defined CAD

As in the general methodology mentioned, common genetic variants were genotyped on Affymetrix 6 genotyping platform, and rare genetic variants were genotyped on Illumina Exome Chip genotyping platform. After SNP and sample QC, there were 162 individuals that had left main disease and 356 individuals who had 3VD. There were totally 435 patients who had either left main disease or three vessel disease (LM + 3VD), and these samples were set as cases; controls were patients from the same cohort, but without LM disease, 1VD, 2VD or 3VD. There were totally 484 individuals that were included as controls in the study (Table 5.1). As in the 484 controls, there were totally 124 individuals that have all 10 branches clear, and 232 individuals that had all 10 branches under 25 percent stenosis. Since the number of controls with clean stenosis was not enough for a proper control size, to find a balance between statistical power and biological meaning, I decided to set patients who had neither left main disease nor 1VD, 2VD, 3VD as controls (Table 5.2).

<table>
<thead>
<tr>
<th>Case/control</th>
<th>3VD (N = 356)</th>
<th>LM (N = 162)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Yes</td>
<td>No</td>
<td>262</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>Yes</td>
<td>27</td>
</tr>
<tr>
<td>Control</td>
<td>No</td>
<td>No</td>
<td>484</td>
</tr>
</tbody>
</table>

3VD: three vessel disease; LM: left main disease; NA: missing.

5.2.2.1. Statistical model

There were a few statistical models that could have been chosen to analyze angiographically defined CAD, such as logistic regression statistical model, which treated
the phenotype as a dichotomous trait, or ordinal logistic regression model with or without proportional odds, which treated the phenotype as an ordered and categorized multi-level trait based on the number of vessels that had occlusions. The reason of choosing logistic regression statistical model rather than ordinal logistic regression model was because this model best represented our data and answered the study question what were the genetic differences between clinically defined vs. angiographically defined CAD. Besides, this model was simple and straight forward, and most of the previous studies on clinically defined CAD were based on logistic regression models (I also did some preliminary studies with ordinal logistic regression model on the number of vessel disease. Results were not shown.).

Covariates that were significantly associated with the phenotype (P-value < 0.05) were selected. Furthermore, forward and backward selection were performed using R package leaps (Lumley 2009) to check for selected covariates. Covariates that passed the significant level as well as forward and backward selections were subsequently included in the multiple regression models. Both local and global ancestry adjustment were performed to control for population substructure.

When adjusting for **global** ancestry:

Angiographic CAD = \( \beta_0 + \beta_1 \text{Genotype} + \beta_2 \text{Age} + \beta_3 \text{Sex} + \beta_4 \text{Anti-cholesterol Medicine} + \beta_5 \text{Diabetes} + \beta_6 \text{Eigenvector 1} + \beta_7 \text{Eigenvector 2} + \epsilon \)

When adjusting for **local** ancestry:

Angiographic CAD = \( \beta_0' + \beta_1' \text{Genotype} + \beta_2' \text{Age} + \beta_3' \text{Sex} + \beta_4' \text{Anti-cholesterol Medicine} + \beta_5' \text{Diabetes} + \beta_6' \text{Local African Ancestry} + \beta_7' \text{Local Native American Ancestry} + \epsilon' \)
5.2.3. Gene-based test

To identify genes that were associated with angiographically defined CAD, gene-based tests were performed. As mentioned in the previous chapter, gene-based test were analyzed using versatile gene-based association study (VEGAS) (Liu et al. 2010, 139-145).

5.2.4. Pathway analyses

As mentioned in previous chapter methodology, pathway analyses were performed by gene-set enrichment analyses (GSEA) (Wang, Li, and Bucan 2007) using the same parameters.

5.2.5. Compare genetic components between clinical and angiographically defined CAD

Angiographically defined CAD results were compared with the most recent publication “A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease” (Nikpay et al. 2015, 1121-1130), which currently identified 65 SNPs that were significantly associated with clinically defined CAD. There were 58 genetic loci that had been identified to be significantly associated with clinically defined CAD. I identified these loci in angiographically defined CAD genetic association results, and reported those that were significant in both angiographically and clinically defined CAD.

5.3. Results

5.3.1. Single genetic variants that were associated with angiographically defined CAD

For common genetic variants, totally 804,235 SNPs were tested in the genetic association studies. I adjusted both global and local ancestry to control for population substructures, in addition to other covariates. However, none of our single genetic variants passed genome-
wide significant threshold. As in Affymetrix 6 genotyping platform, when adjusting for global ancestry, the top association results were: chr2: rs12996713, P-value = 1.23E-06; chr2: rs12991269, P-value = 1.38E-06; chr1: rs4142453, P-value = 2.14E-06 (Figure 5.3.A and 5.4.A).

When adjusting for local ancestry, the top association results were: chr2: rs12996713, beta = 0.11, P-value = 3.02E-07; chr2: rs12991269, beta = 0.11, P-value = 5.10E-07; chr3: rs13096519, beta = -0.12, P-value = 8.17E-07 (Figure 5.3.B and 5.4.B).

For exome chip genotyping platform, which contains mostly rare genetic variants, there were totally 132,158 SNPs included in the analyses. No genetic variants passed genome-wide significant threshold. When adjusting for global ancestry, the top association results were: chr2: rs4662834, P-value = 2.05E-06; chr9: rs55727303, P-value = 8.99E-06; chr5: rs7708070, P-value = 6.31E-05 (Figure 5.3.C and 5.4.C).

When adjusting for local ancestry, the top association results were: chr2: rs4662834, beta = 0.11, P-value = 5.98E-07; chr5: rs7708070, beta = 0.10, P-value = 4.44E-05; chr10: rs4751185, beta = -0.13, P-value = 1.03E-04 (Figure 5.3.D and 5.4.D).
Figure 5.3. QQ Plots for single genetic variants test of Angiographically-defined CAD phenotype. A) Affymetrix 6 genotyping platform when adjusting for global ancestry. B) Affymetrix 6 genotyping platform when adjusting for local ancestry. C) Illumina exome chip genotyping platform when adjusting for global ancestry. D) Illumina exome chip genotyping platform when adjusting for local ancestry. In all figures, x-axis indicated –log10 of expected P-values; y-axis represented –log10 of observed P-values.
Figure 5.4. Manhattan plots for single genetic variants test of angiographically-defined CAD. A) Affymetrix 6 genotyping platform when adjusting for global ancestry. B) Affymetrix 6 genotyping platform when adjusting for local ancestry. C) Illumina exome chip genotyping platform when adjusting for global ancestry. D) Illumina exome chip genotyping platform when adjusting for local ancestry. In all figures, x-axis represented chromosome 1 to 22; y-axis represented –log10 of observed P-values.
5.3.2. Genes that are associated with angiographically defined CAD

Gene-based tests were performed to identify genes that were associated with angiographically defined CAD. As in the Affymetrix 6 genotyping chip, there were 17947 genes tested. When adjusting for global ancestry, the top associated genes were: chr16: MC1R, P-value = 6.51E-05; chr16: TCF25, P-value = 7.00E-05; chr8: MTDH, P-value = 9.72E-05 (Figure 5.5.A).

Figure 5.5. QQ plots showed the P-value distribution for gene-based test of angiographically-defined CAD phenotype. A) Affymetrix 6 genotyping chip results when adjusting for global ancestry; B) Affymetrix 6 genotyping chip results when adjusting for local ancestry; C) Illumina exome chip results when adjusting for global ancestry; D) Illumina exome chip results when adjusting for local ancestry. In all figures, x-axis represented –log10 expected P-values, and y-axis represented –log10 of observed P-values.
When adjusting for local ancestry, the top associated genes were: chr8: MTDH, P-value = 2.37E-05; chr16: MC1R, P-value = 1.44E-04; chr16: TCF25, P-value = 1.55E-04. Totally 17945 genes were tested (Figure 5.5.B).

In Illumina exome chip, there were 17855 genes tested. When adjusting for global ancestry, the top associated genes were: chr16: ZNF720, P-value = 2.37E-04; chr2: INSIG2, P-value = 1.78E-03; chr1: FAM78B, P-value = 4.92E-03 (Figure 5.5.C).

Among 17855 tested genes, when adjusting for local ancestry, the top associated genes were: chr16: ZNF720, P-value = 5.8E-04; chr1: C1orf95, P-value = 8.30E-04; chr3: IL12A, P-value = 9.29E-04 (Figure 5.5.D).

5.3.3. Biological pathways that were associated with angiographically-defined CAD

Gene-set enrichment analyses identified genetic pathways that were associated with angiographically defined CAD. As in Affymetrix 6 genotyping platform, when adjusting global ancestry, totally 178 pathways were tested. Among tested pathways, 1 gene set was significant at FDR < 25%. 2 gene sets were significantly enriched at nominal P-value < 1% and 11 gene sets were significantly enriched at nominal P-value < 5%. The top associated pathways were: glycosaminoglycan degradation (size = 20, NOM P-value = 0.005, FDR q-value = 0.162); glyoxylate and dicarboxylate metabolism (size = 16, NOM P-value = 0.034, FDR q-value = 0.710); renin angiotensin system (size = 15, NOM P-value = 0.070, FDR q-value = 0.837) (Figure 5.6.A).

When adjusting for local ancestry, there were 178 pathways tested. Among them, no gene set was significant at FDR < 25% level; 1 gene set was significantly enriched at nominal P-value < 1%; and 6 gene sets were significantly enriched at nominal P-value < 5%. The top association results were: renin angiotensin system (size =15, NOM p-value =
0.068, FDR q-value = 1.0); ascorbate and aldarate metabolism (size =25, NOM P-value = 0.053, FDR q-value = 1.0); ECM receptor interaction (size =83, NOM P-value = 0.012, FDR q-value = 1.0) (Figure 5.6.B).

As in Illumina exome chip genotyping platform, when adjusting for global ancestry, totally 178 pathways were tested. Among these gene sets, no gene set was significant at FDR <25%; 5 gene sets were significantly enriched at nominal P-value <1%; and 13 gene sets were significantly enriched at nominal P-value <5%. The top association results were: allograft rejection (size = 34, NOM P-value = 0.031, FDR q-value = 1.0); pathogenic escherichia coli infection (size =50, NOM P-value = 0.027, FDR q-value = 1.0); type I diabetes mellitus (size = 41, NOM P-value = 0.037, FDR q-value = 1.0) (Figure 5.6.C).

When adjusting for local ancestry, there were also 178 pathways included in the analyses. No gene set was significant at FDR <25% or at nominal P-value <1%. 3 gene sets were significantly enriched at nominal P-value <5%. The top association results were: riboflavin metabolism (size = 15, NOM P-value = 0.099, FDR q-value = 1.0); arachidonic acid metabolism (size = 58, NOM P-value = 0.031, FDR q-value = 1.0); nicotinate and nicotinamide metabolism (size = 24, NOM P-value = 0.090, FDR q-value = 1.0) (Figure 5.6.D).
Figure 5.6. Pathway analyses showed the top associated pathways with angiographic CAD. Figures showed the top three enriched pathways when genetic variants were genotyped by A) Affymetrix chip adjusting for global ancestry; B) Affymetrix chip adjusting for local ancestry; C) Illumina exome chip adjusting for global ancestry; D) Illumina exome chip adjusting for local ancestry.
5.3.4. Comparing angiographically defined CAD genetic components with previously published clinical CAD genetic components

Currently, 65 SNPs had been identified to be significantly associated with clinically defined CAD (Nikpay et al. 2015, 1121-1130). 6 SNPs among them were also significantly associated with angiographically defined CAD. SNPs that were significantly associated with both angiographically defined CAD vs. clinically defined CAD are listed below (Table 5.3).

There were 48 genetic loci published prior to their paper, and they reported 10 novel loci. Therefore, totally 58 loci had been published for clinically defined CAD, and among these 58 loci, 7 genes were also significantly associated with angiographically defined CAD (Table 5.4).

As seen in the table, among common variants, tribbles pseudokinase 1 (TRIB1) (P-value = 4.14E-02, global adjusted; 4.89E-02, local adjusted), ADAMTS7 (P-value = 9.63E-03, global adjusted; 8.23E-03, local adjusted) and MC4R (P-value = 1.09E-03, global adjusted; 1.09E-03, local adjusted) were significant in our gene-based test (Table 5.4.A).

Among rare variants, ABCG8 (P-value = 0.78, global adjusted; 0.045, local adjusted), ZNF259 (P-value = 0.076, global adjusted; 0.047, local adjusted), ATP2B1 (P-value = 0.042, global adjusted; 0.040, local adjusted), ADAMTS7 (P-value = 0.122, global adjusted; 0.045, local adjusted), APOC1 (P-value = 0.20, global adjusted; 0.049, local adjusted) and MC4R (P-value = 6.13E-03, global adjusted; 0.003, local adjusted) were statistically significantly associated with angiographically defined CAD (Table 5.4.B).
Table 5.3. Comparing previously published 65 SNPs that showed genome-wide significant association with clinically defined CAD to angiographically defined CAD in MCR dataset. Table shows SNPs that were significant in both clinically and angiographically defined CAD.

<table>
<thead>
<tr>
<th>Published SNP</th>
<th>Known locus</th>
<th>Effect/Non-effect allele</th>
<th>OR (95% CI)</th>
<th>P-value</th>
<th>MCR Affy6 - Global Adjusted OR</th>
<th>MCR Affy6 - Local Adjusted P-value</th>
<th>MCR Affy6 - Local Adjusted OR</th>
<th>MCR Affy6 - Local Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2252641</td>
<td>ZEB2-AC074093.1</td>
<td>C/T</td>
<td>1.03 (1.01, 1.05)</td>
<td>5.16E-04</td>
<td>1.27</td>
<td>0.03827</td>
<td>1.05</td>
<td>0.03161</td>
</tr>
<tr>
<td>rs2047009</td>
<td>CXCL12</td>
<td>G/T</td>
<td>1.06 (1.04, 1.08)</td>
<td>2.75E-11</td>
<td>1.39</td>
<td>0.00463</td>
<td>1.05</td>
<td>0.02447</td>
</tr>
<tr>
<td>rs964184</td>
<td>ZNF259-APOA5-APOA1</td>
<td>G/C</td>
<td>1.05 (1.03, 1.08)</td>
<td>5.60E-05</td>
<td>1.47</td>
<td>0.00541</td>
<td>1.08</td>
<td>0.00564</td>
</tr>
<tr>
<td>rs9515203</td>
<td>COL4A1/A2</td>
<td>T/C</td>
<td>1.07 (1.05, 1.10)</td>
<td>9.33E-10</td>
<td>1.31</td>
<td>0.02859</td>
<td>1.05</td>
<td>0.03848</td>
</tr>
<tr>
<td>rs4468572</td>
<td>ADAMTS7</td>
<td>C/T</td>
<td>1.08 (1.06, 1.10)</td>
<td>4.44E-16</td>
<td>1.27</td>
<td>0.02336</td>
<td>1.05</td>
<td>0.02764</td>
</tr>
<tr>
<td>rs4420638</td>
<td>APOE-APOC1</td>
<td>G/A</td>
<td>1.10 (1.07, 1.13)</td>
<td>7.07E-11</td>
<td>1.41</td>
<td>0.01765</td>
<td>1.06</td>
<td>0.02850</td>
</tr>
</tbody>
</table>
Table 5.4. Comparing 58 previously published genome-wide significant loci from meta-analysis additive association with clinically defined CAD results with angiographically defined CAD gene-based test on MCR dataset Affymetrix 6 chip/Exome Chip results. Tables show genetic loci that were significant in both clinical and angiographic CAD. A) Affymetrix 6 chip data. B) Exome chip data.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Known locus</th>
<th>nSNPs</th>
<th>MCR Affy6 - Global Adjusted</th>
<th>MCR Affy6 - Local Adjusted</th>
<th>MCR ExomeChip - Global Adjusted</th>
<th>MCR ExomeChip - Local Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIB1</td>
<td>TRIB1</td>
<td>27</td>
<td>0.04142 rs2980862 4.57E-03</td>
<td>0.04894 rs2980862 5.48E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAMTS7</td>
<td>ADAMTS7</td>
<td>25</td>
<td>0.00963 rs1809420 4.27E-03</td>
<td>0.00823 rs1809420 2.93E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC4R</td>
<td>PMAIP1-MC4R</td>
<td>16</td>
<td>0.00109 rs17066881 8.86E-05</td>
<td>0.00109 rs17066881 8.86E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABCG8</td>
<td>ABCG5-ABCG8</td>
<td>44</td>
<td>0.779 rs9282574 2.52E-02</td>
<td>0.04469 rs113974315 9.56E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZNF259</td>
<td>ZNF259-APOA4-APOA1</td>
<td>39</td>
<td>0.07616 rs145410701 8.62E-03</td>
<td>0.04731 rs145410701 6.06E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP2B1</td>
<td>ATP2B1</td>
<td>4</td>
<td>0.04229 rs2681472 2.06E-02</td>
<td>0.03975 rs2681472 2.10E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAMTS7</td>
<td>ADAMTS7</td>
<td>11</td>
<td>0.122 rs1994016 1.16E-02</td>
<td>0.04471 rs1994016 8.71E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOC1</td>
<td>APOE-APOC1</td>
<td>27</td>
<td>0.198 rs769449 1.02E-02</td>
<td>0.04905 rs769449 9.84E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC4R</td>
<td>PMAIP1-MC4R</td>
<td>4</td>
<td>0.00613 rs11878133 1.43E-04</td>
<td>0.003 rs11878133 1.43E-04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TRIB1 belongs to the protein kinase superfamily. It interacts with MAPK kinases and regulates activation of MAP kinases (resource: GeneCard).

ADAM metallopeptidase with thrombospondin type 1 motif 7 (ADAMTS7), this gene is overexpressed in heart tissue, specifically atrial appendage. The protein encoded by this gene is a member of the ADAMTS (a disintegrin and metalloproteinase with thrombospondin motifs) family. Among its related pathways are transport to the Golgi and subsequent modification and O-linked glycosylation. Go annotation related to this gene include peptidase activity and metallopeptidase activity, which may play a role in the degradation of COMP (resource: GeneCards).

Melanocortin 4 Receptor (MC4R) gene encodes a protein that is a membrane-bound receptor and member of the melanocortin receptor family. The encoded protein interacts with adrenocorticotropic and melanocyte stimulating hormone (MSH) hormones and is mediated by G proteins. This gene is intronless, and defects in this gene are a cause of autosomal dominant obesity (resource: GeneCards).

ATP-Binding Cassette, Sub-Family G (WHITE), member 8 (ABCG8) encodes a protein that is a member of the ATP-binding cassette (ABC) transporter superfamily. These transporters play an indispensable role in the selective transport of the dietary cholesterol in and out of the enterocytes and in the selective sterol excretion by the liver into the bile (resource: GeneCards).

Zinc Finger Protein 259 (ZNF259) encodes a protein that interacts with survival neuron protein (SMN1) to enhance pre-mRNA splicing and to induce neuronal differentiation and axonal growth. Defects in this gene or the SMN1 gene can cause spinal muscular atrophy.
Plasma Membrane Calcium-Transporting ATPase 1 (ATP2B1) coded protein belongs to the family of P-type primary ion transport ATPases characterized by the formation of an aspartyl phosphate intermediate during the reaction cycle. These enzyme remove calcium ions against very large concentration gradient and therefore play a critical role in intracellular calcium homeostasis. Diseases associated with ATP2B1 gene include hypertension (resource: GeneCards).

Apolipoprotein C-I (APOC1) encodes a protein of the apolipoprotein C1 family. This gene product regulates lipoprotein activities by inhibiting lipoprotein binding to the low density lipoprotein (LDL) receptor, LDL receptor-related protein, and very low density lipoprotein (VLDL) receptor. It binds free fatty acids and reduces their intracellular esterification (resource: GeneCards).

5.4. Discussion

5.4.1. Biological functions of identified genetic variants for the phenotype

I identified genetic risk factors for angiographically defined coronary artery disease. Although no genetic variants reached genome-wide significance, some of their functions were cardiovascular disease related and worth further investigation of their impacts on coronary artery disease.

5.4.1.1. Single genetic variants

For Affymetrix 6 genotyping platform, top associated SNPs were identified when adjusting either local or global ancestry. rs4142453 located near ALG6 gene. This gene encodes a member of the ALG6/ALG8 glucosyltransferase family. The encoded protein catalyzes the addition of the first glucose residue to the growing lipid-linked oligosaccharide precursor of N-linked glycosylation. Mutations in this gene are associated with congenital disorders.
of glycosylation type Ic (resource: RefSeq, Jul 2008). SNPs rs12996713, rs12991269 and rs13096519 do not locate near any gene.

For Illumina exome chip platform, rare genetic variants were located nearby or within genes. SNP rs7708070 located near gene ZNF608, which is the zinc finger protein 608. This gene has been found to be associated with familial early onset Type II diabetes and obesity in Hong Kong Chinese (Ma et al. 2014, e84770). SNPs rs4662834, rs55727303 do not locate near any gene; rs4751185 located near gene LOC101927435 with unknown function.

5.4.1.2. Gene-based test identified genes

Transcription Factor 25 (TCF25), also known as NULP1, is important in embryonic development (Steen and Lindholm 2008, 432-437), and may play a role in cell death control. It acts as a transcriptional repressor, and has been shown to repress transcription of serum response factor (SRF) in vitro and so may play a role in heart development (Cai et al. 2006, 973-981).

Insulin Induced Gene 2 (INSIG2): several studies has shown this gene is associated with coronary artery disease (Liu et al. 2010, 421-426; Liu et al. 2008, 886-892; McGeachie et al. 2009, 2448-2454). It mediates feedback control of cholesterol synthesis. It functions by blocking the processing of sterol regulatory element-binding proteins (SREBPs) and is capable of retaining the cleavage-activating protein and sterol regulatory element-binding protein 2 (SCAP-SREBF2) complex in the endoplasmic reticulum thus preventing it from escorting SREBPs to the Golgi.

Interleukin 12A (IL12A) Cytokine can act as a growth factor for activated T and NK cells, enhance the lytic activity of NK/lymphokine-activated Killer cells, and stimulate the
production of interferon-gamma (IFN-γ). Studies had shown IL12 protects against coxsackievirus B3-induced myocarditis by increasing IFN-gamma and macrophage and neutrophil in the heart (Fairweather et al. 2005, 261-269).

Metadherin (MTDH) downregulates SLC1A2/EAAT2 promoter activity when expressed ectopically. It activates the nuclear factor kappa-B (NF-kappa-B) transcription factor. It promotes anchorage-independent growth of immortalized melanocytes and astrocytes which is a key components in tumor cell expansion. It also promotes lung metastasis and has an effect on bone and brain metastasis, possibly by enhancing the seeding of tumor cells to the target organ endothelium. This gene product also induces chemoresistance. Knocking down this gene significantly reduces the adhesion of cancer cells to lung microvascular endothelial cells and the reciprocal effect is observed following overexpression (resource: GeneCards.org).

Melanocortin 1 Receptor (MC1R) is a binding receptor of members of the melanocortin family, and transducing signaling pathways into physiological functions, including pigmentation and inflammation (resources: GeneCards).

Zinc Finger Protein 720 (ZNF720) involves in DNA binding and may play a role in transcriptional regulation. Family with Sequence Similarity 78, Member B (FAM78B) and Chromosome 1 Open Reading Frame 95 (C1orf95) functions are understudied.

5.4.1.3. Functions of biological pathways identified by gene-set enrichment analyses.

Renin angiotensin system is a hormone system that regulates blood pressure and fluid balance. If the renin–angiotensin–aldosterone system is abnormally active, it will cause hypertension (Solomon and Anavekar 2005).
Nicotinate and nicotinamide metabolism is a biological pathway metabolize cofactors and vitamins. Nicotinate is also known as vitamin B3 and nicotinic acid. Pharmaceutical and supplemental nicotinate are primarily used to treat high cholesterol and vitamin B3 deficiency. Vitamin B3 has NOT been found to be useful in decreasing the risk of cardiovascular disease in those already on a statin (Keene et al. 2014, g4379) but appears to be effective in those without statin (Bruckert, Labreuche, and Amarenco 2010, 353-361). Nicotinamide is the amide of nicotinic acid.

Type I diabetes mellitus results from the autoimmune destruction of the insulin-producing beta cells in the pancreas. The subsequent lack of insulin leads to increased glucose in blood and urine. The classical symptoms are frequent urination, increased thirst, increased hunger, and weight loss (Cooke and Plotnick 2008, 374-84; quiz 385). Serious long-term complications related to high blood sugar include heart disease and stroke (WHO March 2016).

Glycosaminoglycan degradation: Glycosaminoglycans (GAGs) are long unbranched polysaccharides consisting of a repeating disaccharide unit. Glycosaminoglycans are highly polar and attract water. They are therefore useful to the body as a lubricant or as a shock absorber (Esko, Kimata, and Lindahl 2009).

Arachidonic acid metabolism is a lipid metabolism pathway. Arachidonic acid is a polyunsaturated fatty acid present in the phospholipids of membranes of the body’s cells, and is abundant in the brain, muscles and liver. It can act in cellular signaling as a lipid second messenger involved in the regulation of signaling enzymes, and it also can act as a key inflammatory intermediate and a vasodilator (Baynes and Marek 2005).
Glyoxylate and dicarboxylate metabolism is related to carbohydrate metabolism. Glyoxylate is the conjugate base of glyoxylic acid; dicarboxylates are the conjugate bases of dicarboxylic acids (KEGG.com).

Ascorbate and aldarate metabolism is another carbohydrate metabolism pathway. Allograft rejection is the consequence of the recipient's alloimmune response to non-self antigens expressed by donor tissues (KEGG.org).

ECM receptor interaction: The extracellular matrix (ECM) consists of a complex mixture of structural and functional macromolecules and serves an important role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Specific interactions between cells and the ECM are mediated by transmembrane molecules, mainly integrins and perhaps also proteoglycans, CD36, or other cell-surface-associated components. These interactions lead to a direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis (KEGG.org).

Pathogenic Escherichia coli infection: Enteropathogenic E. coli (EPEC) and enterohemorrhagic E. coli (EHEC) are closely related pathogenic strains of Escherichia coli. The hallmark of EPEC/EHEC infections is induction of attaching and effacing (A/E) lesions that damage intestinal epithelial cells (KEGG.org).

Riboflavin metabolism: riboflavin is part of the vitamin B group. Riboflavin functions as a coenzyme, and is required for enzymes (proteins) to perform normal physiological actions. The active forms of riboflavin flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD) function as cofactors for diverse flavoprotein enzyme reactions (KEGG.org).
5.5. Summary

In summary, by studying angiographically defined CAD, I identified genetic components that would affect the severity of CAD in a genetically diverse population. Because our controls were from catheterization lab, they were not completely free of CAD symptoms. Therefore, these identified genetic variants indicated association with the severe form of anatomic CAD rather than the presence/absence of CAD.
CHAPTER 6. CONCLUSIONS

In my research, I identified genetic components for angiographic stenosis of CAD in a genetically diverse population. Studying angiographically defined CAD allowed us to clearly define our phenotype in the complex CAD, therefore targeting genetic components that were more specific to the angiographic coronary stenosis. I also identified genetic components and contributing factors for collateralization, a protective mechanism towards CAD in this population. This might help us explain why some atherosclerosis patients had worse clinical outcomes than others, and whether the race-ethnicity differences in CAD prevalence could be partially attributable to collateralization. Clinically, it allowed us to use genetic tools to develop rejuvenated treatment for CAD and promote new opportunities for prevention of patients suffer from severe obstructive atherosclerosis. Last, studying this genetically diverse population allowed us to eliminate health disparities and therefore benefited populations as a whole.
REFERENCES


WHO. March 2016. "Diabetes Fact Sheet N°312.".

