

https://dspace.org/

Genome-wide association study of carotid intima-media thickness (cIMT) in sub-Saharan African populations: An AWI-Gen study

Boua, Palwende Romuald

2020
https://hdl.handle.net/10539/30414

Downloaded from a DSpace repository. https://dspace.org/

Genome-wide association study of carotid intima-media thickness (cIMT) in sub-Saharan African populations: An AWI-Gen study

Palwende Romuald Boua



A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of PhD in Human Genetics

Johannesburg, January 2020

Declaration

I, Palwende Romuald Boua (Students number: 1530955), am a student registered for the degree of Doctor of Philosophy in the academic year 2016.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the thesis has not been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signed this on the 13th day of January 2020 in Johannesburg

Signature: 

To the memory of my beloved mother Seraphine Gana-Zoure Ouena (1955-21 Oct 1991).

To the memory of my beloved father Tanga Narcisse Boua (1949-16 Jan 2017), thank you for your love, education and sacrifice.

To my daughter Yiennongya Marie Yahdiel Raviya and my family

Publications and conference presentations from this research project

- ***Publications***

1. **Boua PR**, Nonterah EA, Choudhury A., Asiki G, Micklesfield LK, Agongo G, Ali SA, Mashinya F, Sorgho H, Nakanabo-Diallo S, Kyobutungi C, Alberts M, Gómez-Olivé FX, Mohamed S. F., Norris S, Tollman S, Tinto H, Soo CC, Mukomana F, Hazelhurst S, Wade AN, Oduro AR, Ramsay M, Crowther NJ; as members and collaborators of AWI-Gen and the H3Africa Consortium. **Regional differences in carotid intima-media thickness reveal higher risk for atherosclerosis in West Africa: An AWI-Gen Study.** (Advanced draft)

2. **Boua PR**, Brandenburg J-T, Choudhury A, Hazelhurst S, Sengupta D, Agongo G, Nontera EA, Oduro AR, Tinto H, Mathew CG, Sorgho H and Ramsay M as members of AWI-Gen and the H3Africa Consortium (2019). ***Novel and known gene-smoking interactions with cIMT identified as potential drivers for atherosclerosis risk in West-African populations of the AWI-Gen Study***. ***Frontiers In Genetics*** (*in press*) . doi: 10.3389/fgene.2019.01354

- ***Conference presentations***

- **Oral presentation**

11th Meeting of the H3Africa Consortium, Uganda (March 2017): "Managing Epidemiological and Genetic Diversity in African GWAS Studies: cIMT in the AWI-Gen Study". **Romuald Palwende Boua**; Ananyo Choudhury, Hermann Sorgho, Halidou Tinto, Christopher Mathew; Michèle Ramsay as members of AWI-Gen and the H3Africa Consortium

- **Poster presentation**

10th Conference of the African Society of Human Genetics, Cairo (November 2017): "Managing Epidemiological and Genetic Diversity in African GWAS Studies". (Poster Prize) **Romuald Palwende Boua**; Ananyo Choudhury, Hermann Sorgho, Halidou Tinto, Christopher Mathew; Michèle Ramsay as members of AWI-Gen and the H3Africa Consortium.

13th Meeting of the H3Africa Consortium, Tunis (April 2019): “Novel and known gene-smoking interactions identified as key drivers for atherosclerosis with cIMT risk in West-African populations of the AWI-Gen Study”. Romuald Palwende Boua, Jean-Tristan Brandenburg, Ananyo Choudhury, Hermann Sorgho, Scott Hazelhurst, Halidou Tinto, Christopher G. Mathew Michèle Ramsay as members of AWI-Gen and the H3Africa Consortium.

69th American Society of Human Genetics (ASHG) annual meeting (October 2019): “African-specific variants reveal novel susceptibility genes for carotid intima-media thickness in Sub-Saharan African populations with sex-differences: An AWI-Gen study”. Palwende Romuald Boua, Ananyo Choudhury, Hermann Sorgho, Jean-Tristan Brandenburg, Scott Hazelhurst, Dhriti Sengupta, Halidou Tinto, Christopher G. Mathew, Michèle Ramsay as members of AWI-Gen and the H3Africa Consortium.

Co-authored publications during the PhD research period

1. George JA, Brandenburg J-T, Fabian J, Crowther NJ, Agongo G, Alberts M, Ali S, Asiki G, **Boua PR**, Gómez-Olivé FX, Mashinya F, Micklesfield L, Mohamed SF, Mukomana F, Norris SA, Oduro AR, Soo C, Sorgho H, Wade A, Naicker S, Ramsay M; AWI-Gen and the H3Africa Consortium. Kidney damage and associated risk factors in rural and urban sub-Saharan Africa (AWI-Gen): a cross-sectional population study. ***Lancet Global Health***. 2019 Dec;7(12):e1632-e1643. doi: 10.1016/S2214-109X(19)30443-7**.**

2. Nonterah EA, **Boua PR**, Klipstein-Grobusch K, Asiki G, Micklesfield LK, Agongo G, Ali SA, Mashinya F, Sorgho H, Nakanabo-Diallo S, Debpuur C, Kyobutungi C, Alberts M, Norris S, Tollman S, Tinto H, Soo CC, Mukomana F, Hazelhurst S, Wade AN, Kahn K, Oduro AR, Grobbee DE, Sankoh O, Ramsay M, Bots ML, Crowther NJ; as members and collaborators of AWI-Gen and the H3Africa Consortium. Classical cardiovascular risk factors and HIV are associated with carotid intima-media thickness in adults from Sub-Saharan Africa: findings from H3Africa AWI-Gen study. ***Journal of the American Heart Association***. **2019 Jul 16;8(14):e011506**. doi: 10.1161/JAHA.118.011506

3. Ali SA, Soo C, Agongo G, Alberts M, Amenga-Etego L, **Boua RP**, Choudhury A, Crowther NJ, Depuur C, Gómez-Olivé FX, Guiraud I, Haregu TN, Hazelhurst S, Kahn K, Khayeka-Wandabwa C, Kyobutungi C, Lombard Z, Mashinya F, Micklesfield L, Mohamed SF, Mukomana F, Nakanabo-Diallo S, Natama HM, Ngomi N, Nonterah EA, Norris SA, Oduro AR, Somé AM, Sorgho H, Tindana P, Tinto H, Tollman S, Twine R, Wade A, Sankoh O, Ramsay M. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. ***Global Health Action***. **2018**;**11(sup2):1507133**. doi: 10.1080/16549716.2018.1507133.

4. **Boua RP**, Sorgho H, Rouamba T, Nakanabo Diallo S, Bognini JD, Konkobo SZ, Valia D, Lingani M, Ouoba S, Tougma AS, Bihoun B, Crowther NJ, Norris SA, Ramsay M, Tinto H; as members of AWI-Gen and the H3Africa Consortium. Gender differences in sociodemographic and behavioural factors associated with BMI in an adult population

in rural Burkina Faso - an AWI-Gen sub-study. ***Global Health Action***. **2018;11(sup2):1527557**. doi: 10.1080/16549716.2018.1527557.

5. Ramsay M, Crowther NJ, Agongo G, Ali SA, Asiki G, **Boua RP**, Gómez-Olivé FX, Kahn K, Khayeka-Wandabwa C, Mashinya F, Micklesfield L, Mukomana F, Nonterah EA, Soo C, Sorgho H, Wade AN, Wagner RG, Alberts M, Hazelhurst S, Kyobutungi C, Norris SA, Oduro AR, Sankoh O, Tinto H, Tollman S & as members of AWI-Gen and the H3Africa Consortium. Regional and sex-specific variation in BMI distribution in four sub-Saharan African countries: The H3Africa AWI-Gen study. ***Global Health Action***. **2018;11(sup2):1556561**. doi: 10.1080/16549716.2018.1556561

6. Gómez-Olivé FX, Ali SA, Made F, Kyobutungi C, Nonterah E, Micklesfield L, Alberts M, **Boua R**, Hazelhurst S, Debpuur C, Mashinya F, Dikotope S, Sorgho H, Cook I, Muthuri S, Soo C, Mukomana F, Agongo G, Wandabwa C, Afolabi S, Oduro A, Tinto H, Wagner RG, Haregu T, Wade A, Kahn K, Norris SA, Crowther NJ, Tollman S, Sankoh O, Ramsay M; as members of AWI-Gen and the H3Africa Consortium. Regional and sex differences in the prevalence and awareness of hypertension: An H3Africa AWI-Gen study across 6 sites in Sub-Saharan Africa. ***Global Heart***. **2017 Jun;12(2):81-90**. doi: 10.1016/j.gheart.2017.01.007

7. Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, Gómez-Olivé X, Jaff N, Sorgho H, Wagner R, Khayeka-Wandabwa C, Choudhury A, Hazelhurst S, Kahn K, Lombard Z, Mukomana F, Soo C, Soodyall H, Wade A, Afolabi S, Agorinya I, Amenga-EtegoL , Ali SA, Bognini JD, **Boua RP**, Debpuur C, Diallo S, Fato E, Kazienga A, Konkobo SZ, Kouraogo PM, Mashinya F, Micklesfield L, Nakanabo-Diallo S, Njamwea B, Nonterah E, Ouedraogo S, Pillay V, Somande AM, Tindana P, Twine R, Alberts M, Kyobutungi C, Norris SA, Oduro AR, Tinto H, Tollman S, Sankoh O. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four Sub-Saharan African countries. ***Global Health, Epidemiology and Genomics*** **(2016), 1, e20, page 1 of 13**. doi:10.1017/gheg.2016.17

Abstract

It is estimated that by 2020 cardiovascular diseases (CVDs) will be the leading cause of death. The demographic and health transition in sub-Saharan Africa (SSA) has shifted the major causes of mortality from communicable and nutritional diseases to non-communicable diseases (NCDs). Atherosclerosis is a complex, progressive disorder affecting large and medium-sized arteries, preceding the development of CVDs such as ischemic heart disease, stroke, heart failure, peripheral arterial disease and rheumatic heart disease. Despite the epidemiological complexity of CVDs in SSA, the genetic susceptibility of the populations has remained understudied. Genome-wide association studies (GWAS) have revolutionized the field of complex disease genetics over the past decade, and provided insights into the genetic architecture of disease susceptibility and to advances in clinical care and personalized medicine. The evolutionary roots and genetic diversity of the continent represent an opportunity to shed new light on the genetic architecture of complex traits such as atherosclerosis.

This thesis used the AWI-Gen (Africa Wits-INDEPTH Partnership for Genomic Studies) data to investigate whether genetic variation in SSA populations is associated with carotid intima-media thickness (cIMT) (a marker for atherosclerosis) in middle-aged adults (40-60 years of age) and to examine interactions with key environmental factors.

First, cIMT distributions in four SSA countries were characterised (**Chapter 2**). West African populations had higher cIMT compared to East and South Africa, despite having a lower prevalence of CVDs risk factors. West Africans also had higher cIMT compared to Asians and Europeans, but were concordant with data in African Americans, a trend not observed for East and South Africans. Our study is the first to report age-specific reference intervals of cIMT in SSA from a population-based study.

Second, genetic susceptibility to atherosclerosis in SSA was investigated by performing a GWAS on 7894 unrelated middle-aged adults (3963 women, 3931 men) from the AWI-Gen study (with participants from East, South and West Africa), with adjustment for age, sex and 8 principal components (PCs) (**Chapter 3**). cIMT was measured by ultrasound and genotyping was performed on the H3Africa SNP Array (~2.3M SNPs). After imputation, we tested for association using BOLT-LMM. We identified two new African-specific genome-wide significant cIMT-associated loci (*SIRPA* (p=4.7E-08) and *FBXL17* (p=2.5E-08)). Sex-

stratification analysis revealed two male-specific loci (*SNX29* (p=6.3E-9) and *MAP3K7* (p=5.3E-8)), and two female-specific loci (*ARNT2* (p=2.4E-09) and *PROK1* (p=1.0E-08)). We also successfully replicated known cIMT-associated loci and identified novel African-specific variant and gene associations.

Finally, we investigated gene-smoking interaction on cIMT in men from West Africa (**Chapter 4**). In Nanoro we identified new variants for the gene-smoking interactions for cIMT within the previously described *RCBTB1* region (p-value=3.08E-08), and identified a new region at *BCHE* (p=2.20E-08). In the combined sample, variants were identified in the regulatory region and in a region of open chromatin of *TBC1D8* (p-value=5.90E-09).

Our study highlighted sex differences in atherosclerosis risk, and the role of gene-environment interaction. There was transferability of signals from studies of non-African populations, providing opportunities for fine-mapping and reducing credible SNP sets using African data. Greater inclusion of African populations in medical genomics is important for accelerating discoveries and identifying new genetic associations with traits for variants absent from other populations. Gene-environment interaction studies provide an opportunity for prevention and a precision public health approach.

Acknowledgement

I would like to express my gratitude to my main supervisor, Prof. Michele Ramsay, for giving me the opportunity to undertake this formidable "adventure called Ph.D". I have learned and I have enjoyed every moment of it. Thank you for your guidance and support.

Thanks to my co-supervisors Dr Ananyo Choudhury, Prof. Christopher G. Mathew and Prof. Hermann Sorgho, for your continuous guidance, support and advice. Your scientific rigour and your attachment to excellence set the bar high.

I am indebted to Prof. Halidou Tinto for his mentorship since I joined his research team at the Clinical Research Unit of Nanoro and mainly throughout this PhD.

I would like to thank Dr Jean-Tristan Brandenburg, Prof. Scott Hazelhurst and Dr Dhriti Sengupta for their scientific and technical support. I owe gratitude to Phelelani Mpangase and Shaun Aron for their support in bioinformatics.

I would like to thank Evans Mathebula, Carl Chen, Jenny Matthew and Andrew Ndlovu, Cassandra Soo, Dr Liesl Hendry and Dr Thandiswa Ngcungcu for all our long discussion hours, doing and undoing the science. To the students, PhD candidates, post-docs and researchers of the Sydney Brenner Institute for Molecular Bioscience (SBIMB), thanks for wonderful environment you provided me with during this PhD.

I am thankful to the Michelle Ungerer, Dr Stuart Ali, Jocelyn Gayenga, Mmatshepo Taunyane and Yusuf Ismail, from the project office of SBIMB, for their administrative support. You made sure that our focus was on science. Thanks to Freedom Mukomana for managing the dataset. I express my gratitude to the Cassandra Soo and Natalie Smyth from the BioBank, your relentless effort allowed the provision of the data in the best timeline.

This work would not have been possible without the support from the AWI-Gen Group. I am indebted to all the study participants for their commitment and to the research teams from the six centres for making this happen. I also thank all my colleagues from the Clinical Research Unit of Nanoro for their support and encouragement.

To my sisters and brother (Marina, Edwige, Tatiana, Stephania, Isaac), Levi Valerie Ki, and to my friends (Rodrigue Sanwidi, Rodrigue Zigani, Juste Thiabonou, Moumouni Sere, Saidou Kabore), thank you for your love, support and patience. You have been a wonderful source of motivation.

I am grateful to the National Research Foundation South Africa (NRF)/The World Academy of Science (TWAS) for granting me the "African Renaissance Doctoral Fellowship" (Grant No:100004).

Tables of Contents

ABSTRACT ---------- **V**
TABLES OF CONTENTS---------- **VIII**
LIST OF FIGURES---------- **XI**
LIST OF TABLES ---------- **XIII**
ABBREVIATIONS ---------- **XIV**

CHAPTER 1: BACKGROUND AND STUDY RATIONALE---------- **1**
1.1. BACKGROUND AND STUDY RATIONALE---------- **1**
1.2. Pathophysiology of atherosclerosis---------- **3**
1.2.1. Definition ---------- 3
1.2.2. Pathophysiology ---------- 4
1.3. Use of carotid Intima-Media Thickness (cIMT) as a marker of atherosclerosis ---------- **5**
1.4. Genetic association studies---------- **7**
1.4.1. Genome-wide association studies (GWASs)---------- 7
1.4.1.1. Definition and concept ---------- 7
1.4.1.2. Steps in genome-wide association studies (GWASs) ---------- 8
1.4.1.2.1. Data quality control---------- 8
1.4.1.2.2. Imputation---------- 11
1.4.1.2.3. Genetic association testing---------- 11
1.4.1.2.4. Multiple testing---------- 13
1.4.1.2.5. Visualization of GWAS results ---------- 14
1.4.1.2.6. Post-GWAS analysis ---------- 14
1.5. GWAS studies in an African context ---------- **16**
1.6. Genetic studies of cIMT ---------- **18**
1.6.1. Candidate gene studies---------- 18
1.6.2. GWAS of atherosclerosis risk using cIMT as the outcome variable---------- 18
1.7. This PhD study---------- **20**
1.7.1. Study Context ---------- 20
1.7.2. Hypothesis, Objective and Aims---------- 21
1.8. Ethical considerations ---------- **24**

CHAPTER 2: REGIONAL DIFFERENCES IN CAROTID INTIMA-MEDIA THICKNESS REVEAL HIGHER RISK FOR ATHEROSCLEROSIS IN WEST AFRICA: AN AWI-GEN STUDY ---------- **27**
ABSTRACT ---------- **27**
2.1. INTRODUCTION---------- **28**
2.2. MATERIALS AND METHODS ---------- **29**
2.2.1. Participants---------- 29
2.2.2. cIMT measurements---------- 29
2.2.3. Other variables ---------- 30
2.2.4. Data analysis ---------- 31
2.2.5. Ethical Considerations ---------- 31
2.3. RESULTS ---------- **32**
2.3.1. Populations characteristics---------- 32
2.3.2. cIMT characterisitics---------- 32
2.3.3. Age-specific reference intervals (RIs) ---------- 36
2.3.4. Multivariate logistic regression of cIMT≥75th percentile---------- 39
2.4. DISCUSSION ---------- **41**

CHAPTER 3: AFRICAN-SPECIFIC VARIANTS REVEAL NOVEL SUSCEPTIBILITY LOCI FOR CAROTID INTIMA-MEDIA THICKNESS IN SUB-SAHARAN AFRICAN POPULATIONS WITH SEX-DIFFERENCES: AN AWI-GEN STUDY ---------- **44**
ABSTRACT ---------- **44**

3.1. Introduction ---- 46
3.2. Method ---- 47
3.2.1 Study population and phenotype assessments ---- 47
3.2.2 cIMT Measurement ---- 48
3.2.3 Association and follow up analysis ---- 48
3.2.3.1. Genotyping and Imputation ---- 48
3.2.3.2. Data Analysis ---- 48
3.2.3.3. Genome-wide association analysis ---- 49
3.2.3.4. GWAS catalog lookup ---- 49
3.2.3.5. Functional analysis ---- 50
3.2.3.6. Functional annotation of mapped genes ---- 50
3.2.3.7. MAGMA Gene-based and gene-sets analysis ---- 51
3.3. Results ---- 52
3.3.1. Association results ---- 52
3.3.2. GWAS Catalogue lookup ---- 61
3.3.3. Functional annotation ---- 64
3.3.4. Gene-based and gene-set analysis ---- 66
3.4. Discussion ---- 68
3.4.1. F-box and leucine rich repeat protein 17 (*FBXL17*) ---- 68
3.4.2. Signal regulatory protein alpha (*SIRPA*) ---- 69
3.4.3. Sorting nexin 29 (*SNX29*) ---- 69
3.4.4. Mitogen-activated protein kinase kinase kinase 7 (*MAP3K7*) ---- 70
3.4.5. La-related protein 6 (*LARP6*) ---- 70
3.4.6. Prokineticin 1 (*PROK1*) ---- 71
3.4.7. Caldesmon 1 (*CALD1*) ---- 71
3.4.8. Fms-related tyrosine kinase (*FLT4*) ---- 72
3.4.9. Relevance of our findings ---- 73
3.4.10. Strength and limitations ---- 74

CHAPTER 4: NOVEL AND KNOWN GENE-SMOKING INTERACTIONS WITH CIMT IDENTIFIED AS POTENTIAL DRIVERS FOR ATHEROSCLEROSIS RISK IN WEST-AFRICAN POPULATIONS OF THE AWI-GEN STUDY ---- 75
ABSTRACT ---- 75
4.1 Introduction ---- 77
4.2 Method ---- 79
4.2.1 Study population and phenotype assessments ---- 79
4.2.2 cIMT Measurement ---- 79
4.2.3 Smoking measurement and other variables ---- 79
4.2.4 Association and follow up analysis ---- 80
4.2.4.1 Genotyping and Imputation ---- 80
4.2.4.2 Data Analysis ---- 80
4.2.4.3 Gene-Environmental interaction (GxE) analysis ---- 80
4.2.4.4 Functional analysis ---- 81
4.2.4.5 Functional annotation of mapped genes ---- 82
4.2.4.6 GWAS Catalog lookup ---- 82
4.2.5 Ethics and consent ---- 82
4.3 Results ---- 83
4.3.1 Characteristics of participants ---- 83
4.3.2 Gene-Smoking interactions ---- 85
4.3.3 Comparison of association signals between sites ---- 100
4.3.4 GWAS Catalog lookup for gene-environment interaction ---- 100
4.3.5 Functional analysis ---- 100
4.3.6 Functional annotation of mapped genes ---- 101
4.3.7 Gene Set analysis ---- 101
4.4 Discussion ---- 104

CHAPTER 5: DISCUSSION AND CONCLUSION ---- 107
5.1 Regional epidemiological differences in Africa ---- 107

5.2 Our GWAS of cIMT linked genetics to biology of atherosclerosis in African populations ------- 108
5.3 African genetic diversity ---------- 108
5.4 Sex-specific genetics associations with African cIMT ---------- 110
5.5 Complex trait genetics as addressed in our study ---------- 112
5.6 Gene-Environment interaction ---------- 113
5.7 Replication and transferability ---------- 113
5.8 Trans-ethnic fine-mapping opportunities from the loci suggestive of replication in African ----- 114
5.9 Strengths and limitations ---------- 116
5.10 Future studies ---------- 118
5.11 Conclusion ---------- 120

REFERENCES ---------- 122
SUPPLEMENTARY CHAPTER 2 ---------- 147
SUPPLEMENTARY CHAPTER 3 ---------- 154
SUPPLEMENTARY CHAPTER 4: ---------- 160

List of Figures

Figure 1.1: Schematic diagram of the proposed determinants of, and risk factors for, cardiovascular diseases 1

Figure 1.2: Choropleth map showing the estimated age-standardized prevalence and mortality of total cvd in 2015 for each country. 2

Figure 1.3: The pathophysiology of atherosclerosis involves interacting systems at multiple levels. 4

Figure 1.4: Estimation of power for the GWAS of cIMT (as a quantitative trait) for the sample size of 12 000 and considering genetic effect β ranging from 0.0067 to 0.156 23

Figure 2.1: Comparison of age-specific reference intervals of median (50th percentile) per site. 37

Figure 3.1: Principal component analysis (PCA) plots showing population sub-structure in the AWI-Gen data and compared to data in selected population from the 1000GP 53

Figure 3.2: QQ and Manhattan plots for cimt association results in AWI-Gen study (7894 participants). 54

Figure 3.3: Miami plot showing female and male-specific associations p-values for cIMT. 55

Figure 3.4: Manhattan plot showing the –log10-transformed two-tailed p-value of each SNP for test of difference between female and male associations for each SNP. 55

Figure 3.5: Regional association plots for the selected top SNPs showing genetic associations 63

Figure 3.6: Manhattan plot for the gene-based test as computed by magma based on our summary statistics. 67

Figure 4.1: Manhattan plots showing the –log10-transformed two-tailed p-value of each SNP from the GWAS gene-smoking interaction for cimt on the Y-axis and base-pair positions along the chromosomes on the X-axis. 86

Figure 4.2: Regional and LD plots for the *RCBTB1* and *TBC1D8* regions. 87

Figure 4.3: Genotypes plots of rs1192824 for smokers *vs* non-smokers. 97

Figure 4.4: Circos plots showing genes on chromosome 2 for combined sample (a) and on chromosome 13 for Nanoro sample (b). 102

Figure 4.5: Gene ontology enrichment for biological processes for gene-smoking interaction in Nanoro, with overlapping genes in gene sets. 103

Figure 5.1: Level plot of allele frequencies by p-values of associations ($p<1e-4$) for the female-specific gwas showing over-representation of suggestive variants in extreme allele frequencies. 109

Figure 5.2: Illustration of the sex-difference in the selected variants using betas and Z. 111

Figure 5.3: Regional plots for genetic associations with cIMT, focusing on signals close to two genes (*CCDC71L* and *CBFA2T3*) previously identified in european GWASs and replicated in our study..... 115

Figure 5.4: LD matrix for selected variants in *CBFA2T3* region showing the LD block distribution in Africans and Europeans, using the 1000G data. 116

List of Tables

Table 2.1: Characteristics of the AWI-Gen study population by site and sex 33

Table 2.2: Characteristics of cIMT measurements per site and for men and women in the combined AWI-Gen cohort 34

Table 2.3: Characteristics of cIMT measurements per site and stratified by cardiovascular risk (Healthy vs Non-healthy) for men and women 35

Table 2.4: Age- and sex-specific (1-year intervals) percentiles of references intervals for carotid artery intima-media thickness (in mm) in the general population and the healthy sub-population for the AWI-Gen study. 36

Table 2.5: Age- and sex-specific (5-year intervals) percentiles of references intervals for carotid artery intima-media thickness (in mm) in the general population and the healthy sub-population 38

Table 2.6: Multiple logistic regression of binary cIMT (cIMT ≥75th percentile) per site and for the entire population 40

Table 3.1: Selected SNPs associated with Mean Max CIMT with p-values <1E-06 for the combined AWI-Gen sample (n=7894). 56

Table 3.2: Selected SNPs with p-values <1E-06 for the Female-specific analysis 57

Table 3.3: Selected SNPs with p-values <1E-06 for the Male-specific analysis 59

Table 3.4: Selected SNPs with p-values <1E-06 for the sex-difference test 60

Table 3.5: Positional annotation of candidate SNPs for SNPs in LD (0.8) with the lead-SNPs which have corresponding functional annotation assigned by ANNOVAR with enrichment relative to all SNPs in the 1000 GP reference African panel. 65

Table 4.1: Descriptive characteristics for participants from Nanoro (Burkina Faso) and Navrongo (Ghana) and the combined sample 84

Table 4.2a: Selected risk loci (p≤1E-05) for SNP-smoking interactions in Nanoro 89

Table 4.2b: Selected risk loci (p≤1E-05) for SNP-smoking interactions in Navrongo 92

Table 4.2c: Selected risk loci (p≤1E-05) for SNP-smoking interactions in combined sample 94

Table 4.3: Comparison of gene-smoking association results for Nanoro and the combined sample (All), based on Model 1 and Model 2 for selected SNPs. 98

Abbreviations

AAA	:	Abdominal Aortic Aneurysm
AAC	:	Abdominal Aortic Calcium
ABI	:	Ankle-Brachial Index
AWI-Gen	:	Africa Wits-INDEPTH partnership for Genomic Studies
BMI	:	Body Mass Index
CAC	:	Coronary Artery Calcification
CADD	:	Combined Annotation-Dependent Depletion
CCA	:	Common Carotid Artery
CED	:	Concordant effect direction
CEU	:	Utah residents of northern and western European ancestry
CHARGE	:	Cohorts for Heart and Aging Research in Genomic Epidemiology
CHD	:	Coronary Heart Disease
chr	:	Chromosome
CI	:	Confidence Interval
cIMT	:	Carotid Intima-Media Thickness
CSF1	:	Colony Stimulating Factor 1
CVD	:	Cardiovascular diseases
DBAC	:	Data and Biospecimen Access Committee
DEPICT	:	Data-driven Expression-Prioritized Integration for Complex Traits
DNA	:	Desoxyribonucleic Acid
EAFD	:	Extreme Allele Frequency Differences
ELSA-Brasil	:	The Brazilian Longitudinal Study of Adult Health
EMMAX	:	Efficient Mixed-Model Association eXpedited
eQTL	:	Expression Quantitative Loci
FaST-LMM	:	Factored Spectrally Transformed Linear Mixed Models
FBAT	:	Family-Based Association Testing
FDR	:	False Discovery Rate
FEV1/FVC	:	Forced expiratory volume in one second over forced vital capacity
FP	:	Fractional Polynomial
FUMA	:	Functional Mapping and Annotation of Genome-Wide Association Studies
GEMMA	:	Genome-wide Efficient Mixed Model Association
GH	:	Ghana
GIF	:	Genomic Inflation Factor

GLS	:	Generalized Least Squares
GMMAT	:	Generalized Linear Mixed Model Association Tests
GRCh37	:	Genome Reference Consortium Human Build 37
GTEx	:	Genotype-Tissue Expression
GWAS	:	Genome-wide association studies
gwasP	:	GWAS p-value
GxE	:	gene-environment interactions
H3Africa	:	Human Heredity and Health in Africa
HDL-C	:	High Density Lipoprotein-Cholesterol
HDSS	:	Health and Demographic Surveillance Systems
HIV	:	Human Immunodeficiency Virus
HLA	:	Human Leukocyte Antigen
HWE	:	Hardy–Weinberg Equilibrium
IL-1	:	Interleukin-1
IMPROVE	:	Carotid Intima Media Thickness (IMT) and IMT-Progression as Predictors of Vascular Events in a High Risk European Population
IndSigSNP	:	Index significant SNP
IQR	:	Interquartile
JNC7	:	The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure
JUH	:	Ju\'hoansi
Kb	:	Kilobases
KE	:	Kenya
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
LD	:	Linkage Disequilibrium
LDL-C	:	Low Density Lipoprotein-Cholesterol
LMM	:	Linear mixed models
LWK	:	Luhya in Webuye/Kenya
MAF	:	Minor Allele Frequency
MAGMA	:	Multi-Marker Analysis of Genomic Annotation
MCP-1	:	Monocyte Chemoattractant Protein 1
MSigDB	:	Molecular Signatures Database
MVPA	:	Moderate to vigorous physical activity in minutes per week
NCD	:	Non-Communicable Diseases
NOMAS	:	Northern Manhattan Study
OED	:	Opposite effect direction

OR	:	Odd Ratio
PAD	:	Peripheral Arterial Disease
PANSS	:	Positive and Negative Syndrome Scale
PC	:	Principal Component
PCA	:	Principal Component Analysis
pos	:	position
PWBT	:	Positional Burrows-Wheeler transform
Q-Q	:	quantile-quantile
QC	:	Quality Control
RA	:	Rheumatoid Arthritis
RDB	:	RegulomeDB
RI	:	Reference Interval
rsID	:	Reference SNP cluster ID
SCA	:	Sub-clinical atherosclerosis
SCAT	:	Sub-Cutaneous Abdominal Adipose Tissue
SD	:	Standard Deviation
SE	:	Standard Error
SMC	:	Smooth Muscle Cell
SNP	:	Single Nucleotide Polymorphism
SNP	:	Single Nucleotide Variant
SSA	:	Sub-Saharan Africa
SDI	:	Socio-Demographic Index
SSE	:	Single sex effect
SWEET	:	Study of Women Entering and Endocrine Transition
TNF-α	:	Tumour-Necrosis Factor-α
UTR3	:	Three prime untranslated region
UTR5	:	Five prime untranslated region
VAT	:	Visceral Adipose Tissue
VCAM-1	:	Vascular Cell Adhesion Molecule 1
VEGFC	:	Vascular endothelial growth factor C
VEGFD	:	Vascular endothelial growth factor D
YRI	:	Yoruba from Ibadan/Nigeria
ZA	:	South Africa

Chapter 1:

Background and Study rationale

Chapter 1: Background and study rationale

1.1. BACKGROUND AND STUDY RATIONALE

It is estimated that by 2020 cardiovascular diseases (CVDs) will be the leading cause of death (Fuster, 2014). In 2015, 19.9 million CVD deaths occurred, accounting for one-third of all global deaths, and 423 million people had prevalent CVD (~1 in 17 of the global population) (Roth *et al.*, 2017). It is estimated that by 2030 the number of deaths due to CVDs will increase to more than 24 million (Fuster, 2014). Sub-clinical atherosclerosis (SCA) (which presents early as an increase in carotid arterial intima-media thickness (cIMT) and has no recognizable clinical findings) and atherosclerosis have been found to be the pillar supporting the burden of CVDs such as ischemic heart disease, stroke, heart failure, peripheral arterial disease and rheumatic heart disease (**Figure 1.1**) (Herrington *et al.*, 2016; Tzoulaki *et al.*, 2016). During the last two decades, the burden of CVDs has considerably increased and the low- and middle-income countries are now experiencing about 80% of the worldwide burden. Sub-Saharan Africa (SSA) is experiencing a health and demographic transition that has shifted the major causes of death from communicable and nutritional diseases to non-communicable diseases (NCDs).

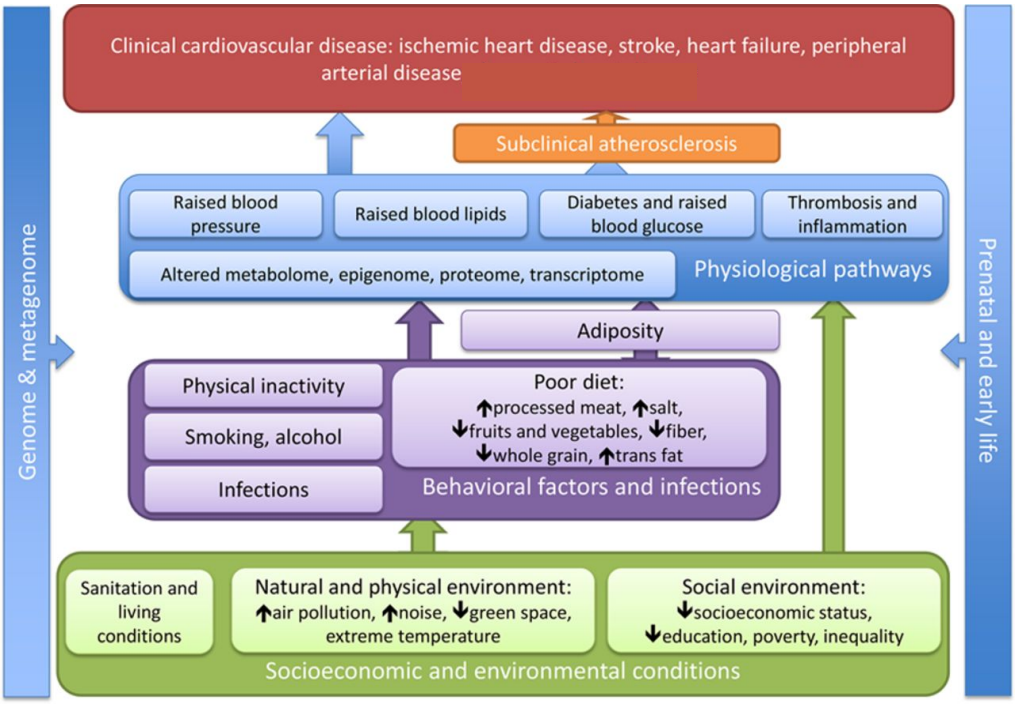


Figure 1.1: Schematic diagram of the proposed determinants of, and risk factors for, cardiovascular diseases
Source: (Tzoulaki et al., 2016)

The mean age of death attributable to CVDs in SSA in 2010 was 64.9 years (95% CI, 64.4–65.4) compared with 67.6-81.2 years for the rest of the world (Roth *et al.*, 2015), making it one of the youngest affected populations globally. In contrast to the conventional wisdom that CVD remains mainly a condition of wealthy nations, far more cases of CVD are now occurring in countries with the lowest socio-economic levels (Mozaffarian, 2017). Effectively, the latest reports on the global obesity epidemic showed that rural areas from low- and middle-income countries were driving the global rise in Body Mass Index (BMI) (NCD Risk Factor Collaboration, 2019). In 2015, countries with the highest age-standardized prevalence of CVDs, all with >9,000 cases per 100,000 persons, included most countries in West Africa, Morocco, Iran, Oman, Zambia, Mozambique, and Madagascar (**Figure 1.2A**). Controversially, this prevalence has not translated into increased CVD mortality (**Figure 1.2B**) (Roth *et al.*, 2017). In an analysis of the effect of demographic transition, despite marked improvement in Socio-Demographic Index (SDI), CVD mortality did not decrease for populations in much of SSA. Worldwide, the prevalence of hypertensive heart disease was highest in Western SSA, followed by Central and Eastern SSA, tropical Latin America, and the Caribbean. Southern Africa was leading the estimates of age-standardised prevalence and mortality of cardiomyopathies. Inversely, the continent suffered less from aortic aneurism and atrial fibrillation. It is important to note that there were considerable regional disparities within SSA, with Southern Africa having the highest mortality from Peripheral Arterial Disease (PAD) and Western Africa having the lowest (Roth *et al.*, 2017).

The epidemiological complexity of CVDs has raised many issues that remain understudied, such as the diagnosis and management of CVDs, the health system readiness to tackle the CVD risk factors, and the genetic susceptibility of different populations. The burden is such that the United Nations' third Sustainable Development Goal includes a target to reduce premature mortality from non-communicable diseases by one-third by 2030 (United Nations, 2015)

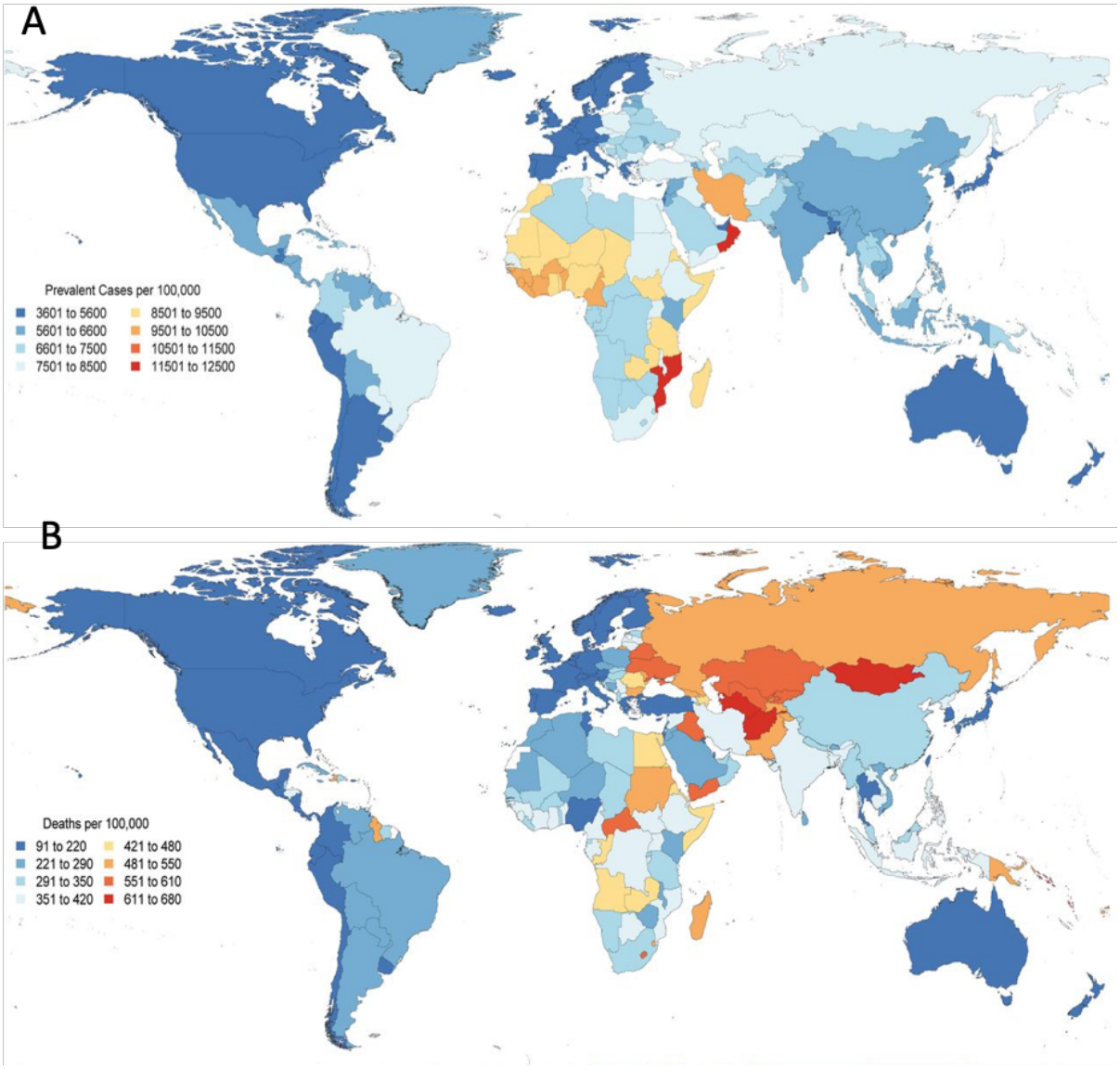


Figure 1.2: Choropleth map showing the estimated age-standardized prevalence and mortality of total CVD in 2015 for each country.

(A) Age-Standardized Prevalence of CVD in 2015. (B) Age-Standardized Death Rate of CVD in 2015 (Roth et al., 2017).

1.2. Pathophysiology of atherosclerosis

1.2.1. Definition

Atherosclerosis is a complex, progressive disorder affecting large and medium-sized arteries. The disease has a silent progression, often with no clinical evidence until the occurrence of a vascular event. Major risk factors for atherosclerosis, include dyslipidaemias, hypertension, diabetes and obesity, and it involves the interaction of several organ systems such as the liver, kidneys, gastrointestinal tract and hormonal systems. Atherosclerosis is a result of the complex interplay between genetic and environmental risk factors at the whole-organism level. Multiple stages characterise the disease evolution (**Figure 1. 3**) (Ramsey, Gold and Aderem, 2010). Ageing and early development of atherosclerotic lesions lead to significant differences between chronic and acute plaque characteristics. The progression of vascular lesions results in the reduction or cessation of blood flow through the vessels to the dependent tissues. These lesions may occur anywhere within the arteries, and are more likely to occur at vessel curves and bifurcations where the smooth flow of blood is disturbed by shear stress forces created by blood flow, resulting in abnormal cell migration and proliferation in the vessel wall (Winkel *et al.*, 2015). The vessels most likely to be affected by this flow-mediated effect include the coronary arteries, the carotid and cerebral arteries as well as the iliofemoral system, resulting in the respective clinical manifestations of myocardial ischaemia and infarction, stroke and peripheral vascular disease (Mota *et al.*, 2017).

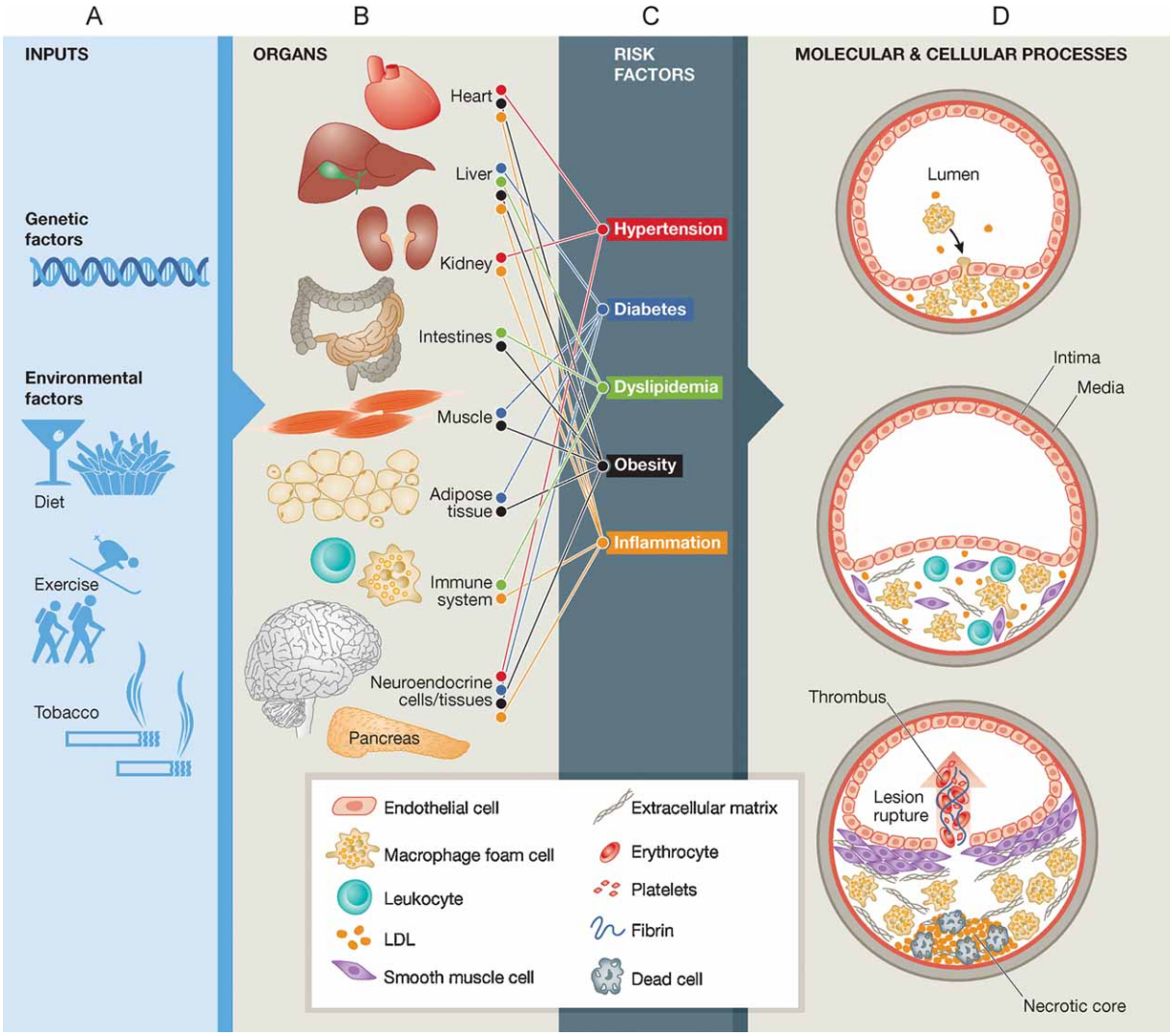


Figure 1.3: The pathophysiology of atherosclerosis involves interacting systems at multiple levels.

(A) Genetic and environmental factors. (B) Multiple organs and organ systems involved in the disease process. (C) Risk factors that are influenced by genetic and environmental inputs and that involve multiple organ systems contribute to the development and progression of atherosclerosis. (D) Atherosclerosis progresses through several stages, each of which involves the interaction of multiple cell types and molecular processes. (Ramsey, Gold and Aderem, 2010)

1.2.2. Pathophysiology

Biologically, the disease involves the formation of plaques in arterial walls and thickening that narrows the arterial passage. This restricts blood flow and increases the risk of occlusion of blood flow resulting in myocardial infarction and other events. Atherosclerosis is a low-grade chronic inflammatory condition characterized by aberrant lipid metabolism and a maladaptive

inflammatory response (Viola and Soehnlein, 2015). Atherosclerosis also involves matrix deposits and smooth muscle cell (SMC) proliferation in the wall of medium- and large-sized arteries. This accumulation is most commonly detected during the second decade of life and develops further with age. There is now a consensus that atherosclerosis represents a state of heightened oxidative stress characterized by lipid and protein oxidation in the vascular wall. The role of lipid oxidation in the initiation and progression of atherosclerosis has been reported in several studies (Ceriello, 2008; Rodford *et al.*, 2008; Alzaid, Patel and Preedy, 2014; Siti, Kamisah and Kamsiah, 2015).

Atherosclerotic plaque development begins with endothelial cell activation, including overexpression of leukocyte adhesion proteins such as vascular cell adhesion molecule 1 (VCAM-1). Chemoattractants such as monocyte chemoattractant protein 1 (MCP-1) promote migration of leukocytes into the intima cell layer (Boring *et al.*, 1998), where macrophage colony stimulating factor 1 (CSF1) promotes the differentiation of monocytes into macrophages (Rajavashisth *et al.*, 1990). These macrophages express scavenger receptors that allow them to engulf and modify lipoproteins and become foam cells which secrete inflammatory mediators such as interleukin-1 (IL-1), tumour-necrosis factor-α (TNF-α), nitric oxide and endothelin. These inflammatory mediators amplify inflammation in the vessel wall and can contribute to additional leukocyte accumulation, SMC proliferation and extracellular matrix remodelling (Goldstein and Brown, 1974; Greaves and Gordon, 2008). Multiple other leukocytes are recruited to the lesion and have been demonstrated to play a critical role in disease development (Weber, Zernecke and Libby, 2008). While foam cell accumulation characterizes fatty streaks, deposition of fibrous tissue defines the more advanced atherosclerotic lesion. SMCs synthesize the bulk of the extracellular matrix that characterizes this phase of plaque evolution (Raines and Ferri, 2005). Plaque rupture resulting from inflammatory activation and the ensuing thrombosis commonly cause the most acute complications of atherosclerosis such as myocardial infarctions or stroke (Fuster *et al.*, 2005).

1.3. Use of carotid Intima-Media Thickness (cIMT) as a marker of atherosclerosis

Atherosclerosis precedes the onset of most of the clinical cardiovascular events and sub-clinical atherosclerosis (SCA) is quite common in young and middle-aged populations. SCA is often measured by non-invasive methods such as the ankle-brachial index (ABI), the internal and common carotid intima-media thickness (cIMT), the coronary artery calcification (CAC) and

the abdominal aortic calcium (AAC). These four measurements have been shown to independently predict future cardiovascular events (Sillesen *et al.*, 2012).

cIMT is a widely-accepted imaging surrogate marker of generalized atherosclerosis. Measured by dual mode ultrasound, cIMT is represented by the distance between the near and the far wall of the carotid artery corresponding to the lumen-intima and the media-adventitia interfaces. Even in the absence of atherosclerosis, the thickness of the intima and the media layer increases with advancing age due to biological changes. Since the molecular and cellular pathways responsible for these changes are also involved in the formation of atherosclerotic plaque, cIMT is related to SCA. The advantage of measuring the cIMT by high resolution B-mode ultrasonography lies in its rapidly applicable and available, non-invasive and cost-effective nature. Therefore, cIMT measurement is an attractive method for use in large-scale research projects (Bartels, Franco and Rundek, 2012). Previous studies have reported that traditional risk factors are not the major contributors to the variance of cIMT, with age (7%), male sex (3%), glucose (<1%), number of cigarettes pack per years of smoking (<1%), and LDL-cholesterol (<1%), cumulatively accounting for only 11% of the variability. They also found that black African ancestry is significantly associated with an increased cIMT (Rundek *et al.*, 2013). Another study reported that parental stroke was associated with a more rapid progression increase of cIMT in young women (Lin *et al.*, 2015). Previous studies have reported that the heritability of cIMT ranges from 18% to 67% (Lange *et al.* 2002; Fox *et al*. 2003; Suh-Hang *et al*. 2004; Xiang *et al*. 2002), and some have reported sex differences (inconclusive) in atherosclerosis (Lin *et al.*, 2015; Spence and Pilote, 2015). The Northern Manhattan Study (NOMAS) which includes families, reported a heritability of 65% (Sacco *et al.*, 2009). cIMT as measured by B-mode ultrasound is associated with future cardiovascular events (Van Den Oord *et al.*, 2013). These findings support the need for an investigation into the genetic association with atherosclerosis, using cIMT as a surrogate marker.

Several populations across the world have been studied for cIMT and reference intervals have been determined for normal variation and higher risk for CVD (Engelen *et al.*, 2013; Lee *et al.*, 2014; Santos *et al.*, 2014; Liu *et al.*, 2017). However, the few reports on cIMT in SSA, come from studies using data from patients affected with specific disorders and usually have small sample sizes (Holland *et al.*, 2009; Okeahialam *et al.*, 2011; Rheeder *et al.*, 2012; Owolabi *et al.*, 2015; Schoffelen *et al.*, 2015; Olt *et al.*, 2016; Zemlin *et al.*, 2017). Therefore, there is a need to fill this gap in data and knowledge on cIMT in African populations.

1.4. Genetic association studies

1.4.1. Genome-wide association studies (GWASs)

1.4.1.1. Definition and concept

Consequent to the failure of candidate-gene studies to shed light on the genetic architecture of complex traits and after the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), Genome-wide association studies (GWAS) have evolved as the "holy grail" of genetic association studies thanks to their non-hypothesis driven approach. In GWAS, hundreds of thousands to millions of genetic variants across the genomes are tested in large numbers of individuals to identify genotype–phenotype associations. This has revolutionized the field of complex disease genetics over the past decade. By January 2019, 3,730 GWASs had been published, with a total of 37,730 SNPs (single nucleotide variants) reporting 52,415 unique SNP–trait associations at a genome-wide significance threshold (Buniello *et al.*, 2019). These associations have led to insights into the genetic architecture of disease susceptibility, to advances in clinical care (by identification of new drug targets and disease biomarkers) and to applications in personalized medicine (risk prediction and optimization of therapies based on genotype).

The goal of a GWAS is to determine which genomic loci are associated with the disease or trait of interest. When the phenotype is a disease, this is a dichotomous variable (present versus absent). The GWAS tests the hypothesis that a SNP's allele frequencies differ between a group of people with disease (cases) and a group of people without the disease (controls). This analysis is typically performed systematically for all common SNPs (minor allele frequency >1%) across the genome (usually detected using a SNP genotyping array with tag SNPs pruned by eliminating loci in close linkage disequilibrium (LD)) (Tam *et al.*, 2019). In a successful GWAS, there will be several SNPs for which the allele frequencies differ significantly between the two groups. These differences may be small, reflecting that each individual SNP has only a small effect on the presence of absence of the disease. Nonetheless, GWAS requires a substantial sample size to be adequately powered, with study samples up to hundreds of thousands of individuals, so that allele frequency differences can be reliably detected at a robust statistical significance threshold. The genome-wide significance threshold is typically $p<5\times10^{-8}$, which is derived from Bonferroni correction for multiple testing (0.05/1 000 000) accounting for the fact that 1 million SNPs are being independently tested for disease

association in a GWAS. When the phenotype is a continuous quantitative trait (e.g. blood pressure or cIMT), the query for any given SNP tests whether there is a statistically significant difference in the phenotype in a large group of people by assessing the distribution of the phenotype across the three different genotypes at the SNP (A/A, A/B, and B/B).

It is important to note that GWAS studies are not assessing causal variant but can only examine associations with SNPs chosen for inclusion in the array. They are usually chosen using linkage disequilibrium to identify independent loci, and the SNP is referred to as a tag SNP or index SNP. It serves as proxy to identify a region (or haplotype) in the genome that may be associated with the phenotype. Identification of causal variants requires follow-up studies. While a causal variant may be a coding variant that alters the amino acid sequence of a protein, thus influencing the disease phenotype, most causal variants are likely to be non-coding and influences a gene's function from a distance through regulation of the gene's expression (Musunuru *et al.*, 2010; Harmston and Lenhard, 2013). Generally, the LD structure dictates that a causal variant itself must lie within the locus or haplotype defined by flanking recombination hot spots. However, a causal gene or variant may lie outside of the locus, as far as hundreds of kilobases away from a causal variant because of, for example, complex chromatin looping that could bring the variant in close physical proximity to the causal gene.

1.4.1.2. Steps in genome-wide association studies (GWASs)

Strategies and approaches for GWASs include a number of steps, from data quality control, to SNP imputation, genetic association analyses, addressing multiple comparisons, visualization of the results, and other post-GWAS analysis such as functional analysis of associated SNPs and regions.

1.4.1.2.1. Data quality control

Successful GWASs require stringent quality control of the data and analyses to identify true genetic associations. Because GWAS typically involves large sample sizes and hundreds of thousands of polymorphisms being studied, spurious associations can emerge due to a variety of factors. These include poor accuracy of phenotyping, poor quality of DNA samples, poor DNA hybridization to the array, poorly performing genotype probes, as well as sample mix-ups or contamination. Therefore, QC of genetic data requires specific steps for individual samples and SNP filtering, broadly as follows there will be data exclusions: (1) high individual sample and SNP missingness, (2) inconsistencies in assigned and genotype-based sex of participants

(sex discrepancy), (3) low minor allele frequency (MAF), (4) highly significant deviations from Hardy-Weinberg equilibrium (HWE), (5) extremes in heterozygosity rate, (6) high relatedness, and (7) extreme population stratification or unexpected recent admixture with distantly related populations.

Missing call rate

The missing call rate of an individual samples is the proportion of SNPs whose genotypes cannot be called in that individual. Individuals with high missing call rates are indicative of poor DNA quality. In general, participants with a missing call rate between 1-5% are removed.

The missing call rate of SNP is the proportion of individuals whose genotypes are not called for a given SNP. SNPs with high missing genotype rate (e.g., >5%) imply problems with the genotyping process or assay, so these SNPs are eliminated from the data. The genotypes of some of these SNPs can be recovered by the imputation process (described below), provided there is access to appropriate reference datasets with well represented haplotypes in the relevant region.

Sex discrepancy

Since sex is an important covariate in GWAS analyses, it is necessary to perform a sex check. Discordant results are also an indicator that samples were mixed-up. The test uses X chromosome data to estimate genetic sex (heterozygosity of markers from the non-recombining region) and compares it to self-reported sex. Discrepancies between reported and genotype-based estimated sex are subject to further examination. In some cases, the original documents can be retrieved and may resolve this issue if data was incorrectly entered into the database. If the sex discrepancy cannot be resolved, those samples should be removed from the dataset.

Minor allele frequency

Because GWAS is testing for association of a phenotype with common SNPs (MAF>1%) to enhance power, SNPs with MAF<1% in the dataset are excluded from subsequent analysis to minimise false positive associations. This is partly because a GWAS would have low power to detect a modest effect size of a rare SNP unless the sample size was very large, or because these SNPs might be genotyping errors and induce spurious association. With the recent increases in the sample sizes of GWASs (over hundreds of thousands on individuals) there is a tendency to lower the MAF cut-off and many SNPs with lower MAFs, but substantial numbers of individuals with those alleles, are retained.

Hardy-Weinberg equilibrium (HWE)

Hardy-Weinberg equilibrium (HWE) describes the relationship between observed and expected genotype frequencies in a population. They should be similar and remain constant across generations in a population of diploid sexually reproducing organisms under the assumptions such as random mating and an infinitely large population size with negligible recurrent mutation rates. Departure from HWE may indicate signals of strong natural selection, recent population admixture, high cryptic relatedness among participants, genotyping errors or strong association with the trait under consideration. A common practice is to flag SNPs that deviate from HWE with a specific p value cut-off and to remove them from further analysis (for GWAS, $p < 0.00001$ is typically applied). HWE should also be re-examined after association analyses have been performed.

Heterozygosity rate

The heterozygosity rate refers to the proportion of heterozygous genotypes carried by an individual. High levels of heterozygosity may indicate genotyping errors and low sample quality, whereas low levels of heterozygosity may be due to inbreeding. The usual practice is to exclude individuals outside ± 3 SD from the mean.

Relatedness

This is an indicator of how strongly a pair of individuals are genetically related. Inclusion of related individual in the study could bias estimations of standard errors of SNP effect sizes. In practice, identity by state (IBS) estimates are used to detect this relatedness by fixing a PIHAT threshold (between 17-25%). Nonetheless, specific methods can be applied for family-based association testing (FBAT). Mixed modelling methods have evolved to allow the inclusion of related individuals in GWAS. These approaches account for relatedness (families, cryptic relatedness) by allowing for kinship adjustment (efficient mixed models and Bayesian mixed models).

Population stratification

As human geographic range expanded within and out of Africa, they split in subgroups and experienced diverse evolutionary stresses that shaped their genetic diversity (Schlebusch *et al.*, 2012). Population stratification occurs when the study participants comprise multiple groups of individuals who differ systematically in genetic ancestry and have different representations of the different population groups across the phenotypes under investigation. Population structure can cause spurious findings if not accounted for. Population structure also refers to population

admixture, family structure, and cryptic relatedness. The population structure can occur in apparently homogenous study populations, and the tendency to combine samples from diverse cohorts is increasing the odds of structure. Even when populations inhabit the same geographical region, population structure can still occur (Hellwege *et al.*, 2017). The use of principal component analyses (PCA) to explicitly detect and adjust for population stratification on a genome-wide scale is an efficient manner to address population structure. The use of mixed models (Linear or Bayesian) is also an efficient strategy to overcome potential bias due to population structure.

1.4.1.2.2. Imputation

Imputation is a computational process by which genotypes from additional SNPs are introduced, into an analytical dataset in addition to the SNPs included on the genotyping array. The principle is to use the interrelationships of SNPs within LD blocks to predict adjacent SNPs from the actual genotyped SNPs, based on haplotype structure. The more similar the reference population datasets are to the participants within the study, the more accurate the imputation will be. The importance of imputation lies in analysing associations of predicted genotypes from non-typed SNPs with the outcome variables, because functional (causal) variants may not be measured but may have been imputed. A major advantage is an approximatively ten-fold increase in the number of SNPs genotyped. This not only increases the chances of detecting significant associations in regions with relatively poor coverage on the array, but also increases confidence of an association and is useful for fine-mapping. Robust methods have been developed to ensure high quality and accuracy of imputation (Marchini *et al.*, 2007; Howie, Donnelly and Marchini, 2009; Howie *et al.*, 2012; Das, Abecasis and Browning, 2018). After the imputation process, additional QC measures are performed before the association analyses are conducted.

1.4.1.2.3. Genetic association testing

In GWASs, tests for genetic association are performed separately for each SNP. In case-control studies, under the null hypothesis of no association with the disease, we expect the relative allele or genotype frequencies to be the same in the case and control groups. The data for each SNP with minor allele “a” and major allele “A” can therefore be represented as a contingency table of counts of disease status by genotype count (a/a, A/a and A/A) or allele count (a and A). When using X^2 analyses, the allelic test is more powerful than the genotype-based test. The

Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955) is used to test different penetrance models (additive, dominant or recessive). In most GWASs, where the underlying genetic models are unknown, the additive model is commonly used. Since complex trait analysis requires the use of additional covariates, logistic regression has been widely used. The regression coefficients fitted in the logistic regression analyses represent the log of the Odd Ratios (ORs) for disease gene association. Furthermore, a logistic mixed model can be used for binary traits and is available in the R package GMMAT (Chen *et al.*, 2016).

In quantitative trait analysis, linear models are used as common practice, allowing control for potential confounding factors such as age and sex. Linear mixed models (LMM) have emerged as the method of choice for association testing in GWASs because they account for both population stratification and cryptic relatedness and achieve increased statistical power by jointly modelling all genotyped markers. LMMs consider the SNP under testing and the traits and covariates as fixed effect variables, and the remaining genotypes as random effect variables (Yang *et al.*, 2014) . The resulting model has the following form:

$$Y = \sum_j jX_j + uG + \sum_k v_k Z_k + \varepsilon,$$

where G is the test SNP, the variables X_j are fixed effect covariates, and Z_k are genotypes with random effects v_k. j is the index of fixed effect covariates and k is the index of random effect genotypes. v_k is assumed to be independently drawn from a normal distribution with mean 0 and variance τ, i.e., $v_k \sim N(0,\tau)$. ε is the error term with $\varepsilon \sim N(0,\sigma^2 I_n)$. I_n is the identity matrix for n individuals and σ^2 is the variance of the error term. The variance of Y can be written as:

$Var(Y) = \tau K + \sigma^2 I_n$

where K is the genetic relationship matrix, or the kinship matrix for related participants. The parameter values are estimated using restricted maximum likelihood (Kang *et al.*, 2010) .Since the LMM allows for the control of extra covariance that may arise from relatedness between participants, it can be used for association testing in family data and for data with population stratification. In fact, the LMM is a generalization of the PCA approach in which all PCs are intrinsically included. The LMM can be implemented through R packages LMM or LME2, or through genetic analysis packages such as EMMAX (Kang *et al.*, 2010), GEMMA (Zhou and Stephens, 2012) or FaST-LMM (Lippert *et al.*, 2011). Recently, BOLT-LMM, a mixed-model association method implementing a Bayesian framework with mixture of prior on marker effect

size, has been introduced. BOLT-LMM was found to increase association power over the standard infinitesimal mixed-model analysis for traits driven by a few thousand causal SNPs.

"Missing heritability" can be defined as the gap between estimated heritability and the proportion of variance explained by GWAS findings (Manolio *et al.*, 2009). This may have resulted from two probable reasons: the highly polygenic nature of complex traits also affected by rare variants that are not captured by genotyping arrays: and the overestimation of heritability by twins studies (Young, 2019). To overcome the issue of "missing heritability" (Young, 2019), testing for gene-gene interactions and gene-environment interactions has rapidly evolved as a valuable contribution to analyses. Nonetheless, issues related to difficulty of replication, computational burden, stringent multiple testing thresholds, and robustness to phenotype scaling or linkage disequilibrium, have complicated the development of consensus methods to detect epistasis and gene-environment interactions. One interesting approach for gene-environment testing was proposed by Aschard and collaborators (Aschard *et al.*, 2010). They showed that meta-analysis of a joint test for genetic main effects and gene-environment interaction effects can be more powerful than classic meta-analytic approaches. Randall *et al*. proposed a determination of difference in beta values between two strata with a correction of the standard error using Spearman rank correlation coefficient for stratum-specific beta values across all SNPs (Randall *et al.*, 2013).

1.4.1.2.4. Multiple testing

Controlling for multiple testing to accurately estimate significance thresholds is a very important aspect of studies involving many genetic markers, particularly in GWASs. The type I error, also called the significance level or false-positive rate, is the probability of rejecting the null hypothesis when it is true. The significance level indicates the proportion of false positives that an investigator is willing to tolerate in a study. Usually a Bonferroni correction is used, but less conservative corrections such as the false discovery rate (FDR) are also employed. Knowing the phenotype and genetic effect and calculating a statistical correction for false positives allows the researcher to determine the required sample size to detect a statistically significant association. The most used threshold to declare statistical evidence of association at a genome-wide level is $p< 5x10^{-8}$, which arises from adjusting a 5% type I error rate (false positives) to the roughly 1–10 million independent tests performed. When FDR is used the threshold of 5% is admissible.

1.4.1.2.5. Visualization of GWAS results

Considering the number of markers tested in a GWAS, quality control metrics are not sufficient for assessing the quality of the results or understanding them. Therefore, visualization tools are useful for sensitivity analyses to identify inconsistencies, probable systematic biases and consistency with findings from previous reports. Manhattan plots are used to visualize GWAS significance levels of individual SNPs by chromosome location. The x-axis represents SNP coordinates with chromosome numbers, the y-axis is the negative of the log p-value, and each dot corresponds to a single SNP. The solid horizontal lines is incorporated to indicate the Bonferonni corrected significance threshold ($-\log(5\times10^{-8})$), and a dotted horizontal line indicates a less stringent suggestive association threshold ($-\log(10^{-5})$). Visual inspection of Manhattan plots allows for identification of SNPs with relatively small p-values which, when they are isolated in a region of non-significant p-values, suggests potential spurious findings. A quick check of signals can help confirm previous findings (e.g. HLA region associations for immunology-related traits). The second visualization tool is the quantile-quantile (Q-Q) plot, which compares expected and observed distributions of the p-values of SNPs statistics. Inflation of significant p-values might be an indication of population structure or an excess of false positives, whereas deflation may indicate overfitted models or potential issues with the phenotype quality. Q-Q plots are often supported by the Genomic Inflation Factor (GIF), commonly referred as Lambda (λ).

λ = median (χ^2)/0.456 (Where χ^2 represent the chi square of p-values)

The closer the value of lambda is to one, the lower the likelihood of inflation. It can be used as an indication of the effects of changes in the association model. After identification of a significantly associated locus, having a closer look at the genomic region might be relevant to understanding the genomic context of the associated SNP. For example, its genic context (intergenic, intronic, exonic etc.), its surrounding genes and its position relative to recombination hotspots. This can be achieved by generating regional plots, which can be complemented by LD plots.

1.4.1.2.6. Post-GWAS analysis

Salvatore *et al*. summarized the post GWAS analysis in three steps:

1) Using several computational biology approaches to provide *in silico* support for the functional significance of associated SNP.

2) Aggregating the effects across the SNPs (especially when genome-wide significance is not achieved at some loci,) can help discoveries (e.g. gene-based analysis or gen-set analysis) allowing to assess heritability (variance explained) or to identify genes and networks of biological importance.
3) Mining expression data is of high relevance, because the importance of single loci, genes or networks identified using *in silico* approaches lies in validating their role in regulating differences in gene expression or some other functional feature (Salvatore *et al.*, 2018).

Functional annotation

Functional annotation of GWAS results is a critical step in understanding the functional consequences of the associated SNPs and to accurately elucidate the biological mechanism by which these genes and SNPs act. Whereas Mendelian diseases result from the inheritance of a single variant primarily in a coding region, the vast majority of associated SNPs that have been identified for common complex diseases map to non-coding intergenic and intronic or regulatory regions (Maurano *et al.*, 2012). Many of these regions are enriched for expression quantitative trait loci (eQTLs) (GTex Consortium *et al.*, 2017), DNase I hypersensitive sites sequencing (DHSseq) peaks, and chromatin immunoprecipitation sequencing (ChIPseq) peaks (ENCODE Project Consortium *et al.*, 2012). Numerous tools are available for functional follow-up of GWAS results, but two of them stand out due to their extensive inclusion of methods using summary statistics as inputs. DEPICT (Data-driven Expression-Prioritized Integration for Complex Traits) is an integrative tool that, based on predicted gene functions, systematically prioritizes the most likely causal genes at associated loci, highlighting enriched pathways, and identifying tissues/cell types where genes from associated loci are highly expressed (Pers *et al.*, 2015). The second tool is FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies), a platform that can be used to annotate, prioritize, visualize and interpret GWAS results (Watanabe *et al.*, 2017).

Replication

Replication can be defined as the repeated observation of associations between marker and outcomes in different populations under the same investigation design. Replication aims primarily to provide convincing statistical evidence of association, and to exclude associations due to biases. It also serves to improve effect estimation and to provide data in diverse populations. Guidelines for replication in GWASs suggest that exact replication should be

attained by genotyping the same marker (or a near-perfect proxy) across studies, analysed using the same genetic model, and assessed for the same phenotype. Large sample sizes are required to identify and replicate increasingly small effect sizes for common variants. To achieve this, collaboration between consortia and data sharing are invaluable tools for well-powered replication studies.

1.5. GWAS studies in an African context

Populations of African descent have been under-represented by the genomic revolution, representing only about 3% of the participants worldwide used for GWAS up to 2016 (Popejoy and Fullerton, 2016). This proportion is decreasing due to the huge European studies (including the UK Biobank) that are now available, and although African studies are being conducted, they remain of modest sample size. The increased genetic diversity in Africans affords ample opportunity to improve fine-mapping resolution for associated loci, discover novel genetic associations with phenotypes, build more generalizable genetic risk prediction models, and better understand the genetic architecture of complex traits and diseases subject to varying environmental pressures. Greater genetic diversity comes with greater ancestral heterogeneity, and this higher level of understudied diversity has the potential to yield novel genetic associations.

This is a gap that needs to be filled, considering that Africa is the continent with the highest genetic diversity due to its deep evolutionary roots and that the genomes of African individuals generally have lower linkage disequilibrium (LD). The more rapid LD decay in Africa also means there is greater fine-mapping resolution to pinpoint causal variants influencing traits than will be discovered in any global population (Wu *et al.*, 2013). Additionally, the investigation of genetic variants with extreme allele frequency differences (EAFD) illustrates the interest in investigations of African genomes. In their study, Sulovari and collaborators, who worked on 2504 genomes from 26 world populations, found that African populations were more diverse and had significantly more EAFD involved in genes and pathways (Sulovari *et al.*, 2017). The LD and EAFD together justify that the African genome is highly relevant for the discovery of new genetic associations and better understanding of human disease mechanisms (Tekola-ayele and Rotimi, 2015). Greater inclusion of African populations in medical genomics is important for accelerating genomic discoveries, enabling reconstruction of modern human origins, producing results that can be translated across populations more accurately, and identifying genetic associations with traits for variants absent in non-African populations. Nonetheless,

methods that assume homogeneous population structure and work well in European populations may not work as well in the presence of greater heterogeneity in African populations. These differences in African genetics also raises the issue of the genotyping technology.

Genotyping arrays are a common tool for genetic association studies primarily due to their cost-efficiency and the simplicity in the storage and analysis of array-based data. The arrays that are currently being used for population genetic studies and GWASs in African populations were predominantly designed on the basis of common variants in Eurasian whole genome sequences and therefore have been shown to have limited power in capturing the enormous genetic diversity in Africans (Peprah, Tekola-ayele and Royal, 2015). Because of intrinsic differences in the LD architecture between African and non-African populations the current arrays possess a limited ability to identify association signals (Choudhury *et al.*, 2014).

To address these concerns, the H3Africa genotyping array was designed as an African-centric, unbiased cost-efficient solution for genomic research in populations from the continent. In contrast to a few hundred African genomes that formed the basis of the design of previous genotyping arrays, the H3Africa array is based on an extensive study of genomic variants and genetic architecture in more than 3607 African genomes (many of which are unpublished datasets) from more than 20 different populations across Eastern, Western and Southern Africa (Botha, 2016). This ensures a much better representation of African genetic diversity and LD architecture in the array. To achieve cost efficiency, the array, containing a total of approximately 2.5 million SNPs, was designed to include 1.8 million SNPs from existing Illumina bead pools and 0.7 million custom variants. Owing to restrictions imposed by this design, rigorous analysis to discard potentially monomorphic SNPs has been performed. Therefore, we expect the genotyping studies based on this array to be considerably more informative. Moreover, being designed for a consortium that has a focus on studying NCDs, the H3Africa array has a dense coverage of single nucleotide variants (SNPs) that have previously been found to be associated with NCDs such as hypertension and kidney disease in both GWAS and gene-centric studies. The designed array has been shown to outperform most of the current arrays for the majority of genomic regions both in terms of imputation accuracy and coverage during the testing phase (Botha, 2016).

1.6. Genetic studies of cIMT

1.6.1. Candidate gene studies

Over the past decades, a specific area of research has focused on accurately defining the genetic components that modulate susceptibility to atherosclerosis and its risk factors. Since the purpose of these studies was to identify specific genes and gene variants that would allow development of better diagnostic tests and improved disease therapies, they focused specifically on candidate genes known to be involved in the pathophysiology of atherosclerosis. These included genes in the renin-angiotensin system, lipoprotein metabolism and inflammation. The selected genes of interest were generally investigated in small case–control cohorts, which have a high risk of false-negative and false-positive findings (Morgan, Krumholz and Lifton, 2007). Nonetheless, several candidate gene studies have identified some associations with atherosclerosis for genetic variants in the following genes: *GPX1*, *GCLM*, *MPO*, *PON*, *NADPH, NRF2*, *SOD2* (Hamanishi *et al.*, 2004; Yang *et al.*, 2004; Katakami *et al.*, 2009, 2010; Yang and Wang, 2012; Zhang *et al.*, 2014). It is important to consider that the design of candidate gene studies means that they cannot identify disease pathways previously unknown in the pathogenesis of atherosclerosis. The results of most candidate gene studies could not be replicated in subsequent studies in other populations. These disappointing findings of the candidate gene approach initially fostered scepticism and led some to conclude that "genomics will never be useful in arterial disease"(Reitsma, 2007).

1.6.2. GWAS of atherosclerosis risk using cIMT as the outcome variable

Several genetic association studies for atherosclerosis have been performed in the major world populations, with the exception of sub-Saharan Africans. The first GWAS for atherosclerosis was performed in the Framingham Heart Study and did not reveal any SNPs reaching the genome-wide significance level, but identified several variants with modestly significant associations ($p<10^{-5}$) (O'Donnell *et al.*, 2007). This study included 984 individuals of European origin and data from the Affymetrix 100K chip array, with maximum common cIMT as the outcome variable and adjusting for age and sex. In 2010, a small GWA study of carotid atherosclerosis in HIV-infected Caucasian men from USA found that two SNPs located in the ryanodine receptor gene (*RYR3*) on chromosome 15 were significantly associated with common cIMT in a group of 177 HIV-positive individuals, using the Illumina HumanCNV370-quad beadchip genotyping array (p-value $< 3.4\times10^{-8}$) (Shrestha, Irvin and Grunfeld, 2010). The variants were further replicated in men and in women (Shrestha *et al.*, 2012; Shendre *et al.*,

2014). The *RYR* gene family has been shown to play a role in the aetiology of cardiovascular disease and has been shown to be regulated by the HIV TAT protein (Köhler *et al.*, 2001; Paladugu *et al.*, 2003; Norman *et al.*, 2008). It was in 2011 that the first genome-wide significant associations with cIMT were found through a meta-analysis of 16 studies (n > 42000 participants) performed in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, which identified genetic associations with SNP markers in three loci *(ZHX2, APOC1, PINX1*) ($p < 5\times10^{-8}$) that explained 1.1% of cIMT variance (Bis *et al.*, 2012). The result from the IMPROVE (Carotid Intima Media Thickness (IMT) and IMT-Progression as Predictors of Vascular Events in a High Risk European Population) study revealed that the *BCAR1-CFDP1-TMEM170A* locus was associated with cIMT, reaching the array-wide significance ($p = 7.2\times10^{-6}$) (Gertow *et al.*, 2012). GWASs performed in Mexican-Americans (Melton *et al.*, 2013) and in the Chinese (Xie *et al*. 2015) revealed only suggestive signals.

A genome-wide interaction study identified *RCBTB1* as a modifier for a smoking effect on cIMT ($p = 2.9\text{x}10^{-9}$) (Wang *et al*. 2014). The genome-wide interaction of sex study in NOMAS identified *LEKR1* and *GALNT10* (genes associated with control of adiposity and weight) as modulators of a sex-difference in cIMT (Dong *et al.*, 2015).

More recently, in Mexican American and European American patients with rheumatoid arthritis (RA), Arya *et al.* identified novel associations between cIMT and variants in *SLC17A2* and *PPCDC*, and between arterial plaque lesions and variants in *COL4A1* and *SLCA13* that may point to new candidate risk loci for subclinical atherosclerosis associated with RA (Arya *et al.*, 2018). In a Spanish study, a 3'-untranslated genetic variant of *RARB* was associated with cIMT in rheumatoid arthritis patients (López-Mejías *et al.*, 2019).

The most recently published meta-analysis of GWAS in European ancestry populations for cIMT (71,128 individuals), identified eight novel susceptibility loci for cIMT (1q32.2 intergenic (rs201648240), *ATP6AP1L* (rs224904), *AIG1* (rs6907215), *PIK3CG* (rs13225723), *MCPH1* (rs2912063), *SGK223* (rs11785239), *VTI1* (rs1196033), and *CBFA2T3* (rs844396)), and one independent association at the previously-identified *PINX1* locus (Franceschini *et al.*, 2018). They also reported significant genetic correlations between cIMT with CHD, stroke subtypes and ischemic stroke (Franceschini *et al.*, 2018).

These studies together provide insights into genes and tissue-specific regulatory mechanisms linking atherosclerosis both to its functional genomic origins and its clinical consequences in humans. They also highlight the complexity of atherosclerosis. To date, a total of 76 SNPs have

been found to be associated at genome-wide significance (from $8x10^{-6}$ to 10^{-14}) with cIMT (GWAS Catalogue – accessed 29 May 2019) (Buniello *et al.*, 2019).

1.7. This PhD study

1.7.1. Study Context

Genetic variation and its role in studies of disease risk is necessary to enhance the comprehensive understanding of cardiometabolic diseases in sub-Saharan Africans. The Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen) is a National Institutes of Health (NIH) funded Collaborative Centre of the Human Heredity and Health in Africa (H3Africa) Consortium investigating the genomic and environmental risk factors for cardiometabolic diseases in Africans. The study is cross-sectional in nature and is investigating populations from six sub-Saharan African sites in four countries (Burkina Faso, Ghana, Kenya and South Africa) in East, West and Southern Africa (Ramsay *et al.*, 2016). Study participants complying with the study criteria and providing informed consent were enrolled for data collection from 2012 to 2016. The participants included approximately 12 000 sub-Saharan Africans form rural and urban settings and aged from 40 to 60 years with a parity between male and female individuals (Ramsay *et al.*, 2016). All participants completed a questionnaire with questions on demography, health history and behaviour. Anthropometric measurements were taken and fasting venous blood samples were collected for genotyping (H3Africa array) and phenotyping (biomarkers) purposes as described in the AWI-Gen resource paper and the detailed AWI-Gen methods paper (Ramsay *et al.*, 2016; Ali *et al.*, 2018). Ultrasound scans were performed to assess visceral adipose tissue thickness (VAT), sub-cutaneous abdominal adipose tissue thickness (SCAT) and the right and left common carotid intima-Media thickness (cIMT).

This PhD project is nested in the AWI-Gen study and used demographic, anthropometric, biomarker and GWAS array data to study genetic association with cIMT as a proxy for to atherosclerosis in sub-Saharan African populations.

1.7.2. Hypothesis, Objective and Aims

Hypothesis:

Genetic variation in sub-Saharan African populations, together with specific environmental factors, are associated with cIMT as a surrogate marker for atherosclerosis and demonstrate both universal and African-specific associations.

Main objective:

To investigate the contribution of genetic factors to cIMT in approximately 12 000 sub-Saharan African individuals aged between 40 and 60 years.

Specifics aims:

1. Characterise the distribution and relationship of associated risk factors for cIMT in the AWI-Gen data.
2. Perform a GWAS for cIMT in middle-aged adults in sub-Saharan Africa, firstly in the entire dataset and then stratified by sex.
3. Investigate the interaction of genetic variants and smoking in susceptibility to atherosclerosis in West African men, using cIMT as the outcome variable.

Prior to performing GWAS analyses, an understanding of the phenotype is critical. Therefore, we first undertook a characterization of cIMT in the populations of the AWI-Gen study: we characterised cIMT measurements and defined age-sex-site specific reference values for each population and for low-risk subsets of individuals. We then assessed the relationship of cIMT with risk factors to determine several potential confounders: using regression models to investigate factors associated with cIMT. Subsequently we performed a GWAS for cIMT as a quantitative trait using Bayesian linear mixed models with imputed genotypes, first for the whole cohort and then for the sex-stratified data sets. Finally, we investigated the gene-smoking interaction by performing a genome-wide interaction study using the West-African male subset from our study.

The power of the study was assessed using Quanto software by simulating data from the GWAS catalogue for cIMT (**Figure 1.4**). A model that assesses the cIMT in independent individuals with an additive genetic inheritance and an allele frequency of 0.04 will be > 93% powered ($\alpha = 5x10^{-8}$) to detect a βG (genetic effect) of 0.0147. Likewise, an allele frequency of 0.48 will have 98% power to detect even a very small genetic effect ($\beta = 0.0067$). Analyses per site, with only 2000 participants, would reduce the power to 81% to detect a $\beta G = 0.0147$ given an allele

frequency of 0.04. A threshold of $p<5x10^{-8}$ will be used for a genome-wide significance of associated signals.

This PhD is a bioinformatics research study focused on complex genetic association analyses using questionnaire, anthropometric, ultrasound and GWAS data generated for the AWI-Gen project.

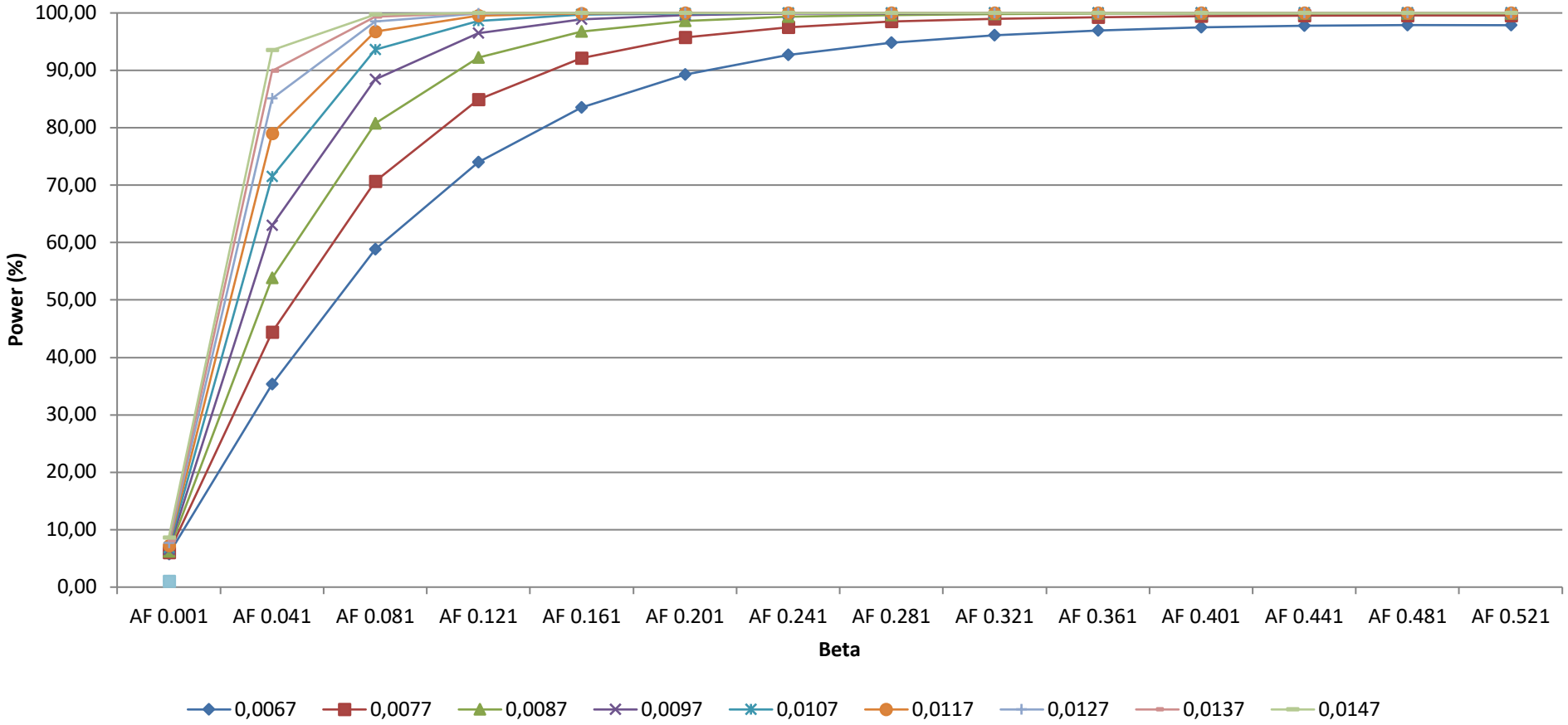


Figure 1.4: Estimation of power for the GWAS of cIMT (as a quantitative trait) for the sample size of 12 000 and considering genetic effect β ranging from 0.0067 to 0.156.

1.8. Ethical considerations

The AWI-Gen study has been approved by the Wits Human Research Ethics Committee (Medical) and the clearance number is **M121026.** Additionally, this PhD project has been approved by the Wits Human Research Ethics Committee (Medical) under the clearance number **M1706110** the same ethics committee for approval. Broad informed consent was provided by participants, to ensure that the data and samples can be utilised in future studies. Data will be submitted to EGA and access will be managed by the H3Africa database access committee (DBAC) according to the H3Africa guidelines and policies.

Chapter 2:

Regional differences in carotid intima-media thickness reveal higher risk for atherosclerosis in West Africa: An AWI-Gen Study

Chapter 2: Regional differences in carotid intima-media thickness reveal higher risk for atherosclerosis in West Africa: An AWI-Gen Study

ABSTRACT

Common carotid artery intima-media thickness (cIMT) is a widely used surrogate marker of atherosclerosis, considering its predictive association with cardiovascular disease (CVD). However, there is a paucity of data on cIMT in Sub-Saharan African (SSA) populations. Our aim is to describe cIMT and determine age-specific reference intervals in populations from SSA.

Methods

AWI-Gen participants aged 40 to 60 years were from Agincourt (rural South Africa), Dikgale (rural South Africa), Nairobi (urban Kenya), Nanoro (rural Burkina Faso), Navrongo (rural Ghana) and Soweto (urban South Africa). Ultrasound measurements were performed for cIMT and questionnaire data were collected.

Results

We analysed cIMT data from 8872 (4404 males, 4468 females) participants. In the combined sample, left carotid was higher than right carotid in females (0.643, CI = 0.639-0.646 vs 0.626, CI = 0.622-0.629; $p < 0.001$) and males (0.648, CI = 0.645-0.652 *vs* 0.624, CI = 0.621-0.628; $p < 0.001$). The differences between minimum and maximum cIMT were higher in West Africa. We identified regional differences with West Africans having overall higher cIMT than East and South Africans (mean cIMT respectively 0.660, IQR = 0.155 and 0.590, IQR = 0.127; $p > 0.001$). Sex differences were observed for mean cIMT in Nanoro (male higher than female ($p < 0.001$)) and Nairobi (female higher than male ($p = 0.0191$)). We determined site and sex stratified age-specific reference intervals (25th, 50th, 75th and 90th percentiles) for the general population and for a healthy (CVD risk-free) subset. The risk of increased cIMT (cIMT ≥ 75th percentile) was highest in West Africa and was influenced by traditional risk factors.

Conclusion

Our study determined age-specific reference intervals of cIMT, illustrating regional differences for atherosclerosis risk in SSA.

2.1. INTRODUCTION

It is estimated that by 2020 cardiovascular diseases (CVDs) will be the leading cause of death (Fuster, 2014), and the number of deaths due to CVDs will increase from 17.5 million in 2016 (Tzoulaki *et al.*, 2016) to more than 24 million in 2030 (Fuster, 2014). Sub-clinical atherosclerosis (SCA) (which presents at its early stages with no recognizable clinical findings) and atherosclerosis have been found to be the pillar supporting the burden of CVDs such as ischemic heart disease, stroke, heart failure, peripheral arterial disease and rheumatic heart disease. CVDs have increased considerably during the last two decades and low- and middle-income countries are now experiencing about 80% of the worldwide burden. Sub-Saharan Africa (SSA) is facing a health and demographic transition that has shifted the major causes of death from communicable and nutritional diseases to non-communicable diseases (NCDs). The mean age of death attributable to CVDs in SSA in 2010 was 64.9 years (95% CI, 64.4–65.4) compared with 67.6–81.2 years for the rest of the world (Roth *et al.*, 2015), making the region one of the youngest affected populations globally. The context of limited resources is raising challenges such as accurate diagnosis and disease management of CVDs, with poor health system readiness to tackle this epidemic.

Atherosclerosis precedes the onset of most of the clinical cardiovascular events and SCA is relatively common in young and middle-aged populations. Common carotid intima-media thickness (cIMT) is a widely-accepted imaging surrogate marker of generalized atherosclerosis. Measured by dual mode ultrasound, cIMT is represented by the distance between the near and the far wall of the carotid artery corresponding to the lumen-intima and the media-adventia interfaces. Even in the absence of atherosclerosis, the thickness of the intima and the media layer increases with advancing age because of biological changes. Because the molecular and cellular pathways responsible for these changes are also involved in the formation of atherosclerotic plaque, cIMT is a predictor for SCA. Therefore, measurement of cIMT is an attractive method for use in large-scale research (Bartels, Franco and Rundek, 2012). As a surrogate marker of atherosclerosis, there is a need for reference intervals of cIMT across different populations since this measure is important for diagnostic, therapeutic and prognostic purposes. Moreover, interpreting results at individual or population level relies on the accuracy of reference intervals.

Numerous populations across the world have been characterised for cIMT and reference intervals had been determined (Engelen *et al.*, 2013; Lee *et al.*, 2014; Santos *et al.*, 2014; Liu

et al., 2017), but few studies has been reported for SSA, most of them focused on patient data and/or with small sample sizes (Holland *et al.*, 2009; Okeahialam *et al.*, 2011; Rheeder *et al.*, 2012; Owolabi *et al.*, 2015; Schoffelen *et al.*, 2015; Olt *et al.*, 2016; Zemlin *et al.*, 2017). The Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen) is a Collaborative Centre of the Human Heredity and Health in Africa (H3Africa) Consortium investigating the genomic and environmental risk factors for cardiometabolic diseases in Africans. The study was cross-sectional and investigated populations from six SSA sites in East, West and South Africa including participants from four countries (Burkina Faso, Ghana, Kenya and South Africa) (Ramsay *et al.*, 2016). In this study, we report cIMT characteristics (minimum, maximum and mean for left and right cIMT) in these populations and present age-specific reference intervals for middle-aged adults (40-60 years) for mean cIMT (average of left and right).

2.2. MATERIALS AND METHODS

2.2.1. Participants

The study included SSAs from Agincourt (rural South Africa), Dikgale (rural South Africa), Nairobi (urban Kenya), Nanoro (rural Burkina Faso), Navrongo (rural Ghana) and Soweto (urban South Africa). They were randomly selected from Health and Demographic Surveillance Systems (HDSS) or by convenience sampling (Ali *et al.*, 2018). Inclusion criteria were that they reside in the area of the HDSS, are aged between 40–60 years, not closely related (i.e. first-degree relatives) and women should not be pregnant (Ramsay *et al.*, 2016). All participants completed a questionnaire with questions on demography, health history and behaviour, and anthropometric measurements were taken. More details on data collection procedures are described elsewhere (Ramsay *et al.*, 2016; Ali *et al.*, 2018).

2.2.2. cIMT measurements

cIMT is measured in Dual B-mode ultrasound images of the carotid tree as a typical double line of the arterial wall. cIMT is best visible in the measurement segment of the distal common carotid artery with lowest measurement variability. The measurement is most reliable over a one centimeter-segment and has been performed by semi-automatic reading methods, which minimises reading errors. The far wall of both the left and right common carotid artery were imaged using a linear-array 12L-RS transducer with a GE Healthcare B-mode LOGIQe ultrasound machine (GE, Healthcare, CT, USA). The participant was in a supine position for the measurements, head turned towards the left/right at a 45 degree angle. Operators used

anatomical landmarks to identify the common carotid artery (CCA) on a longitudinal plane and the image was frozen. The operator then identified a continuous one-centimeter segment (10 mm) of the CCA far wall. The operator then placed a cursor between two points (10 mm apart) on this identified segment of the far wall with the proximal starting point 1cm from the bulb of the CCA. The inbuilt software then automatically detected the intima-lumen and the media-adventitia interfaces and calculated the minimum, maximum and mean common cIMT in mm and to three decimal places. To measure the left carotid, the participant's head was turned to the opposite side, and the process was repeated.

Plaque was defined as a focal structure that encroaches into the arterial lumen of at least 0.5 mm or 50% of the surrounding cIMT value or demonstrates a thickness >1.5 mm as measured from the media-adventitia interface to the intima-lumen interface.

2.2.3. Other variables

Smokers were categorized as former smokers when they had a history of smoking at least once during their lifetimes, as never smokers when they were not actively smoking at the time inclusion into the study, or as current smokers. Hypertension was defined according to the JNC7 definition (Joint National Committee on Prevention Detection Evaluation and Treatment of High Blood Pressure, 2003): reporting the use of anti-hypertensive medications or having been diagnosed as hypertensive, or having a measured systolic blood pressure ≥ 140 mmHg or a diastolic blood pressure ≥ 90 mmHg. In the study blood pressure was measured three time and the average of the last two values used for analysis. Diabetes was defined as a medical history of diabetes mellitus, a fasting serum glucose ≥ 7 mmol/L, or use of medications for diabetes. High cholesterol was defined as total cholesterol level > 5 mmol/L. Abdominal obesity was defined as a waist circumference >88 cm in women and >102 cm in men. Previous cardiovascular disease (CVD) was defined by a previous medical diagnosis of myocardial infarction, transient ischemic stroke, stroke and/or angina pectoris. "Problematic drinking" was defined according to the CAGE questionnaire when a participant answered "yes" to at least two of the four questions (Cut-down, Annoyed, Guilty, Eyes-opener). Variables and measurements were described previously (Ali *et al.*, 2018).

We defined healthy individuals as those free of hypertension (JNC7 definition), diabetes or high cholesterol, who did not have previous CVDs, without abdominal obesity, were not current smokers, and had a BMI between 18.5 and 25 kg/m^2.

2.2.4. Data analysis

The distributions of continuous variables are reported as median and interquartile and categorical variables as percentages. To compare data across groups (e.g. sex differences), we used χ^2 and/or Fisher's test for categorical variables and Mann-Whitney-Wilcoxon test for continuous variables. The paired t-test was performed to compare right and left cIMT measurements, and ANOVA was performed to assess site differences. Mean cIMT was defined as the average of the mean right cIMT and mean left cIMT measurements. Mean Max cIMT was generated as the average of the maximum cIMT from the left and right. All analyses were site and sex stratified.

Calculation of age-specific reference intervals (RIs) for cIMT was conducted in two stages. First, we generated age-specific RIs for the study population (referred to as the general population). Second, we calculated age-specific RIs for the CVD risk-free population, referred to as the healthy sub-population. All RIs were sex and site specific. We used a parametric regression method based on fractional polynomials (FPs) as described by Royston and Wright (Royston and Wrightt, 1998) and implemented in STATA software (version 15.1 Stata Corp., College Station, TX, USA) (*XRIGLS* package). Briefly, cIMT data were assumed to be normally distributed, conditional on age. The best fitting FPs for the age-specific mean and standard deviation (SD) regression curves were defined using an iterative procedure (Generalized Least Squares - GLS). The advantages of using FPs is that the method involves transformations of the data towards normality which remove non-normal skewness and/or kurtosis. The method also accommodates parsimonious modelling of age- and sex-specific centile curves.

To assess risk factors associated with increased cIMT, we further generated a binary cIMT for participants with cIMT greater than or equal to the age-sex-site specific 75th percentile. Those participants were considered at increased risk of developing CVDs (Stein *et al.*, 2008). We performed a multiple logistic regression for increased cIMT.

2.2.5. Ethical Considerations

The AWI-Gen study was approved by the Wits Human Research Ethics Committee (Medical) under clearance number is **M121026**, this sub-study was approved by Wits Human Research Ethics Committee (Medical) (**M1706110**). All enrolled participants signed informed consent and study procedures were compliant with the Declaration of Helsinki.

2.3. RESULTS

2.3.1. Populations characteristics

Of the 10 702 participants, 8872 (4404 males, 4468 females) were analysed in this study after implementing stringent quality control. A summary of the population's characteristics (site and sex stratified) is presented in **Table 2.1**. Median ages of our population were 50.0 years (IQR=10) for men and women. Sex difference were observed for all variables except for total cholesterol, hypertension and diabetes, when analysing the whole cohort. Visceral fat was higher in men whereas abdominal subcutaneous fat was higher in women. The combined healthy subset was including 1303 (29.16%) women and 1198 (26.41%) men (**Supplementary Table 2.1**). The prevalence of healthy individuals per site was 47.4%, 41.6%, 22.7%, 15.0%, 10.5% and 6.5% respectively for Nanoro, Navrongo, Nairobi, Agincourt, Dikgale and Soweto.

2.3.2. cIMT characterisitics

There was a significant difference between the values of cIMT from the left and right carotid, with the left carotid being higher than right carotid in both women (0.621, IQR = 0.152 vs 0.610, IQR = 0.160; $p < 0.001$) and men (0.630, IQR = 0.170 vs 0.600, IQR = 0.165; $p<0.001$) at all sites, except in Nanoro ($p = 0.1412$) and Dikgale ($p = 0.4458$) (**Table 2.2**). The sex difference was not significant for the combined cohort, but men in Nanoro had higher mean cIMT ($p < 0.001$) whereas women in Nairobi had higher mean cIMT ($p = 0.0191$). In South Africa there were no differences between cIMT measurements of rural (Agincourt and Dikgale) and urban (Soweto) male populations ($p = 0.2335$). Differences between minimum cIMT and maximum cIMT were higher in the West African sites and highest in Navrongo (**Supplementary Figure 2.1a, b**). In our cohort, mean cIMT values were 0.618 (IQR = 0.142) for women and 0.620 (IQR = 0.157) for men without sex-differences ($p = 0.9127$). When analyses where stratified per site, sex and health status, we found significantly higher values of mean cIMT in non-healthy women compared to healthy women at every site. Healthy men displayed lower mean cIMT values than non-healthy men in all sites, except in West Africa (Nanoro, Navrongo) (**Table 2.3**). When sites were combined, no significant differences were observed for mean cIMT between healthy vs non-healthy women and men (respectively $p=0.6104$ and $p=0.2397$). There was a regional difference between West vs East-South Africa (with mean cIMT respectively 0.660, IQR=0.155 and 0.590, IQR=0.127; $p>0.001$).

Table 2.1: Characteristics of the AWI-Gen study population by site and sex

	Agincourt					Digkale					Nairobi					Nanoro					Navrongo					Soweto	All males		All Females		
	Male (N=504)		Female (N=723)		p-value	Male (N=348)		Female (N=767)		p-value	Male (N=856)		Female (N=1,034)		p-value	Male (N=1,029)		Female (N=1,024)		p-value	Male (N=775)		Female (N=920)		p-value	Male (N= 892)	N=4404		N=4468		p-value
Age (years)	51.00	11.00	52.00	10.00	0.7999	50.00	10.00	51.00	10.00	0.1446	48.00	9.00	47.00	8.00	**0.0479**	50.00	11.00	50.00	9.00	0.9773	51.00	10.00	52.00	9.00	**0.0019**	49.00	10.00	50.00	10.00	50.00	10.00	**0.0145**
Bmi	23.32	6.51	28.57	9.14	**p<0.001**	20.58	5.24	29.91	9.89	**p<0.001**	22.25	4.96	26.84	8.60	**p<0.001**	21.08	4.19	19.75	3.47	**p<0.001**	20.55	3.16	21.52	4.09	**p<0.001**	23.95	7.81	21.66	5.20	23.81	9.29	**p<0.001**
Systolic blood pressure	129.50	27.00	130.50	28.00	0.0821	122.25	25.50	123.00	27.50	0.1736	119.00	24.00	113.00	26.50	**p<0.001**	117.50	22.00	108.00	21.50	**p<0.001**	121.50	25.00	119.50	27.25	**0.0235**	129.50	26.25	122.0(	25.00	118.0(	28.50	**p<0.001**
Diastolic blood pressure	79.00	17.00	79.50	16.50	0.2670	77.50	16.00	81.00	15.50	**p<0.001**	77.00	14.50	76.50	17.00	0.6991	75.00	14.00	70.00	11.50	**p<0.001**	75.50	17.00	75.50	16.00	0.4023	88.25	16.50	78.50	17.00	76.00	16.50	**p<0.001**
Total cholesterol	3.70	1.30	4.12	1.31	**p<0.001**	3.90	1.29	4.21	1.40	**p<0.001**	4.16	1.40	4.34	1.31	**0.002**	3.54	1.27	3.09	1.03	**p<0.001**	3.11	1.08	3.19	1.14	**0.0424**	4.01	1.38	3.71	1.41	3.70	1.52	0.7375
Triglycerides	0.78	0.51	0.78	0.46	0.6120	0.95	0.64	0.96	0.60	0.8048	0.93	0.62	0.92	0.54	0.7489	0.70	0.44	0.67	0.36	**p<0.0015**	0.56	0.31	0.55	0.30	0.8815	0.83	0.57	0.76	0.53	0.76	0.50	**0.0440**
Ldl (Friedwald)	2.08	1.12	2.49	1.17	**p<0.001**	2.18	1.08	2.49	1.11	**p<0.001**	2.41	1.16	2.58	1.12	**p<0.001**	1.96	1.01	1.69	0.83	**p<0.001**	1.68	0.85	1.86	0.87	**p<0.001**	2.43	1.15	2.10	1.16	2.16	1.19	**0.001**
Hdl	1.10	0.46	1.16	0.41	0.2697	1.17	0.51	1.14	0.39	0.1107	1.17	0.52	1.17	0.47	0.9877	1.12	0.44	1.00	0.37	**p<0.001**	1.14	0.49	1.08	0.43	**0.0116**	1.15	0.53	1.14	0.49	1.11	0.42	**p<0.001**
Glucose	4.75	1.12	4.71	0.87	0.4590	4.66	1.03	4.82	0.95	**0,010**	5.05	0.85	5.06	0.96	0.2845	4.92	0.92	4.90	0.82	0.8494	4.41	0.89	4.53	0.81	**0.0062**	5.04	0.82	4.87	0.97	4.81	0.92	**0.0487**
Serum Creatinin	72.75	19.78	64.39	17.16	**p<0.001**	73.80	19.53	66.61	19.55	**p<0.001**	71.04	17.45	61.62	16.10	**p<0.001**	69.65	16.52	57.32	13.20	**p<0.001**	72.27	17.77	61.62	13.41	**p<0.001**	71.29	20.31	71.43	18.42	61.49	16.51	**p<0.001**
MVPA (min/week)	720	1260	600	1230	0.6972	1200	2085	1260	1395	0.9444	1260	2640	900	1860	**p<0.001**	2160	2910	2940	3060	**p<0.001**	2430	2160	1620	2332	**p<0.001**	645	1590	1200	2520	1260	2430	**0.0225**
Visceral fat	6.29	2.62	5.81	2.97	**0.0012**	5.88	2.64	6.58	3.46	**p<0.001**	4.94	2.32	4.60	2.22	**p<0.001**	4.33	1.68	4.26	1.54	0.8113	4.07	1.45	3.32	1.30	**p<0.001**	6.22	2.84	4.97	2.49	4.53	2.55	**p<0.001**
Subcutaneous fat	1.10	0.85	2.13	1.34		0.82	0.69	2.19	1.32	**p<0.001**	1.01	0.77	1.95	0.97	**p<0.001**	0.81	0.60	0.89	0.58	**0.0075**	0.70	0.31	1.01	0.69	**p<0.001**	1.38	1.04	0.91	0.75	1.49	1.32	**p<0.001**
Htn																																
Yes	222	44.05	419	57.95	**p<0.001**	111	31.90	373	48.63	**p<0.001**	191	22.31	304	29.40	**p<0.001**	206	20.02	124	12.11	**p<0.001**	178	22.97	220	23.91	0.647	460	51.57	1368	31.06	1440	32.23	0.238
Smoking status																																
Never Smoked	264	52.59	715	98.89	**p<0.001**	53	15.27	713	93.08	**p<0.001**	406	47.49	955	92.36	**p<0.001**	766	74.59	1018	99.80	**p<0.001**	276	35.61	889	96.74	**p<0.001**	267	30.00	2032	46.22	4290	96.15	**p<0.001**
Current Smoker	132	26.29	2	0.28		220	63.40	24	3.13		198	23.16	27	2.61		141	13.73				330	42.58	18	1.96		483	54.27	1504	34.21	71	1.59	
Former Smoker	106	21.12	6	0.83		74	21.33	29	3.79		251	29.36	52	5.03		120	11.68	2	0.20		169	21.81	12	1.31		140	15.73	860	19.56	101	2.26	
Diabetes																																
Yes	32	6.50	42	5.94	0.690	27	7.85	79	10.51	0.167	33	3.89	90	8.78	**p<0.001**	52	5.11	23	2.27	**0.001**	11	1.47	12	1.35	0.836	55	6.25	210	4.85	246	5.61	0.112
Socio-economical status					**p<0.001**					**0.003**					**p<0.001**					**p<0.001**					**p<0.001**						**p<0.001**
1st Quintile	103	20.44	86	11.89		64	18.39	80	10.44		92	10.75	137	13.25		133	12.95	199	19.53		117	15.10	188	20.46		23	2.58	532	12.09	690	15.47	
2nd Quintile	133	26.39	157	21.72		83	23.85	168	21.93		161	18.81	262	25.34		211	20.55	183	17.96		131	16.90	175	19.04		84	9.42	803	18.24	945	21.18	
3rd Quintile	59	11.71	94	13.00		42	12.07	108	14.10		203	23.71	236	22.82		190	18.50	209	20.51		146	18.84	192	20.89		135	15.13	775	17.61	839	18.81	
4th Quintile	116	23.02	179	24.76		65	18.68	179	23.37		166	19.39	218	21.08		182	17.72	201	19.73		192	24.77	211	22.96		241	27.02	962	21.85	988	22.15	
5th Quintile	93	18.45	207	28.63		94	27.01	231	30.16		234	27.34	181	17.50		311	30.28	227	22.28		189	24.39	153	16.65		409	45.85	1330	30.21	999	22.39	
Hiv status																																
Yes	168	33.40	255	35.32	0.487	71	20.94	169	22.32	0.609	64	8.79	170	17.84	**p<0.001**	4	0.39	4	0.39	0.994	6	0.77	6	0.65	0.765	176	20.78	489	11.58	604	13.80	**0.002**
Highest level of education					**0.001**					0.147					**p<0.001**					**p<0.001**					**p<0.001**						**p<0.001**
No Formal education	113	22.47	239	33.06		22	6.32	69	9.01		33	3.86	110	10.64		749	73.07	945	93.10		484	62.61	714	78.03		8	0.90	1,41	32.04	2,08	46.64	
Primary	193	38.37	256	35.41		108	31.03	259	33.81		434	50.70	650	62.86		177	17.27	58	5.71		170	21.99	145	15.85		104	11.66	1,19	26.97	1,37	30.72	
Secondary	159	31.61	181	25.03		202	58.05	416	54.31		369	43.11	271	26.21		83	8.10	10	0.99		97	12.55	47	5.14		646	72.42	1,56	35.39	925	20.77	
Tertiary	38	7.55	47	6.50		16	4.60	22	2.87		20	2.34	3	0.29		16	1.56	2	0.20		22	2.85	9	0.98		134	15.02	246	5.59	83	1.86	

Continuous variables are presented as median and interquartile. Mann-Whitney Wilcoxon test was used to assess sex-specific differences

Categorical variables are presented in number and percentage. Chi-squared test was used to assess sex-specific differences

Table 2.2: Characteristics of cIMT measurements per site and for men and women in the combined AWI-Gen cohort

	Agincourt					Digkale					Nairobi					Nanoro					Navrongo					Soweto					All Men		All Women		
	Men (N=504)		Women (N=723)		p-value (a)	Men (N=348)		Women (N=767)		p-value (a)	Men (N=856)		Women (N=1,034)		p-value (a)	Men (N=1,029)		Women (N=1,024)		p-value (a)	Men (N=775)		Women (N=920)		p-value (a)	Men (N= 892)		Women (N=1,006)	p-value (a)	N=4404			N=4468		p-value (a)
	Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR			Medians	IQR	Medians	IQR	
min cimt right carotid	0.480	0.120	0.480	0.120	0.495	0.520	0.140	0.550	0.160	0.268	0.470	0.140	0.480	0.150	**0.002**	0.560	0.170	0.520	0.130	**p<0.001**	0.480	0.160	0.480	0.160	0.265	0.470	0.150	-----		0.510	0.160	0.510	0.150	0.078
min cimt left carotid	0.520	0.130	0.520	0.110	0.853	0.550	0.145	0.520	0.120	0.282	0.480	0.130	0.510	0.120	**0.001**	0.580	0.180	0.520	0.150	**p<0.001**	0.520	0.160	0.520	0.160	0.805	0.480	0.130	-----		0.520	0.160	0.520	0.160	0.602
p-value (b)	**p<0.001**		**p<0.001**			0.8428		**p<0.001**			**p<0.001**		**p<0.001**			**p<0.001**		**0.026**			**p<0.001**		**p<0.001**			**p<0.001**				**p<0.001**		**p<0.001**		
mean cimt right carotid	0.575	0.130	0.580	0.135	0.463	0.615	0.137	0.620	0.155	0.227	0.545	0.135	0.560	0.140	**0.015**	0.660	0.160	0.640	0.160	**p<0.001**	0.655	0.175	0.655	0.180	0.522	0.575	0.150	-----		0.600	0.165	0.610	0.160	0.197
mean cimt left carotid	0.600	0.140	0.610	0.120	0.196	0.608	0.140	0.600	0.150	0.780	0.580	0.145	0.590	0.130	**0.048**	0.675	0.185	0.640	0.160	**p<0.001**	0.695	0.180	0.695	0.170	0.919	0.600	0.160	-----		0.630	0.170	0.620	0.152	0.178
p-value (b)	**p<0.001**		**p<0.001**			0.4458		**p<0.001**			**p<0.001**		**p<0.001**			**p<0.001**		0.141			**p<0.001**		**p<0.001**			**p<0.001**				**p<0.001**		**p<0.001**		
max cimt right carotid	0.670	0.150	0.670	0.150	0.114	0.680	0.165	0.710	0.160	0.221	0.640	0.160	0.640	0.170	0.112	0.760	0.170	0.750	0.170	**0.035**	0.820	0.220	0.830	0.230	0.627	0.680	0.200	-----		0.720	0.200	0.720	0.190	0.342
max cimt left carotid	0.680	0.130	0.710	0.150	**0.022**	0.680	0.145	0.680	0.180	0.698	0.680	0.160	0.680	0.160	0.223	0.760	0.200	0.750	0.170	**p<0.001**	0.870	0.200	0.845	0.200	0.775	0.720	0.200	-----		0.730	0.200	0.720	0.200	0.081
p-value (b)	**p<0.001**		**p<0.001**			**0.0344**		**p<0.001**			**p<0.001**		**p<0.001**			**p<0.001**		0.588			**p<0.001**		**p<0.001**			**p<0.001**				**p<0.001**		**p<0.001**		
mean cIMT (c)	0.587	0.120	0.595	0.113	0.321	0.610	0.135	0.618	0.137	0.571	0.567	0.130	0.580	0.127	**0.019**	0.668	0.158	0.637	0.147	**p<0.001**	0.670	0.157	0.670	0.156	0.823	0.592	0.146	-----		0.620	0.157	0.618	0.142	0.913
mean Max cIMT (d)	0.660	0.125	0.680	0.125	**0.036**	0.680	0.135	0.695	0.150	0.353	0.660	0.145	0.675	0.140	0.159	0.760	0.175	0.740	0.165	**p<0.001**	0.840	0.185	0.840	0.185	0.919	0.700	0.160	-----		0.725	0.195	0.720	0.175	0.508

(a) Mann-Whitney Wilcoxon test has been used to assess sex-specific differences

(b) Paired t-test was used to compare measurements on left and right carotid

(c) Mean cIMT : average of means from left and right carotids

(d) Mean Max cIMT: average of maximum cIMT from left and right carotids

IQR: Interquartile

Table 2.3: Characteristics of cIMT measurements per site and stratified by cardiovascular risk (Healthy vs Non-healthy) for men and women

Men	Agincourt					Digkale					Nairobi					Nanoro					Navrongo					Soweto					All Men				
	N=504				p-value (a)	N=348				p-value (a)	N=856				p-value (a)	N=1,029				p-value (a)	N=775				p-value (a)	N= 892				p-value (a)	N=4404				p-value (a)
	Healthy=97		non-Healthy=407			Healthy=39		non-Healthy=309			Healthy=263		non-Healthy=593			Healthy=455		non-Healthy=574			Healthy=251		non-Healthy=524			Healthy=58		non-Healthy=834			Healthy=1163		non-Healthy=3241		
	Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR	
min cimt right carotid	0.480	0.090	0.480	0.120	**0.023**	0.500	0.050	0.550	0.150	**0.001**	0.440	0.120	0.480	0.160	**0.001**	0.560	0.180	0.560	0.170	0.743	0.480	0.160	0.480	0.190	**0.041**	0.440	0.110	0.470	0.150	0.244	0.510	0.160	0.510	0.160	0.799
min cimt left carotid	0.520	0.120	0.520	0.120	**0.005**	0.480	0.080	0.550	0.160	**p<0.001**	0.480	0.120	0.480	0.120	**0.004**	0.560	0.200	0.590	0.170	0.353	0.520	0.160	0.520	0.160	0.819	0.475	0.150	0.480	0.130	0.058	0.520	0.160	0.520	0.160	0.582
mean cimt right carotid	0.555	0.115	0.580	0.140	**0.010**	0.570	0.065	0.620	0.140	**p<0.001**	0.520	0.135	0.560	0.140	**p<0.001**	0.660	0.160	0.660	0.155	0.616	0.630	0.195	0.660	0.170	**0.018**	0.555	0.120	0.575	0.145	0.059	0.600	0.165	0.600	0.165	0.534
mean cimt left carotid	0.580	0.105	0.620	0.125	**p<0.001**	0.560	0.085	0.620	0.140	**p<0.001**	0.560	0.135	0.595	0.150	**0.004**	0.665	0.185	0.680	0.185	0.248	0.690	0.180	0.695	0.175	0.375	0.577	0.120	0.605	0.160	**0.006**	0.620	0.175	0.630	0.165	0.280
max cimt right carotid	0.640	0.090	0.680	0.160	**0.008**	0.640	0.080	0.710	0.180	**p<0.001**	0.620	0.150	0.640	0.160	**0.001**	0.760	0.160	0.760	0.190	0.669	0.800	0.240	0.830	0.200	**0.033**	0.665	0.120	0.680	0.190	**0.024**	0.720	0.190	0.720	0.200	0.440
max cimt left carotid	0.640	0.120	0.680	0.140	**p<0.001**	0.640	0.120	0.680	0.160	**p<0.001**	0.640	0.120	0.680	0.160	**0.001**	0.760	0.190	0.790	0.200	0.199	0.870	0.200	0.870	0.215	0.149	0.680	0.130	0.720	0.200	**0.002**	0.720	0.200	0.740	0.200	0.244
mean cIMT (b)	0.567	0.103	0.598	0.123	**p<0.001**	0.567	0.068	0.623	0.147	**p<0.001**	0.548	0.103	0.572	0.135	**p<0.001**	0.665	0.155	0.674	0.160	0.358	0.655	0.157	0.680	0.150	0.096	0.574	0.100	0.595	0.142	**0.010**	0.618	0.155	0.620	0.155	0.239
mean Max cIMT (c)	0.640	0.100	0.680	0.135	**p<0.001**	0.640	0.075	0.695	0.145	**p<0.001**	0.640	0.115	0.675	0.155	**p<0.001**	0.760	0.165	0.772	0.170	0.289	0.820	0.205	0.840	0.178	0.052	0.680	0.115	0.707	0.170	**0.003**	0.720	0.180	0.730	0.195	0.178
Women	Agincourt					Digkale					Nairobi					Nanoro					Navrongo					Soweto					All Women				
	N=723				p-value (a)	N=767				p-value (a)	N=1,034				p-value (a)	N=1,024				p-value (a)	N=920				p-value (a)						N=4468				p-value (a)
	Healthy=87		non-Healthy=636			Healthy=78		non-Healthy=689			Healthy=166		non-Healthy=868			Healthy=518		non-Healthy=506			Healthy=454		non-Healthy=466								Healthy=1303		non-Healthy=3165		
	Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR		Medians	IQR	Medians	IQR							Medians	IQR	Medians	IQR	
min cimt right carotid	0.480	0.090	0.480	0.120	**0.027**	0.520	0.160	0.560	0.160	**p<0.001**	0.445	0.120	0.480	0.145	**0.036**	0.520	0.160	0.520	0.150	**0.028**	0.480	0.160	0.480	0.190	0.119	------		------			0.480	0.130	0.520	0.160	**p<0.001**
min cimt left carotid	0.480	0.110	0.520	0.120	**p<0.001**	0.505	0.120	0.520	0.140	**p<0.001**	0.480	0.150	0.510	0.120	**p<0.001**	0.520	0.160	0.520	0.160	0.057	0.520	0.160	0.530	0.160	0.287	------		------			0.520	0.160	0.520	0.130	**p<0.001**
mean cimt right carotid	0.550	0.100	0.580	0.140	**0.003**	0.580	0.125	0.630	0.160	**p<0.001**	0.540	0.120	0.575	0.150	**0.015**	0.623	0.140	0.640	0.165	**0.005**	0.642	0.170	0.660	0.200	0.059	------		------			0.610	0.160	0.610	0.160	0.653
mean cimt left carotid	0.575	0.110	0.620	0.125	**p<0.001**	0.580	0.100	0.610	0.150	**p<0.001**	0.560	0.120	0.600	0.130	**p<0.001**	0.635	0.160	0.640	0.170	0.063	0.683	0.155	0.700	0.190	**0.037**	------		------			0.630	0.165	0.620	0.145	0.567
max cimt right carotid	0.630	0.160	0.680	0.155	**p<0.001**	0.680	0.160	0.720	0.160	**p<0.001**	0.630	0.160	0.670	0.160	**0.017**	0.750	0.180	0.760	0.180	**0.003**	0.800	0.200	0.840	0.240	**0.029**	------		------			0.740	0.200	0.710	0.180	**p<0.001**
max cimt left carotid	0.660	0.120	0.720	0.150	**p<0.001**	0.640	0.100	0.680	0.170	**p<0.001**	0.640	0.130	0.680	0.160	**p<0.001**	0.740	0.200	0.760	0.160	0.102	0.840	0.200	0.870	0.210	**0.001**	------		------			0.750	0.230	0.720	0.190	**p<0.001**
mean cIMT (b)	0.565	0.103	0.600	0.110	**p<0.001**	0.570	0.105	0.620	0.142	**p<0.001**	0.550	0.123	0.585	0.127	**p<0.001**	0.630	0.142	0.640	0.148	**0.010**	0.659	0.150	0.680	0.167	**0.016**	------		------			0.620	0.150	0.618	0.140	0.610
mean Max cIMT (c)	0.650	0.125	0.690	0.125	**p<0.001**	0.652	0.105	0.700	0.150	**p<0.001**	0.647	0.135	0.675	0.150	**0.001**	0.740	0.165	0.755	0.160	**0.008**	0.825	0.185	0.858	0.185	**0.002**	------		------			0.740	0.185	0.720	0.175	**p<0.001**

(a) Mann-Whitney Wilcoxon test has been used to assess sex-specific differences

(b) Mean cIMT : average of means from left and right carotids

(c) Mean Max cIMT: average of maximum cIMT from left and right carotids

IQR: Interquartile range

2.3.3. Age-specific reference intervals (RIs)

We generated percentiles (25th, 50th,75th and 90th) for sex-age-specific RIs for the whole cohort (1-year intervals) considering the general population and healthy population (**Table 2.4**) and for each site (**Supplementary Tables 2.2-2.7**). We reported the site-sex-age-specific references intervals (5-year intervals) (**Table 2.5**). The regional differences between West vs East-South Africa were more pronounced at the 50th RI percentiles by site and when subset to the healthy population (**Figure 2.1**). The age-related increase of mean cIMT was higher in the West African sites for men. In healthy women from Nanoro, the age-related increase was higher than in Navrongo.

Table 2.4: Age- and sex-specific (1-year intervals) percentiles of references intervals for carotid artery intima-media thickness (in mm) in the general population and the healthy sub-population for the AWI-Gen study.

	Total references populations										Healthy reference populations									
	Male					Female					Male					Female				
	Sample					Sample					Sample					Sample				
age	size	25th	50th	75th	90th	size	25th	50th	75th	90th	size	25th	50th	75th	90th	size	25th	50th	75th	90th
40	136	0.507	0.565	0.623	0.676	90	0.509	0.563	0.617	0.666	39	0.507	0.565	0.623	0.675	27	0.508	0.566	0.624	0.676
41	224	0.513	0.572	0.632	0.686	198	0.515	0.570	0.625	0.675	77	0.513	0.573	0.632	0.685	77	0.514	0.573	0.632	0.685
42	257	0.519	0.580	0.641	0.696	252	0.521	0.577	0.634	0.684	84	0.519	0.580	0.641	0.695	82	0.521	0.581	0.640	0.694
43	224	0.525	0.587	0.649	0.705	226	0.527	0.584	0.642	0.694	65	0.526	0.588	0.650	0.705	77	0.528	0.588	0.649	0.703
44	219	0.531	0.595	0.658	0.715	193	0.532	0.591	0.650	0.704	68	0.532	0.595	0.659	0.716	65	0.534	0.596	0.657	0.713
45	233	0.537	0.602	0.667	0.725	247	0.538	0.599	0.659	0.713	62	0.538	0.603	0.668	0.726	76	0.540	0.603	0.666	0.722
46	226	0.543	0.609	0.676	0.735	237	0.544	0.606	0.667	0.723	69	0.544	0.610	0.677	0.736	64	0.547	0.611	0.674	0.732
47	207	0.549	0.617	0.684	0.745	221	0.550	0.613	0.676	0.732	46	0.550	0.618	0.686	0.747	89	0.553	0.618	0.683	0.741
48	232	0.555	0.624	0.693	0.756	205	0.556	0.620	0.684	0.742	58	0.556	0.625	0.695	0.757	59	0.560	0.626	0.691	0.751
49	202	0.561	0.631	0.702	0.766	218	0.562	0.627	0.693	0.752	53	0.562	0.633	0.704	0.768	68	0.566	0.633	0.700	0.760
50	206	0.567	0.639	0.711	0.776	225	0.567	0.635	0.702	0.762	42	0.568	0.640	0.713	0.778	51	0.572	0.641	0.709	0.770
51	233	0.572	0.646	0.720	0.786	227	0.573	0.642	0.710	0.772	63	0.574	0.648	0.722	0.789	65	0.579	0.648	0.717	0.780
52	239	0.578	0.653	0.729	0.797	264	0.579	0.649	0.719	0.782	58	0.580	0.655	0.731	0.799	76	0.585	0.656	0.726	0.789
53	212	0.584	0.661	0.738	0.807	227	0.585	0.656	0.728	0.792	67	0.585	0.663	0.740	0.810	59	0.591	0.663	0.735	0.799
54	181	0.590	0.668	0.747	0.818	203	0.590	0.663	0.736	0.802	43	0.591	0.670	0.750	0.821	61	0.598	0.671	0.743	0.809
55	223	0.595	0.676	0.756	0.828	264	0.596	0.670	0.745	0.812	48	0.597	0.678	0.759	0.832	71	0.604	0.678	0.752	0.819
56	210	0.601	0.683	0.765	0.839	203	0.602	0.678	0.754	0.822	50	0.603	0.685	0.768	0.843	64	0.610	0.685	0.761	0.829
57	204	0.607	0.690	0.774	0.849	205	0.607	0.685	0.763	0.832	46	0.608	0.693	0.778	0.854	43	0.616	0.693	0.770	0.839
58	172	0.612	0.698	0.783	0.860	180	0.613	0.692	0.771	0.843	41	0.614	0.701	0.787	0.865	46	0.622	0.700	0.778	0.849
59	171	0.618	0.705	0.792	0.871	203	0.618	0.699	0.780	0.853	38	0.620	0.708	0.796	0.876	46	0.629	0.708	0.787	0.859
60	193	0.623	0.712	0.802	0.882	180	0.624	0.706	0.789	0.864	46	0.625	0.716	0.806	0.887	37	0.635	0.715	0.796	0.869

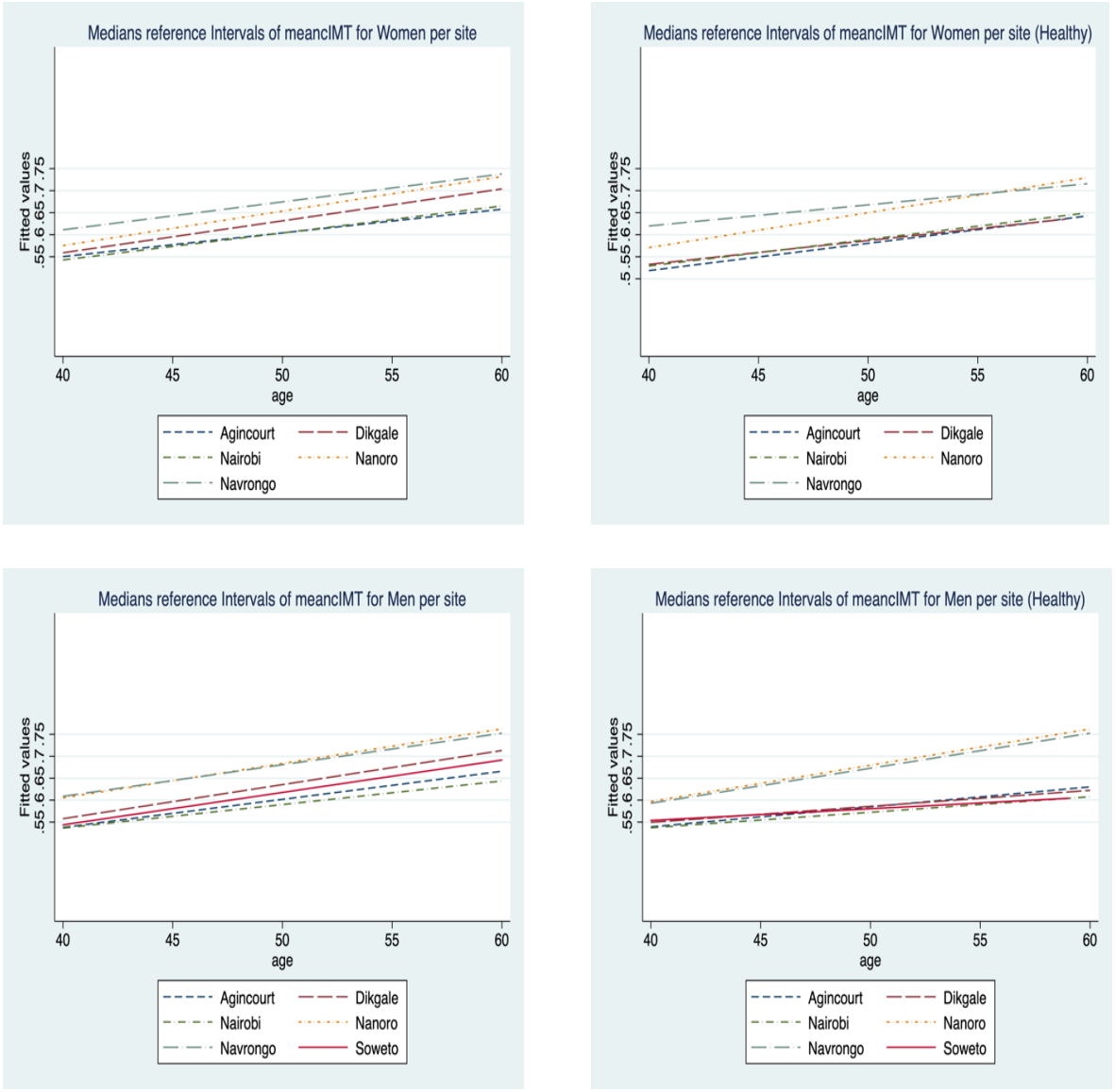


Figure 2.1: Comparison of age-specific reference intervals of median (50th percentile) per site.

(A) Comparison of age-specific medians RIs for women in the general population. (B) Comparison of age-specific medians RIs for women in the healthy population. (C) Comparison of age-specific medians RIs for men in the general population. (D) Comparison of age-specific medians RIs for women in the healthy population.

Table 2.5: Age- and sex-specific (5-year intervals) percentiles of references intervals for carotid artery intima-media thickness (in mm) in the general population and the healthy sub-population

		Male					Female				
	age group	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
Agincourt											
	45	132	0.509	0.554	0.599	0.640	165	0.518	0.565	0.613	0.655
Total	50	85	0.536	0.588	0.640	0.686	147	0.542	0.593	0.643	0.689
population	55	152	0.563	0.622	0.681	0.734	227	0.566	0.620	0.674	0.723
	60	135	0.589	0.656	0.722	0.782	184	0.590	0.648	0.706	0.758
CVD risk	45	44	0.511	0.550	0.588	0.623	31	0.495	0.535	0.574	0.610
free	50	17	0.529	0.574	0.620	0.661	18	0.524	0.566	0.608	0.646
reference	55	14	0.546	0.599	0.653	0.701	24	0.553	0.597	0.642	0.682
populations	60	22	0.562	0.624	0.686	0.742	14	0.581	0.629	0.676	0.719
Digkale											
	45	99	0.528	0.578	0.629	0.675	185	0.526	0.579	0.631	0.678
Total	50	80	0.556	0.619	0.681	0.737	190	0.556	0.616	0.676	0.730
population	55	94	0.584	0.659	0.733	0.801	198	0.584	0.653	0.722	0.784
	60	75	0.610	0.699	0.787	0.867	194	0.612	0.690	0.769	0.839
CVD risk	45	16	0.519	0.558	0.598	0.633	25	0.506	0.550	0.595	0.634
free	50	8	0.532	0.566	0.600	0.631	21	0.530	0.575	0.620	0.661
reference	55	10	0.546	0.574	0.602	0.627	20	0.554	0.600	0.646	0.688
populations	60	5	0.560	0.582	0.603	0.623	12	0.578	0.625	0.673	0.715
Nairobi											
	45	307	0.498	0.554	0.610	0.660	387	0.504	0.560	0.616	0.667
Total	50	214	0.517	0.579	0.641	0.696	303	0.529	0.591	0.653	0.710
population	55	202	0.536	0.604	0.672	0.734	226	0.553	0.622	0.691	0.753
	60	133	0.554	0.629	0.705	0.772	118	0.576	0.653	0.730	0.799
CVD risk	45	109	0.494	0.547	0.599	0.646	84	0.496	0.544	0.593	0.636
free	50	68	0.510	0.565	0.619	0.668	47	0.517	0.576	0.635	0.688
reference	55	60	0.526	0.583	0.640	0.691	28	0.538	0.608	0.678	0.741
populations	60	26	0.541	0.601	0.661	0.715	7	0.557	0.640	0.722	0.797
Nanoro											
	45	296	0.556	0.625	0.693	0.754	291	0.540	0.596	0.652	0.702
Total	50	261	0.594	0.666	0.738	0.803	254	0.573	0.636	0.700	0.757
population	55	247	0.631	0.707	0.784	0.853	275	0.605	0.677	0.749	0.814
	60	225	0.667	0.749	0.830	0.904	204	0.636	0.717	0.799	0.872
CVD risk	45	132	0.552	0.619	0.685	0.745	167	0.535	0.592	0.649	0.701
free	50	97	0.590	0.660	0.731	0.794	138	0.569	0.633	0.697	0.755
reference	55	124	0.627	0.702	0.776	0.844	122	0.602	0.674	0.746	0.811
populations	60	102	0.663	0.743	0.823	0.894	91	0.635	0.715	0.795	0.868
Navrongo											
	45	176	0.566	0.628	0.690	0.746	178	0.565	0.629	0.693	0.751
Total	50	208	0.597	0.665	0.734	0.795	212	0.592	0.662	0.731	0.794
population	55	196	0.626	0.702	0.778	0.846	259	0.618	0.694	0.770	0.839
	60	195	0.656	0.739	0.823	0.898	271	0.644	0.727	0.810	0.884
CVD risk	45	71	0.553	0.613	0.674	0.729	97	0.571	0.635	0.699	0.757
free	50	62	0.588	0.655	0.723	0.784	107	0.591	0.659	0.727	0.789
reference	55	59	0.622	0.697	0.772	0.840	138	0.610	0.683	0.756	0.821
populations	60	59	0.656	0.739	0.822	0.897	112	0.628	0.706	0.785	0.855
Soweto											
	45	283	0.509	0.562	0.615	0.662					
Total	50	225	0.539	0.600	0.662	0.718					
population	55	197	0.568	0.639	0.711	0.776					
	60	187	0.596	0.678	0.761	0.835					
CVD risk	45	23	0.515	0.566	0.618	0.664					
free	50	16	0.526	0.575	0.623	0.666					
reference	55	12	0.538	0.583	0.628	0.668					
populations	60	7	0.551	0.592	0.633	0.669					
All											
	45	1,293	0.522	0.584	0.647	0.703	1,206	0.525	0.582	0.640	0.692
Total	50	1,073	0.553	0.623	0.692	0.754	1,106	0.555	0.619	0.684	0.742
population	55	1,088	0.583	0.661	0.738	0.808	1,185	0.585	0.656	0.728	0.792
	60	950	0.613	0.699	0.785	0.862	971	0.613	0.693	0.773	0.845
CVD risk	45	395	0.522	0.583	0.645	0.700	404	0.526	0.586	0.646	0.701
free	50	268	0.554	0.623	0.692	0.755	331	0.558	0.624	0.690	0.749
reference	55	279	0.585	0.663	0.740	0.810	332	0.590	0.662	0.734	0.799
populations	60	221	0.615	0.702	0.789	0.867	236	0.622	0.701	0.779	0.850

2.3.4. Multivariate logistic regression of cIMT ≥75th percentile

We evaluated the risk factors associated with a cIMT value greater than or equal to the 75th percentile of the general population. This threshold is known to indicate increased risk for a cardiovascular event. Prevalences of increased cIMT were 23.7% for women and 23.6% for men. Our results showed that participants from the South African sites (Agincourt, Dikgale and Soweto) were at lower risk compared to East Africa (Nairobi) (OR = 1.26, 95%CI = 1.03-1.54) and the West African sites, respectively Nanoro (OR = 1.44, 95%CI = 1.14-1.82) and Navrongo (OR = 1.46, 95%CI = 1.15-1.84) (**Table 2.6**). Visceral fat, overweight-obese, hypertension and high total cholesterol were associated with an increased risk for atherosclerosis with OR = 1.05 (95%CI = 1.01-1.08), OR = 1.44 (95%CI = 1.23-1.69), OR=1.57 (95%CI=1.39-177) and OR = 1.12 (95%CI = 1.05-1.19) respectively. We also found that HIV status, high triglycerides, and high HDL-C were associated with reduced risk of cIMT greater or equal to the 75th percentile. Non-problematic drinking was associated with a reduction of the risk (OR = 0.85, 95%CI = 0.72-0.99).

Table 2.6: Multiple logistic regression of binary cIMT (cIMT ≥75th percentile) per site and for the entire population

	Agincourt				Digkale				Nairobi				Nanoro				Navrongo				Soweto				All			
	N=	1 118			N=	1062			N=	1662			N=	1998			N=	1597			N=	589			N=	8034		
	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]	Odds Ratio	p-value	[95% Conf.	Interval]
age group (5 years interval) (Ref:45)																												
50	0,82	0.39	0,51928	1,29165	1,11	0.647	0,71785	1,70444	0,84	0.261	0,62907	1,13386	1,05	0.731	0,78611	1,40917	1,01	0.971	0,70679	1,43348	0,79	0.419	0,44110	1,40554	0,96	0.631	0,83354	1,11671
55	0,67	0.059	0,43691	1,01567	0,91	0.69	0,58205	1,43069	0,82	0.23	0,60060	1,13064	1,05	0.747	0,78198	1,40901	0,96	0.83	0,67701	1,36773	0,99	0.962	0,53150	1,82568	0,93	0.358	0,80509	1,08151
60	0,79	0.306	0,50554	1,23902	0,73	0.196	0,45496	1,17561	0,82	0.312	0,55392	1,20748	1,06	0.708	0,77535	1,45413	0,95	0.772	0,67082	1,34528	0,78	0.43	0,41513	1,45428	0,90	0.206	0,77238	1,05734
site																												
2	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	1,00	0.977	0,80392	1,23619
3	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	**1,26**	**0.026**	**1,02770**	**1,54359**
4	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	**1,44**	**0.002**	**1,14441**	**1,82131**
5	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	**1,46**	**0.002**	**1,15067**	**1,84197**
6	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	0,85	0.271	0,63285	1,13723
sex (Ref:Women)																												
Men	**1,65**	**0.016**	**1,09749**	**2,47395**	1,57	0.146	0,85537	2,86519	0,90	0.575	0,63402	1,28797	0,89	0.365	0,68636	1,14850	1,14	0.49	0,78824	1,64296	------	------	------	------	1,13	0.088	0,98176	1,30403
Socio-economical status (Ref: 1rst Quintile)																												
2nd Quintile	0,88	0.591	0,55382	1,40001	0,81	0.458	0,47168	1,40310	1,01	0.949	0,65963	1,55941	**1,57**	**0.017**	**1,08354**	**2,26126**	0,69	0.056	0,46671	1,00927	1,36	0.785	0,14659	12,67484	0,98	0.817	0,81446	1,17562
3rd Quintile	**0,47**	**0.011**	**0,26058**	**0,84120**	0,67	0.193	0,36029	1,22879	1,03	0.896	0,67053	1,57862	1,22	0.325	0,82432	1,79175	**0,60**	**0.011**	**0,41006**	**0,89071**	5,27	0.132	0,60540	45,84254	0,85	0.105	0,70519	1,03377
4th Quintile	**0,49**	**0.004**	**0,30422**	**0,79866**	0,91	0.728	0,52950	1,55946	0,92	0.722	0,59282	1,43679	**1,51**	**0.035**	**1,02951**	**2,22249**	0,78	0.184	0,54832	1,12251	2,60	0.38	0,30815	21,95095	0,91	0.313	0,75811	1,09271
5th Quintile	0,63	0.056	0,38679	1,01182	0,93	0.792	0,54995	1,57755	0,91	0.676	0,57811	1,42684	**1,64**	**0.01**	**1,12581**	**2,37969**	0,86	0.427	0,58128	1,25830	4,33	0.176	0,51866	36,20950	1,00	0.976	0,83594	1,20290
Highest level of education (Ref:No formal education)																												
Primary	0,94	0.767	0,64748	1,37786	0,80	0.472	0,44367	1,45730	1,03	0.901	0,63767	1,66639	0,74	0.098	0,51915	1,05662	0,89	0.458	0,64819	1,21588	0,88	0.913	0,08283	9,27097	**0,82**	**0.017**	**0,69975**	**0,96604**
Secondary	1,07	0.766	0,70098	1,62030	0,87	0.65	0,48879	1,56349	1,22	0.439	0,73568	2,02826	0,58	0.063	0,32941	1,02915	0,87	0.552	0,56006	1,36321	0,50	0.554	0,04868	5,05554	0,89	0.19	0,73807	1,06205
Teriary	1,34	0.363	0,71522	2,50052	0,59	0.311	0,20803	1,64972	0,70	0.57	0,20857	2,36963	0,80	0.69	0,26211	2,42612	0,80	0.635	0,32282	1,99376	0,58	0.652	0,05229	6,34501	0,14	0.624555	1,19790	
Overweight/Obese (Ref: No)																												
Yes	**1,67**	**0.016**	**1,09978**	**2,52147**	**2,78**	**p<0.001**	**1,69720**	**4,56432**	**1,45**	**0.027**	**1,04310**	**2,02781**	**1,53**	**0.043**	**1,01400**	**2,31212**	1,05	0.814	0,68641	1,61416	**2,06**	**0.013**	**1,16195**	**3,64371**	**1,44**	**p<0.001**	**1,23134**	**1,69449**
Hypertension (Ref: No)																												
Yes	**2,06**	**p<0.001**	**1,49537**	**2,83993**	**1,86**	**p<0.001**	**1,34786**	**2,57944**	**1,43**	**0.008**	**1,09584**	**1,85514**	1,49	0.2142	1,12593	1,97705	**1,40**	**0.016**	**1,06489**	**1,82912**	**1,80**	**0.012**	**1,13709**	**2,85926**	**1,57**	**p<0.001**	**1,39611**	**1,77400**
Smoking status (Ref:Never smoked)																												
Current Smoker	0,81	0.524	0,42927	1,53814	1,24	0.528	0,63893	2,39656	1,14	0.57	0,73043	1,76940	1,13	0.51	0,61300	0,71217	0,91	0.65	0,61664	1,35235	1,33	0.316	0,76392	2,30269	1,04	0.671	0,86416	1,25451
Former Smoker	0,82	0.508	0,45103	1,48280	0,71	0.342	0,35747	1,42833	0,95	0.774	0,64337	1,38872	1,08	0.32	0,74800	0,67931	1,37	0.16	0,88377	2,11584	1,14	0.723	0,56125	2,29956	1,01	0.915	0,83044	1,23016
Alocohol consumption (Ref:Never Consumed)																												
Current non-problematic	1,35	0.223	0,83161	2,20754	0,87	0.649	0,47592	1,58817	0,82	0.418	0,51767	1,31428	**0,66**	**0.001**	**0,51057**	**0,85008**	1,05	0.795	0,72687	1,51644	------	------	------	------	**0,85**	**0.041**	**0,72282**	**0,99338**
Current problematic	0,64	0.589	0,12771	3,21959	0,74	0.298	0,41838	1,30612	0,59	0.069	0,33737	1,04173	0,74	0.112	0,50550	1,07424	1,15	0.493	0,77145	1,71397	------	------	------	------	0,87	0.177	0,70952	1,06517
Former consumer	0,94	0.778	0,60831	1,45103	1,07	0.763	0,69841	1,63202	1,03	0.864	0,75983	1,38705	0,94	0.738	0,64394	1,36588	0,98	0.908	0,65050	1,46575	------	------	------	------	0,99	0.887	0,84348	1,15861
Visceral fat	1,01	0.796	0,92963	1,09983	**1,11**	**0.021**	**1,01587**	**1,21542**	**1,18**	**p<0.001**	**1,09084**	**1,27637**	**0,91**	**0.042**	**0,83188**	**0,99660**	0,94	0.307	0,83818	1,05705	1,09	0.183	0,96138	1,22980	**1,05**	**0.009**	**1,01162**	**1,08376**
Subcutaneous fat	1,04	0.499	0,92837	1,16498	0,81	0.061	0,64585	1,01007	**0,78**	**0.047**	**0,61814**	**0,99650**	1,04	0.754	0,80795	1,34218	1,15	0.38	0,84259	1,56717	0,92	0.57	0,69997	1,21690	0,98	0.577	0,91057	1,05354
MVPA (min/week)	0,99	0.155	0,99975	1,00004	1,00	0.613	0,99985	1,00009	1,00	0.826	0,99993	1,00006	0,99	**0.007**	**0,99983**	**0,99997**	0,99	**0.026**	**0,99984**	**0,99999**	0,99	0.293	0,99970	1,00009	0,99	**0.004**	**0,99991**	**0,99998**
Hdl	**0,46**	**0.003**	**0,27994**	**0,76280**	0,77	0.316	0,46027	1,28517	0,99	0.955	0,73352	1,33966	0,85	0.37	0,58708	1,21930	0,94	0.72	0,65644	1,33753	0,81	0.413	0,48156	1,34977	**0,81**	**0.009**	**0,69369**	**0,94813**
Total cholesterol	**1,54**	**p<0.001**	**1,29052**	**1,84033**	0,99	0.873	0,82725	1,17468	1,14	0.063	0,99313	1,30952	1,16	0.059	0,99440	1,36072	1,00	0.955	0,86272	1,16935	1,16	0.178	0,93489	1,43859	**1,12**	**0.001**	**1,04899**	**1,19083**
Triglycerides	**0,50**	**p<0.001**	**0,34259**	**0,74008**	0,74	0.064	0,54508	1,01772	0,86	0.255	0,67330	1,11053	0,94	0.674	0,69882	1,26087	1,06	0.75	0,73906	1,52118	0,99	0.966	0,68973	1,42737	**0,87**	**0.021**	**0,76656**	**0,97877**
Serum Creatinin	1,00	0.828	0,99524	1,00383	1,00	0.806	0,99364	1,00498	**1,01**	**0.001**	**1,00474**	**1,01814**	1,00	0.0026	0,99455	1,00505	1,00	0.383	0,98894	1,00427	0,99	0.227	0,97849	1,00517	1,00	0.689	0,99831	1,00257
Glucose	0,98	0.725	0,88506	1,08869	**1,12**	**p<0.001**	**1,05176**	**1,19125**	0,98	0.513	0,91970	1,04268	**0,90**	**0.045**	**0,80906**	**0,99739**	0,96	0.643	0,81817	1,13195	0,91	0.203	0,78225	1,05352	1,01	0.638	0,97556	1,04121
Hiv status (Ref: No)																												
Yes	0,86	0.37	0,62362	1,19223	0,78	0.223	0,52405	1,16299	**0,61**	**0.013**	**0,41440**	**0,90216**	------	------	------	------	0,28	0.22	0,03504	2,16565	0,76	0.359	0,42291	1,36559	**0,71**	**p<0.001**	**0,59210**	**0,86185**

2.4. DISCUSSION

Population data on cIMT reference intervals for adults with and without CVDs are rare in Africa. In this study, we described the characteristics of cIMT measurements in several SSA populations using population-based data from the AWI-Gen cross-sectional study of men and women aged 40 to 60 years. We defined age- and sex-specific cIMT reference intervals for populations from West, East and South Africa, showing regional differences, with West Africans having higher values than East and South Africans. Our study is, to date, the largest population-based study describing cIMT in SSA.

Comparisons with previous studies among Africans is limited due to differences in study designs and participant recruitment strategies, as well as age ranges. In South Africa we showed that the population means were lower in black Africans, compared to those reported for the mixed ancestry populations, although the comparison needs to be assessed in light of the mean age of participants in the latter being higher (58 ± 10 years) (Zemlin *et al.*, 2017). Our findings were however comparable to the mean cIMT (0.61 ± 12) described in normotensive black South Africans from the Sympathetic Activity and Ambulatory Blood Pressure in Africans study (Schutte *et al.*, 2009), and the mean cIMT reported by Schoffelen and collaborators from a cohort of HIV positive patients (Schoffelen *et al.*, 2015), despite the younger ages of those cohorts. Ranges from Nanoro and Navrongo were similar to those reported in Nigerian normotensives controls (Owolabi *et al.*, 2015). Generally, our study reported comparable values of mean cIMT to previous studies when geographically matched.

Carotid atherosclerosis positively correlates with the severity of coronary atherosclerosis, and cIMT independently correlates with the risk for major cardiovascular diseases such as transient cerebral ischemia, stroke and coronary events. Several large epidemiological studies from North and South America, Asia and Europe, have reported cIMT RIs in percentiles by sex, age, and ethnicity, whereas there is a paucity of studies reporting references intervals for cIMT in SSA. Our study is the first to report age-specific reference intervals in this region. Our combined sample reported values comparable to black men and women from the Atherosclerosis Risk in Communities Study cohort (Howard *et al.*, 1993) (USA), Multi Ethnic Study of Atherosclerosis (USA) (O'Leary *et al.*, 1999) and The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) (Santos *et al.*, 2014). The mean cIMT distributions are also similar to the Argentinian (Diaz *et al.*, 2018) and Korean populations (Youn *et al.*, 2011).

We observed regional differences in our study. In West African populations, the values for women (0.650, IQR=0.155) and for men (0.670, IQR=0.155) were often higher than those reported in European populations (Engelen *et al.*, 2013), Asian populations (Lee *et al.*, 2014; Liu *et al.*, 2017) and Latin American populations (Santos *et al.*, 2014). Whereas in East and South African populations they appeared similar or, in some cases, lower (0.598, IQR=0.125 for female; 0.587, IQR=0.133 for male). Interestingly, the findings in West Africans are not consistent with the prevalence of other CVD risk factors such as overweight/obesity and hypertension, in comparison to East and South Africa (Gómez-Olivé *et al.*, 2017; Ramsay *et al.*, 2018). However, previous studies in European and American populations reported that traditional risk factors were not the major contributors to the variance of cIMT, but that age (7%), male sex (3%), fasting glucose levels (<1%), number of cigarettes pack per years of smoking (<1%), and LDL-cholesterol (<1%), together accounted for 11% of the variability. They further reported that black African ancestry was significantly associated with an increased cIMT (Rundek *et al.*, 2013). This racial difference was consistent over numerous studies that reported higher cIMT values in populations with African ancestry (Bennett *et al.*, 2011; Whincup *et al.*, 2012; Rundek *et al.*, 2013; Santos *et al.*, 2014; Erqou *et al.*, 2015; Villines *et al.*, 2017). Considering that high genetic influences on cIMT were detected in previous studies, with heritability (h^2) estimates ranging from 0.35 to 0.65 (Lange *et al.*, 2002; Xiang *et al.*, 2002; Fox *et al.*, 2003; Suh-Hang *et al.*, 2004; Sacco *et al.*, 2009; Kuipers *et al.*, 2013; Medda *et al.*, 2014), and that demographic history with West African ancestry (e.g. African Americans) suggest higher cIMT ranges, there is likely a genetic component that would explain the regional differences we have observed. The regional differences were more pronounced when examining RIs among the healthy subsets.

Our study showed that despite the high prevalence of traditional CVD risk factors, South African populations were less likely to have increased cIMT ($\geq 75^{th}$ percentile) than populations from East and West African. When analysing the combined cohort, hypertension, overweight/obesity and high cholesterol were strong correlates with increased cIMT. In the regional analysis, it is important to note that the odds ratio for being in the top 25th cIMT centile was higher for individuals with hypertension and being overweight/obese in the South African sites (Agincourt, Dikgale and Soweto) compare to the site in West and East Africa. Non-problematic drinking was found to be protective in the combined cohort, and this association was mainly driven by the data from the Nanoro site, with an OR = 0.66 (95%CI = 0.51-0.85). Similar reports on the protective effect of moderate alcohol consumption suggest that it is linked

to the benefits of the polyphenol content of beer and to the contribution of alcohol to a more favourable lipid profile (Britton *et al*. 2016; Britton *et al*. 2017; Vu *et al*. 2016; Chiva-Blanch *et al*. 2015). Effectively, sorghum, which is rich in polyphenols, is one of the main staple foods in the Nanoro population and sorghum beer is the most commonly consumed (Boua *et al.*, 2010; Abdoul-latif *et al.*, 2012) with suggestive protective effect in moderate users.

Limitations of our study include the absence of longitudinal data, the fact that in some countries we do not have data on both urban and rural communities and the relatively small number of healthy participants from the South African sites and the absence of cIMT measurements from the Soweto women. At a mean age of 50 years, our cohort was relatively young compared to other cohorts that report on cIMT ranges, making comparisons with other cohorts challenging, and potentially underestimating the risk of atherosclerosis in our study communities. However, the data provide an opportunity to investigate early risk for atherosclerosis in Africans. Our study reported the age-specific reference intervals for cIMT in six African communities. The percentile estimates in the population can be used for clinical screening for CVDs as well as reference for future studies, in addition to the classical risk factors.

Chapter 3:

African-specific variants reveal novel susceptibility loci for carotid intima-media thickness in sub-Saharan African populations with sex-differences: An AWI-Gen study

Chapter 3: African-specific variants reveal novel susceptibility loci for carotid intima-media thickness in sub-Saharan African populations with sex-differences: An AWI-Gen study

ABSTRACT

Atherosclerosis precedes the onset of many clinical manifestations of cardiovascular diseases (e.g. stroke, myocardial infarction and peripheral arterial disease), which are the leading causes of death worldwide. Given this disease burden, genetic susceptibility to atherosclerosis has been widely investigated, but not yet in sub-Saharan Africans (SSA). Considering the deep evolutionary roots and the genetic diversity of the African continent, GWASs in SSA represent an opportunity to shed new light on understanding the genetic architecture of complex traits such as atherosclerosis. We used carotid intima-media thickness (cIMT), to investigate genetic susceptibility to atherosclerosis in SSA. Our sample included 7894 unrelated middle-aged adults (3963 women, 3931 men) aged 40 to 60 years, from the AWI-Gen study (from East, South and West Africa). cIMT was measured by ultrasound and genotyping was performed on the H3Africa SNP Array (~2.3M SNPs). After imputation to ~14M SNPs using the Sanger Imputation Server African reference panel, we tested for association using BOLT-LMM, with adjustment for age, sex and 8 Principal Components (PCs). In sex-stratified analyses, we adjusted for age and 5 PCs. In the combined sample we identified two new African-specific genome-wide significant loci, *SIRPA* (p = 4.7E-08), and *FBXL17* (p = 2.5E-08). Sex-stratified analysis revealed two male-specific loci, *SNX29* (p = 6.3E-9) and *MAP3K7* (p = 5.3E-8), and two female-specific loci, *LARP6* (p = 2.4E-09) and *PROK1* (p = 1.0E-08). We replicated regional associations with known risk loci for atherosclerosis and CVDs, but with different lead SNPs. In a gene-based analysis of the entire dataset there were significant associations with *CALD1* (p = 5.9E-07 < 2.6E-06) and *FLT4* (p = 4.3E-07 < 2.6E-06). We found significant enrichment for oestrogen response genes for female-specific signals. The genes identified in our study showed previous biological evidence of involvement with atherosclerosis and/or CVDs (*SIRPA*, *FBXL17*, *MAP3K*, *LARP6*, *PROK1*) and sex-difference (*SNX29*, *PROK1*). Our study successfully replicated loci and identified novel African-specific variant and gene associations for cIMT. Sex-stratified analyses highlighted sex differences in atherosclerosis risk. Transferability of signals from non-African studies provide opportunities for fine-mapping

signals using African data. Greater inclusion of African populations in medical genomics is important for accelerating discoveries and identifying new genetic associations.

3.1. Introduction

Atherosclerosis is a complex heritable trait with an enigmatic genetic aetiology. Despite GWAS discoveries, little is known about the genetic contribution to of atherosclerosis, raising the question of "missing heritability". Meanwhile, the worldwide epidemic of cardiovascular diseases (CVDs), including clinical manifestation of atherosclerosis, is growing and has become the leading cause of deaths worldwide (Fuster, 2014; Roth *et al.*, 2017). Moreover, the undergoing health and demographic transition in sub-Saharan Africa (SSA) has shifted the major causes of death from communicable and nutritional diseases to non-communicable diseases (NCDs), with sickle cell trait aggravating the risk for CVDs (Caughey *et al.*, 2014). Atherosclerosis results from injury to the arterial endothelium, resulting in an inflammatory response in the vessel wall. The location and morphology of the atherosclerotic lesions predict the nature of the resulting vascular disease. Whereas family and twins studies provided evidence of high heritability of cIMT (20-65%) (Fox *et al.*, 2003; Sacco *et al.*, 2009; Fagnani *et al.*, 2013; Medda *et al.*, 2014), the GWAS studies reported association that account for only 1.1% of the variance of cIMT (Bis *et al.*, 2012).

The genetic diversity of African populations and their deep evolutionary roots represent opportunities for novel genetic discoveries. Because haplotypes blocks are shorter in Africans compared to other populations (average haplotype block ~8.8 kb in Africans, ~20.7 kb in Europeans, and ~25.2 kb in Han Chinese), identification of causal variants is facilitated (Lonjou *et al.*, 2003; Hinds *et al.*, 2005; McPherson and Tybjaerg-Hansen, 2016). The role of ancestry in atherosclerosis risk has been established from studies in multi-ethnic settings and admixture studies for atherosclerosis (Wassel *et al.*, 2009; Shendre, Irvin, *et al.*, 2017; Shendre, Wiener, *et al.*, 2017). African ancestry was reported to be associated with a higher risk of atherosclerosis compared to Europeans, Hispanic and Asians.

Since phenotypic differences between men and women are a pervasive feature of quantitative traits, studies of sex interactions for complex human traits may shed light on the molecular mechanisms that lead to biological differences between men and women. Sex has been found to play a role in variations between gene expression and genotype across a range of human complex traits (Rawlik *et al.*, 2016). Sex-differences in cells involved in the atherosclerotic process have been previously reported (Franconi *et al.*, 2017) and are supported by sex-stratified GWAS analyses (Liao *et al.*, 2010; Dong *et al.*, 2015; Lin *et al.*, 2015). Sex provides

two different environmental contexts determined by the hormonal milieu and differential gene expression between the sexes.

Several genetic association studies of cIMT have been performed in the major world populations and provided insights into genes and tissue-specific regulatory mechanisms linking atherosclerosis both to its functional genomic origins and its clinical consequences in humans. To date, about 76 SNPs have been found to be associated with cIMT (GWAS Catalogue) (Buniello *et al.*, 2019). However, none of these focused on sub-Saharan African populations.

This study aimed to investigating the genetic susceptibility to atherosclerosis in sub-Saharan Africans, using cIMT in the AWI-Gen study as an endophenotype, with further investigation of sex-differences in these populations.

3.2. Method

3.2.1 Study population and phenotype assessments

This is a cross-sectional study that investigated populations from six sub-Saharan African sites in West Africa (Burkina Faso (Nanoro) and Ghana (Navrongo)), East Africa (Kenya (Nairobi)) and South Africa (Agincourt, Dikgale and Soweto) as part of the AWI-Gen study (Richter *et al.*, 2007; Derra *et al.*, 2012; Kahn *et al.*, 2012; Oduro *et al.*, 2012; Alberts *et al.*, 2015; Beguy *et al.*, 2015; Ramsay *et al.*, 2016; Ali *et al.*, 2018). The participants for this study include 10703 black African men and women from two urban settings (Nairobi and Soweto) and four rural settings (Agincourt, Dikgale, Nanoro and Navrongo), aged 40 to 60 years. Participants completed a questionnaire requesting information on demography, health history and behaviour. Anthropometric measurements were taken and blood collected for genotyping (H3Africa SNP array) and phenotyping (biomarkers) (Ali *et al.*, 2018). Ultrasound scans were performed to assess cIMT of the right and left carotid arteries. No cIMT data was collected for female participants from Soweto because they were drawn from the Study of Women Entering and Endocrine Transition (SWEET) study, and they were therefore not included in the subsequent GWAS. This study received approval from the Human Research Ethics Committee (Medical), University of the Witwatersrand, South Africa (M121029, M1706110). All the participants signed an Informed Consent form before any study procedures were performed.

3.2.2 cIMT Measurement

cIMT was measured using Dual B-mode ultrasound images of the carotid tree showing a typical double line for the arterial wall. Details of the method for measurement were provided in **Chapter 2**. The cIMT values were QCed according to the Mannheim Consensus defining the use of cIMT in population-based studies. To allow comparison and meta-analysis with other studies in European ancestry populations, the Mean Max cIMT was generated as the average of the maximum cIMT from the left and right, and this value was used for the GWAS analyses.

3.2.3 Association and follow up analysis

3.2.3.1. Genotyping and Imputation

The H3Africa genotyping array, designed as an African-common-variant-enriched GWAS array (Illumina) with ~2.3 million SNPs, was used to genotype genomic DNA using the Illumina FastTrack Sequencing Service[1]. The following pre-imputation QC steps were applied to the entire AWI-Gen genotype data set. Individuals with a missing SNP calling rate greater than 0.05 were removed. SNPs with a genotype missingness greater than 0.05, MAF less than 0.01 and Hardy-Weinberg equilibrium (HWE) P-value less than 0.0001 were removed. Non-autosomal and mitochondrial SNPs, and ambiguous SNPs that did not match the GRCh37 references alleles or strands were also removed. Imputation was performed on the cleaned dataset (with 1,729,661 SNPs and 10,903 individuals) using the Sanger Imputation Server and the African Genome Resources as reference panel. We selected EAGLE2 (Loh *et al.*, 2016) for pre-phasing and the default PBWT algorithm was used for imputation. After imputation, poorly imputed SNPs with info scores less than 0.6, MAF less 0.01, and HWE P-value less than 0.00001 were excluded. The final QC-ed imputed data had 13.98 M SNPs, and only participants with both good quality cIMT and genotyping data (n = 7894) were used for the GWAS analyses.

3.2.3.2. Data Analysis

Descriptive statistics were used to summarise the population characteristics. Continuous variables were reported in median and interquartile ranges and categorical variables were reported in percentages. To compare data across groups (e.g. sex differences), we used χ^2 and/or

[1] https://www.illumina.com/services/sequencing-services.html

Fisher's test for categorical variables and Mann-Whitney-Wilcoxon test for continuous variables.

3.2.3.3. Genome-wide association analysis

Linear regression of Mean Max cIMT was performed with covariates in R[2]. Residuals were extracted from the linear regression analyses and used for the GWAS analysis. We used as covariates age, sex and 8 principal components (PCs) computed on genetics data (to account for genetic structure). In our sex stratified analysis, the covariates were age and 5 PCs. The number of PCs to include in each model were determined using a stepwise regression. We performed all association testing with the residuals in BOLT-LMM, which implement testing using a Linear Mixed Model (LMM). To run efficiently, BOLT-LMM required three important components: the (imputed) genotypic data for association testing; a reference panel of LD scores per SNP, calculated using 1000 Genomes Project African samples; and genotype data used to approximate a genetic relationship matrix (GRM) (Using a subset of the SNP Array genotypes with LD filtering). This method is expected to account for all forms of relatedness, ancestral heterogeneity in the samples and other (potentially hidden) structure in the data. The analyses were run on the automated workflow of H3abionet/H3agwas[3] (Baichoo *et al.*, 2018). We screened the output for a genome-wide significance threshold (p-values < 5.E-08). To assess genomic inflation, we compared our observed distribution of −log10(P) values to that expected in the absence of association (Lambda) and illustrated the results in QQ plots. The same process was applied for sex-stratified analyses.

We used EasyStrata (Winkler, Kutalik, *et al.*, 2015) to test for the joint effect calculated from sex strata results (Aschard *et al.*, 2010) and to test for the difference between two strata results as a means to test for sex effects (Randall *et al.*, 2013). The joint and stratified frameworks were found to be the most efficient way to test for gene-environment interactions (Yun J. Sung *et al.*, 2016). Power calculations were performed with Quanto[4] (Version 1.2.4).

3.2.3.4. GWAS catalog lookup

[2] https://www.R-project.org/
[3] http://github.com/h3abionet/h3agwas/
[4] http://biostats.usc.edu/Quanto.html

The GWAS Catalog database was downloaded from the website[5] (accessed on 12 Jan 2019) and a subset of the data generated using the following key words relevant to our study: coronary artery disease, carotid atherosclerosis, cIMT, coronary artery calcification and abdominal artery aneurism. The marker co-ordinates from the GWAS Catalog are given in build 38. Since our dataset was in build 37, we performed lift-over of GWAS Catalog to build 37 in order to allow accurate comparison. In order to look whether our study was replicating previous findings, we searched for the same marker or any markers within 25 kb (considering the highest mean size of LD blocks) We then searched for SNPs in a 25 kb region of all SNPs with suggestive associations (p-value < 1E-04) found in our study. We further defined loci by grouping SNPs with p-value < 1E-04 within 250 kb of each other. These loci were used for lookup for regional replication and transferability analysis.

3.2.3.5. Functional analysis

The FUMA online platform[6] (Watanabe *et al.*, 2017) was used to annotate, prioritize, visualize and interpret GWAS results. GWAS summary statistics (p<1E-05) from out study was used as the input. FUMA provided extensive functional annotation for all SNPs in genomic areas identified by lead SNPs. From the list of gene IDs (as identified by SNP2GENE option in FUMA) FUMA annotated genes in a biological context (Watanabe *et al.*, 2017). We selected all candidate SNPs in the associated genomic region having $r^2 \geq 0.6$ (with 1000 Genome Project African references) with one of the independently significant SNPs, with a suggestive P-value ($p < 1E-05$) and MAF > 0.01 for annotation. Predicted functional consequences for these SNPs were obtained by matching the SNP's chromosome base-pair position, and reference and alternate alleles, to databases containing known functional annotations, including ANNOVAR (Wang, Li, and Hakonarson 2010), combined annotation-dependent depletion (CADD) scores (Kircher *et al.*, 2014), and RegulomeDB (RDB) (Boyle *et al*. 2012) scores. Additionally, eQTLs scans (The GTExArd Consortium *et al.*, 2015; GTex Consortium *et al.*, 2017) were performed.

3.2.3.6. Functional annotation of mapped genes

[5] https://www.ebi.ac.uk/gwas/
[6] http://fuma.ctglab.nl/

Genes implicated by mapping of significant GWAS SNPs were further investigated using the GENE2FUNC procedure in FUMA (Watanabe *et al.*, 2017), which provides hypergeometric tests of enrichment of the list of mapped genes in 53 GTEx tissue-specific gene expression sets (The GTExArd Consortium *et al.*, 2015; GTex Consortium *et al.*, 2017), 7,246 MSigDB gene sets[7], and chromatin states (Ernst and Kellis, 2013; Consortium Roadmap Epigenomics *et al.*, 2015).

3.2.3.7. MAGMA Gene-based and gene-sets analysis

Multi-marker analysis of genomic annotation (MAGMA, v1.6) gene analysis were performed using summary statistics of our association results as input. Gene-based analysis enabled summarization of SNPs associations at the gene level and association of the set of genes to biological pathways. MAGMA employs multiple linear regression to obtain gene-based p-values (de Leeuw *et al.*, 2015; Watanabe *et al.*, 2017). The window for gene annotation was set for 25kb and genome-wide significance was set at 0.05/number of tested genes. MAGMA gene-set analysis used a competitive testing framework, with gene-sets from MsigDB (v6.2, 10678 gene sets (curated gene sets: 4761, GO terms: 5917)) (Liberzon *et al.*, 2011, 2015). MAGMA analysis was implemented within FUMA.

[7] http://software.broadinstitute.org/gsea/msigdb

3.3. Results

3.3.1. Association results

Results of the analyses are presented for 13.9M SNPs. Despite the population sub-structure demonstrated by principal component analysis in the study sample (**Figure 3.1**), our results did not show evidence of genomic inflation ($\lambda = 0.997$). The genome-wide association results for the combined set are illustrated in the Manhattan plots and the genomic inflation by the QQ-plots (**Figure 3.2a, 3.2b**). In the combined sample, we identified two new genome-wide significant loci in *SIRPA* (rs6045318, p = 4.7E-08) and *FBXL17* (rs552690895, p = 2.5E-08). Other suggestive association signals had lead variants located in an intergenic region on chromosome 8 (rs11781274, p = 1.8E-07), an intronic region in SORCS1 (rs11193156, p = 2.1E-07), an intronic region in *ANKK1* (rs11214599, p = 5.4E-07), an exonic region in *CTBP2* (rs3781409, p = 6.6E-07), and an intronic region in *SMARCA2* (rs1324201, p = 8.6E-07) (**Table 3.1, Supplementary Table 3.1**).

Sex-specific analyses revealed four genome-wide significant loci (as illustrated in the Miami plots, **Figure 3.3**): two male-specific loci led by intronic variants in *SNX29* (rs190770959, p = 6.3E-9) and *MAP3K7* (rs284509, p = 5.3E-8), and two female-specific loci located in an intergenic region between *UACA* and *LRP6* (a downstream variant located in the promoter flanking region) (rs150840489, p = 2.4E-09) and a variant in a transcription factor binding site near *PROK1*/*CYMP* (rs115473055, p = 1.0E-08). Loci with suggestive associations (p < 1E-06) included variants in *AK2P2*, *RBMS3*, *FIP1L1*:*LNX1*, *CDH18*, *FLT4*, *FOXK1*, *CDH17*, and *C16orf45* in female-specific analysis (**Table 3.2, Supplementary Table 3.1**); whereas for male-specific suggestive associations were identified in *U3*, *OR2T35*, *FBXL17*, *SORCS1*, and *USP12* (**Table 3.3, Supplementary Table 3.1**).

Our analysis for sex-differences assessed effects that were limited to men or women (**Figure 3.4**). We found suggestive signals in intergenic region of *IGFBPL*/*FAM95C* (rs12350396, p = 3.4E-07), *UACA*/*LARP6* (rs150840489, p = 4.4E-07), *MYOD1*/*KCNC1* (rs150481830, p = 4.9E-07), and in an intronic region of *CROCC* (rs11585710, p = 4.8E-07). These sex-difference were displayed on a table below (**Table 3.4, Supplementary Table 3.1**). Regional plots of significant loci are shown in **Figure 3.5**.

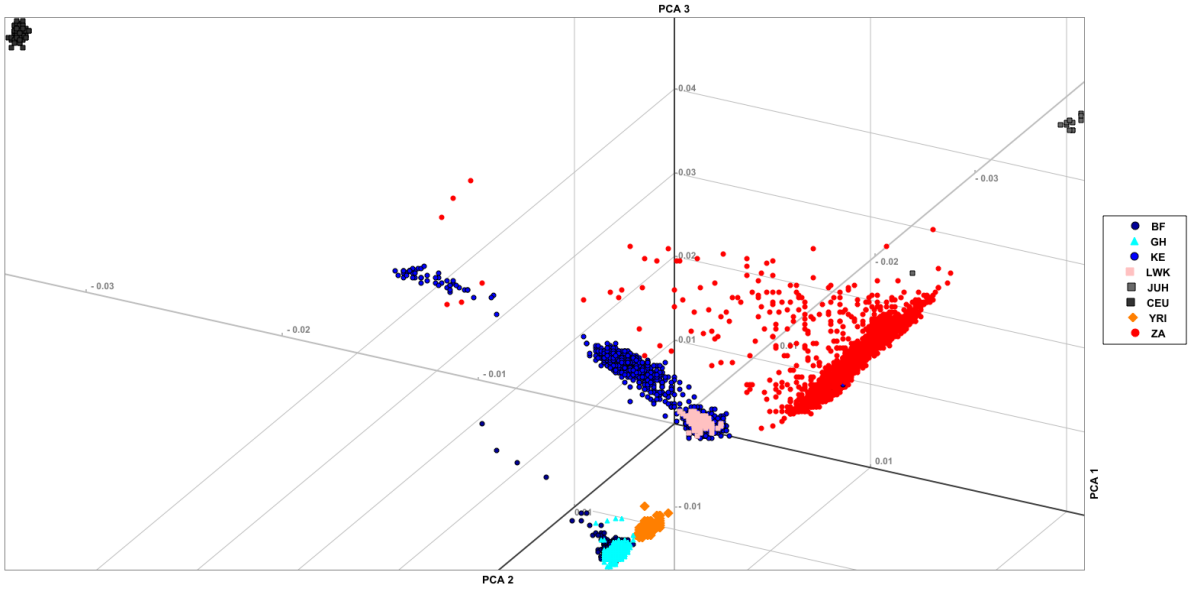


Figure 3.1: Principal Component Analysis (PCA) plots showing population sub-structure in the AWI-Gen data and compared to data in selected population from the 1000GP.

Each dot represents a participant and xxx SNPs were used in the analysis. AWI-Gen data - BF: Burkina Faso; GH: Ghana; KE: Kenya; ZA: South Africa. 1000GP data - LWK: Luhya in Webuye/Kenya; JUH: Ju\'hoansi; CEU: Utah residents of northern and western European ancestry; YRI: Yoruba from Ibadan/Nigeria.

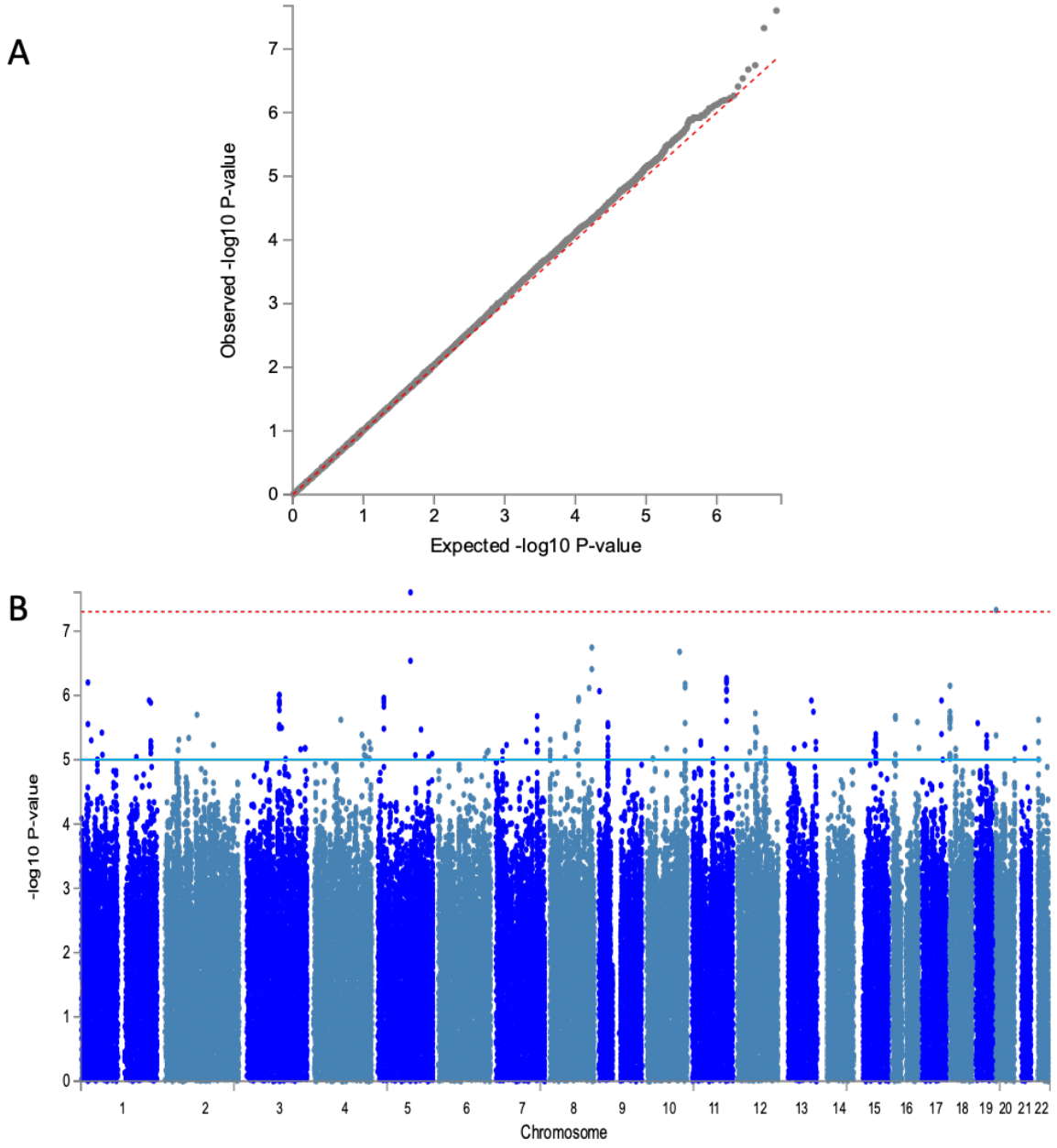


Figure 3.2: QQ and Manhattan plots for cIMT association results in AWI-Gen study (7894 participants).

(A) QQ plot for combined set GIF=0.997. (B) Manhattan plot showing the –log10-transformed two-tailed P-value of each SNP from the GWAS for Mean Max cIMT on the Y axis and base-pair positions along the chromosomes on the X axis. (Adjustment was made for age, sex and 8 PCs) The red line indicates Bonferroni-corrected genome-wide significance ($p < 5 \times 10^{-8}$); the blue line indicates the threshold for suggestive association ($p < 1 \times 10^{-5}$).

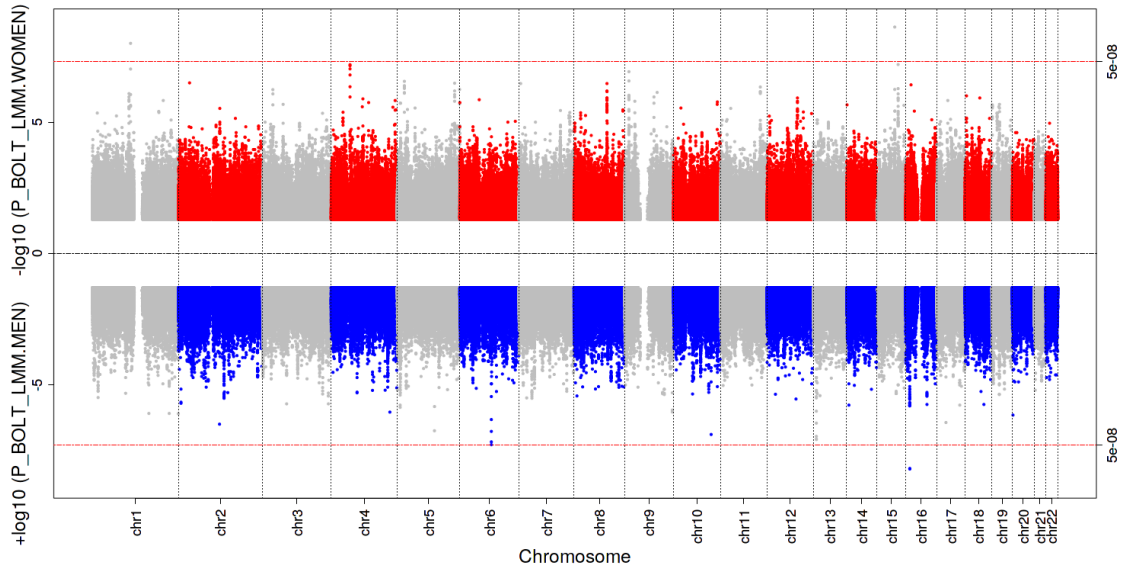


Figure 3.3: Miami plot showing female and male-specific associations p-values for cIMT.

The –log10-transformed two-tailed P-value of each SNP from the GWAS for Mean Max cIMT on the Y axis (Adjustment was made for age and 5 PCs for each sex) and base-pair positions along the chromosomes on the X axis. On top are the results for female-specific analysis and on the bottom the results for the male-specific analysis. The red line indicates Bonferroni-corrected genome-wide significance ($p < 5 \times 10^{-8}$)

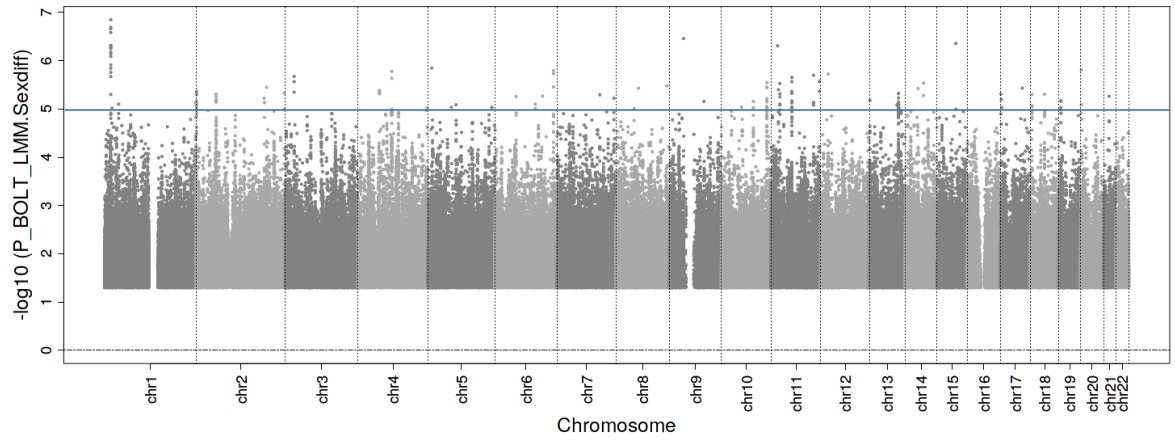


Figure 3.4: Manhattan plot showing the –log10-transformed two-tailed P-value of each SNP for test of difference between female and male associations for each SNP.

The blue line indicates the threshold for suggestive association ($p < 1 \times 10^{-5}$).

Table 3.1: Selected SNPs associated with Mean Max CIMT with p-values <1E-06 for the combined AWI-Gen sample (n=7894).

rsID	chr	pos	non effect allele	effect allele	MAF	gwasP	Beta	SE	IndSigSNP	Nearest Gene	Genic position
rs12048810	1	22690547	A	G	0,083	6,30E-07	-0,015	0,003	rs12048810	*RP11-415K20.1*	intergenic
rs547840497	5	107469179	C	T	0,014	2,90E-07	-0,041	0,008	rs552690895	*FBXL17*	intronic
rs552690895	**5**	**107570359**	**A**	**G**	**0,013**	**2,50E-08**	**-0,043**	**0,008**	**rs552690895**	***FBXL17***	**intronic**
rs116216559	8	129492156	A	G	0,026	7,70E-07	-0,026	0,005	rs116216559	*RP11-89M16.1*	ncRNA_intronic
rs28731472	8	138092782	C	A	0,090	3,90E-07	-0,016	0,003	rs11781274	*RNU6-144P*	intergenic
rs11781274	8	138097368	A	G	0,082	1,80E-07	-0,017	0,003	rs11781274	*RNU6-144P*	intergenic
rs1324201	9	2183330	C	T	0,256	8,60E-07	-0,010	0,002	rs1324201	*SMARCA2*	intronic
rs11193156	10	108759228	A	T	0,134	2,10E-07	-0,016	0,003	rs11193156	*SORCS1*	intronic
rs10444192	10	126714166	T	C	0,100	7,50E-07	-0,017	0,003	rs10444192	*CTBP2*	intronic
rs3781409	10	126715629	T	C	0,100	6,60E-07	-0,017	0,003	rs3781409	*CTBP2*	exonic
rs11214599	11	113271360	T	C	0,020	5,40E-07	-0,039	0,008	rs11214599	*ANKK1*	downstream
rs2511520	11	113274771	T	C	0,020	5,90E-07	-0,039	0,008	rs11214599	*RP11-159N11.3*	intergenic
rs2242591	11	113279921	T	C	0,020	8,10E-07	-0,039	0,008	rs11214599	*RP11-159N11.3*	ncRNA_intronic
rs6278	11	113280724	A	C	0,020	8,60E-07	-0,038	0,008	rs11214599	*RP11-159N11.3:DRD2*	ncRNA_intronic
rs1124491	11	113282090	A	G	0,020	6,40E-07	-0,039	0,008	rs11214599	*RP11-159N11.3:DRD2*	ncRNA_intronic
rs685058	18	2146395	G	A	0,143	7,10E-07	0,014	0,003	rs685058	*RP11-161I6.2*	ncRNA_intronic
rs6045318	**20**	**1883451**	**G**	**A**	**0,024**	**4,70E-08**	**-0,031**	**0,006**	**rs6045318**	***SIRPA***	**intronic**

chr: Chromosome; pos: position; MAF: Minor allele frequency; gwasP: GWAS p-value; SE: Standard Error; IndSigSNP: Index significant SNP
Test performed with BOLT-LMM with adjustment for age, sex and 8 PCs

Table 3.2: Selected SNPs with p-values <1E-06 for the Female-specific analysis

rsID	chr	pos	Non effect allele	Effect allele	MAF	gwasP	beta	se	IndSigSNP	Nearest Gene	Genic position
rs116672087	1	106514615	G	A	0,036	8,20E-07	0,053	0,011	rs116672087	*RP11-24P14.1*	ncRNA_intronic
rs4839398	1	111006986	G	A	0,154	8,90E-07	-0,018	0,004	rs4839398	*PROK1:RP11-470L19.5*	intergenic
rs115473055	**1**	**111020548**	**T**	**C**	**0,080**	**1,00E-08**	**-0,026**	**0,005**	**rs115473055**	***CYMP***	**intergenic**
rs140337389	1	111033882	A	G	0,061	9,50E-08	-0,037	0,007	rs140337389	*CYMP*	ncRNA_exonic
rs139454995	2	32062127	C	A	0,014	3,20E-07	-0,060	0,012	rs139454995	*AK2P2*	intergenic
rs9830575	3	29486586	G	A	0,026	5,80E-07	-0,041	0,008	rs9830575	*RBMS3*	intronic
rs28513301	3	29569688	G	C	0,022	8,40E-07	-0,040	0,008	rs9830575	*RBMS3*	intronic
rs79028487	4	54487185	C	T	0,022	7,10E-08	-0,050	0,009	rs79028487	*FIP1L1:LNX1*	intronic
rs145752200	4	54488817	G	A	0,022	6,50E-08	-0,051	0,010	rs145752200	*FIP1L1:LNX1*	intronic
rs139178248	4	54495071	A	T	0,022	9,40E-08	-0,050	0,010	rs139178248	*FIP1L1:LNX1*	intronic
rs78315295	4	54508985	G	C	0,023	1,60E-07	-0,049	0,010	rs78315295	*FIP1L1:LNX1*	intronic
rs184055133	4	54511431	T	C	0,022	4,50E-07	-0,048	0,010	rs184055133	*FIP1L1:LNX1*	intronic
rs150744905	5	20536559	C	T	0,012	7,50E-07	-0,049	0,010	rs190082809	*CDH18*	intronic
rs146793722	5	20542546	T	C	0,012	4,50E-07	-0,052	0,010	rs190082809	*CDH18*	intronic
rs145849737	5	20544251	G	C	0,012	4,10E-07	-0,052	0,010	rs190082809	*CDH18*	intronic
rs150340920	5	20549408	G	A	0,012	5,20E-07	-0,051	0,010	rs190082809	*CDH18*	intronic
rs190082809	5	20552942	A	G	0,013	2,80E-07	-0,053	0,010	rs190082809	*CDH18*	intronic
rs11959032	5	165602212	T	C	0,033	3,30E-07	-0,048	0,009	rs11959032	*AC122720.1*	intergenic
rs77635228	5	165603117	A	C	0,033	9,00E-07	-0,047	0,010	rs11959032	*AC122720.1*	intergenic
rs116437926	5	165614455	A	T	0,036	6,50E-07	-0,046	0,009	rs11959032	*AC122720.1*	intergenic
rs112967731	5	180083479	A	G	0,090	5,70E-07	-0,023	0,005	rs112967731	*FLT4*	intergenic
rs140285788	7	4779605	T	A	0,016	3,40E-07	0,056	0,011	rs140285788	*FOXK1*	intronic

rsID	chr	pos	Non effect allele	Effect allele	MAF	gwasP	beta	se	IndSigSNP	Nearest Gene	Genic position
rs7013153	8	95092061	C	A	0,165	6,50E-07	-0,018	0,004	rs114684581	*KB-1184D12.1*	intergenic
rs114684581	8	95100302	C	T	0,166	3,40E-07	-0,018	0,004	rs114684581	*KB-1184D12.1*	intergenic
rs80279158	8	95174434	T	C	0,073	6,90E-07	-0,024	0,005	rs80279158	*CDH17*	intronic
rs10960551	9	12260408	G	C	0,274	1,20E-07	0,015	0,003	rs10960551	*RP11-71E22.1*	intergenic
rs4740521	9	12261388	A	G	0,274	2,70E-07	0,015	0,003	rs10960551	*RP11-71E22.1*	intergenic
rs1416577	9	12264458	C	G	0,291	9,50E-07	0,014	0,003	rs10960551	*RP11-71E22.1*	intergenic
rs74744940	9	93865035	A	G	0,023	7,40E-07	-0,051	0,010	rs74744940	*RP11-305L7.1*	intergenic
rs2515324	11	115446096	A	G	0,487	6,90E-07	-0,013	0,003	rs922205	*RP11-136I14.4*	intergenic
rs2515326	11	115446460	A	T	0,487	7,20E-07	-0,013	0,003	rs922205	*RP11-136I14.4*	intergenic
rs2446890	11	115447446	G	A	0,488	8,10E-07	-0,013	0,003	rs922205	*RP11-136I14.4*	intergenic
rs922205	11	115448115	A	T	0,440	4,60E-07	-0,013	0,003	rs922205	*RP11-136I14.4*	upstream
rs150840489	**15**	**71088277**	**A**	**G**	**0,037**	**2,40E-09**	**-0,051**	**0,009**	**rs150840489**	***RPL29P30***	**downstream**
rs77298235	15	80852328	T	G	0,030	6,40E-08	-0,044	0,008	rs77298235	*ARNT2:RP11-379K22.2*	ncRNA_intronic
rs76017996	15	80852338	C	T	0,064	5,30E-07	-0,032	0,006	rs76017996	*ARNT2:RP11-379K22.2*	ncRNA_intronic
rs141882243	15	80882112	T	C	0,019	8,40E-07	-0,046	0,009	rs141882243	*ARNT2*	intronic
rs116132984	16	15690120	C	T	0,088	3,80E-07	-0,026	0,005	rs116132984	*C16orf45:KIAA0430*	UTR3

chr: Chromosome; pos: position; MAF: Minor allele frequency; gwasP: GWAS p-value; SE: Standard Error; IndSigSNP: Index significant SNP;

Table 3.3: Selected SNPs with p-values <1E-06 for the Male-specific analysis

rsID	chr	pos	Non effect allele	Effect allele	MAF	gwasP	beta	se	IndSigSNP	Nearest Gene	Genic position
rs114119990	1	163865550	C	A	0,043	8,20E-07	0,032	0,007	rs114119990	*U3*	intergenic
rs67416193	1	222666445	C	G	0,322	8,20E-07	0,015	0,003	rs67416193	*RNU6-791P*	intergenic
rs146962840	1	248797787	G	C	0,047	8,00E-07	-0,049	0,010	rs146962840	*OR2T35*	intergenic
rs145287839	4	169571145	G	T	0,025	9,20E-07	-0,052	0,011	rs145287839	*PALLD*	intronic
rs547840497	5	107469179	C	T	0,014	1,80E-07	-0,062	0,012	rs547840497	*FBXL17*	intronic
rs806282	6	91197382	T	G	0,264	4,80E-07	0,015	0,003	rs284509	*MAP3K7*	intergenic
rs284509	**6**	**91208376**	**G**	**T**	**0,231**	**5,30E-08**	**0,017**	**0,003**	**rs284509**	***MAP3K7***	**intergenic**
rs284511	6	91208542	C	T	0,237	6,80E-08	0,017	0,003	rs284509	*MAP3K7*	intergenic
rs182259	6	91209111	T	C	0,231	1,70E-07	0,016	0,003	rs284509	*MAP3K7*	intergenic
rs11193156	10	108759228	A	T	0,134	1,30E-07	-0,024	0,005	rs11193156	*SORCS1*	intronic
rs112120989	13	27636336	T	C	0,023	1,10E-07	-0,041	0,008	rs56940748	*USP12*	intergenic
rs542629237	13	27661417	C	T	0,030	3,50E-07	-0,038	0,007	rs56940748	*USP12*	intronic
rs56940748	13	27662100	A	G	0,023	8,40E-08	-0,045	0,008	rs56940748	*USP12*	intronic
rs73497361	13	27663613	C	T	0,023	8,70E-08	-0,045	0,008	rs56940748	*USP12*	intronic
rs190770959	**16**	**12158574**	**T**	**C**	**0,015**	**6,30E-09**	**-0,056**	**0,010**	**rs190770959**	***SNX29***	**intronic**
rs147978408	**16**	**12171475**	**C**	**T**	**0,016**	**6,60E-09**	**-0,055**	**0,010**	**rs190770959**	***SNX29***	**intronic**
rs4239212	17	26309966	G	A	0,092	3,70E-07	-0,025	0,005	rs4239212	*RP11-218F4.1*	intergenic
rs6045318	20	1883451	G	A	0,024	7,10E-07	-0,040	0,008	rs6045318	*SIRPA*	intronic

chr: Chromosome; pos: position; MAF: Minor allele frequency; gwasP: GWAS p-value; SE: Standard Error; IndSigSNP: Index significant SNP

Table 3.4: Selected SNPs with p-values <1E-06 for the sex-difference test

SNP	CHR	BP	A1	A0	A1FREQ W	BETA W	SE W	P W	A1FREQ M	BETA M	SE MEN	P M	p Sex diff	Genic position	Nearest Gene
rs12350396	9	38487113	T	C	0,860	0,009	0,004	1,30E-02	0,852	-0,018	0,004	8,20E-06	3,46E-07	intergenic	*IGFBPL1;FAM95C*
rs150840489	15	71088277	G	A	0,975	-0,051	0,009	2,40E-09	0,977	0,012	0,009	1,70E-01	4,38E-07	intergenic	*UACA;LARP6*
rs11585710	1	17276223	C	T	0,617	0,007	0,003	4,90E-03	0,612	-0,012	0,003	2,70E-05	4,84E-07	intronic	*CROCC*
rs11586910	1	17276357	T	C	0,617	0,007	0,003	4,90E-03	0,612	-0,012	0,003	2,70E-05	4,84E-07	intronic	*CROCC*
rs150481830	11	17754454	T	C	0,458	0,012	0,003	7,70E-05	0,469	-0,009	0,003	3,30E-03	4,90E-07	intergenic	*MYOD1;KCNC1*

SNP: Single Nucleotide Polymorphism;
CHR: chromosome;
A1: alternative allele;
A0: reference Allele;
A1FREQW: alternative allele frequency women;
A1FREQM: alternative allele frequency men;
PW: p-value women;
PM: p-value men;
BETA W: beta women;
BETA W: beta men;
SE W: standard error women ;
SE M: standard error men;
p Sex diff: p-value sex-difference

3.3.2. GWAS Catalogue lookup

Because of the restricted number of genome-wide significant SNPs previously reported, our look-up also included screening for phenotypes similar to cIMT (coronary artery calcification (CAC), abdominal aortic aneurysm (AAA)). Our study replicated the locus for association with cIMT in the *CBFA2T3* region with two SNPs are 18979 bp apart from one another: rs9934287 (p = 6.6E-06) was found suggestive of association in our study whereas rs844396 (p = 6.00E-09) was previously reported by Franceschini and collaborators in a European ancestry population (Franceschini *et al.*, 2018). The previously reported variant rs844396 in Europeans did not show any evidence of association in our study (p = 0.85). The rs9934287 SNP (MAF=0.047) is monomorphic in populations of European ancestry from 1000GP. A previously reported locus for association with carotid plaque in European populations (Pott *et al.*, 2017) at *GEM* (rs72672639, p = 4.0E-06) was replicated in our female-specific subset with two SNPs (rs78571209, rs76489670, p= 7.8E-05) approximatively 2200 bp away from the SNP reported for plaque in Europeans. Equally, the association with the *MRPL37* locus (rs11206301, p= 8.00E-06) for plaque in European populations was replicated in our male-specific analysis for cIMT (rs13374450, p = 3.0E-05). The two SNPs in *MRPL37* locus were not in LD despite their proximity (201 bp). The suggestive variant in our study rs4773141 (p = 4.7E-05, in the combined set), located in *COL4A1*, was previously reported for CAD (p = 4.0E-17) in European populations (Van der Harst and Verweij, 2017).

For the combined analysis, a total 10 SNPs replicated loci previously reported for CAC (a surrogate marker of atherosclerosis as cIMT) (Inouye *et al*. 2012; O'Donnell *et al*. 2011) and for CAC in Type 2 Diabetes African patients (Divers *et al.*, 2017). Fourteen SNPs replicated loci for coronary heart diseases, coronary artery disease (Lettre *et al.*, 2011; Schunkert *et al.*, 2011; Dichgans *et al.*, 2014; Nikpay *et al.*, 2015; Van der Harst and Verweij, 2017; Charmet *et al.*, 2018) (**Supplementary Table 3.2-3.3**).

In female-specific GWAS, 13 SNPs were suggestive of replication for CAC (Inouye *et al.*, 2012; Wojczynski *et al.*, 2013; Divers *et al.*, 2017), 16 SNPs were replicating loci for coronary heart diseases, coronary artery disease, coronary aneurism and coronary atherosclerosis (Kim *et al.*, 2013; Van der Harst and Verweij, 2017; Ward-Caviness *et al.*, 2017; Siewert and Voight, 2018). Five variants (p < 3.1E-05) replicated a locus on *MIR100HG* reported for association with stroke in African Americans (Carty *et al.*, 2015), and rs114299344 (p = 2.2E-05, beta = -

0.043) replicated the *ADAMTS2* locus in paediatric stroke (rs469568, p = 8.0E-06) (Arning *et al.*, 2012) (**Supplementary Table 3.2-3.3**).

For male-specific analyses, 18 SNPs replicated 6 loci for CAC (Inouye *et al.*, 2012; Wojczynski *et al.*, 2013; Divers *et al.*, 2017). For coronary heart disease, coronary artery disease, and coronary atherosclerosis, 13 SNPs replicated loci (*ZNF652*, *ZFPM2-AS1/ZFPM2*, *PKD2L1*, *CFDP1*, *AC027506.1/AC007948.1*, *AC096558.2/ARHGAP15/AC096558.1*, *C1GALT1*, *SLC22A3*, *LPAL2*, *LPA*) (Schunkert *et al.*, 2011; Dichgans *et al.*, 2014; Van der Harst and Verweij, 2017; Ward-Caviness *et al.*, 2017). rs75601989 (p = 2.0E-05) replicated *RN7SL363P/FURIN* locus for stroke (Malik *et al.*, 2018) (**Supplementary Table 3.2-3.3**).

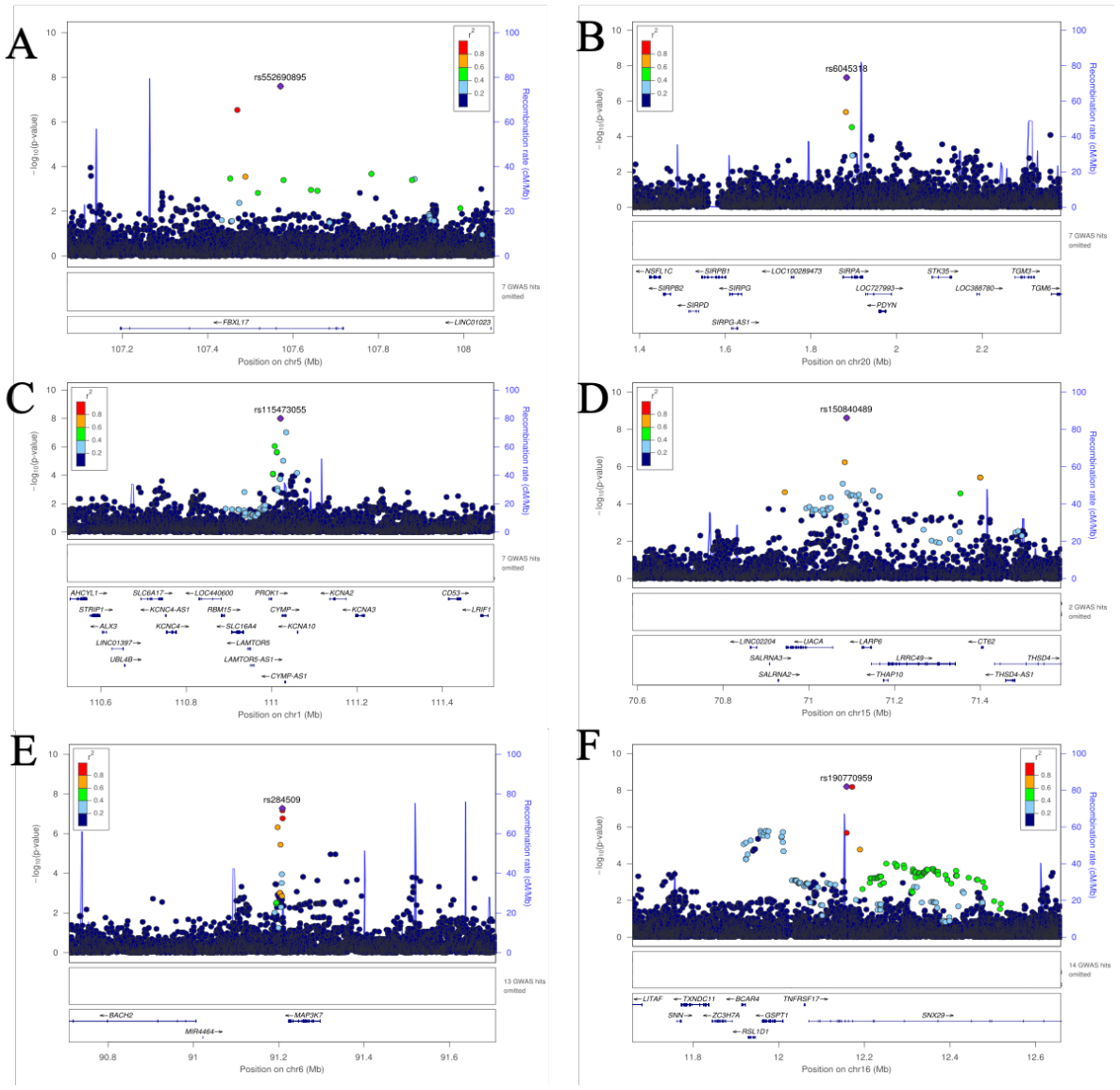


Figure 3.5: Regional association plots for the selected top SNPs showing genetic associations.

(A) Regional association plot of the *FBXL17* region in the combined dataset. (B) Regional association plot of the *SIRPA* region in the combined dataset. (C) Regional association plot of the *PROK1* region in the female-specific dataset. (D) Regional association plot of the *LARP6/UACA* region female-specific dataset. (E) Regional association plot of the *MAP3K7* region in the male-specific dataset. (F) Regional association plot of the *SNX29* region in the male-specific dataset. For each locus, the plots show the –log10 transformed p–value of each SNP on the y-axis and base pair positions along the chromosomes on the x-axis. Genes overlapping the locus are displayed below the plot. SNPs are coloured by their LD value (generated from the study populations) with the lead SNP in the region shown as purple diamonds.

3.3.3. Functional annotation

Annotation of the genic position of the 467, 515 and 581 SNPs respectively from combined, female-specific and male-specific analysis with significant and suggestive associations (p<1E-05) showed that these were mostly intronic or intergenic (**Tables 3.4a, 3.4b, 3.4c**). 50 SNPs displayed a CADD score above 12.37 suggestive of being potentially deleterious (19 in the combined; 18 in female-specific; 13 in male-specific datasets) (**Supplementary Table 3.4a, 3.4b, 3.4c**). In the female-specific sample, the lead SNP on *CYMP* (rs115473055) had a RDB score of 2a suggesting the variant was likely affecting transcription binding sites (CTCF). Positional mapping, eQTL mapping (matched cis-eQTL SNPs) and chromatin interaction mapping (on the basis of 3D DNA–DNA interactions) is reported (**Supplementary Table 3.5a, 3.5b, 3.5c**). We found that rs78172571 in high LD with rs150840489 (the top SNP associated in our female-specific) was involved in HiC type chromatin interactions in multiple tissues including aorta, in which the variant act as an enhancer of *THAP10* (FDR = 2.03E-17).

Table 3.5: Positional annotation of candidate SNPs for SNPs in LD (0.8) with the lead-SNPs which have corresponding functional annotation assigned by ANNOVAR with enrichment relative to all SNPs in the 1000 GP reference African panel.

	Annotation	Ref.Count	Ref.Prop	Count	Prop	Enrichment	Fisher P
Combined :	UTR3	409561	0,009370	4	0,008565	0,91	1,0E+00
	UTR5	123855	0,002834	3	0,006424	2,27	1,5E-01
	downstream	490010	0,011211	6	0,012848	1,15	6,6E-01
	exonic	429504	0,009827	8	0,017131	1,74	1,5E-01
	intergenic	20334166	0,465229	150	0,321199	0,69	2,9E-10
	intronic	15981769	0,365650	160	0,342612	0,94	3,1E-01
	ncRNA_exonic	453334	0,010372	5	0,010707	1,03	8,2E-01
	ncRNA_intronic	5020829	0,114872	123	0,263383	2,29	9,5E-19
	ncRNA_splicing	2186	0,000050	1	0,002141	42,81	2,3E-02
	splicing	4299	0,000098	0	0,000000	0,00	1,0E+00
	upstream	458349	0,010487	7	0,014989	1,43	3,5E-01
Female :	UTR3	409561	0,009370	3	0,006012	0,64	6,4E-01
	UTR5	123855	0,002834	0	0,000000	0,00	4,1E-01
	downstream	490010	0,011211	6	0,012024	1,07	8,3E-01
	exonic	429504	0,009827	9	0,018036	1,84	6,8E-02
	intergenic	20334166	0,465229	225	0,450902	0,97	5,3E-01
	intronic	15981769	0,365650	192	0,384770	1,05	3,8E-01
	ncRNA_exonic	453334	0,010372	10	0,020040	1,93	4,3E-02
	ncRNA_intronic	5020829	0,114872	43	0,086172	0,75	4,9E-02
	ncRNA_splicing	2186	0,000050	0	0,000000	0,00	1,0E+00
	splicing	4299	0,000098	0	0,000000	0,00	1,0E+00
	upstream	458349	0,010487	11	0,022044	2,10	2,3E-02
Male :	UTR3	409561	0,009370	4	0,006885	0,73	8,3E-01
	UTR5	123855	0,002834	5	0,008606	3,04	2,6E-02
	downstream	490010	0,011211	5	0,008606	0,77	6,9E-01
	exonic	429504	0,009827	6	0,010327	1,05	8,3E-01
	intergenic	20334166	0,465229	204	0,351119	0,75	2,8E-08
	intronic	15981769	0,365650	304	0,523236	1,43	1,4E-14
	ncRNA_exonic	453334	0,010372	7	0,012048	1,16	6,8E-01
	ncRNA_intronic	5020829	0,114872	37	0,063683	0,55	3,9E-05
	ncRNA_splicing	2186	0,000050	0	0,000000	0,00	1,0E+00
	splicing	4299	0,000098	0	0,000000	0,00	1,0E+00
	upstream	458349	0,010487	9	0,015491	1,48	2,2E-01

Annot: annotated region; Ref.count: count of SNPs in the reference population; Ref.prop: proportion in the reference population; Count: count of SNPs in the sample; Prop: proportion of SNPs in the sample

3.3.4. Gene-based and gene-set analysis

In a gene-based analysis of the combined dataset analysis there was a significant association with *CALD1* (p = 5.9E-07 < 2.6E-06) (**Figure 3.6A**) with Mean-Max cIMT, whereas in female-specific analysis *FLT4* (p = 4.3E-07 < 2.6E-06) was significantly associated (**Figure 3.6B**). The results from gene-set analysis in the combined dataset showed significant enrichment for "Chemical and Genetic perturbation" in RAMASWAMY METASTASIS UP gene-set (adjP = 3.9E-05). The female-specific analysis revealed significant enrichment of gene-sets (**Supplementary Table 3.6a, 3.6b, 3.6c**), with among them "Hallmark gene-sets for Oestrogen response", with "Early Oestrogen response" and "Late Oestrogen response" both being significant (2.2E-6).

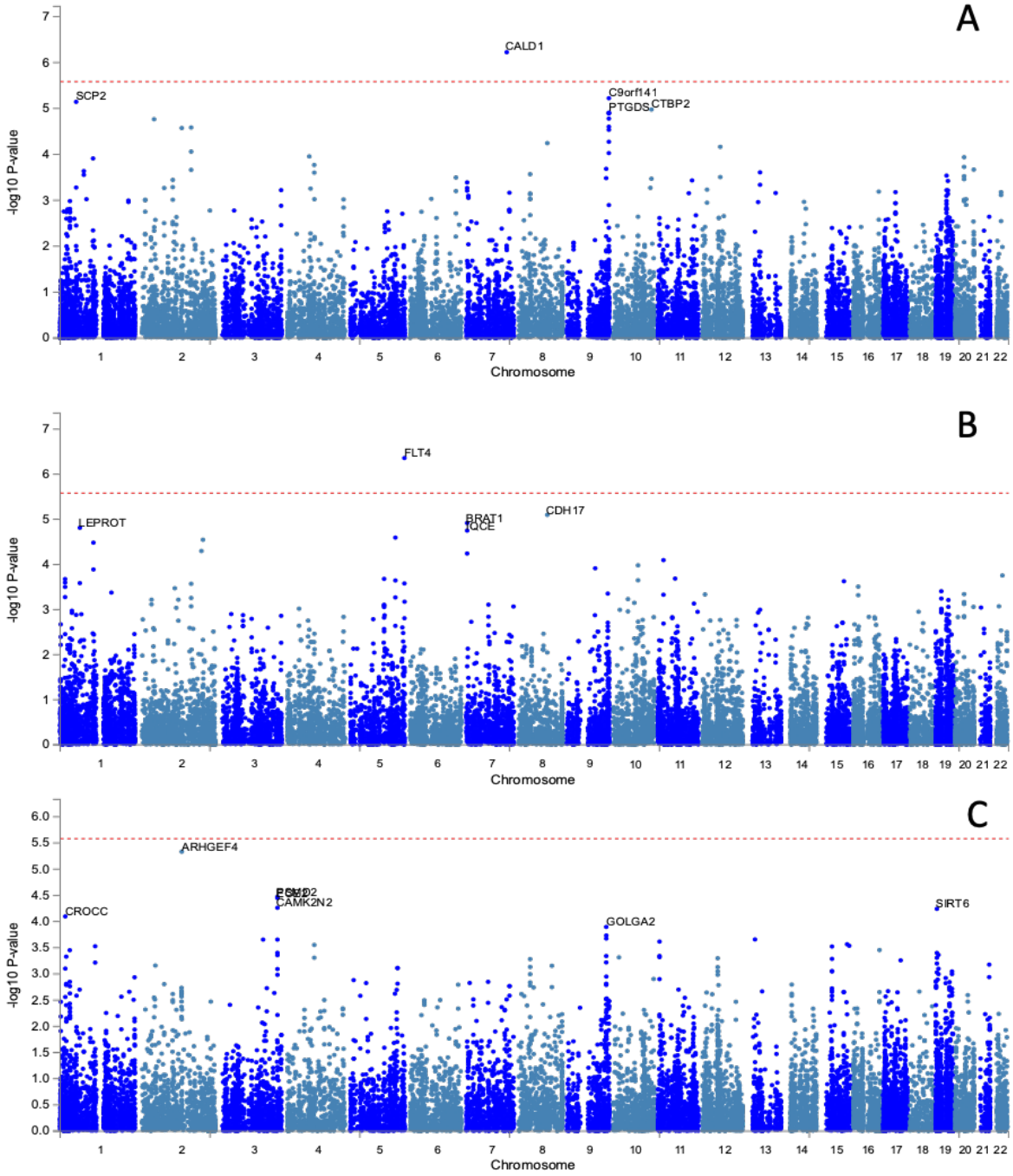


Figure 3.6: Manhattan plot for the gene-based test as computed by MAGMA based on our summary statistics.

Input SNPs were mapped to 19152 protein coding genes. Genome wide significance (red dashed line in the plot) was defined at P = 0.05/19152 = 2.611e-6. (A) combined dataset. (B) Female-specific. (C) Male-specific

3.4. Discussion

In this GWAS for cIMT as the outcome variable, as a proxy for atherosclerosis in an African population, we replicated regional associations with known loci associated with atherosclerosis and CVDs, but with different lead SNPs at almost all loci. In total, 54 loci were replicated with SNPs at p<1E-04 within 25 kb of previously reported genome-wide significantly associated variants, those loci were associated with the following: atherosclerosis phenotypes (cIMT, carotid plaque, coronary artery calcification, and abdominal aortic aneurism) and outcomes (coronary artery disease, coronary heart disease, myocardial infarctus, and stroke). We also identified new loci associated with cIMT in our African population at p < 5E-08.

We discussed below the identified loci in our study with regard to their potential functions, biological evidence from previous studies and previous reports from GWAS.

3.4.1. F-box and leucine rich repeat protein 17 (*FBXL17*)

Our study identified rs552690895 (p = 2.5E-08) in *FBXL17* in the combined set. F-box and leucine rich repeat protein 17 (*FBXL17*) is characterized by an approximately 40-amino acid F-box motif. SCF complexes, formed by *SKP1* (S-phase kinase-associated protein 1), cullin (*CUL1*), and F-box proteins, act as protein-ubiquitin ligases. F-box proteins interact with SKP1 through the F-box, and they interact with ubiquitination targets through other protein interaction domains (Jin *et al.*, 2004). Evidence suggest that the SCF is a key complex in the ubiquitine-proteasome system (UPS) that is involved in 70.0-90.0% of protein degradation processes (Depre, Powell and Wang, 2010). It has been found that protein degradation by the UPS play a central role in cardiovascular physiology and disease: from endothelial function, the cell cycle, atherosclerosis, myocardial ischaemia, cardiac hypertrophy, inherited cardiomyopathies, and heart failure (Fasanaro, Capogrossi and Martelli, 2010; Herrmann, Lerman and Lerman, 2010; Powell and Divald, 2010; Stangl and Stangl, 2010; Yu and Kem, 2010). A GWAS in Lithuanian families found that variants in *FBXL17* were associated with coronary heart diseases (Domarkiene *et al.*, 2013). Other reports on genetic associations for *FBXL17* include studies for educational attainment and mathematical ability (Lee *et al.*, 2018), intelligence (Davies *et al.*, 2018), and pulse pressure (Evangelou *et al.*, 2018).)

3.4.2. Signal regulatory protein alpha (*SIRPA*)

Our combined analysis identified also genetic association of rs6045318 (p = 4.7E-08) with cIMT in *SIRPA* gene. Signal regulatory protein alpha (*SIRPA*) is a regulatory membrane glycoprotein from the SIRP family, which inhibits the cytoskeleton-intensive process of phagocytosis by the macrophage. *SIRPA* is activate cancer cells (through high expression of CD47) and upregulated *SIRPA* inhibits macrophage-mediated destruction. This mediation of phagocytosis and polarization of macrophages is important in the pathophysiology of atherosclerosis (Chen *et al.*, 2019). There is evidence that *SIRPA* is involved in discrete stages of cardiovascular cell lineage differentiation (Skelton *et al.*, 2014) and that defects in the gene (knock out) reduces atherosclerosis in mice (Szilagyi *et al.*, 2014). *SIRPA* expression has been found as a signature of inflamed atherosclerotic plaque (Puig *et al.*, 2011). Previous reports on GWAS studies associate *SIRPA* with blood protein (Emilsson *et al.*, 2018; Sun *et al.*, 2018), and the percentage of basophil in granulocytes (Astle *et al.*, 2016).

3.4.3. Sorting nexin 29 (*SNX29*)

On the chromosome 16, rs147978408 (p = 6.3E-09) was the top cIMT associated variant in *SNX29* for the male-specific analysis. The sorting nexin (*SNX*) family is a diverse group of cytoplasmic- and membrane-associated phosphoinositide-binding proteins that play pivotal roles in the regulation of protein trafficking. *SNX* gene variants are associated with CVDs, and dysfunction of the *SNX* pathway is involved in several forms of cardiovascular disease (CVD) (Yang *et al.*, 2019). In a study of genes that regulate smooth muscle cell differentiation and disease risk, *SNX29* was involved in pathways for occlusion of blood vessels and atherosclerosis (Iyer *et al.*, 2018). Ito and collaborators identified sex-dependent differentially methylated regions close to *SNX29* in mouse liver and found that this methylation status was influenced by testosterone and contributed to sex-dimorphic chromatin decondensation (Ito *et al.*, 2015). This might explain the sex-specific effect observed in our study. A study in children with sickle cell disease, identified *SNX29* variants as suggestive of association with systolic blood pressure (Bhatnagar *et al.*, 2013). Also, variants in *SNX29* were found in suggestive association with subcutaneous adipose tissue in women (Sung *et al*. 2016). In patients with pulmonary arterial hypertension, *SNX29* variation was reported for differential responses to vasodilator treatment (Karnes *et al.*, 2017). Further GWAS analysis stratified by hypertensive status showed that the association was driven by the hypertensive group (effect three times higher in hypertensives compared to the non-hypertensives), therefore demonstrating that the association of *SNX29*

with cIMT might be mediated by the vascular remodeling caused by hypertension. GWA studies reported *SNX29* variants for association with educational attainment, mathematical ability and cognitive function measurement (Lee *et al*. 2018), intelligence (Davies *et al.*, 2018; Savage *et al.*, 2018), bone mineral density (Morris *et al.*, 2019), and smoking (Liu *et al*. 2019).

3.4.4. Mitogen-activated protein kinase kinase kinase 7 (*MAP3K7*)

On the chromosome 6, we found rs284509 (p=5.3E-08) in *MAP3K7* region associated with cIMT in the male-specific analysis. Mitogen-activated protein kinase kinase kinase 7 (*MAP3K7*) also called *TAK1* encodes a serine/threonine protein kinase family member, with a central role as regulator of cell death. Because of its role in kinase pathway, and regulation of transforming growth factor beta (TGF-β), *MAP3K7* plays a role in growth inhibition in vascular smooth muscle cells and can be atheroprotective or atherogenic (Toma and MacCaffrey, 2012). More biological evidence of the contribution of *MAP3K7* to atherosclerosis is through its regulation by micro-RNAs (Njock *et al.*, 2015; Paone *et al.*, 2019). In a study of women receiving hormone replacement therapy, variants in *MAP4K4,* a gene targeting *MAP3K7* (Xie *et al.*, 2007) were associated to cIMT (Miller *et al.*, 2013). The sex-specific association observed might be related to the fact that *MAP3K7IP3* (located on the X chromosome), which is known to form a ternary complex with *MAP3K7* in response to inflammatory stimuli, has shown sex-differential expression in ischemic stroke (Stamova *et al.*, 2012; Rocha *et al.*, 2016). In a study on expression of androgen-modulated micro-RNAs, it has been reported that *MAP3K7* was a target of mmu-miR-467h and mmu-miR-669i in the angiogenesis and transforming growth factor beta receptor signalling pathways (Bouhaddioui, Provost and Tremblay, 2016). Despite biological relevance to atherosclerosis, GWA studies reported variants in gene region to be associated with cancer progression (Penney *et al.*, 2019), anti-TNF response in rheumatoid arthritis (Honne *et al.*, 2016), attention deficit hyperactivity disorder (Lasky-Su *et al.*, 2008), adolescent idiopathic scoliosis (Liu *et al*. 2018), and sporadic amyotrophic lateral sclerosis (Xie *et al*. 2014). Our study is the first to report *MAP3K7* association with a CVD phenotype.

3.4.5. La-related protein 6 (*LARP6*)

LARP6 (La-related protein 6) is a ribonucleoprotein domain family member 6. Studies showed that it has a role in collagen regulation by targeting mRNA encoding Type I collagen (Cai *et al*. 2010; Zhang and Stefanovic 2016; Glenn, Wang, and Schwartz 2009; Stefanovic *et al*. 2019).

Sukhanov and collaborators found that *IGF1R* deficiency downregulated collagen mRNA-binding protein *LARP6* and vascular collagen, and showed an atheroprotective effect (Sukhanov *et al.*, 2018; Danchuk, Delafontaine and Higashi, 2019). Collagen is a hallmark of atherosclerotic plaque stability, thus alteration of the collagen balance may lead to an instability of atherosclerotic lesions, and therefore promote plaque formation and rupture (Puig *et al.*, 2011; Higashi *et al.*, 2016). In the Taiwanese population, the *LARP6* locus was found to be associated with coronary artery disease (Assimes *et al.*, 2016). In European ancestry populations, *LARP6* was found associated with insulin measurement (Strawbridge *et al.*, 2011). However, Mendelian randomization for cIMT found that despite the limited effects of proinsulin-increasing SNP scores on cIMT, proinsulin was unlikely to have causal effects on cIMT (Strawbridge *et al.*, 2017). Myocardial gene expression in non-ischemic human heart failure found *LARP6* to be differentially expressed between men and women (1.36 fold) (Fermin *et al.*, 2008). The female-specific effect of the loci may find its explanation in the enhancer function of rs78172571 in high LD with rs150840489 (the top SNP associated in our female-specific) on *THAP10* gene (FDR = 2.03E-17) known to be regulated by oestrogen.

3.4.6. Prokineticin 1 (*PROK1*)

Prokineticin 1 (*PROK1*), also called endocrine gland derived vascular endothelial growth factor (*EG-VEGF*), is a specific placental angiogenic factor which play a role in the control of normal (e.g endometrial decidualization) and pathological placental angiogenesis (Hoffmann, Feige and Alfaidy, 2006). It is involved in pathologies such as recurrent pregnancy loss, gestational trophoblastic diseases, foetal growth restriction, and preeclampsia (LeCouter *et al.*, 2001; Kisliouk *et al.*, 2003; Samson *et al.*, 2004; Alfaidy *et al.*, 2014; Corlan *et al.*, 2017). The gene is known to be predominantly expressed in the steroidogenic glands, such as ovary, testis, and adrenal cortex, and is often complementary to the expression of vascular endothelial growth factor (*VEGF*), suggesting that these molecules function in a coordinated manner. The function and particular pattern of this gene's activity might explain why we identified the locus only in our female-specific analysis. Our study is the first to report *PROK1* for any trait in a GWAS.

3.4.7. Caldesmon 1 (*CALD1*)

Our gene-based analysis identified caldesmon 1 significantly associated with cIMT in our combined set led by rs7781307 (p = 2.1E-06) on 7q33. Caldesmon 1 is calmodulin binding protein encoding for a calmodulin-and actin-binding protein that play a major role in the

regulation of smooth muscle contraction, cell migration and cell invasion (Mayanagi and Sobue, 2011). *CALD1* was identified as key gene in the "regulation of actin cytoskeleton" module from protein-protein interaction network resulting from a bioinformatics analysis of key pathways and genes in advanced coronary atherosclerosis (Tan *et al.*, 2017). A study screening for keys genes for abdominal aortic aneurysm found that *CALD1* was leading a KEGG enrichment signal pathways (Vascular smooth muscle contraction) of differential expressed genes (DEGs) (Wan *et al.*, 2018). Under expression of *CALD1* was found to be a key feature of calcification of vascular smooth muscle cells from atherosclerotic plaque (Puig *et al.*, 2011; Goikuria *et al.*, 2018; Trillhaase *et al.*, 2018). Additionally, studies on epigenetic modifications reported *CALD1* to exhibit differential methylation in atherosclerosis (Zaina *et al.*, 2014; Nazarenko *et al.*, 2015; Fernández-Sanlés *et al.*, 2017). Previous GWAS reported CALD1 for phenotypes such as Lung function (FEV1/FVC) (Kichaev *et al.*, 2019), Response to paliperidone in schizophrenia (PANSS score) (Li *et al*. 2017), Attention deficit hyperactivity disorder symptom score (Middeldorp *et al.*, 2019), Diverticular disease (Maguire *et al.*, 2018)

3.4.8. Fms-related tyrosine kinase (*FLT4*)

FLT4 or Vascular endothelial growth factor receptor 3 (*VEGFR3*) is a major signalling protein involved in angiogenesis, vasculogenesis and maintenance of the endothelium. By acting as receptor to VEGFC and VEGFD, it plays an essential role in lymphangiogenesis in adults and in the development cardiovascular system during embryonic phase. Defect and/or downregulation of *VEGR3* was found to lead to cardiovascular failure in embryonic stage and to higher mortality after myocardial infarction in mice models (Dumont *et al.*, 1998; Vuorio *et al.*, 2018). Biological studies have highlighted the role of *FLT4* in atherosclerosis in major pathological processes. The gene has been reported to be involved in plaque instability by two process: the mediation of monocytes/macrophages apoptosis and consequently alteration plaque stability (Schmeisser *et al.*, 2006) ; and the modulation of vascular remodelling and shear stress resulting in plaques haemorrhages and calcification in carotids (Baeyens *et al.*, 2015; Tuenter *et al.*, 2016; Camaré *et al.*, 2017). Our study is the first to report association of *FLT4* locus (rs112967731, p = 5.7E-07, female-specific) with cIMT or any cardiovascular phenotype in GWAS studies. Previous studies reported the locus for association with folic acid measurement (Deng *et al.*, 2018) and blood protein measurements (Emilsson *et al.*, 2018; Sun *et al.*, 2018).

3.4.9. Relevance of our findings

Relevant to the evidence for sex-specific genetic effects, 50 to 150 genes were found to be differentially expressed by sex in tissues of the cardiovascular system (Gershoni and Pietrokovski, 2017), and approximatively 15% of expression quantitative trait loci (eQTLs) are sex-biased in human (Dimas *et al.*, 2012). Analysis of sex-dimorphism requires both a significant SNP association with cIMT in at least in one sex and a significant sex-difference for the SNP association (P-value testing for difference in sex-specific effect estimates). Several scenarios can describe sexual dimorphism for SNP associations: (i) concordant effect direction (CED), if the association is present in one sex (P-men or P-women at p < 1E-04) and at least nominally significant and directionally concordant in the other (P-women or P-men,0.05, respectively); (ii) single sex effect (SSE), if the association is present in one sex and not significant in the other; or (iii) opposite effect direction (OED), if the association is present in one sex and at least nominally significant in the opposite direction in the other sex (Randall *et al.*, 2013). In our study, we identified all three types of sexual dimorphism: the *LARP6* locus was a case of a single sex effect (rs150840489: p-female = 2.4E-09, beta- female = -0.051; p-male = 0.17, beta- male = 0.012); the *CROCC* locus showed opposite effect direction (rs11585710: p- female = 490E-03, beta- female = 0.007; p- male = 2.7E-05, beta- male = -0.012), and the *FBXL17* variant showed a concordant effect direction was identified (rs547840497, p- female = 0,037, beta- female = -0.022; p- male = 1.8E-07, beta-male= -0.062). In total 177 SNPs showed CED, 89 SNPs had OED and 3213 SNPs showed SSE.

Measurements of cIMT are used clinically to assess vascular pathophysiology and to reflect the atherosclerosis process. Our study identified cIMT-associated loci relevant to genes related to macrophage activity and polarisation (*SIRPA*), to vascular smooth muscle cells (*MAP3K7, CALD1*), to vascular endothelial growth (*PROK1, FLT4*), to collagen synthesis and plaque stability (*LARP6*), and a pathway of blood vessel occlusion (*SNX29*). Our sex-specific analyses revealed loci which support the hypothesis that sex-dimorphism may be due to sex-specific epigenetic modification, that is, regulation of gene expression in a sex-specific manner, independent of sex hormone levels. When analysing sex-specific or gene-sex interactions, it is important to keep in mind that we are also reflecting the influences of non-genetic factors such as behaviour, environmental exposure, anatomical differences, and sex hormone environment,

which create systemic differences between males and females for trait expression, and which in turn affect disease risk and heritability (Gilks, Abbott and Morrow, 2014).

Our study identified significant enrichment of oestrogen pathway genes in our female-specific analysis. Oestrogen-dependent regulation of vascular gene expression and vascular physiology encompasses complex processes involving both nuclear and membrane-associated oestrogen signalling pathways. In recent years we have witnessed major progress in understanding how these regulatory processes contribute to the atheroprotective effects exerted by oestrogens. Animal models of atherosclerosis provided compelling evidence that physiological oestrogen levels potently attenuate both early and advanced stages of atherosclerosis lesion development in females, and suggested similar protective effects in males. The effect of oestrogens on atherosclerosis can target metabolism (lipid, glucose), macrophage function or smooth muscle cells. Nonetheless, hormone replacement therapy during menopause has not been shown to conclusively reduce atherosclerosis risk, suggesting that more studies are needed to fully decipher the biological mechanisms.

3.4.10. Strength and limitations

Our study is the first and the largest population-based study to investigate the genetic architecture of cIMT in sub-Saharan African populations. In addition to our sex-specific analyses, we tested for sex-difference between the two strata using a minimal model (adjustment for age and PCs) to avoid the "collider bias". We used an analysis framework allowing us to identify genetic effects that point in opposite directions in men and women and to detect genetic effects that are only (or more pronounced) in one stratum without filtering overall association test filtering prior to a difference test, this method has been shown to have better power to identify qualitative gene-sex interactions (Winkler *et al.*, 2017). The use of a new SNP genotyping array with better representation of common African variants and imputation reference panels from African participants have allowed us to improve the SNP coverage in ethnically diverse African samples significantly. The lack of an ethnically matched replication cohort is a limitation in our study, and it will be important to replicate these findings in additional suitable cohorts. We identified African-specific variants in new loci and replicated previously reported loci, revealing opportunities for trans-ancestry fine-mapping.

Chapter 4:

Novel and known gene-smoking interactions with cIMT identified as potential drivers for atherosclerosis risk in West-African populations of the AWI-Gen Study

Chapter 4: Novel and known gene-smoking interactions with cIMT identified as potential drivers for atherosclerosis risk in West-African populations of the AWI-Gen Study

ABSTRACT

Atherosclerosis is a key contributor to the burden of cardiovascular diseases (CVDs). Many epidemiological studies have reported on the effect of smoking on carotid intima-media thickness (cIMT) and its subsequent effect on CVD risk. Gene-environment interaction studies have contributed towards understanding some of the missing heritability of genome-wide association studies. Gene-smoking interactions on cIMT have been studied in non-African populations (European, Latino-American and African American), but no comparable African research has been reported. Our aim was to investigate smoking-SNP interactions on cIMT in two West African populations by genome-wide analysis. Male participants were recruited from Burkina Faso (Nanoro = 993) and Ghana (Navrongo = 783). Smoking was rare among women, therefore they were not included. Phenotype and genotype data underwent stringent QC and genotype imputation was performed using the Sanger African Imputation Panel. Smoking prevalence among men was 13.3% in Nanoro and 42.5% in Navrongo. We analysed gene-smoking interactions with PLINK after adjusting for covariates: age and 6 PCs (Model 1); age, BMI, blood pressure, fasting glucose, cholesterol levels and 6 PCs (Model 2). All analyses were performed at site level and for the combined dataset.

In Nanoro, we identified new variants for the gene-smoking interactions for cIMT within the previously described *RCBTB1* region (rs112017404, rs144170770 and rs4941649) (Model 1 p values = 1.35E-07; Model 2 p-value = 3.08E-08). In the combined sample, two novel intergenic interacting variants were identified, rs1192824 in the regulatory region of *TBC1D8* (p-value = 5.90E-09) and rs77461169 (p-value = 4.48E-06) located in an upstream region of open chromatin. In silico functional analysis suggests the involvement of genes implicated in biological processes related to cell or biological adhesion and regulatory processes in gene-smoking interactions with cIMT (as evidenced by chromatin interactions and eQTLs), possibly increasing the risk for atherosclerosis.

The current study is the first on gene-smoking interactions for cIMT as a risk factor for atherosclerosis in sub-Saharan African populations, and in addition to replicating previously

known signals for *RCBTB1*, we identified two novel genomic regions (*TBC1D8*, near *BCHE*) involved in this GxE.

Keywords: GWIS; atherosclerosis, smoking, cIMT, gene-environment interactions

4.1 Introduction

During the last two decades, the burden of CVDs has increased considerably, and low- and middle-income countries are now experiencing about 80% of the worldwide burden. Sub-Saharan Africa (SSA) is undergoing a health and demographic transition that has shifted the major causes of death from communicable and nutritional diseases to non-communicable diseases (NCDs). The mean age of death attributable to CVDs in SSA in 2010 was 64.9 years (95% CI, 64.4–65.4) compared with 67.6-81.2 years for the rest of the world (Roth *et al.*, 2015), making it one of the youngest affected populations globally. Populations of African descent have been under-represented in genomic studies, representing only about 3% of the participants worldwide used for genome-wide association studies up to 2016 (Popejoy and Fullerton, 2016). This is a gap that needs to be filled, considering that Africa is the continent with the highest genetic diversity owing to its deep evolutionary roots, and African genomes generally have lower linkage disequilibrium (LD). A previous study reported that African populations were more diverse and had significantly more genes and pathways involved in extreme allele frequency differences (EAFD) (Sulovari *et al.*, 2017). The African genome is therefore highly relevant for the discovery of new genetic associations and a better understanding of human disease mechanisms (Tekola-Ayele and Rotimi, 2015; Choudhury *et al.*, 2018; Martin *et al.*, 2018).

Genetic understanding of complex traits has developed immensely over the past decade but remains hampered by the missing heritability. Hence, a large proportion of the phenotypic variance of many traits remains ill-defined. A major contributor to this phenotypic variance includes gene-environment interactions (GxE) (Duncan *et al.*, 2014). G×E can be defined broadly as the interplay between the product of a genetic variant and an environmental factor as they affect a specific trait. GxE therefore refers to modification, by an environmental factor, of the effect of a genetic variant on a phenotypic trait. The phenotypic flexibility resulting from adjustments due to GxE could determine or modulate health or disease by modulating the adverse effects of a risk allele, or exacerbating the genotype-phenotype relationship to increase risk. Environmental stimuli, acting over hundreds of generations, have been promoting adaptation that is observed in current populations for traits and disease risk.

Identifying GxE represents the cornerstone in the path to “Precision Public Health” that will allow individuals to adjust exposure to a particular environmental factor involved in GxE interactions for the benefit of lessening disease risk according to a fixed genotype. However, if

GxE testing represents an opportunity, challenges remain in interaction studies: interaction can be dependent on scale, SNP-based analyses can lack power, exposure measurements can be inconsistent and imperfect, and optimal software for efficient GxE analysis is lacking. Fortunately, recent advances in methodology development have boosted GxE interaction analysis and more studies are being published.

Smoking is an important risk factor for coronary heart disease (CHD) and cardiovascular disease (CVD) (Schroeder, 2013). Despite improved understanding, the pathophysiological mechanisms underpinning the association between smoking and CVD have yet to be elucidated fully. Nonetheless smoking is known to have an effect on endothelial cells, inflammatory states, platelet activation, pro-coagulant factors and anti-fibrinolytic factors (Barua and Ambrose, 2013). Several studies have reported the effect of smoking on subclinical atherosclerosis (Liang *et al.*, 2009; Yang *et al.*, 2015; Hansen *et al.*, 2016; Kianoush *et al.*, 2017). Atherosclerosis is a complex, progressive disorder affecting large and medium-sized arteries. The disease has a silent progression, often with no clinical evidence until the occurrence of a vascular event.

Gene-smoking interactions for atherosclerosis have been reported. In the Bogalusa Heart Study, a variant located in the region of *EDNRA* was found to be associated with the status of the left cIMT (Li *et al*. 2015), and in the Northern Manhattan Study (NOMAS), *RCBTB1* was reported as a modifier of the smoking effect on cIMT, and *MXD1-JPH1* for carotid plaque burden in the presence of smoking (Wang *et al*. 2014; Della-Morte *et al*. 2014). Genes involved in inflammatory pathways mediated by the NF-kB axis (*TBC1D4* and *ADAMTS9*) have been identified displaying a gene-smoking interaction (Polfus *et al.*, 2013). In 2017, a case-control study on coronary heart disease identified variants in *ADAMTS7* responsible for a loss of cardio-protective effects resulting from gene-smoking interactions (Saleheen *et al.*, 2017). The Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen) is a Collaborative Centre of the Human Heredity and Health in Africa (H3Africa) Consortium investigating the genomic and environmental risk factors for cardio-metabolic diseases in Africans that provided the context and data for this study.

In this paper, we report on a genome-wide analysis of gene-environment interactions to explore the role of smoking on cIMT (a measure of atherosclerosis) in male participants from Nanoro (Burkina Faso) and Navrongo (Ghana) in West-Africa, as part of the AWI-Gen study.

4.2 Method

4.2.1 Study population and phenotype assessments

The study was cross-sectional in nature and investigated populations from two sub-Saharan African sites in West Africa (Burkina Faso and Ghana) as a sub-set of the AWI-Gen study (Derra *et al.*, 2012; Oduro *et al.*, 2012; Ramsay *et al.*, 2016; Ali *et al.*, 2018). The participants for this study include 1776 West African men from two rural settings, Nanoro (Burkina Faso) and Navrongo (Ghana), aged 40 to 60 years. Only men were included in this study as smoking rates in women are very low in these communities. Participants completed a questionnaire with questions on demography, health history and behaviour. Anthropometric measurements were taken and blood collected for genotyping (H3Africa array) and phenotyping (biomarkers) (Ali *et al.*, 2018). Ultrasound scans were performed to assess right and left common cIMT. The reason for performing the analysis on the two West African study sites, and not the other AWI-Gen study sites, is that they were comparable in terms of environmental exposures and genetic background (Derra *et al.*, 2012; Oduro *et al.*, 2012).

4.2.2 cIMT Measurement

cIMT was measured using Dual B-mode ultrasound images of the carotid tree showing a typical double line for the arterial wall. Details of the method for measurement were provided in **Chapter 2.** Mannheim Consensus defining the use of cIMT in population-based studies has been applied to QC cIMT values. The mean cIMT was calulated as the average of the mean common right and left cIMT.

4.2.3 Smoking measurement and other variables

Smokers (current) and non-smokers (never and former) were classified based on self-reporting. A dichotomous categorization of smoking status was chosen over a quantitative measure (e.g., pack-years) owing to the inherent high dimensionality of GWAS analysis (13.98 million SNPs with two-fold main effects and interaction variables per SNP). Additionally, previous studies reported the reversal of the effect of smoking after several years of cessation. Smoking intensity was assessed using pack/years calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked.

Other variables recorded (BMI, blood pressure, glucose, total cholesterol, physical activity, medical history), have been previously described (Ali *et al.*, 2018).

4.2.4 Association and follow up analysis

4.2.4.1 Genotyping and Imputation

The H3Africa genotyping array, designed as an African-common-variant-enriched GWAS array (Illumina) with ~2.3 million SNPs, was used to genotype genomic DNA using the Illumina FastTrack Sequencing Service[8]. The following pre-imputation QC steps were applied to the entire AWI-Gen genotype data set. Individuals with a missing SNP calling rate greater than 0.05 were removed. SNPs with a genotype missingness greater than 0.05, MAF less than 0.01 and Hardy-Weinberg equilibrium (HWE) P-value less than 0.0001 were removed. Non-autosomal and mitochondrial SNPs, and ambiguous SNPs that did not match the GRCh37 references alleles or strands were removed. Imputation was performed on the cleaned dataset (with 1,729,661 SNPs and 10,903 individuals) using the Sanger Imputation Server and the African Genome Resources as reference panel. We selected EAGLE2 (Loh *et al.*, 2016) for pre-phasing and the default PBWT algorithm was used for imputation. After imputation, poorly imputed SNPs with info scores less than 0.6, MAF less 0.01, and HWE P-value less than 0.00001 were excluded. The final QC-ed imputed data had 13.98 M SNPs, and data from male participants from the Nanoro and Navrongo study sites were extracted for analysis.

4.2.4.2 Data Analysis

Descriptive statistics were used to summarise the population characteristics. Continuous variables were reported in median and interquartile ranges and categorical variables were reported in percentages. All the data were analysed per site before reporting for the combined group. We examined group differences using the Mann-Whitney test for continuous variables and Pearson Chi-square test for categorical variables.

4.2.4.3 Gene-Environmental interaction (GxE) analysis

Linear regression of mean cIMT was performed with covariates using R[9]. Residuals were extracted from the linear regression analyses and used for the GWAS analysis. Model 1 used covariate ages and 6 principal components (PCs) computed on genetics data (to account for genetic structure). To check the consistency of our association, a second model (Model 2) with further adjustment was implemented (model 1 + BMI + systolic blood pressure + diastolic blood

[8] https://www.illumina.com/services/sequencing-services.html
[9] https://www.R-project.org/

pressure + fasting glucose + total cholesterol + physical activity (moderate to vigorous physical activity in minutes per week (MVPA)) to analyse SNPs with p-values lower than 1e-06 in Model 1.

GxE testing was performed using the PLINK "-gxe" option (Purcell *et al.*, 2007; Chang *et al.*, 2015) on the 'awigen' branch of the automated workflow[10] (Baichoo *et al.*, 2018). We screened the output for a genome-wide significance threshold (p-values < 5.E-08).

To assess genomic inflation, we compared our observed distribution of $-\log10(P)$ values to that expected in the absence of association (Lambda) and illustrated the results in QQ plots. We used a cross-replication approach between the two sites, suggestive signals (p-values < 1E-04) in one site were checked in the other site and vice-versa.

Power calculations were performed with Quanto[11] (Version 1.2.4). The study was powered at 80% to identify SNPs with MAF ≤0.05 and interaction effect size of > 4, given our sample size and smoking prevalence in Nanoro and power would be higher for Navrongo and the combined dataset.

4.2.4.4 Functional analysis

The FUMA online platform[12] (Watanabe *et al.*, 2017) was used to annotate, prioritize, visualize and interpret GWAS results. From GWAS summary statistics as an input, it provided extensive functional annotation for all SNPs in genomic areas identified by lead SNPs. From the list of gene IDs (as identified by SNP2GENE option in FUMA) FUMA annotated genes in biological context (Watanabe *et al.*, 2017). We selected all candidate SNPs in the associated genomic region having $r^2 \geq 0.6$ with one of the independently significant SNPs, with a suggestive P-value (P < 1E-05) and MAF > 0.05 for annotation. Predicted functional consequences for these SNPs were obtained by matching the SNP's chromosome base-pair position, and reference and alternate alleles, to databases containing known functional annotations, including ANNOVAR (Wang, Li, and Hakonarson 2010), combined annotation-dependent depletion (CADD) scores (Kircher *et al.*, 2014), and RegulomeDB (RDB) (Boyle *et al.*, 2012) scores**.**

[10] http://github.com/h3abionet/h3agwas/
[11] http://biostats.usc.edu/Quanto.html
[12] http://fuma.ctglab.nl/

4.2.4.5 Functional annotation of mapped genes

Genes implicated by mapping of significant GWAS SNPs were further investigated using the GENE2FUNC procedure in FUMA (Watanabe *et al.*, 2017), which provides hypergeometric tests of enrichment of the list of mapped genes in 53 GTEx tissue-specific gene expression sets (The GTExArd Consortium *et al.*, 2015; GTex Consortium *et al.*, 2017), 7,246 MSigDB gene sets[13], and chromatin states (Ernst and Kellis, 2013; Consortium Roadmap Epigenomics *et al.*, 2015).

4.2.4.6 GWAS Catalog lookup

The GWAS Catalog database was downloaded from the website[14] (Accessed on 12 Jul 2018) and a subset the dataset generated using the following key words relevant to our study: genome-wide interaction, gene-environment interactions atherosclerosis, coronary artery diseases, carotid atherosclerosis, cIMT, coronary artery calcification, abdominal artery aneurism. Since our dataset was in build 37, lift-over was performed to allow comparison. We searched for SNPs in a 100 kb region of all suggestive SNPs (p-value≤1E-04) found in this study and further filtered the list using key-words pertaining to coronary artery diseases and gene-environment interactions.

4.2.5 Ethics and consent

This study received the approval of Human Research Ethics Committee (Medical), University of the Witwatersrand/South Africa (M121029), the approval of the Centre Muraz Institutional Ethics Committee/Burkina Faso (015-2014/CE- CM) and the approval of the National Ethics Committee For Health Research/Burkina Faso (2014-08-096), the Ghana Health Service Ethics Review Committee (ID No: GHS-ERC:05/05/2015) and the Navrongo Institutional Review Board (ID No: NHRCIRB178). All the participants signed an Informed Consent Form before any study procedure was performed.

[13] http://software.broadinstitute.org/gsea/msigdb
[14] https://www.ebi.ac.uk/gwas/

4.3 Results

4.3.1 Characteristics of participants

The participant characteristics are presented for Nanoro (993), Navrongo (783) and for the combined data (1776) in **Table 4.1**. Only males were included. Smoking prevalence was 13.3%, 42.5% and 26.2%, respectively for Nanoro, Navrongo and the combined dataset. In Nanoro, smokers were younger than non-smokers. Interestingly, although the prevalence was lower, the smoking intensity was much higher (in packs/year) in the Nanoro individuals, 6.6 (3.7-11.6) in Nanoro vs 1.7 (0-4.2) in Navrongo. Moreover, although there were no differences in mean cIMT values between smokers and non-smokers, significant differences were observed for the following risk factors, BMI, systolic and diastolic blood pressure, for both sites and overall. Total cholesterol and low-density lipid cholesterol (LDL-C) were lower in smokers compared to non-smokers in Nanoro and the combined sample, whereas fasting glucose was lower for smokers in Navrongo and the combined sample. Smokers in our study displayed lower BP and LDL-C measurements compared to non-smokers, further investigation beyond the scope of this thesis are needed to understand these trends. In spite of the geographic and genetic proximity of the two groups, the differences in the prevalence and smoking intensity, as well as that of other risk factors, suggest that the GxE interaction mechanism might not be the same in the two centres.

Table 4.1: Descriptive characteristics for participants from Nanoro (Burkina Faso) and Navrongo (Ghana) and the combined sample

	Nanoro			Navrongo					
	Non-Smokers (861)	**Smokers (132)**	**P-values♦**	**Non-Smokers (450)**	**Smokers (333)**	**P-values♦**	**Non-Smokers (1311)**	**Smokers (465)**	**P-values♦**
Age	50 (45-55)	48 (43-52)	**0.0024**	50 (45-56)	51 (46-55)	0.5053	50 (45-55)	50 (45-55)	0.6104
BMI (kg/m^2)	21.31 (19.47-23.61)	19.50 (17.99-21.54)	**<0.001**	20.99 (19.42-22.95)	20.03(18.58-21.44)	**<0.001**	21.22 (19.46-23.37)	19.90 (18.38-21.45)	**<0.001**
Systolic BP (mmHg)	118 (109-132)	113 (101-124)	**<0.001**	123 (112-136)	119 (108-133)	**0.0033**	120 (110-133)	117 (107-131)	**0.0016**
Diastolic BP (mmHg)	76 (69-83)	72 (64-81)	**0.0012**	76 (68-85)	74 (66-83)	**0.0233**	76 (69-84)	74 (65-82)	**<0.001**
Fasting Glucose (mmol/l)	4.90 (4.46-5.35)	4.94 (4.52-5.42)	0.5672	4.49 (4.05-5.00)	4.29 (3.94-4.73)	**<0.001**	4.77 (4.30-5.24)	4.47 (4.06-4.96)	**<0.001**
Total cholesterol (mmol/l)	3.59 (2.96-4.26)	3.3 (2.88-3.84)	**0.0063**	3.10 (2.57-3.66)	3.18 (2.61-3.67)	0.4888	3.37 (2.78-4.10)	3.21 (2.67-3.70)	**<0.001**
Triglycerides (mmol/l)	0.70 (0.53-0.97)	0.70 (0.55-0.94)	0.7984	0.55 (0.42-0.73)	0.57 (0.43-0.73)	0.6104	0.64 (0.48-0.89)	0.60 (0.45-0.77)	**<0.001**
LDL-C (mmol/l)	1.98 (1.51-2.55)	1.82 (1.44-2.14)	**0.0031**	1.70 (1.22-2.19)	1.67 (1.28-2.06)	0.6141	1.88 (1.39-2.44)	1.72 (1.31-2.08)	**<0.001**
HDL-C (mmol/l)	1.12 (0.94-1.36)	1.13 (0.88-1.41)	0.8254	1.12 (0.89-1.34)	1.16 (0.94-1.45)	**0.0352**	1.12 (0.92-1.36)	1.16 (0.92-1.43)	0.1209
MVPA (min/week))	1485 (80-3055)	2520 (420-3155)	**0.0185**	2408 (1035-3060)	2475 (1080-3120)	0.7610	2220 (250-3060)	2490 (720-3120)	**<0.001**
Cigarette (Pack/Year)ϒ	----	6.6 (3.7-11.6)	**----**	----	1.7 (0-4.2)	**----**	----	2.8 (0-6)	
Mean cIMT (mm)	0.673 (0.598-0.758)	0.656 (0.585-0.736)	0.0952	0.678 (0.610-0.768)	0.680 (0.607-0.753)	0.7332	0.675 (0.603-0.760)	0.675 (0.598-0.750)	0.4297

♦ Mann-Whitney rank-sum test

ϒ Number of pack-years = (number of cigarettes smoked per day/20) × number of years smoked

Abbreviations: Body Mass Index (BMI); Blood pressure (BP); Moderate to vigorous physical activity (MVPA)

4.3.2 Gene-Smoking interactions

The association results of SNPs (MAF ≤ 0.05) are illustrated by Manhattan plots for each site and the combined dataset in **Figures 4.1A, 4.1B, 4.1C**. Genomic inflation factors were 0.994, 0.993 and 1.007, respectively for Nanoro (7,828,913 SNPs), Navrongo (7,842,446 SNPs) and the combined sample (7,839,440 SNPs) (**Figures 4.1D, 4.1E, 4.1F**).

The strongest signal in the Nanoro sample was found for the GxE where the C allele of rs7649061 (allele frequency (AF) = 0.25, p-value = 7.37E-08); this was associated with a decrease of mean-cIMT in the presence of smoking (**Table 4.2a, Supplementary Table 4.1**). This variant is located in an intergenic region between *BCHE* (butyrylcholinesterase) and *ZBBX* (zinc finger B-box domain containing). The next strongest signal in Nanoro was rs7095209 (p-value = 1.23E-07), an intergenic SNP close to the *SORCS3* gene (sortilin related VPS10 domain containing receptor 3), the G allele of which (AF = 0.42) was associated with a decrease of mean-cIMT in the presence of smoking (**Supplementary Figure 4.1 & 4.2**). We identified three variants located on chromosome 13 (rs112017404, rs144170770 and rs4941649) in high LD with the lowest p-value at 1.35E-07 (AF = 0.08), associated with an increase of mean-cIMT in smokers and located between *RCBTB1* (Regulator of chromosome condensation (RCC1) and BTB (POZ) domain containing protein 1) and *ARL11* (ADP-ribosylation factor-like 11) (**Figure 4.2A**). There was a signal in the *CNBD1* (cyclic nucleotide binding domain containing 1) region led by rs13268575 (p = 2.77E-07) which was associated with an increase of mean-cIMT in smokers for the A allele (AF = 0.23). A missense variant, rs17844302 (p-value = 4.07E-07), was identified in *PCDHA6* (protocadherin alpha 6), and found to be associated with an increase of mean-cIMT in smokers.

In the Navrongo sample (**Table 4.2b, Supplementary Table 4.1**), the strongest signal was found for rs4869800 (p-value = 1.92E-06, AF = 0.92), an intergenic variant between *RGS17* (regulator of G-protein signaling 17) and *OPRM1* (opioid receptor, mu 1). Other suggestive signals included rs12743305 (2.72E-06) between *LINC01307* and *OLFM3* (olfactomedin 3), rs452108 (p-value = 3.17E-06) between *OR6K3* (olfactory receptor, family 6, subfamily K, member 3) and *OR6K6* (olfactory receptor, family 6, subfamily K, member 6), and rs11043124 (p-value = 3.36E-06) between *ZBED5-AS1* (Zinc finger, BED-type containing antisense RNA 1) and *GALNT18* (Polypeptide N-acetylgalactosaminyltransferase 18) (Supplementary Figure 1 & 2).

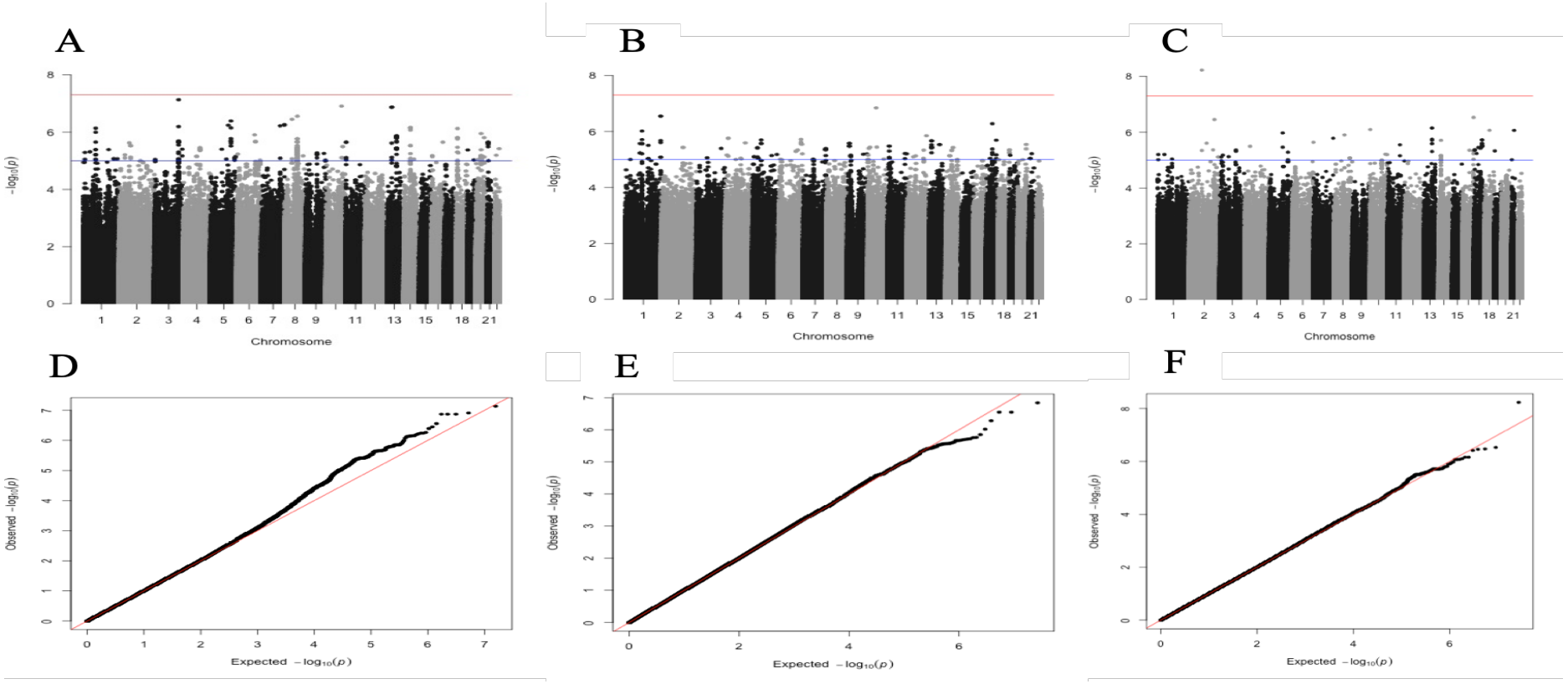


Figure 4.1: Manhattan plots showing the –log10-transformed two-tailed P-value of each SNP from the GWAS gene-smoking interaction for cIMT on the y-axis and base-pair positions along the chromosomes on the x-axis.

The red line indicates Bonferroni-corrected genome-wide significance (P < 5E-08); the blue line indicates the threshold for suggestive association (P < 1E-04). Manhattan plot for Nanoro, 993 participants, 78289913 SNPs (A). Manhattan plot for Navrongo, 783 participants, 7842446 SNPs (B). Manhattan plot for combined set, 1776 participants, 7839440 SNPs (C). QQ plot for Nanoro, GIF = 0.9944 (D). QQ plot for Navrongo, GIF = 0.9934 (E). QQ plot for combined set, GIF=1.0079 (F).

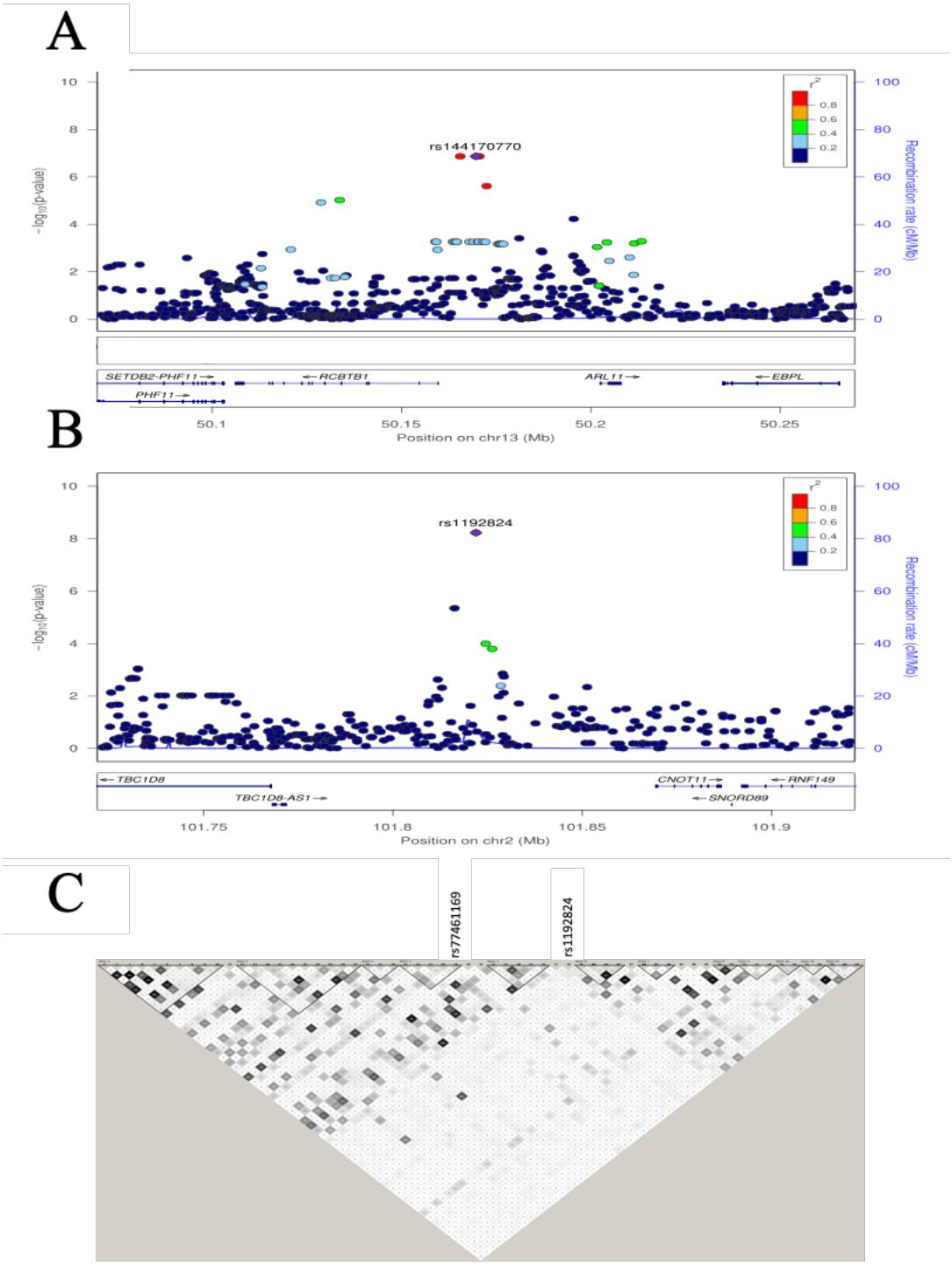


Figure 4.2: Regional and LD plots for the *RCBTB1* and *TBC1D8* regions.

Regional plots for the *RCBTB1* **in Nanoro** (A). Regional association plots of *TBC1D8* region in the *combined* dataset (B). Distinct genomic risk loci were defined as LD-independent regions (r^2 separated by 100 kb and containing one or more SNPs with a suggestive association (p-values < 1E-05). For each locus, the plots show the –log10 transformed p-value of each SNP on the y-axis and base pair positions along the chromosomes on the x-axis. Genes overlapping the locus are displayed below the plot. SNPs are colored by their LD value with the lead SNP in the region, and

those LD values has been generated from the study populations. Haplotypes blocks showing that rs1192824 and rs77461169 are not in LD. Haplotype blocks was built using Haploview with LD values calculated from the two study populations together (C).

Table 4.2a: Selected risk loci (p≤1E-05) for SNP-smoking interactions in Nanoro

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
1	rs205482	1	10813675	5,16E-06	10805454	10813675	rs205482	intronic	*CASZ1*
2	rs11161464	1	85298111	7,34E-07	84737302	85312366	rs616472; rs11161464	intronic	*LPAR3*
3	rs1057305	1	175126223	4,08E-06	175118895	175126423	rs1057305	UTR3	*KIAA0040*
4	rs79936010	1	210977179	8,55E-06	210924696	210992946	rs79936010	intronic	*KCNH1*
5	rs148751085	2	38310007	6,27E-06	38310007	38310986	rs148751085; rs187829394	intergenic	*CYP1B1, CYP1B1-AS1*
6	rs11126201	2	69084017	2,39E-06	69084017	69100735	rs11126201	intergenic	*ARHGAP25, BMP10*
7	rs6747064	2	78141022	2,98E-06	78134360	78164757	rs6747064	ncRNA_intronic	*LOC101927967*
8	rs4854101	2	241139141	3,90E-06	241139141	241139141	rs4854101	intergenic	*OTOS, GPC1*
9	rs4684453	3	4975879	9,01E-06	4975451	4976785	rs4684453	ncRNA_intronic	*BHLHE40-AS1*
10	rs7649061	3	166457671	7,37E-08	166453919	166541341	rs7649061	intergenic	*BCHE, ZBBX*
11	rs779272	3	191506748	3,94E-06	191493008	191506748	rs779272	intergenic	*LINCR-0002, FGF12*
12	rs79243722	4	93634354	7,08E-06	93634354	93634921	rs79243722; rs146860029	intronic	*GRID2*
13	rs6822279	4	99745705	9,88E-06	99722261	99745705	rs6822279	intergenic	*TSPAN5, EIF4E*
14	rs6819524	4	116593891	3,44E-06	116574034	116687428	rs6819524	intergenic	*NDST4, MIR1973*
15	rs73764616	5	76294836	3,96E-06	76294836	76294836	rs73764616	intergenic	*CRHBP, AGGF1*
16	rs79414964	5	107950880	8,74E-06	107938080	107952834	rs79414964	intergenic	*FBXL17, LINC01023*
17	rs2407218	5	121909381	5,78E-07	121876491	121909381	rs2407218	intergenic	*MGC32805, LOC101927379*
18	5:140208927	5	140208927	4,07E-07	140120573	140223827	5:140208927	exonic	*PCDHA6*
19	rs180973	5	165988671	8,08E-06	165984684	166020904	rs180973	intergenic	*LOC102546299, CTB-7E3.1*
20	5:171912171	5	171912171	7,15E-06	171903310	171916696	5:171912171	intergenic	*SH3PXD2B, NEURL1B*
21	rs6934349	6	6899063	5,89E-06	6899063	6899063	rs6934349	intergenic	*LY86, RREB1*
22	rs699402	6	122588540	1,24E-06	122427846	122588540	rs699402	intergenic	*GJA1, HSF2*
23	rs59322395	6	128914510	2,14E-06	128853214	128943383	rs76515539; rs76865199; rs59322395	intergenic	*PTPRK, LAMA2*

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
24	rs7450411	6	161010354	9,52E-06	160982051	161015406	rs7450411	intronic	*LPA*
25	rs4728159	7	128856192	6,07E-07	128855457	128980332	rs4728159	intergenic	*SMO, AHCYL2*
26	rs6978311	7	158665920	5,47E-07	158660582	158743970	rs6978311	intronic	*WDR60*
27	rs113394492	8	9817684	2,13E-06	9817065	9817684	rs113394492	intergenic	*MIR124-1, MSRA*
28	rs28890775	8	52207622	3,56E-07	52158173	52207622	rs28890775	intergenic	*SNTG1, PXDNL*
29	rs13268575	8	88391237	2,77E-07	88309814	88461836	rs13268575	intronic	*CNBD1*
30	rs74815096	8	128577675	6,79E-06	128571759	128577675	rs74815096	intergenic	*CASC8, CASC11*
31	rs28687781	9	78496674	5,33E-06	78492036	78503618	rs28687781	intergenic	*MIR548H3, PCSK5*
32	rs58733878	9	126968674	5,59E-06	126965953	126968674	rs58733878	intergenic	*LHX2, NEK6*
34	rs112473634	9	136780781	9,52E-06	136780781	136786697	rs112473634	intronic	*VAV2*
35	rs7095209	10	106781929	1,23E-07	106781929	106789694	rs7095209; rs7083262	intronic	*SORCS3*
36	rs55729345	10	115390091	3,18E-06	115390091	115392919	rs55729345; rs56273432; rs2270182	intronic	*NRAP*
37	11:2307694	11	2307694	3,82E-06	2307694	2307694	11:2307694	intergenic	*ASCL2, C11orf21*
38	rs116449999	11	3506177	2,27E-06	3499256	3523720	rs116449999	intergenic	*TSSC2, LOC101927708*
39	rs11054696	12	12271739	3,95E-06	12271739	12272466	rs11054696	UTR3	*LRP6*
41	rs112017404	13	50165471	1,35E-07	50107718	50358384	rs112017404	intergenic	*RCBTB1, ARL11*
42	rs2819241	13	83709519	1,33E-06	83620970	83734452	rs2819241	intergenic	*NONE, SLITRK1*
43	rs2357001	14	64352814	7,01E-07	64347625	64361496	rs2357001	intronic	*SYNE2*
44	rs11851487	14	88284233	8,78E-06	88273402	88286529	rs11851487	intergenic	*LOC283585, GALC*
45	rs77506474	16	6068251	6,68E-06	6067412	6069280	rs77506474	upstream	*RBFOX1*
46	rs74032742	16	78725080	2,26E-06	78725080	78725080	rs74032742	intronic	*WWOX*
47	rs11663276	18	8653957	7,53E-07	8640869	8656961	rs11663276	intergenic	*RAB12, GACAT2*
48	rs60637140	18	13808717	5,36E-06	13808717	13808717	rs60637140	intergenic	*RNMT, MC5R*
49	rs67256262	19	1778023	4,20E-06	1778023	1778023	rs67256262	intergenic	*ONECUT3, ATP8B3*
50	rs36010337	19	44530696	9,41E-06	44530696	44530696	rs36010337	intronic	*ZNF222*
51	rs113096000	20	35308843	1,12E-06	35300135	35433702	rs113096000	intronic	*NDRG3*

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
52	rs6025645	20	56157341	1,56E-06	56157341	56159276	rs6025645	intergenic	*PCK1, ZBP1*
53	rs9982779	21	20459913	2,28E-06	20455408	20467128	rs9982779	intergenic	*LOC101927797, LINC00320*
54	rs9680265	21	22433521	9,94E-06	22433492	22433521	rs9680265	intronic	*NCAM2*
55	rs111752108	22	21442319	3,00E-06	21442319	21442319	rs111752108	intergenic	*LRRC74B, BCRP2*
56	rs13058149	22	27831347	6,49E-06	27831347	27831347	rs13058149	intergenic	*LINC01422, MN1*
57	rs73425956	22	48984237	3,79E-06	48984237	48984237	rs73425956	intronic	*FAM19A5*

Table 4.2b: Selected risk loci (p≤1E-05) for SNP-smoking interactions in Navrongo

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
1	rs12743305	1	102039804	2,72E-06	102039804	102039858	rs12743305	intergenic	*LINC01307, OLFM3*
2	rs452108	1	158716549	3,17E-06	158643513	158716549	rs452108	intergenic	*OR6K3, OR6K6*
3	rs78380843	2	150390155	3,73E-06	150390155	150390760	rs78380843; rs77831436	intergenic	*LYPD6, MMADHC*
4	rs59483727	4	108724898	8,92E-06	108724898	108724898	rs59483727	intergenic	*PAPSS1, SGMS2*
5	rs62352637	5	38912972	9,02E-06	38907814	38922398	rs62352637	intronic	*OSMR*
6	rs6908852	6	130781638	4,59E-06	130781638	130787061	rs6908852	intergenic	*TMEM200A, SMLR1*
7	rs4869800	6	154002533	1,92E-06	153994951	154081678	rs4869800	intergenic	*RGS17, OPRM1*
8	rs2849519	6	163132372	6,23E-06	163128672	163142618	rs2849519	intronic	*PARK2*
9	rs79668927	7	154114710	4,15E-06	154107706	154120424	rs79668927	intronic	*DPP6*
10	rs59522150	9	12043812	2,68E-06	12043812	12043812	rs59522150	intergenic	*PTPRD-AS2, TYRP1*
11	rs10964835	9	21099991	7,03E-06	20914355	21100743	rs116544389; rs10964835	intergenic	*IFNB1, IFNW1*
12	rs11793060	9	133407956	6,28E-06	133407956	133407956	rs11793060	intergenic	*ASS1, LOC100272217*
13	rs76252315	10	367258	6,98E-06	306421	589928	rs76252315	intronic	*DIP2C*
14	rs11043124	11	11184095	3,36E-06	11177862	11189537	rs11043124	intergenic	*ZBED5-AS1, GALNT18*
15	rs891290	11	119245010	9,33E-06	119245010	119249355	rs891290	intronic	*USP2*
16	rs12797802	11	132031224	6,24E-06	132031224	132031224	rs12797802	intronic	*NTM*
17	rs7358623	12	59151544	7,42E-06	59151544	59151544	rs7358623	ncRNA_intronic	*LOC100506869, LOC101927653*
18	rs11067175	12	115013760	4,64E-06	115013760	115013760	rs11067175	intergenic	*TBX5-AS1, TBX3*
19	rs201818410	14	47817368	7,39E-06	47817368	47817372	rs201818410	intronic	*MDGA2*
20	rs11158732	14	68825191	3,77E-06	68825191	68825191	rs11158732	intronic	*RAD51B*
21	rs10873486	14	98256752	9,90E-06	98238493	98257638	rs10873486	intergenic	*LOC100129345, LINC01550*
22	rs116720528	17	40505531	4,24E-06	40505531	41070574	rs116720528; rs7225801	intronic	*STAT3*

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
23	rs74656678	17	67987435	2,06E-06	67967563	68023539	rs74656678	intergenic	*LINC01497, LINC01028*
24	rs7245961	19	4843612	7,04E-06	4843612	4843612	rs7245961	intronic	*PLIN3*
25	rs16981530	20	56537517	2,94E-06	56535706	56537631	rs16981530	intergenic	*MIR4532, C20orf85*

Table 4.2c: Selected risk loci (p≤1E-05) for SNP-smoking interactions in combined sample

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
1	rs6685095	1	6502548	6,23E-06	6502548	6505671	rs6685095	intronic	*ESPN*
2	rs11588151	1	55487648	6,22E-06	55487648	55487648	rs11588151	intergenic	*BSND, PCSK9*
3	rs74115213	1	115054290	9,02E-06	115052507	115054290	rs74115213	upstream	*TRIM33*
4	rs1192824	2	101821934	5,90E-09	101816286	101821934	rs77461169; rs1192824	intergenic	*TBC1D8, CNOT11*
5	rs12991332	2	138748302	2,49E-06	138748302	138748302	rs12991332	intronic	*HNMT*
6	rs73047627	2	189842214	4,27E-06	189842214	189842214	rs73047627	intronic	*COL3A1*
7	rs11695675	2	200607333	3,48E-07	200578867	200626678	rs11695675	intergenic	*LOC101927641, FTCDNL1*
8	rs12465362	2	230587448	7,63E-06	230569546	230588734	rs12465362	intergenic	*DNER, TRIP12*
9	rs7619102	3	12728045	7,65E-06	12511478	12833280	rs7619102	intergenic	*RAF1, TMEM40*
10	rs869784	3	24348008	6,57E-06	24347800	24355634	rs869784	intronic	*THRB*
11	rs9834968	3	69552204	7,45E-06	69552204	69573754	rs9834968	intergenic	*FRMD4B, MITF*
12	rs782444	3	127409941	4,32E-06	127315788	127411436	rs782444	UTR3	*MGLL*
13	rs76169119	4	40556609	3,20E-06	40556609	40556609	rs76169119	intronic	*RBM47*
14	rs77655815	5	107952834	1,06E-06	107938080	107952834	rs77655815	intergenic	*FBXL17, LINC01023*
15	rs10454990	5	148838117	5,17E-06	148831133	148838117	rs10454990	intergenic	*MIR143HG, CSNK1A1*
16	rs62435247	6	169725827	2,30E-06	169725827	169729242	rs62435247	intergenic	*THBS2, WDR27*
17	rs62475193	7	154476372	1,65E-06	154476372	154476372	rs62475193	intronic	*DPP6*
18	rs16900766	8	126649041	8,30E-06	126648227	126649582	rs16900766	intergenic	*TRIB1, LINC00861*
19	rs74509340	10	480774	7,97E-07	452477	589928	rs74509340	intronic	*DIP2C*
20	rs114020277	10	71741171	3,02E-06	71741171	71815919	rs114020277; rs144552991	intergenic	*COL13A1, H2AFY2*
21	rs1854462	10	91991941	9,95E-06	91991941	91995070	rs1854462; rs114248843	intergenic	*LINC01375, LOC101926942*
22	rs112148169	10	133210477	6,26E-06	133210477	133210477	rs112148169	intergenic	*TCERG1L, LINC01164*
23	rs1939309	11	100421331	2,84E-06	100421331	100421428	rs1939309	intergenic	*CNTN5, ARHGAP42*
24	rs7492237	13	84258581	7,01E-07	84251980	84279302	rs7492237	intergenic	*NONE, SLITRK1*
25	rs1546993	14	40153203	1,96E-06	40120082	40166542	rs1546993	intergenic	*FBXO33, LOC644919*

Locus	rsID	chr	pos	p	start	end	LeadSNPs	Func.refgene	Gene.refgene
26	rs60889063	16	54371831	4,39E-06	54359850	54392615	rs60889063	intergenic	*IRX3, CRNDE*
27	rs934166	16	84631300	4,63E-06	84631268	84631300	rs934166	intronic	*COTL1*
28	rs75985084	16	87142452	9,24E-06	87142452	87164603	rs75985084	intergenic	LOC440390, LOC101928708
29	rs11653290	17	6983529	4,22E-06	6983315	6987355	rs11653290	UTR5	*CLEC10A*
30	17:21509395	17	21509395	5,73E-06	21455337	21525990	17:21509395	intergenic	*C17orf51, FAM27L*
31	rs8067970	17	35145662	3,52E-06	35113090	35145662	rs8067970	intergenic	*MRM1, LHX1*
32	rs10401037	17	52235347	3,76E-06	52118993	52243229	rs10401037	intergenic	*KIF2B, TOM1L1*
33	rs115990591	17	55703525	1,30E-06	55703464	55704447	rs115990591	intronic	*MSI2*
34	rs4791039	17	64787600	1,90E-06	64782722	64790745	rs4791039	intronic	*PRKCA*
35	rs551160836	18	41413084	8,49E-07	41413084	41413084	rs551160836	intergenic	*SYT4, LINC01478*
36	rs8111212	19	6166665	4,65E-06	6142535	6166665	rs8111212	intronic	*ACSBG2*
37	rs113561658	20	40336290	4,47E-06	40334493	40345717	rs113561658	intergenic	*CHD6, PTPRT*
38	rs389332	21	18270964	9,65E-06	18270964	18283289	rs389332	intergenic	*MIR99AHG, LINC01549*
39	rs9974620	21	36097794	8,56E-07	36097794	36097794	rs9974620	ncRNA_intronic	*LINC00160*

In the combined sample, one SNP (rs1192824) reached the genome-wide significance level (**Table 4.2c, Supplementary Table 4.1**). The C allele of rs1192824 (AF = 0.69) showed a SNP-smoking interaction (p = 5.90E-09). This variant was located in the intergenic region between *TBC1D8* (TBC1 domain family, member 8) and *CNOT11* (CCR4-NOT Transcription Complex Subunit 11) (**Figure 4.2B**). The variant is in the promoter-flanking region of *TBC1D8*. rs77461169 is 5648 bp away from rs1192824, and showed a suggestive interaction (p = 4.48E-06), and was located in an open chromatin region. The two variants were not in LD (**Figure 4.2C**). The distribution of mean cIMT for the three rs1192824 genotypes showed no difference in the non-smokers, but there was a significant decrease of mean-cIMT for homozygote (C/C) and heterozygote (C/T) carriers among the smokers, when compared to the T/T genotype (**Figure 4.3**). This suggests a recessive mode of action for the risk allele (T). The second strongest signal was observed for rs12444312 (p-value = 1.27E-07) located in a non-coding RNA exon of *LOC440390*. This SNP is a regulatory region variant located in a CTCF binding site. A signal was found with rs11695675 (p-value = 3.48E-07) near *FTCDNL1* (formiminotransferase cyclodeaminase N-terminal like). Near *PCSK9*, rs1158815 was found suggestive of the GxE interaction for cIMT in the combined sample at p-value=6.22E-06 (**Supplementary Figure 4.1, 4.2**).

When adjustment was applied for additional covariates, Model 1 + BMI + systolic blood pressure + diastolic blood pressure + fasting glucose + cholesterol + physical activity (MVPA), in Model 2, four variants in 2 loci reached genome-wide significance (rs7649061, p = 2.20E-08; rs112017404 - rs144170770 - rs4941649, p=3.08E-08) in Nanoro. In the combined sample, rs1192824 (p = 2.70E-08) remained the single locus below the genome-wide significance threshold (**Table 4.3**). The direction of the allelic effect was the same in all cases where significant associations were identified in Nanoro and the combined sample.

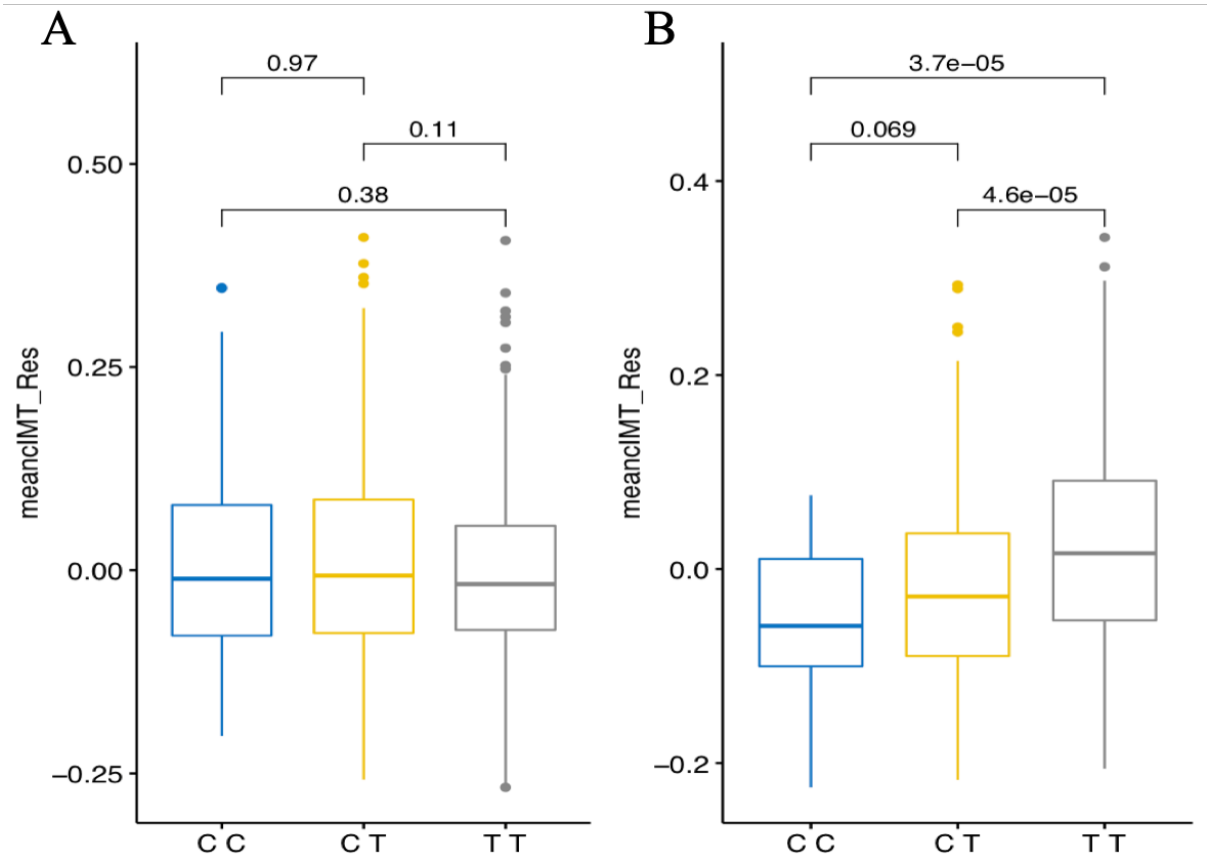


Figure 4.3: Genotypes plots of rs1192824 for smokers vs non-smokers.

Distributions of mean cIMT residuals (adjusted for age, 6 principal components (PCs) for population structure) in non-smokers (A) and smokers groups (B). Mean_cIMT_Res are in mm. Displayed p-values reports between genotype groups comparison performed using Kruskal-Wallis test.

Table 4.3: Comparison of gene-smoking association results for Nanoro and the combined sample (All), based on Model 1 and Model 2 for selected SNPs.

SNP	CHR	POS	A1	A2	MODEL 1				MODEL 2			
					All.P_GXE	Beta Interaction*	Nanoro.P_GXE	Beta Interaction*	All.P_GXE	Beta Interaction*	Nanoro. P_GXE	Beta Interaction*
rs11161464	1	85298111	G	A	3,29E-03	2,94	7,34E-07	4,95	4,61E-03	2,83	2,14E-07	5,19
rs1192824	**2**	**101821934**	**T**	**C**	**5,90E-09**	**-5,82**	**5,01E-04**	**-3,48**	**2,70E-08**	**-5,56**	**3,10E-03**	**-2,96**
rs11695675	2	200607333	A	G	3,48E-07	5,10	5,48E-03	2,78	3,27E-07	5,11	9,66E-03	2,59
rs7649061	**3**	**166457671**	**A**	**C**	**6,85E-04**	**-3,40**	**7,37E-08**	**-5,38**	**8,97E-04**	**-3,32**	**2,20E-08**	**-5,59**
rs7615817	3	166528151	G	A	1,70E-04	-3,76	6,45E-07	-4,98	2,01E-04	-3,72	1,27E-06	-4,84
rs2407218	5	121909381	G	A	4,05E-02	2,05	5,78E-07	5,00	3,20E-02	2,14	3,58E-07	5,09
5:140208927	5	140208927	A	C	7,98E-03	2,65	4,07E-07	5,06	1,08E-02	2,55	4,18E-07	5,06
5:140223827	5	140223827	A	G	1,15E-02	2,53	7,16E-07	4,96	1,71E-02	2,39	1,23E-06	4,85
rs4728159	7	128856192	C	T	1,84E-02	2,36	6,07E-07	4,99	2,66E-02	2,22	4,96E-06	4,57
rs115650684	7	158660582	T	C	7,74E-02	1,77	5,77E-07	5,00	1,49E-01	1,44	4,04E-06	4,61
rs6978311	7	158665920	A	G	7,49E-02	1,78	5,47E-07	5,01	1,46E-01	1,45	3,93E-06	4,61
rs28890775	8	52207622	T	C	9,83E-03	2,58	3,56E-07	5,09	8,92E-03	2,62	1,27E-06	4,84
rs13268575	8	88391237	C	A	2,96E-02	2,18	2,77E-07	5,14	4,14E-02	2,04	1,54E-06	4,81
rs74509340	10	480774	C	T	7,97E-07	-4,94	1,04E-01	-1,63	2,05E-06	-4,75	1,41E-01	-1,47
rs7095209	10	106781929	A	G	5,46E-02	-1,92	1,23E-07	-5,29	8,11E-02	-1,74	1,90E-07	-5,21
rs112017404	**13**	**50165471**	**T**	**C**	**5,95E-03**	**2,75**	**1,35E-07**	**5,27**	**8,83E-03**	**2,62**	**3,08E-08**	**5,54**

rs144170770	**13**	**50169651**	**G**	**A**	**5,95E-03**	**2,75**	**1,35E-07**	**5,27**	**8,83E-03**	**2,62**	**3,08E-08**	**5,54**
rs4941649	**13**	**50170562**	**C**	**A**	**6,64E-03**	**2,71**	**1,35E-07**	**5,27**	**1,06E-02**	**2,56**	**3,08E-08**	**5,54**
rs9546479	13	84256394	C	T	7,01E-07	-4,96	1,45E-03	-3,19	1,25E-06	-4,85	1,10E-03	-3,26
rs7492237	13	84258581	T	C	7,01E-07	-4,96	1,45E-03	-3,19	1,25E-06	-4,85	1,10E-03	-3,26
rs12435958	14	64352150	C	G	5,93E-03	-2,75	7,01E-07	-4,96	2,46E-03	-3,03	1,86E-06	-4,77
rs2357001	14	64352814	G	A	6,35E-03	-2,73	7,01E-07	-4,96	2,78E-03	-2,99	1,86E-06	-4,77
rs12895879	14	64353738	C	T	6,43E-03	-2,72	7,19E-07	-4,96	2,82E-03	-2,99	1,90E-06	-4,76
rs8004989	14	64354638	T	C	6,08E-03	-2,74	8,43E-07	-4,92	2,65E-03	-3,00	2,16E-06	-4,74
rs12444312	16	87097284	A	G	2,95E-07	5,13	1,15E-03	3,25	1,46E-06	4,82	2,52E-03	3,02
rs11663276	18	8653957	T	G	5,48E-02	1,92	7,53E-07	4,95	8,04E-02	1,75	1,46E-06	4,82
rs551160836	18	41413084	G	T	8,49E-07	4,92	3,96E-02	2,06	7,24E-07	4,96	3,78E-02	2,08
rs9974620	21	36097794	C	T	8,56E-07	4,92	8,42E-04	3,34	6,48E-07	4,98	6,67E-04	3,40

*For beta interaction, positive values indicate association of allele A1 and increased cIMT; negative values indicate association of allele A1 with decreased cIMT

In bold are the variant reaching genome-wide significance threshold (5E-08) in Model 2.

4.3.3 Comparison of association signals between sites

We found 8 SNPs from Nanoro with some evidence for interaction (p-values < 1E-04) which showed nominal replication (p-value < 0.05) in Navrongo, with the strongest being rs77655815 (Nanoro p-value = 4.01E-05), replicated at p-value of 6.83E-05, and rs79419964 (Nanoro, p = 8.74E-06; Navrongo, p = 8.23E-04), and with p = 1.06E-06 (2.66E-06 for Model 2) in the combined analysis. From Navrongo, 19 SNPs (p < 1E-04) were nominally replicated in Nanoro, of which rs12444312 (Nanoro p = 1.15E-03), was found associated at a p-value of 2.95E-07 in the combined sample (**Supplementary Table 4.1**). When significant, the direction of the allelic effect was the same in all cases.

4.3.4 GWAS Catalog lookup for gene-environment interaction

We replicated previously described gene-environment associations loci for smoking interactions with body composition (BMI and waist circumference) (Justice *et al.*, 2017), smoking or alcohol interaction with blood pressure (Feitosa *et al*. 2018; Sung *et al*. 2018; Taylor *et al*. 2016), coronary artery calcified plaque in type 2 diabetes (Divers *et al.*, 2017) and peripheral arterial disease interaction with air pollution (Ward-Caviness *et al.*, 2017) (**Supplementary Table 4.5**). No SNPs showed evidence of transference to African populations in our study. In the Nanoro sample, our study identified 15 SNPs located within 100 kb of previously reported gene-environment interaction loci. The loci include a gene-alcohol interaction on blood pressure (Feitosa *et al.*, 2018), gene-smoking interaction on waist circumference (Justice *et al.*, 2017), gene-smoking interaction on lung cancer (Park *et al.*, 2015) and gene-smoking interaction on blood pressure (Sung *et al*. 2018). For the Navrongo sample, 13 SNPs replicated previous loci for gene-smoking interaction on BMI (Justice *et al.*, 2017), lung cancer (McKay *et al.*, 2017) and blood pressure (Taylor *et al.*, 2016). Seventeen SNPs in the combined sample replicated previously reported interaction loci for gene-diabetes interaction for atherosclerotic plaque (Divers *et al.*, 2017), gene-alcohol interaction for blood pressure (Feitosa *et al.*, 2018), gene-smoking interaction for BMI (Justice *et al.*, 2017) and gene-smoking interaction for blood pressure (Sung *et al*. 2018; Taylor *et al*. 2016).

4.3.5 Functional analysis

Functional annotation of SNPs with suggestive associations showed that these were mostly intronic or intergenic (**Supplementary Tables 4.2a, 4.2b, 4.2c**). 30 SNPs displayed a CADD

score above 12.37 (17 in Nanoro; 3 in Navrongo; 10 in Combined), suspected to be deleterious. In the Nanoro sample, two SNPs (rs6701037, rs6677097), in high LD with rs10573305 (*KIAA0040* region), had a RDB score of 1f suggesting they were likely affecting binding sites and gene expression. Equally, for the combined sample two variants (rs10409209 and rs4807840), in LD with rs8111212, displayed a RDB score of 1f.

4.3.6 Functional annotation of mapped genes

Genes implicated by mapping of significant SNPs were further investigated using the GENE2FUNC procedure in FUMA. Positional mapping, eQTL mapping (matched cis-eQTL SNPs) and chromatin interaction mapping (on 3D DNA–DNA interactions) is reported (**Supplementary Tables (4.3a, 4.3b, 4.3c**), **Supplementary Figure 4.3**). We found that rs1192824 and rs77461169 in *TBC1D8* were implicated with eQTLs influencing the expression of *TBC1D8*, *SNORD89* and *RNF149* and also displaying chromatin interactions (**Figure 4.4A**). The *RCBTB1* locus SNPs were eQTLs for *RCBTB1*, *ARL11*, *CAB39L* and *PSME2P2*; their chromatin interactions included surrounding genes such as *RCBTB1*, *KPNA3*, *CAB39L*, *SETDB2*, *MLNR* and *CDADC1* (**Figure 4.4B**). A lookup into gene expression datasets, revealed expression of genes of interest in specific tissues such as arteries (**Supplementary Figure 4.4**).

4.3.7 Gene Set analysis

Gene set analysis was only reported when at least five genes were implicated. In Nanoro, significant Gene Ontology Biological Processes were found for biological adhesion (AdjP = 2.69E-11), cell-cell adhesion (AdjP = 8.03E-11), homophilic cell adhesion via plasma membrane adhesion molecules (AdjP = 1.29E-09) and cell-cell adhesion via plasma membrane adhesion molecules (AdjP = 5.52E-09) (**Figure 4.5; Supplementary Table 4.4a, 4.4b, 4.4c**).

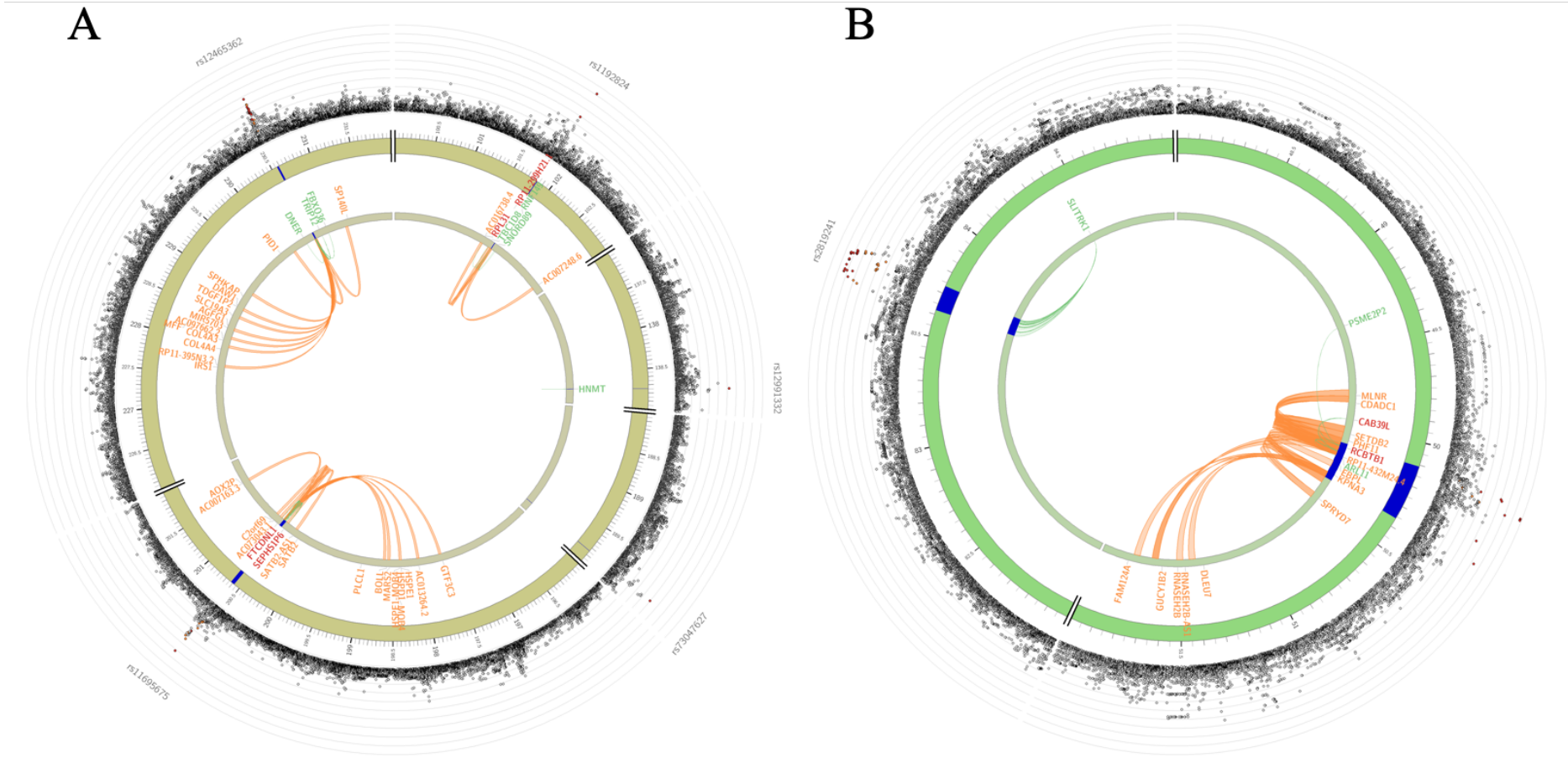


Figure 4.4: Circos plots showing genes on chromosome 2 for combined sample (A) and on chromosome 13 for Nanoro sample (B).

Blue regions are risk loci (p<1E-05). Green lines are connecting an eQTL SNP to its associated gene. Orange lines are connecting two interacting regions for chromatin interactions. Genes implicated by eQTLs are in green, by chromatin interactions are in orange, and by both eQTLs and chromatin interactions are in red. The outer layer shows a Manhattan plot containing the –log10-transformed two-tailed P-value of each SNP from the gene-environment interaction analysis, with SNPs coloured according to LD patterns with the lead SNP. Higher-resolution Circos plots for all chromosomes are provided in Supplementary Figure 4.3.

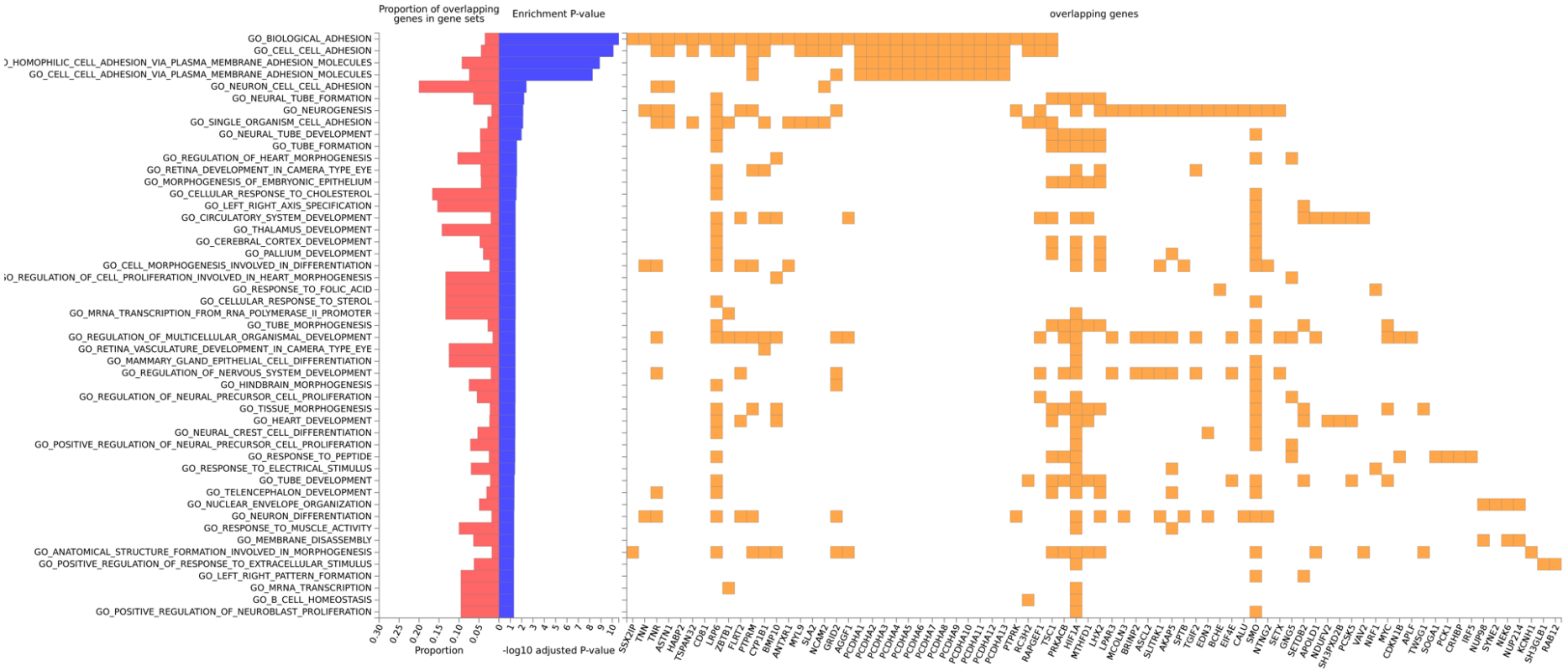


Figure 4.5: Gene Ontology enrichment for biological processes for gene-smoking interaction in Nanoro, with overlapping genes in gene sets.

4.4 Discussion

Atherosclerosis is a low-grade chronic inflammatory condition characterized by aberrant lipid metabolism and a maladaptive inflammatory response. Biologically, the disease involves the formation of plaques in arterial walls and thickening that narrows the arterial passage, restricting blood flow and increasing the risk of occlusion resulting in a myocardial infarction and other events. Although environmental factors such as diet and/or smoking play an important role in the development of atherosclerosis, genetic factors represent important determinants of atherosclerotic cardiovascular disease risk. Key gene-environment interactions may increase the risk for adverse outcomes and lead to an increase of cIMT; this was the aim of this study.

The gene-smoking interaction signals identified for mean cIMT were with loci where the associated allele frequencies ranged from low to common, with most of them displaying effect sizesin smokers over 4 time the one in non-smokers. This suggests that the effects were unlikely to be attributed to additive independent effects of genetic-association and smoking, but rather to the interaction. The high effects provided us with sufficient power to discover the gene-smoking interactions, given our sample size. Moreover, many previously reported loci for GxE with cardiovascular-related traits were replicated, and this study identified new GxE variants for cIMT. The loci identified are biologically relevant in terms of the pathophysiology of atherosclerosis and involve genes implicated in macrophage activation and recruitment in the endothelial layer, cholesterol metabolism at the cellular level, inflammation processes and signalling, and cell membrane activity.

This study is the first to report an association of *TBC1D8* in the GxE interaction with cIMT and consequent risk for atherosclerosis. *TBC1D8* (also called Vascular Rab-GAP/TBC-containing protein) is a gene involved in blood circulation, intracellular protein transport and positive regulation of cell proliferation. The gene is regulated by vascular genes like *VEGF*, *ACKR1*, *VEGFA*, *SIRT1* and *TNF*. Previous GWASs found variants in *TBC1D8* associated with bone mass (Kiel *et al.*, 2007), cognitive decline (Li *et al*. 2015) and osteoporosis (Hsu *et al.*, 2010) and found that *TBC1D8* expression was subject to change under environmental stress. A study on the effect of smoking on gene expression found that *TBC1D8* was differentially expressed in lymphocytes of smokers and non-smokers (Charlesworth *et al.*, 2010), as well as in macrophages from atherosclerotic plaques (Puig *et al.*, 2011), depending on the inflammatory status of patients. Later Verdugo and colleagues (Verdugo *et al.*, 2013) reported that *TBC1D8* expression in monocytes was subject to a gene interaction with smoking among atherosclerotic

patients, and that *TBC1D8* was involved in one of the shortest gene paths between smoking and atherosclerotic plaques (smoking and plaques were separated by a relatively low number of genes). Their analysis of causality models provided evidence of gene expression partially mediating the relationship between smoking and atherosclerosis. This study is therefore confirming the importance of *TBC1D8* gene-environment interaction in atherosclerosis pathophysiology.

RCBTB1, previously identified as a modifier for smoking on cIMT in multi-ethnic northern American populations (Wang *et al.*, 2014), has been replicated in our study. The *RCBTB1* gene encodes a protein with an N-terminal RCC1 domain and a C-terminal BTB (broad complex, tramtrack and bric-a-brac) domain. In rat, over-expression of this gene in vascular smooth muscle cells induced cellular hypertrophy. In rat, the C-terminus of *RCBTB1* interacts with the angiotensin II receptor-1A (Guo *et al.*, 2004). In humans, this gene maps to a region of chromosome 13q that is frequently deleted in B-cell chronic lymphocytic leukaemia and other lymphoid malignancies. These results suggest that gene-smoking interaction for atherosclerosis might be acting through an intensification of monocyte activation and recruitment under the endothelial layers before their differentiation onto macrophages, a process known to trigger foam cell formation and subsequent plaques (Jia, Gao and Zhao, 2017). The results from the functional analyses suggest that the gene-smoking interaction for cIMT is likely acting through a regulatory process, explaining the involvement of multiple loci displaying chromatin interactions and acting as eQTLs. We were able, in this study, to reproduce the results on *RCBTB1*, meaning the associations are more generalizable. This study is the first independent validation of the involvement of *RCBTB1* in gene-smoking interaction for cIMT.

There were differences in the association results between the two sites with low replication. These differences may be partly explained by differences in the prevalence of smoking, sample sizes and smoking intensity. Effectively, the median smoking intensity in Nanoro was three times higher than in Navrongo (6.6 pack/year *vs* 1.7 pack/year) (**Table 4.1**), whereas the number of smokers in Nanoro (n = 132) was less than half of those in Navrongo (n = 333). A previous study using a systems biology approach revealed that cigarette smoke induced a concentration-dependent (direct and indirect) biological mechanisms that promotes monocyte–endothelial cell adhesion (Poussin *et al.*, 2015). Hence, the influence of smoking intensity on the detection of gene-smoking interaction was previously reported in a study of gene-smoking interaction for blood pressure in the Framingham Heart Study (Basson *et al.*, 2015). They found different associated loci in the light smokers and the heavy smokers group (> 10 cigarettes per day). In

the study by Wang *et al*. on gene-smoking interaction, the strongest association was among heavy smokers (≥ 20 pack/year) (Wang *et al*. 2014). This might explain why the *RCBTB1* region was only replicated in Nanoro, where the smoking intensity was higher than in Navrongo.

We reported a substantial number of suggestive GxE signals that may be African-specific as they have not yet been observed in non-African studies with larger cohorts. Since African genetic diversity is generally higher, it is possible that there are more novel gene variants that are related to pathways involved in complex diseases like atherosclerosis (Sulovari *et al.*, 2017). Our study is restricted to men and is limited by the sample size and relatively low prevalence of smokers in Nanoro. There was, however, sufficient power to detect the effect sizes we observed, and our sample size exceeded several previously published studies.

Probable contributors to the heterogeneity of signals between the two geographical groups include differences in the patterns of smoking exposure and the simplistic measure of smoking status that we used in this study (current smokers vs non-smokers), over the use of a continuous measure (pack/year).

Our study provides the first report of gene-smoking interactions for cIMT in sub-Saharan African populations. Novel genome-wide significant variants in *TBC1D8* for interactions with smoking for cIMT, were identified. The replication of eight previous signals identified in non-African populations, demonstrates that these signals are transferable to West Africa. The discovery of the novel signals, on the other hand, indicates the possibility of African-specific associations. The strategies of functional annotation and gene mapping using biological data resources provided useful information on the likely consequences of relevant genetic variants and identified plausible gene targets and biological mechanisms for functional follow-up. Gene set analyses contributed novel insight into underlying pathways, confirming the importance of gene-environment interactions in atherosclerosis and pointing toward the involvement of specific cell types. Gene-environment GWAS studies will benefit from colocalization analyses for interpreting the biological and clinical relevance of the GWAS results. When variants associated with GxE are present at high frequency in target populations, this provides an opportunity for precision public health. Future studies based on populations from other African regions may provide validation of transferability to SSA more generally, and to identify further novel signals and to generate more insights into the relationship between these associations and disease pathophysiology.

Chapter 5:

Discussion and conclusion

Chapter 5: Discussion and conclusion

The work of this PhD aims to uncover the genetic architecture of cIMT in sub-Saharan African middle-aged adults. We used data collected in the population-based study of AWI-Gen to investigate loci associated with cIMT in an exploratory framework. In order to ensure the robustness of our findings, our investigations followed state of the art practices in GWA studies. This includes a thorough investigation of cIMT epidemiology, the use of an adequate genotyping technology with a highly informative output, and the application of current GWAS statistical methods. The post-GWAS analysis provided a better understanding of our results and their potential biological relevance to cIMT and, to some extent, to atherosclerosis. Considering the paucity of such studies in Africa, our study is an important contribution to the genetics of cIMT.

5.1 **Regional epidemiological differences in Africa**

Our study demonstrated differences in the distribution of cIMT in sub-Saharan African populations, with the two West African countries in the AWI-Gen study showing, on average, higher cIMT than those in the East and South. The regional differences could not be explained only by classical cardiovascular risk factors previously shown to be associated with cIMT (Nonterah *et al.*, 2019). Studies on hypertension reported a prevalence of 15.1% in Nanoro and 24.5% in Navrongo, while the three south African sites had a higher mean prevalence of over 40% (Gómez-Olivé *et al.*, 2017). Equally, mean BMI was lower in the West African sites compared to South and East African sites (Ramsay *et al.*, 2018). It is important to note that besides the differences in the income levels of the study sites, they were also at different stages of the demographic and health transition (Defo, 2014). Considering all the risk factors above, the characteristics of the distributions of cIMT as found in our study may not find its full explanation only from an epidemiological perspective, but would benefit from investigating potential genetic contributions to cIMT. Our study presents a unique opportunity to explore both epidemiological and genetic factors that contribute to cIMT, and has the advantage of being a population-based cross-sectional study where individuals have not been selected based on specific diseases of interest, thus minimising selection bias and enabling the estimation of a population prevalence (Lieb and Vasan, 2018). In addition, the low awareness and general lack of treatment due to under-developed health services and relative high prevalence of CVDs risk

factors such as hypertension and dyslipidaemia (Gómez-Olivé *et al.*, 2017; Reiger *et al.*, 2017; Noubiap *et al.*, 2018) in these regions allow for studies that can measure unbiased genetic effects.

5.2 Our GWAS of cIMT linked genetics to biology of atherosclerosis in African populations

This genetic study successfully identified 11 loci robustly associated with cIMT, of which 10 were novel. These loci were located nearby or within regions of genes with biological relevance to atherosclerosis and CVDs. In Chapters 3 and 4, we discussed the potential biological relevance of the loci and how they might be linked to the phenotype. Most importantly, our results confirmed the importance of known biological processes and pathways involved in atherosclerosis. Throughout our study, the results confirmed the involvement of genes in atherosclerosis pathophysiology: genes related to macrophage activity and polarisation (*SIRPA*), to vascular smooth muscle cells (*MAP3K7*, *CALD1*), to vascular endothelial growth (*PROK1*, *FLT4*), to collagen synthesis and plaque stability (*LARP6*), and a pathway of blood vessel occlusion (*SNX29*). Our results also linked the phenotypes to processes of regulation of gene expression (eQTLs (*RCBTB1*, *TBC1D8*, *BCHE*), chromatin interactions (*UACA-LARP6*, *RCBTB1*, *TBC1D8*), to the role of hormones in regulation (oestrogen (*UACA-LARP6*, *PROK1*), testosterone (*SNX29*)), to epigenetics (*SNX29*, *CALD1*).

We found evidence of gene set enrichment for biological processes. Our study is the first GWAS to report significant enrichment of genes in the oestrogen pathway for cIMT in our female-specific analysis. In our study of gene-smoking interaction, genes were significantly enriched for biological adhesion, cell-cell adhesion, homophilic cell adhesion via plasma membrane adhesion molecules and cell-cell adhesion via plasma membrane adhesion molecules, all processes known to be important in the pathophysiology of atherosclerosis.

The findings from our study support the notion that genomics studies in Africa are likely to contribute to the understanding of complex traits, such as atherosclerosis.

5.3 African genetic diversity

African populations are characterized by greater levels of genetic diversity, extensive population substructure, and lower linkage disequilibrium (LD) among loci compared to non-African populations (Campbell and Tishkoff, 2008; Li and Keating, 2014). A consequence of the lower LD is that more markers (tag-SNPs) are necessary to allow adequate genomic

coverage to detect associations in GWASs, and for that reason we used the customized-African-variant-enriched H3Array to overcome this challenge. The participants in this study included individuals from multiple populations with genetic diversity derived from populations in East, South and West Africa, as illustrated in the PC plot showing population sub-structure (Chapter 3). Studies comparing ethnically diverse populations from Africa are important for understanding the genetic basis of phenotypic adaptation and complex disease. Our study reported suggestive and strongly associated variants across the spectrum of allele frequencies, with a predominance of high non-reference genome allele frequencies variants in the loci identified as illustrated in **Figure 5.1**.

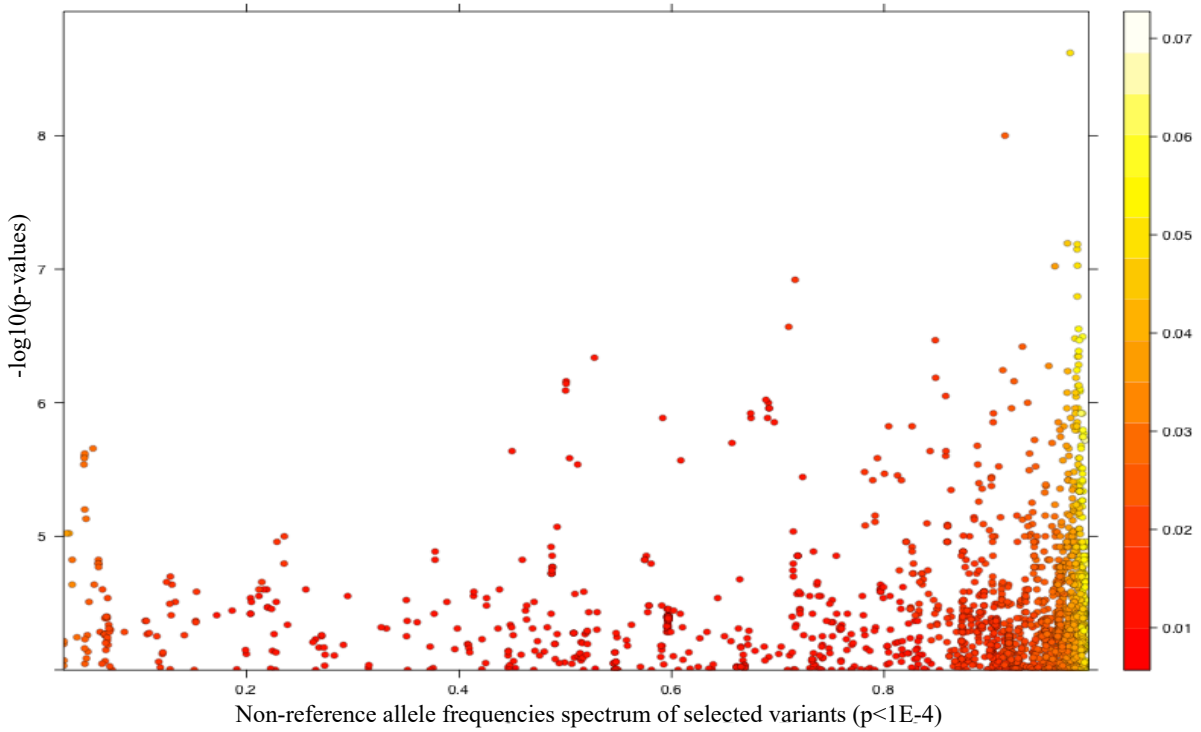


Figure 5.1: Level plot of allele frequencies by p-values of associations (p<1e-4) for the female-specific GWAS showing over-representation of suggestive variants in extreme allele frequencies.

Each dot represents a variant and the colour is function of the effect size (betas in mm). x-axis allele frequencies. Y-axis p-values -log10(p).

Extreme allele frequency differences represent an opportunity for identifying population-specific functional variants. The new biological insights identified throughout this study support the power of studies on African populations, even with modest sample sizes (Choudhury *et al.*, 2014, 2018; Gurdasani *et al.*, 2015; Busby *et al.*, 2016). Sulovari found that

African populations had more extreme allele frequencies difference (EAFD) variants involved in disease-associated genes and pathways compared to other world populations (Sulovari *et al.*, 2017). It has been shown that despite a consistency of eQTLs across populations, many eQTLs in one population were poorly correlated (not significant) in another mainly because of allele frequency differences between populations, therefore affecting the power to detect associations (Mogil *et al.*, 2018). Finally, African populations evolved in a unique environment and may have experienced local adaptation, some due to infectious disease (e.g. sickle cell trait and malaria), that would have shaped population-specific susceptibility to disease.

5.4 Sex-specific genetics associations with African cIMT

Despite the fact that cIMT did not show evidence of sex differences in the overall distribution (**Chapter 2**), we carried out sex stratified analyses given previous evidence of sex-differences in susceptibility to atherosclerosis and CVDs (Appelman *et al.*, 2015; Spence and Pilote, 2015). Our results highlighted the importance of sex-differences in atherosclerosis by pinpointing relevant biological genes or pathways. In genomic studies, sex-biased gene expression is a fundamental characteristic in humans, and sex should be seen from diverse perspectives including both biological and environmental factors (by influencing exposures). First, sex-biased gene expression is a common characteristic of genes encoded both on the sex chromosomes and on the autosomes. Second, most changes between male and female expression levels are typically small for significant sex-differentially expressed genes. Third, genes exhibiting sex-biased expression include those implicated in relevant human phenotypes and disease prevalence that differ between the sexes. Finally, sex-biased gene expression varies across tissues, developmental stages and environmental conditions (endogenous or exogenous) at both the individual gene level and on the genome-wide scale (Dimas *et al.*, 2012; Gershoni and Pietrokovski, 2017; Dias *et al.*, 2019; Khramtsova, Davies and Stranger, 2019). Gilks and collaborators explained sex differences in complex trait genetics from the perspective of evolutionary genetics, with the hypothesis that intra-locus sexual conflict could be maintaining harmful sex-dependent alleles in the population if they are beneficial to the opposite sex (Gilks, Abbott and Morrow, 2014). It is thought that sexual antagonism occurs when genetically correlated traits have opposite effects on male and female fitness (Gilks, Abbott and Morrow, 2014), with the possibility that a negative intersexual genetic correlation could occur when the same gene product is incorporated in competing alternative pathways. In our study, we reported *IGFBPL1-FAM95C*, *UACA-LARP6*, *CROCC* and *MYOD1-KCNC1* as typical examples of

genes involved in sexual antagonism. Our study reported 2.6% of OED SNPs, this low proportion of OED variants is expected considering that CED (4.3%) and SSE (94%) are the most predominant forms of sex-dimorphism. This is in line with several previous studies, emphasising the role of biological sex in affecting specific phenotypes such as cIMT, Alzheimer's disease, body composition, paediatric bone mineral content (Randall *et al.*, 2013; Winkler, Justice, *et al.*, 2015; Chesi *et al.*, 2017; Hübel *et al.*, 2018; Nazarian, Yashin and Kulminski, 2019) (**Fig. 5.2**).

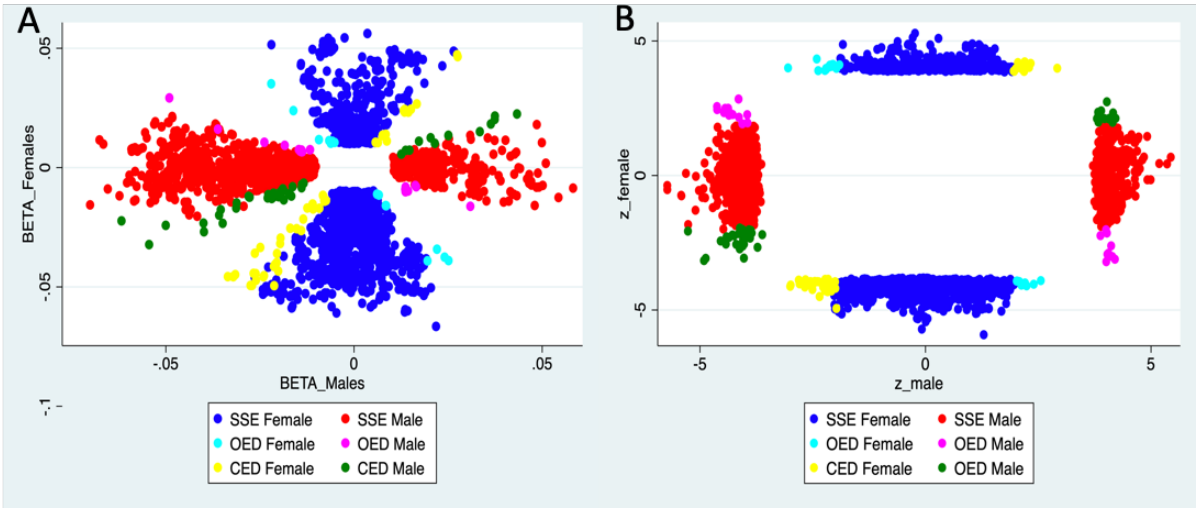


Figure 5.2: Illustration of the sex-difference in the selected variants using betas and z.

(A) Correlation of variants effects from male and female according to the type of sexual dimorphism.

(B) Correlation of z (beta/se) of variants from male and female according to the type of sexual dimorphism.

SSE Female: variant with single sex-effect in our female-specific (P-female < 1E-04);

SSE Male: variant with single sex-effect in our male-specific (P-male < 1E-04);

OED Female: variant with opposite effect direction in our female-specific (P-female < 1E-04);

OED Male: variant with opposite effect direction in our male-specific (P-male < 1E-04);

CED Female: variant with concordant effect direction in our female-specific (P-female < 1E-04),

CED Female: variant with concordant effect direction in our male-specific (P-male < 1E-04).

5.5 Complex trait genetics as addressed in our study

Association mapping for complex disease relies on having some a priori knowledge of genetic diversity, population structure, and LD in populations to identify polymorphisms associated with traits. In the early GWAS era, two models dominated the debate about genetic architecture: the 'Common Disease, Common Variant (CDCV)' hypothesis which argues that genetic variations with appreciable frequency in the population (MAF>1%), but relatively low 'penetrance', are the major contributors to genetic susceptibility to common diseases. The second 'Common Disease, Rare Variant (CDRV)' hypothesis, on the contrary, argues that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases (Wright and Hastie, 2001; Bodmer and Bonilla, 2008; Iles, 2008; Schork *et al.*, 2009). Since then, analyses showed that both common SNPs with effect sizes well below genome-wide statistical significance account for most of the "missing heritability" of many traits, and rare variants with larger effect sizes also contribute genetic variance, supporting the "infinitesimal model" pioneered by Fisher and revisited by Barton and collaborators (Barton, Etheridge and Véber, 2017). The recent "omnigenic" model tends to shed more light on understanding complex trait genetics, and implies that considering that most large effect variants are modestly enriched in specific genes or pathways directly involved in diseases, SNPs that contribute to the bulk of the heritability tends to be spread across the genome in enriched regions that are transcriptionally active in relevant tissues (Boyle, Li and Pritchard, 2017). Although a multitude of cardiovascular risk loci have been identified by GWASs, variants in only a few of these have been provisionally connected to function and potential causality. This is in part because the SNPs on genotyping arrays are chosen for different reasons, i.e. as tag SNPs for markers in LD on a haplotype block. Nonetheless, broad new insights into disease can be achieved even without a detailed molecular understanding of variant to function relationships, fine-mapping and identification of the causal variant. The annotation of our results in **Table 3.5** illustrated the low rendering of GWAS studies in causal variants.

In our GWAS we only assessed SNPs, and the majority did not have predicted functional impact. We have not as yet investigated structural variation, such as insertions, deletions, and copy number variants, which may contribute to the genetic architecture of cIMT. This omission could be addressed by whole genome sequencing, which would be very costly and require large

bioinformatic resources for the analysis. Moreover, we did not investigate variation on the X chromosome, thereby leaving out 5% of the haploid human genome and ~800 genes out of a total of ~20,000.

5.6 Gene-Environment interaction

Testing gene-environment (GxE) interactions has evolved as a useful tool for understanding the genetic underpinning of complex diseases. GxE interactions have been investigated in several populations for environmental exposures such as air pollution and smoking, and traits such as atherosclerosis and type 2 diabetes, but not in sub-Saharan Africans. Our study investigated the SNP-smoking interaction for cIMT (a surrogate marker of atherosclerosis) in West African men from Nanoro (Burkina Faso) and Navrongo (Ghana). The results showed SNPs from three genomic regions to be significantly associated with interactions with smoking. In Nanoro, the previously described smoking interaction with markers in *RCBTB1* was replicated in addition to one newly identified region near *BCHE* on chromosome 3. In the combined sample, common variants in *TCB1D8* were found to interact with smoking. Our positive signals for gene-smoking interactions depended on the availability of high-quality exposure assessment with both qualitative and quantitative measurement of smoking. The study populations presented a wide range of exposure levels, and the inclusion of ethnically and geographically diverse populations. This allowed us to identify gene-smoking interactions that seemed to be activated from a certain level of smoking intensity (i.e. number of cigarettes smoked per week).

5.7 Replication and transferability

The genetic architecture of specific complex traits may be variable across different human ancestries, making replication challenging, specifically for studies in non-European ancestry populations since most discoveries were made in European ancestry populations. Whether defined as the same genetic marker being consistently associated with the same phenotype in independent cohorts, or positive association at variants in strong linkage disequilibrium (LD) with the original marker SNP, replication has been set as a gold standard for reporting GWAS associations, especially when the primary association does not reach genome-wide significance levels. This is a difficult criterion to meet in African studies, not only because of natural differences in genomic LD across ethnically diverse populations (Li and Keating, 2014), but also because the continent is at an early stage in genomics studies and few studies have suitable data for investigating replication. Instead, the concept of local replication across a region close

to the previously associated locus (e.g within 100 kb), was preferred in our study, since no European index SNPs for cIMT showed evidence of association. Our study locally replicated at suggestive level (p-values at least 1E-05) more than 80 of over 150 loci previously described for CVDs phenotypes and endophenotypes in the different analyses of sex-combined, stratified GWAS or gene-smoking interaction. Given the relatively small sample size of our study (7894 individuals) compared to over 50,000 reported by Franceshini *et al*. (Franceschini *et al.*, 2018) in Europeans, our study showed some evidence of transferability of the loci identified from other ancestries. We showed that a suggestive variant in *CBFA2T3* (rs9934287) was in perfect LD (r^2=1, in 1000GP Africans) with a previously identified variant associated with cIMT in European populations (rs844396), The rs9934287 variant is extremely rare (MAF 0.0002) in Europeans. At this locus, our study overcame the "winner's curse by reporting a beta higher than in the discovery study for cIMT association (0.021 vs 0.005). Overall, in our analyses we did not find exact replication of association for SNPs identified in European populations. Cross-ancestry replication is known to be challenging notably because of low correlation of effect size across ethnicities and differences in the phenotypic variance explained by the variants.

5.8 Trans-ethnic fine-mapping opportunities from the loci suggestive of replication in African

Throughout the thesis I have described variants located in loci previously reported to be associated with cIMT both in GWASs and the gene-smoking interaction analyses. Fine-mapping of causal variants is the process of refinement of a signal by analysing densely placed SNPs in studies with large sample sizes in discovery populations to detect rare ancestral recombination events and thus prioritise potential causal variants, or by leveraging differences in LD structure between diverse populations with the aim to identify causal variants. It is recognised that the inclusion of African ancestry in multi-ethnic studies may yield substantial improvements in localisation of causal variants because of low LD and high genetic heterogeneity in Africans (Gurdasani *et al.*, 2015). Unfortunately, the lack of publicly available datasets (summary statistics) from GWA studies from Africans made this analysis challenging. Nonetheless, we tried to understand the potential for fine-mapping of our results by comparing our study's findings with the results from the widest GWAS of cIMT published to date. The SNP rs116517341, which leads the association with the *CCDC71L* locus in our male-specific analysis, is located over 100 kb from the lead-SNP found in European, but our lead locus was closer to *CCDC71L* than that found by the study of cIMT in Europeans (**Fig. 5.3A-B**).

Therefore, different variants may influence the expression of the *CCDC71L* gene in cIMT. When analysing the variants in *CBFA2T3*, our lead-SNP (rs9934287) was located 18,979 bp away from the SNP reported by Franceschini et al. (2018). The regional plot showed more dense signals with numerous SNPs in LD with the lead-SNP in European, whereas our lead-SNP had fewer SNPs in high LD (**Fig. 5.3 C-D**). A lookup with the LD structure using 1000G CEU compared to 1000G African populations (YRI-LWK-GWD-MSL-ESN) showed that LD blocks were smaller in the African populations (**Fig. 5.4A**) compared to European populations, where the SNPs were in a single block (**Fig. 5.4B**). These examples may not be ideal, however they are providing ideas of the value of fine-mapping using ethnically different populations. Larger sample sizes of African studies will realise of the potential of reducing the credible set and identifying causal variants.

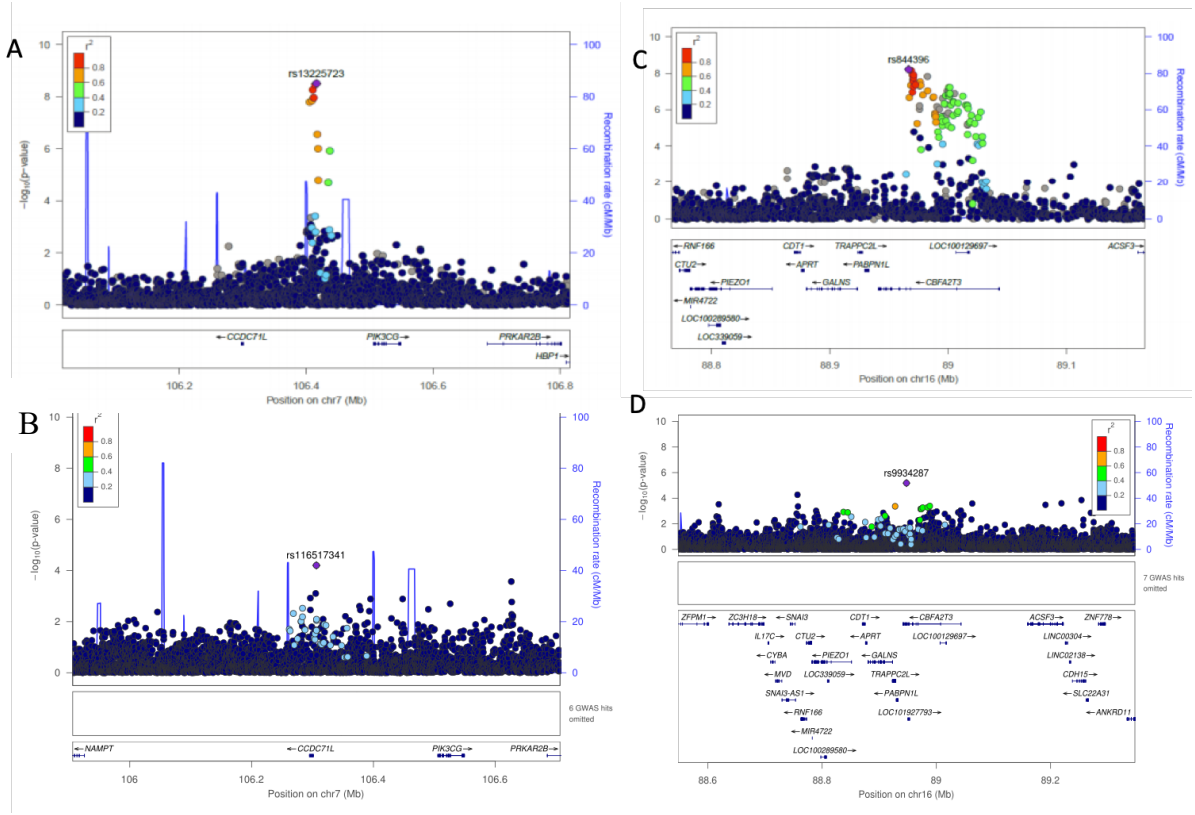


Figure 5.3: Regional plots for genetic associations with cIMT, focusing on signals close to two genes (*CCDC71L* and *CBFA2T3*) previously identified in European GWASs and replicated in our study.

(A) and (B) show the plots for the *CCDC71L* region with the lead SNP, rs13225723, for cIMT in Europeans (A) and the lead SNP, rs116517341, for cIMT in our African male-specific analysis (B). The signals are in different intervals between recombination hot spots (C) and (D) show the plots for the *CBFA2T3* region with the lead SNP rs844396 for cIMT in Europeans (C) and rs9934287 for cIMT in our African sex-combined analysis (D). In Europeans the signal overlaps the gene whereas

the African signal is downstream of the gene. Panels A and C were copied from Franceschini et al. (Franceschini *et al.*, 2018)

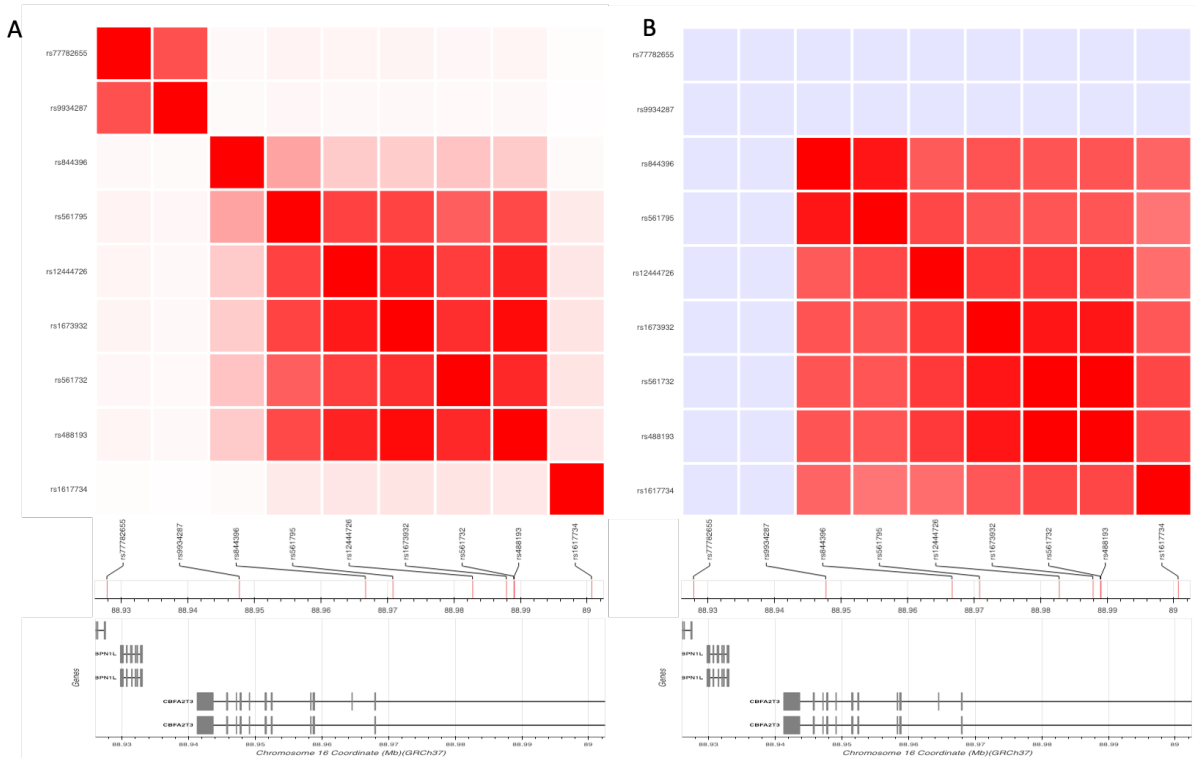


Figure 5.4: LD matrix for selected variants in *CBFA2T3* region showing the LD block distribution in Africans and Europeans, using the 1000GP data.

(A) LD matrix using 1000 GP African samples (YRI-LWK-GWD-MSL-ESN). (B) LD matrix using 1000 GP European samples (CEU).

5.9 Strengths and limitations

The strength of our study lies in the design. The population-based nature of the study allowed us to study a population that was not selected based on a disease phenotype and that was not over-medicated (low access to health services), enhancing its ability to assess real genetic effects. The participants were drawn from three regions in Africa (East, South and West). This was both a strength and a potential weakness and it was necessary to take into account genetic diversity across the groups as well as diverse regional environments. The wealth of phenotypes available in the AWI-Gen study (over 300) allowed the use of different data mining approaches, thereby augmenting the chances of discoveries. Additionally, all the phenotypes were collected by following a single standardised procedure across all sites. Our study is the largest study performed for cIMT in African ancestry populations with epidemiological and genetic data.

Only eligible and consenting participants were enrolled in the study and the refusal rate was low, but was not quantified by the different recruitment sites.

The use of cIMT as a quantitative trait, which is an endophenotype of CVDs with a clear genetic component, improved the power of our interaction studies. In some situations, such endophenotypes may be a better measurement and more representative of the mechanistic pathway toward disease risk. For example, fasting lipid profiles or glucose tolerance tests may be more appropriate for certain genetic association tests because they are closer to gene action than the disease outcome of CVDs or diabetes (McAllister *et al.*, 2017; Ritz *et al.*, 2017). There is increased chance to identify functional impact of a genetic locus by including a wide diversity of functional omics data, including epigenomes, methylomes, proteomes, and transcriptomes. Integrative genomics is therefore a powerful tool to decipher the role of functional loci across human tissues (Ritchie *et al.*, 2017), and allow us to link genetic associations with the biology of our findings.

Our study samples remain small compared to the sample sizes in European ancestry studies that can be over a million participants. The sample size for this study was significantly lower than the entire AWI-Gen sample size because there was no CIMT data from women at the Soweto site, and the quality of the data varied between recruitment sites. The stringent QC procedure applied after data collection led to further reduction in the data, but ensured that only good quality data were used in this study. The lack of comparable studies in African populations precluded attempts of replication or look up in a similar ancestral population.

We used the African reference panel of the Sanger Imputation Server to impute our genotype data. The reference panel included African and non-African populations from 1000 Genomes Phase 3 and additionally ~2000 samples from Uganda (Baganda, Banyarwanda, Barundi and others) and ~100 samples from each of a set of populations from Ethiopia (Gumuz, Wolayta, Amhara, Oromo, Somali), Egypt, Namibia (Nama/Khoesan) and South Africa (Zulu). This did not include our populations from West Africa (from Burkina Faso and Ghana representing 3439 participants), therefore the accuracy of the imputation might not be optimal for our study. Furthermore, the LD patterns in African populations, which are generally lower may have an impact on the imputation accuracy.

Considering the regional differences in cIMT distribution and the differences in ancestry between populations from East, West and Southern Africa, another approach may have been to analyse regions separately, followed by meta-analysis. This approach is especially appropriate

when data from the individual studies cannot be analysed together because of differences in ethnicity and phenotype distribution, or when constraints on sharing of individual level data imposed. Meta-analysis usually results in little or no loss of efficiency compared to analysis of a combined dataset including data from all individual studies and adjusting for ethnicity and phenotypic differences between populations.

In our analysis of the effect of gene-smoking interaction on cIMT, we showed that the intensity of the exposure influenced the capacity to detect gene-environment interactions. For this reason, we did not identify any significant signal from our Navrongo subset who were exposed to lower levels of tobacco.

In our GWAS analysis we used a reduced complexity model, only adjusting for age, sex and PCs. This approach may not allow the best estimation of the genetic effect of the variants investigated, because it does not account for the effect of other confounders, such as known CVD risk factors. We showed in our analysis of gene-smoking interaction that addition of covariates could result in changes in effect sizes (beta values) and the p-values.

5.10 Future studies

Because of its unique analytical challenge, the X chromosome has long been excluded from GWAS studies. Despite recent developments in analytical methods (König *et al.*, 2014; Gao *et al.*, 2015; Graffelman and Weir, 2016), X chromosome SNP associations have not been reported in GWAS of cIMT. This is a gap that needs to be addressed in future studies, as well as associations with markers on the Y chromosome and mitochondrial DNA that need to be studied in African populations for associations with atherosclerosis. Finally it will be important to include epigenetic investigations in future studies of atherosclerosis, considering the importance of the developmental origins of health and diseases theory, and possible molecular mechanisms for susceptibility to chronic noncommunicable diseases may be attributed to epigenetic modifications which play an important role in foetal programming, with consequences in later life (Barker, 2007; Waterland and Michels, 2007; Hobbs and Ramsay, 2015).

Considering the burden of HIV infection in South and East Africa, and previous reports of gene-HIV interaction for cIMT (Shrestha *et al.*, 2010, 2012; Shendre *et al.*, 2014), it would be relevant to perform a study of gene-HIV interaction for cIMT in our African populations.

Regarding the loci where we have suggestive evidence of replication, future collaboration allowing access to summary statistics or raw data from other African studies would benefit fine-mapping initiatives.

5.11 Conclusion

Our study of genetic susceptibility to atherosclerosis in middle-aged sub-Saharan Africans has contributed insights into susceptibility for atherosclerosis. We found that cIMT is not evenly distributed among regions within African population. Whether considering sex-specific susceptibility or gene-smoking interaction, confirmation of our discoveries could lead to public health interventions to prevent atherosclerosis, and may pioneer precision-guided public health. Future studies will benefit from temporal considerations such the development across the lifecourse using birth cohorts and longitudinal studies in adults. In addition, refined measurement of exposures will provide more accurate insights into genetic susceptibility and gene-environment interactions. Studies in genomics in Africa will benefit from integrating medically relevant information with evolutionary approaches such as environmental adaptation. The study of the trajectories of genetic variant frequencies over time and across geographic regions will help to explore ways in which population genomics can contribute in a multi-disciplinary environment together with statistical genetics, molecular biology, physiology, and anthropology in elucidating the origin and evolution of human disease.

References

References

Abdoul-latif, F. M. *et al.* (2012) 'Comparison of phenolic compounds and antioxidant capacities of traditional sorghum beers with other alcoholic beverages', *African Journal of Biotechnology*, 11(81), pp. 14671–14678. doi: 10.5897/AJB12.2153.

Alberts, M. *et al.* (2015) 'Health & Demographic Surveillance System Profile: The Dikgale Health and Demographic Surveillance System', *International Journal of Epidemiology*, 44(5), pp. 1565–1571. doi: 10.1093/ije/dyv157.

Alfaidy, N. *et al.* (2014) 'The Multiple Roles of EG-VEGF/PROK1 in Normal and Pathological Placental Angiogenesis', *BioMed Research International*, 2014, pp. 1–10. doi: 10.1155/2014/451906.

Ali, S. A. *et al.* (2018) 'Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study', *Global Health Action*. Taylor & Francis, 11(sup2). doi: 10.1080/16549716.2018.1507133.

Alzaid, F., Patel, V. B. and Preedy, V. R. (2014) *Cardiovascular Disease in Aging and the Role of Oxidative Stress*, *Aging*. Elsevier. doi: 10.1016/B978-0-12-405933-7.00003-2.

Appelman, Y. *et al.* (2015) 'Sex differences in cardiovascular risk factors and disease prevention.', *Atherosclerosis*. Elsevier Ltd, 241(1), pp. 211–218. doi: 10.1016/j.atherosclerosis.2015.01.027.

Armitage, P. (1955) 'Tests for Linear Trends in Proportions and Frequencies', *Biometrics*, 11(3), pp. 375–386.

Arning, A. *et al.* (2012) 'Agenome-wide association study identifies a gene network of ADAMTS genes in the predisposition to pediatric stroke', *Blood*, 120(26), pp. 5231–5236. doi: 10.1182/blood-2012-07-442038.

Arya, R. *et al.* (2018) 'A genetic association study of carotid intima-media thickness (CIMT) and plaque in Mexican Americans and European Americans with rheumatoid arthritis', *Atherosclerosis*. Elsevier Ltd, 271, pp. 92–101. doi: 10.1016/j.atherosclerosis.2017.11.024.

Aschard, H. *et al.* (2010) 'Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects', *Human Heredity*, 70(4), pp. 292–300. doi:

10.1159/000323318.

Assimes, T. L. *et al.* (2016) 'Genetics of Coronary Artery Disease in Taiwan: A Cardiometabochip Study by the Taichi Consortium', *PloS one*, 11(3), p. e0138014. doi: 10.1371/journal.pone.0138014.

Astle, W. J. *et al.* (2016) 'The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease', *Cell*, 167(5), pp. 1415-1429.e19. doi: 10.1016/j.cell.2016.10.042.

Baeyens, N. *et al.* (2015) 'Vascular remodeling is governed by a VEGFR3-dependent fluid shear stress set point', *eLife*, 4, pp. 1–16. doi: 10.7554/elife.04645.

Baichoo, S. *et al.* (2018) 'Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics', *BMC Bioinformatics*. BMC Bioinformatics, 19(1), p. 457. doi: 10.1186/s12859-018-2446-1.

Barker, D. J. P. (2007) 'The origins of the developmental origins theory', *Journal of Internal Medicine*, 261(5), pp. 412–417. doi: 10.1111/j.1365-2796.2007.01809.x.

Bartels, S., Franco, A. R. and Rundek, T. (2012) 'Carotid intima-media thickness (cIMT) and plaque from risk assessment and clinical use to genetic discoveries', *Perspectives in Medicine*. Elsevier GmbH, 1–12(1), pp. 139–145. doi: 10.1016/j.permed.2012.01.006.

Barton, N. H., Etheridge, A. M. and Véber, A. (2017) 'The infinitesimal model: Definition, derivation, and implications', *Theoretical Population Biology*. Elsevier Inc., 118, pp. 50–73. doi: 10.1016/j.tpb.2017.06.001.

Barua, R. S. and Ambrose, J. A. (2013) 'Mechanisms of coronary thrombosis in cigarette smoke exposure', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 33(7), pp. 1460–1467. doi: 10.1161/ATVBAHA.112.300154.

Basson, J. *et al.* (2015) 'Influence of Smoking Status and Intensity on Discovery of Blood Pressure Loci Through Gene-Smoking Interactions', *Genetic Epidemiology*, 39(6), pp. 480–488. doi: 10.1002/gepi.21904.

Beguy, D. *et al.* (2015) 'Health & Demographic Surveillance System Profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS) Donatien', *International*

Journal of Epidemiology, 44(5), pp. 1565–1571. doi: 10.1093/ije/dyv157.

Bennett, P. C. *et al.* (2011) 'Ethnic differences in common carotid intima – media thickness , and the relationship to cardiovascular risk factors and peripheral arterial disease : the Ethnic-Echocardiographic Heart of England Screening Study', *QJM: An International Journal of Medicine*, 104(3), pp. 245–254. doi: 10.1093/qjmed/hcq187.

Bhatnagar, P. *et al.* (2013) 'Genome-Wide Meta-Analysis of Systolic Blood Pressure in Children with Sickle Cell Disease', *PLoS ONE*, 8(9), pp. 1–11. doi: 10.1371/journal.pone.0074193.

Bis, J. C. *et al.* (2012) 'Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque', *Nat Genet*, 43(10), pp. 940–947. doi: 10.1038/ng.920.Meta-analysis.

Bodmer, W. and Bonilla, C. (2008) 'Common and rare variants in multifactorial susceptibility to common diseases', *Nature Genetics*, 40(6), pp. 695–701. doi: 10.1038/ng.f.136.

Boring, L. *et al.* (1998) 'Decreased lesion formation in CCR2−/− mice reveals a role for chemokines in the initiation of atherosclerosis', *Nature*, 394(6696), pp. 894–897. doi: 10.1038/29788.

Botha, G. (2016) 'Custom genotyping chip for African populations', in *2016 Blue Waters Symposium: Abstracts and Presentations*. Available at: https://bluewaters.ncsa.illinois.edu/liferay-content/document-library/2016 Symposium/botha_h3africa.pdf.

Boua, R. P. *et al.* (2010) 'Phenolic compounds and antioxidant activity of 10 West African Sorghum Varieties [Etudes des composes phenoliques et de l'activite antioxydante de 10 varietes de sorgho en Afrique de l'Ouest]', in *Actes FRSIT*, pp. 21–31.

Bouhaddioui, W., Provost, P. R. and Tremblay, Y. (2016) 'Expression profile of androgen-modulated microRNAs in the fetal murine lung', *Biology of Sex Differences*. Biology of Sex Differences, 7(1), pp. 1–13. doi: 10.1186/s13293-016-0072-z.

Boyle, A. P. *et al.* (2012) 'Annotation of functional variation in personal genomes using RegulomeDB', *Genome Research*, 22(9), pp. 1790–1797. doi: 10.1101/gr.137323.112.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) 'An Expanded View of Complex Traits: From Polygenic to Omnigenic', *Cell*. Elsevier, 169(7), pp. 1177–1186. doi: 10.1016/j.cell.2017.05.038.

Britton, A. *et al.* (2016) 'Twenty-year trajectories of alcohol consumption during midlife and atherosclerotic thickening in early old age: Findings from two British population cohort studies', *BMC Medicine*. BMC Medicine, 14(1), pp. 1–8. doi: 10.1186/s12916-016-0656-9.

Britton, A. R. *et al.* (2017) 'Alcohol consumption and common carotid intima-media thickness: The USE-IMT study', *Alcohol and Alcoholism*, 52(4), pp. 483–486. doi: 10.1093/alcalc/agx028.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D1005–D1012. doi: 10.1093/nar/gky1120.

Busby, G. B. *et al.* (2016) 'Admixture into and within sub-Saharan Africa', *eLife*, 5, p. e15266. doi: 10.7554/eLife.15266.

Cai, L. *et al.* (2010) 'Binding of LARP6 to the Conserved 5′ Stem-Loop Regulates Translation of mRNAs Encoding Type I Collagen', *Journal of Molecular Biology*. Elsevier Ltd, 395(2), pp. 309–326. doi: 10.1016/j.jmb.2009.11.020.

Camaré, C. *et al.* (2017) 'Angiogenesis in the atherosclerotic plaque', *Redox Biology*. Elsevier, 12(January), pp. 18–34. doi: 10.1016/j.redox.2017.01.007.

Campbell, M. C. and Tishkoff, S. A. (2008) 'African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping', *Annual Review of Genomics and Human Genetics*, 9(1), pp. 403–433. doi: 10.1146/annurev.genom.9.081307.164258.

Carty, C. L. *et al.* (2015) 'Meta-Analysis of Genome-Wide Association Studies Identifies Genetic Risk Factors for Stroke in African Americans.', *Stroke; a journal of cerebral circulation*, 46(8), pp. 2063–8. doi: 10.1161/STROKEAHA.115.009044.

Caughey, M. C. *et al.* (2014) 'Sickle cell trait and incident ischemic stroke in the Atherosclerosis Risk in Communities study.', *Stroke; a journal of cerebral circulation*, 45(10),

pp. 2863–7. doi: 10.1161/STROKEAHA.114.006110.

Ceriello, A. (2008) 'Possible role of oxidative stress in the pathogenesis of hypertension.', *Diabetes care*, 31 Suppl 2(24). doi: 10.2337/dc08-s245.

Chang, C. C. *et al.* (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience*, 4(1), p. 7. doi: 10.1186/s13742-015-0047-8.

Charlesworth, J. C. *et al.* (2010) 'Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes', *BMC Medical Genomics*, 3(1), p. 29. doi: 10.1186/1755-8794-3-29.

Charmet, R. *et al.* (2018) 'Novel risk genes identified in a genome-wide association study for coronary artery disease in patients with type 1 diabetes', *Cardiovascular diabetology*. BioMed Central, 17(1), p. 61. doi: 10.1186/s12933-018-0705-0.

Chen, H. *et al.* (2016) 'Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models', *American Journal of Human Genetics*, 98, pp. 653–666. doi: 10.1016/j.ajhg.2016.02.012.

Chen, W. *et al.* (2019) 'miR-378a Modulates Macrophage Phagocytosis and Differentiation through Targeting CD47-SIRPα Axis in Atherosclerosis', *Scandinavian Journal of Immunology*, 90(e12766), pp. 1–10. doi: https://doi.org/10.1111/sji.12766.

Chesi, A. *et al.* (2017) 'A Genomewide Association Study Identifies Two Sex-Specific Loci, at SPTB and IZUMO3, Influencing Pediatric Bone Mineral Density at Multiple Skeletal Sites', *Journal of Bone and Mineral Research*, 32(6), pp. 1274–1281. doi: 10.1002/jbmr.3097.

Chiva-Blanch, G. *et al.* (2015) 'Effects of alcohol and polyphenols from beer on atherosclerotic biomarkers in high cardiovascular risk men: A randomized feeding trial', *Nutrition, Metabolism and Cardiovascular Diseases*. Elsevier B.V, 25(1), pp. 36–45. doi: 10.1016/j.numecd.2014.07.008.

Choudhury, A. *et al.* (2014) 'Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance', *BMC Genomics*, 15(1), p. 437. doi: 10.1186/1471-2164-15-437.

Choudhury, A. *et al.* (2018) 'African genetic diversity provides novel insights into evolutionary

history and local adaptations', *Human Molecular Genetics*, 0(0), pp. 1–10. doi: 10.1093/hmg/ddy161.

Cochran, W. G. (1954) 'Some Methods for Strengthening the Common χ2 Tests', *Biometrics*, 10(4), pp. 417–451.

Consortium Roadmap Epigenomics *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–330. doi: 10.1038/nature14248.

Corlan, A. S. *et al.* (2017) 'Endocrine Gland-Derived Vascular Endothelial Growth Factor/Prokineticin-1 in Cancer Development and Tumor Angiogenesis', *International Journal of Endocrinology*, 2017. doi: 10.1155/2017/3232905.

Danchuk, S., Delafontaine, P. and Higashi, Y. (2019) 'Mir-1976 Downregulation and Phosphorylation of LARP6 are Associated with Insulin-like Growth Factor-1 (IGF1)-induced Collagen Upregulation in Vascular Smooth Muscle Cells', *The FASEB Journal*. Federation of American Societies for Experimental Biology, 33(1_supplement), pp. 522.9-522.9. doi: 10.1096/fasebj.2019.33.1_supplement.522.9.

Das, S., Abecasis, G. R. and Browning, B. L. (2018) 'Genotype Imputation from Large Reference Panels.', *Annual Review of Genomics and Human Genetics*, 19, pp. 73–96. doi: 10.1146/annurev-genom-083117-021602.

Davies, G. *et al.* (2018) 'Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function', *Nature Communications*, 9(1), pp. 1–16. doi: 10.1038/s41467-018-04362-x.

Defo, B. K. (2014) 'Demographic, epidemiological, and health transitions: are they relevant to population health patterns in Africa?', *Global health action*, 7, p. 22443. doi: 10.3402/gha.v7.22443.

Della-Morte, D. *et al.* (2014) 'Novel genetic variants modify the effect of smoking on carotid plaque burden in Hispanics', *Journal of the Neurological Sciences*. Elsevier B.V., 344(1–2), pp. 27–31. doi: 10.1016/j.jns.2014.06.006.

Deng, C. *et al.* (2018) 'Identi fi cation of three novel loci of ALDH2 Gene for Serum Folate levels in a Male Chinese Population by Genome-Wide Association Study', *Gene*. Elsevier,

674(June), pp. 121–126. doi: 10.1016/j.gene.2018.06.080.

Depre, C., Powell, S. R. and Wang, X. (2010) 'The role of the ubiquitin-proteasome pathway in cardiovascular disease', *Cardiovascular Research*, 85(2), pp. 251–252. doi: 10.1093/cvr/cvp362.

Derra, K. *et al.* (2012) 'Profile: Nanoro health and demographic surveillance system', *International Journal of Epidemiology*, 41(5), pp. 1293–1301. doi: 10.1093/ije/dys159.

Dias, C. *et al.* (2019) 'Sex-specific effects of a microsatellite polymorphism on human growth hormone receptor gene expression', *Molecular and Cellular Endocrinology*. Elsevier, 492(110442), pp. 1–11. doi: 10.1016/j.mce.2019.05.001.

Diaz, A. *et al.* (2018) 'Carotid Intima Media Thickness Reference Intervals for a Healthy Argentinean Population Aged 11 – 81 Years', *International Journal of Hypertension*, 2018. doi: https://doi.org/10.1155/2018/8086714.

Dichgans, M. *et al.* (2014) 'Shared genetic susceptibility to ischemic stroke and coronary artery disease : A genome-wide analysis of common variants', *Stroke*, 45(1), pp. 24–36. doi: 10.1161/STROKEAHA.113.002707.

Dimas, A. S. *et al.* (2012) 'Sex-biased genetic effects on gene regulation in humans', *Genome Research*, 22(12), pp. 2368–2375. doi: 10.1101/gr.134981.111.

Divers, J. *et al.* (2017) 'Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes', *BMC Genetics*. BMC Genetics, 18(1), pp. 1–13. doi: 10.1186/s12863-017-0572-9.

Domarkiene, I. *et al.* (2013) 'RTN4 and FBXL17 genes are associated with coronary heart disease in genome-wide association analysis of Lithuanian families', *Balkan Journal of Medical Genetics*, 16(2), pp. 17–22. doi: 10.2478/bjmg-2013-0026.

Dong, C. *et al.* (2015) 'Genetic variants in LEKR1 and GALNT10 modulate sex-difference in carotid intima-media thickness: A genome-wide interaction study', *Atherosclerosis*. Elsevier Ltd, 240(2), pp. 462–467. doi: 10.1016/j.atherosclerosis.2015.04.019.

Dumont, D. J. *et al.* (1998) 'Cardiovascular Failure in Mouse Embryos Deficient in VEGF Receptor-3', *Science*, 282(October), pp. 946–950.

Duncan, E. *et al.* (2014) 'Unlocking the Genetics of Complex Diseases : THE GWAS and Beyond', *Bioinformatics*, 5(AUG), pp. 1–14. doi: 10.3389/fgene.2014.00250.

Emilsson, V. *et al.* (2018) 'Co-regulatory networks of human serum proteins link genetics to disease', *Science*, 361(6404), pp. 769–773. doi: 10.1126/science.aaq1327.

ENCODE Project Consortium *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome.', *Nature*, 489(7414), pp. 57–74. doi: nature11247 [pii]\n10.1038/nature11247.

Engelen, L. *et al.* (2013) 'Reference intervals for common carotid intima-medi thickness measured with echotracking: Relation with risk factors', *European Heart Journal*, 34(30), pp. 2368–2380. doi: 10.1093/eurheartj/ehs380.

Ernst, J. and Kellis, M. (2013) 'ChromHMM: automating chromatin state discovery and characterization Supplementary Material', *Nat Methods*, 9(3), pp. 215–216. doi: 10.1038/nmeth.1906.

Erqou, S. *et al.* (2015) 'Racial differences in the burden of coronary artery calcium and carotid intima media thickness between Blacks and Whites', *Neth Heart J*, 23, pp. 44–51. doi: 10.1007/s12471-014-0610-4.

Evangelou, E. *et al.* (2018) 'Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits', *Nature Genetics*, 50(10), pp. 1412–1425. doi: 10.1038/s41588-018-0205-x.

Fagnani, C. *et al.* (2013) 'Genetic and environmental components of carotid artery elasticity: An Italian twin study', *European Journal of Internal Medicine*. European Federation of Internal Medicine., 24(4), pp. e53–e54. doi: 10.1016/j.ejim.2013.03.001.

Fasanaro, P., Capogrossi, M. C. and Martelli, F. (2010) 'Regulation of the endothelial cell cycle by the ubiquitin-proteasome system', *Cardiovascular Research*, 85(2), pp. 272–280. doi: 10.1093/cvr/cvp244.

Feitosa, M. F. *et al.* (2018) 'Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries', *PLoS ONE*, 13(6), pp. 1–36. doi: 10.1371/journal.pone.0198166.

Fermin, D. R. *et al.* (2008) 'Sex and age dimorphism of myocardial gene expression in nonischemic human heart failure.', *Circulation. Cardiovascular genetics*, 1(2), pp. 117–125. doi: 10.1161/CIRCGENETICS.108.802652.

Fernández-Sanlés, A. *et al.* (2017) 'Association between DNA methylation and coronary heart disease or other atherosclerotic events: a systematic review', *Atherosclerosis*, 263, pp. 325–333. doi: 10.1111/ijlh.12426.

Fox, C. S. *et al.* (2003) 'Genetic and environmental contributions to atherosclerosis phenotypes in men and women: Heritability of carotid intima-media thickness in the Framingham heart study', *Stroke*, 34(2), pp. 397–401. doi: 10.1161/01.STR.0000048214.56981.6F.

Franceschini, N. *et al.* (2018) 'GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes', *Nature Communications*, 9(1), p. 5141. doi: 10.1038/s41467-018-07340-5.

Franconi, F. *et al.* (2017) 'Human cells involved in atherosclerosis have a sex', *International Journal of Cardiology*. Elsevier Ireland Ltd, 228, pp. 983–1001. doi: 10.1016/j.ijcard.2016.11.118.

Fuster, V. *et al.* (2005) 'Atherothrombosis and high-risk plaque: Part I: Evolving concepts', *Journal of the American College of Cardiology*. Elsevier Masson SAS, 46(6), pp. 937–954. doi: 10.1016/j.jacc.2005.03.074.

Fuster, V. (2014) 'Global burden of cardiovascular disease: Time to implement feasible strategies and to monitor results', *Journal of the American College of Cardiology*, 64(5), pp. 520–522. doi: 10.1016/j.jacc.2014.06.1151.

Gao, F. *et al.* (2015) 'XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome', *Journal of Heredity*, 106(5), pp. 666–671. doi: 10.1093/jhered/esv059.

Gershoni, M. and Pietrokovski, S. (2017) 'The landscape of sex-differential transcriptome and its consequent selection in human adults', *BMC Biology*. BMC Biology, 15(1), pp. 1–15. doi: 10.1186/s12915-017-0352-z.

Gertow, K. *et al.* (2012) 'Identification of the BCAR1-CFDP1-TMEM170A Locus as a

Determinant of Carotid Intima-Media Thickness and Coronary Artery Disease Risk', *CircGenetics*. doi: 10.1161/CIRCGENETICS.112.963660.

Gilks, W. P., Abbott, J. K. and Morrow, E. H. (2014) 'Sex differences in disease genetics: Evidence, evolution, and detection', *Trends in Genetics*. Elsevier Ltd, 30(10), pp. 453–463. doi: 10.1016/j.tig.2014.08.006.

Glenn, H. L., Wang, Z. and Schwartz, L. M. (2009) 'Acheron, a Lupus antigen family member, regulates integrin expression, adhesion, and motility in differentiating myoblasts', *American Journal of Physiology-Cell Physiology*, 298(1), pp. C46–C55. doi: 10.1152/ajpcell.00387.2009.

Goikuria, H. *et al.* (2018) 'Characterization of Carotid Smooth Muscle Cells during Phenotypic Transition', *Cells*, 7(3), p. 23. doi: 10.3390/cells7030023.

Goldstein, J. L. and Brown, M. S. (1974) 'Expression of the familial hypercholesterolemia gene in heterozygotes: model for a dominant disorder in man.', *Transactions of the Association of American Physicians*, 87, pp. 120–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4376290.

Gómez-Olivé, F. X. *et al.* (2017) 'Regional and Sex Differences in the Prevalence and Awareness of Hypertension: An H3Africa AWI-Gen Study Across 6 Sites in Sub-Saharan Africa', *Global Heart*, 12(2). doi: 10.1016/j.gheart.2017.01.007.

Graffelman, J. and Weir, B. S. (2016) 'Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome', *Heredity*. Nature Publishing Group, 116(6), pp. 558–568. doi: 10.1038/hdy.2016.20.

Greaves, D. R. and Gordon, S. (2008) 'The macrophage scavenger receptor at 30 years of age: current knowledge and future challenges: TABLE 1.', *Journal of Lipid Research*, 50(Supplement), pp. S282–S286. doi: 10.1194/jlr.r800066-jlr200.

GTex Consortium *et al.* (2017) 'Genetic effects on gene expression across human tissues', *Nature*, 550(7675), pp. 204–213. doi: 10.1038/nature24277.

Guo, D. F. *et al.* (2004) 'A novel angiotensin II type 1 receptor-associated protein induces cellular hypertrophy in rat vascular smooth muscle and renal proximal tubular cells', *Journal of Biological Chemistry*, 279(20), pp. 21109–21120. doi: 10.1074/jbc.M401544200.

Gurdasani, D. *et al.* (2015) 'The African Genome Variation Project shapes medical genetics in Africa.', *Nature*. Nature Publishing Group, 517(7534), pp. 327–332. doi: 10.1038/nature13997.The.

Hamanishi, T. *et al.* (2004) 'Functional variants in the glutathione peroxidase-1 (GPx-1) gene are associated with increased intima-media thickness of carotid arteries and risk of macrovascular diseases in Japanese type 2 diabetic patients', *Clinical Laboratory*, 53(9), pp. 2455–2460. doi: 10.2337/diabetes.53.9.2455.

Hansen, K. *et al.* (2016) 'European Journal of Internal Medicine The effect of smoking on carotid intima – media thickness progression rate and rate of lumen diameter reduction', *European Journal of Internal Medicine*. European Federation of Internal Medicine., 28, pp. 74–79. doi: 10.1016/j.ejim.2015.10.018.

Harmston, N. and Lenhard, B. (2013) 'Chromatin and epigenetic features of long-range gene regulation', *Nucleic Acids Research*, 41(15), pp. 7185–7199. doi: 10.1093/nar/gkt499.

Van der Harst, P. and Verweij, N. (2017) 'Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease', *Circulation Research*, (122), pp. 433–443. doi: 10.17632/2zdd47c94h.1.

Hellwege, J. N. *et al.* (2017) 'Population Stratification in Genetic Association Studies', *Current Protocols in Human Genetics*, 95(1), pp. 1.22.1-1.22.23. doi: 10.1002/cphg.48.

Herrington, W. *et al.* (2016) 'Epidemiology of Atherosclerosis and the Potential to Reduce the Global Burden of Atherothrombotic Disease', *Circulation Research*, 118(4), pp. 535–546. doi: 10.1161/CIRCRESAHA.115.307611.

Herrmann, J., Lerman, L. O. and Lerman, A. (2010) 'On to the road to degradation: Atherosclerosis and the proteasome', *Cardiovascular Research*, 85(2), pp. 291–302. doi: 10.1093/cvr/cvp333.

Higashi, Y. *et al.* (2016) 'Abstract 13854: Micro-RNA Regulation of Collagen Production by Vascular Smooth Muscle Cells Mediated by La Ribonucleoprotein Domain Family Member 6: Potential Mechanisms Underlying Stable Phenotype of Atherosclerotic Plaque by Insulin-like Growth Factor 1', *Circulation*. American Heart Association, 134(suppl_1), pp. A13854–A13854. doi: 10.1161/circ.134.suppl_1.13854.

Hinds, D. A. *et al.* (2005) 'Whole-Genome Patterns of Common DNA Variation in Three Human Populations', *Science*, 307(5712), pp. 1072–1079. doi: 10.1126/science.1105436.

Hobbs, A. and Ramsay, M. (2015) 'Epigenetics and the burden of noncommunicable disease: a paucity of research in Africa', *Epigenomics*, 7(4), pp. 627–639. doi: 10.2217/epi.15.17.

Hoffmann, P., Feige, J. J. and Alfaidy, N. (2006) 'Expression and oxygen regulation of endocrine gland-derived vascular endothelial growth factor/prokineticin-1 and its receptors in human placenta during early pregnancy', *Endocrinology*, 147(4), pp. 1675–1684. doi: 10.1210/en.2005-0912.

Holland, Z. *et al.* (2009) 'Carotid intima-media thickness is a predictor of coronary artery disease in South African black patients.', *Cardiovascular journal of Africa*, 20(4), pp. 237–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19701535%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3721310.

Honne, K. *et al.* (2016) 'A longitudinal genome-wide association study of anti-tumor necrosis factor response among Japanese patients with rheumatoid arthritis', *Arthritis Research and Therapy*. Arthritis Research & Therapy, 18(1), pp. 1–10. doi: 10.1186/s13075-016-0920-6.

Howard, G. *et al.* (1993) 'Carotid Artery Intimal-Medial Thickness Distribution in General Populations As Evaluated by B-Mode Ultrasound', *Stroke*, 24(9), pp. 1297–1304.

Howie, B. *et al.* (2012) 'Fast and accurate genotype imputation in genome-wide association studies through pre-phasing', *Nature Genetics*. Nature Publishing Group, 44(8), pp. 955–959. doi: 10.1038/ng.2354.

Howie, B. N., Donnelly, P. and Marchini, J. (2009) 'A flexible and accurate genotype imputation method for the next generation of genome-wide association studies', *PLoS Genetics*, 5(6). doi: 10.1371/journal.pgen.1000529.

Hsu, Y. H. *et al.* (2010) 'An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits', *PLoS Genetics*, 6(6), pp. 1–16. doi: 10.1371/journal.pgen.1000977.

Hübel, C. *et al.* (2018) 'Genomics of body fat percentage may contribute to sex bias in anorexia

nervosa', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, (November 2018), pp. 428–438. doi: 10.1002/ajmg.b.32709.

Iles, M. M. (2008) 'What can genome-wide association studies tell us about the genetics of common disease?', *PLoS Genetics*, 4(2). doi: 10.1371/journal.pgen.0040033.

Inouye, M. *et al.* (2012) 'Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis', *PLoS Genetics*, 8(8). doi: 10.1371/journal.pgen.1002907.

International Human Genome Sequencing Consortium (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(February), pp. 860–921.

Ito, S. *et al.* (2015) 'Novel sex-dependent differentially methylated regions are demethylated in adult male mouse livers', *Biochemical and Biophysical Research Communications*. Elsevier Ltd, 462(4), pp. 332–338. doi: 10.1016/j.bbrc.2015.04.137.

Iyer, D. *et al.* (2018) 'Coronary artery disease genes SMAD3 and TCF21 promote opposing interactive genetic programs that regulate smooth muscle cell differentiation and disease risk', *PLoS Genetics*, 14(10), pp. 1–28. doi: 10.1371/journal.pgen.1007681.

Jia, S. J., Gao, K. Q. and Zhao, M. (2017) 'Epigenetic regulation in monocyte/macrophage: A key player during atherosclerosis', *Cardiovascular Therapeutics*, 35(3), pp. 1–11. doi: 10.1111/1755-5922.12262.

Jin, J. *et al.* (2004) 'Systematic analysis and nomenclature of mammalian F-box proteins', *Genes & Development*, 18(1), pp. 2573–2580. doi: 10.1016/j.matchemphys.2015.10.013.

Joint National Committee on Prevention Detection Evaluation and Treatment of High Blood Pressure (2003) 'Prevention , Detection , Evaluation , and Treatment of High Blood Pressure', *Blood Pressure*, 289(19), pp. 1206–52. doi: 10.1161/01.HYP.0000107251.49515.c2.

Justice, A. E. *et al.* (2017) 'Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits', *Nature Communications*, 8, pp. 1–19. doi: 10.1038/ncomms14977.

Kahn, K. *et al.* (2012) 'Profile: Agincourt health and socio-demographic surveillance system', *International Journal of Epidemiology*, 41(4), pp. 988–1001. doi: 10.1093/ije/dys115.

Kang, H. M. *et al.* (2010) 'Variance component model to account for sample structure in genome-wide association studies', *Nature Genetics*. Nature Publishing Group, 42(4), pp. 348–354. doi: 10.1038/ng.548.

Karnes, J. H. *et al.* (2017) 'Abstract 24015: Genome-Wide Association Study of Vasodilator Response in Pulmonary Arterial Hypertension', *Circulation*. American Heart Association, 136(suppl_1), pp. A24015–A24015. doi: 10.1161/circ.136.suppl_1.24015.

Katakami, N. *et al.* (2009) 'Combined effect of oxidative stress-related gene polymorphisms on atherosclerosis', *Biochemical and Biophysical Research Communications*. Elsevier Inc., 379(4), pp. 861–865. doi: 10.1016/j.bbrc.2008.12.154.

Katakami, N. *et al.* (2010) 'Accumulation of gene polymorphisms related to oxidative stress is associated with myocardial infarction in Japanese type 2 diabetic patients.', *Atherosclerosis*. Elsevier Ireland Ltd, 212(2), pp. 534–8. doi: 10.1016/j.atherosclerosis.2010.06.010.

Khramtsova, E. A., Davies, L. K. and Stranger, B. E. (2019) 'The role of sex in the genomics of human complex traits', *Nature Reviews Genetics*. Springer US, 20, pp. 173–190. doi: 10.1038/s41576-018-0083-1.

Kianoush, S. *et al.* (2017) 'Associations of Cigarette Smoking With Subclinical Inflammation and Atherosclerosis: ELSA-Brasil (The Brazilian Longitudinal Study of Adult Health)', *Journal of the American Heart Association*, 6(6), p. e005088. doi: 10.1161/JAHA.116.005088.

Kichaev, G. *et al.* (2019) 'Leveraging Polygenic Functional Enrichment to Improve GWAS Power', *The American Journal of Human Genetics*. ElsevierCompany., 104(1), pp. 65–75. doi: 10.1016/j.ajhg.2018.11.008.

Kiel, P. D. *et al.* (2007) 'Genome-wide association with bone mass and geometry in the Framingham Heart Study', *BMC Medical Genetics*, 4(8), pp. 1–13. doi: 10.1186/1471-2350-8-S1-S14.

Kim, J. J. *et al.* (2013) 'Identification of KCNN2 as a susceptibility locus for coronary artery aneurysms in Kawasaki disease using genome-wide association analysis', *Journal of Human Genetics*. Nature Publishing Group, 58(8), pp. 521–525. doi: 10.1038/jhg.2013.43.

Kircher, M. *et al.* (2014) 'A general framework for estimating the relative pathogenicity of

human genetic variants', *Nature Genetics*. Nature Publishing Group, 46(3), pp. 310–315. doi: 10.1038/ng.2892.

Kisliouk, T. *et al.* (2003) 'Presence and regulation of endocrine gland vascular endothelial growth factor/prokineticin-1 and its receptors in ovarian cells', *Journal of Clinical Endocrinology and Metabolism*, 88(8), pp. 3700–3707. doi: 10.1210/jc.2003-030492.

Köhler, R. *et al.* (2001) 'Expression of ryanodine receptor type 3 and TRP channels in endothelial cells: Comparison of in situ and cultured human endothelial cells', *Cardiovascular Research*, 51(1), pp. 160–168. doi: 10.1016/S0008-6363(01)00281-4.

König, I. R. *et al.* (2014) 'How to Include Chromosome X in Your Genome-Wide Association Study', *Genetic Epidemiology*, 38(2), pp. 97–103. doi: 10.1002/gepi.21782.

Kuipers, A. L. *et al.* (2013) 'Genetic epidemiology and genome-wide linkage analysis of carotid artery ultrasound traits in multigenerational African ancestry families', *Atherosclerosis*. Elsevier Ltd, 231(1), pp. 120–123. doi: 10.1016/j.atherosclerosis.2013.09.005.

Lange, L. A. *et al.* (2002) 'Heritability of carotid artery intima-medial thickness in type 2 diabetes', *Stroke*, 33(7), pp. 1876–1881. doi: 10.1161/01.STR.0000019909.71547.AA.

Lasky-Su, J. *et al.* (2008) 'Genome-wide association scan of attention deficit hyperactivity disorder', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 147(8), pp. 1337–1344. doi: 10.1002/ajmg.b.30866.

LeCouter, J. *et al.* (2001) 'Identification of an angiogenic mitogen selective for endocrine gland endothelium', *Nature*, 412(6850), pp. 877–884. doi: 10.1038/35091000.

Lee, J. J. *et al.* (2018) 'Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals', *Nature Genetics*, 50(8), pp. 1112–1121. doi: 10.1038/s41588-018-0147-3.

Lee, Y. H. *et al.* (2014) 'Normative and mean carotid intima-media thickness values according to metabolic syndrome in Koreans: The Namwon Study', *Atherosclerosis*, 234(1), pp. 230–236. doi: 10.1016/j.atherosclerosis.2014.02.023.

de Leeuw, C. A. *et al.* (2015) 'MAGMA: Generalized Gene-Set Analysis of GWAS Data', *PLoS Computational Biology*, 11(4). doi: 10.1371/journal.pcbi.1004219.

Lettre, G. *et al.* (2011) 'Genome-Wide association study of coronary heart disease and its risk factors in 8,090 african americans: The nhlbi CARe project', *PLoS Genetics*, 7(2). doi: 10.1371/journal.pgen.1001300.

Li, C. *et al.* (2015) 'Genetic association and gene-smoking interaction study of carotid intima-media thickness at five GWAS-indicated genes: The Bogalusa Heart Study', *Gene*. Elsevier B.V., 562(2), pp. 226–231. doi: 10.1016/j.gene.2015.02.078.

Li, Q. *et al.* (2017) 'Genome-wide association study of paliperidone efficacy', *Pharmacogenetics and genomics*, (27), pp. 7–18. doi: 10.1097/FPC.0000000000000250.

Li, Q. S. *et al.* (2015) 'Variations in the FRA10AC1 fragile site and 15q21 are associated with Cerebrospinal fluid Aβ1-42 level', *PLoS ONE*, 10(8), pp. 1–17. doi: 10.1371/journal.pone.0134000.

Li, Y. R. and Keating, B. J. (2014) 'Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations.', *Genome medicine*, 6(10), p. 91. doi: 10.1186/s13073-014-0091-5.

Liang, L. *et al.* (2009) 'Cross-sectional and longitudinal association of cigarette smoking with carotid atherosclerosis in Chinese adults', *Preventive Medicine*. Elsevier Inc., 49(1), pp. 62–67. doi: 10.1016/j.ypmed.2009.05.006.

Liao, Y. C. *et al.* (2010) 'Sex-differential genetic effect of phosphodiesterase 4D (PDE4D) on carotid atherosclerosis', *BMC Medical Genetics*, 11(1). doi: 10.1186/1471-2350-11-93.

Liberzon, A. *et al.* (2011) 'Molecular signatures database (MSigDB) 3.0', *Bioinformatics*, 27(12), pp. 1739–1740. doi: 10.1093/bioinformatics/btr260.

Liberzon, A. *et al.* (2015) 'The Molecular Signatures Database Hallmark Gene Set Collection', *Cell Systems*, 1(6), pp. 417–425. doi: 10.1016/j.cels.2015.12.004.

Lieb, W. and Vasan, R. S. (2018) 'Scientific Contributions of Population-Based Studies to Cardiovascular Epidemiology in the GWAS Era', *Frontiers in Cardiovascular Medicine*, 5(June), pp. 1–6. doi: 10.3389/fcvm.2018.00057.

Lin, H.-F. *et al.* (2015) 'Age and sex differences in the effect of parental stroke on the progression of carotid intima-media thickness.', *Atherosclerosis*. Elsevier Ltd, 241(1), pp. 229–

233. doi: 10.1016/j.atherosclerosis.2015.02.025.

Lippert, C. *et al.* (2011) 'FaST linear mixed models for genome-wide association studies', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 8, pp. 833–835. Available at: https://doi.org/10.1038/nmeth.1681.

Liu, B. *et al.* (2017) 'Carotid Intima-media Thickness and its Association with Conventional Risk Factors in Low-income Adults: A Population-based Cross-Sectional Study in China', *Scientific Reports*. Nature Publishing Group, 7(September 2016), p. 41500. doi: 10.1038/srep41500.

Liu, J. *et al.* (2018) 'The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease', *Human Genetics*. Springer Berlin Heidelberg, 137(6–7), pp. 553–567. doi: 10.1007/s00439-018-1910-3.

Liu, M. *et al.* (2019) 'Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use', *Nature Genetics*, 51(2), pp. 237–244. doi: 10.1038/s41588-018-0307-5.

Loh, P. *et al.* (2016) 'Reference-based phasing using the Haplotype Reference Consortium panel', *Nature Genetics*, 48(11), pp. 1443–1448. doi: 10.1038/ng.3679.

Lonjou, C. *et al.* (2003) 'Linkage disequilibrium in human populations', *Proceedings of the National Academy of Sciences*, 100(10), pp. 6069–6074.

López-Mejías, R. *et al.* (2019) 'Identification of a 3′-Untranslated Genetic Variant of RARB Associated With Carotid Intima-Media Thickness in Rheumatoid Arthritis: A Genome-Wide Association Study', *Arthritis and Rheumatology*, 71(3), pp. 351–360. doi: 10.1002/art.40734.

Maguire, L. H. *et al.* (2018) 'Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease', *Nature Genetics*. Springer US, 50(10), pp. 1359–1365. doi: 10.1038/s41588-018-0203-z.

Malik, R. *et al.* (2018) 'Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes', *Nature Genetics*, 50(4), pp. 524–537. doi: 10.1038/s41588-018-0058-3.

Manolio, T. A. *et al.* (2009) 'Finding the missing heritability of complex diseases.', *Nature*, 461(7265), pp. 747–53. doi: 10.1038/nature08494.

Marchini, J. *et al.* (2007) 'A new multipoint method for genome-wide association studies by imputation of genotypes.', *Nature genetics*, 39(7), pp. 906–13. doi: 10.1038/ng2088.

Martin, A. R. *et al.* (2018) 'The critical needs and challenges for genetic architecture studies in Africa', *Current Opinion in Genetics & Development*, (53), pp. 113–120. doi: 10.1016/j.gde.2018.08.005.

Maurano, M. T. *et al.* (2012) 'Systematic Localization of Common Disease-Associated Variation in Regulatory DNA', *Science*, 337(6099), pp. 1190–1195. doi: 10.1126/science.1222794.

Mayanagi, T. and Sobue, K. (2011) 'Diversification of caldesmon-linked actin cytoskeleton in cell motility', *Cell Adhesion and Migration*, 5(2), pp. 150–159. doi: 10.4161/cam.5.214398.

McAllister, K. *et al.* (2017) 'Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases', *American Journal of Epidemiology*, 186(7), pp. 753–761. doi: 10.1093/aje/kwx227.

McKay, J. d *et al.* (2017) 'Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes', *Nat Genet*, 49(7), pp. 1126–1132. doi: 10.1038/ng.3892.

McPherson, R. and Tybjaerg-Hansen, A. (2016) 'Genetics of Coronary Artery Disease', *Circulation Research*, 118(4), pp. 564–578. doi: 10.1161/CIRCRESAHA.115.306566.

Medda, E. *et al.* (2014) 'Heritability of arterial stiffness and carotid intima-media thickness: An Italian twin study', *Nutrition, Metabolism and Cardiovascular Diseases*. Elsevier Ltd, 24(5), pp. 511–517. doi: 10.1016/j.numecd.2013.10.031.

Melton, P. E. *et al.* (2013) 'Genetic architecture of carotid artery intima-media thickness in Mexican Americans', *Circulation: Cardiovascular Genetics*, 6(2), pp. 211–221. doi: 10.1161/CIRCGENETICS.113.000079.

Middeldorp, C. M. *et al.* (2019) 'A Genome-Wide Association Meta-Analysis of Attention-De fi cit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts', *J Am Acad*

Child Adolesc Psychiatry, 55(10), pp. 896–905. doi: 10.1016/j.jaac.2016.05.025.

Miller, V. M. *et al.* (2013) 'Genetic polymorphisms associated with carotid artery intima-media thickness and coronary artery calcification in women of the Kronos Early Estrogen Prevention Study', *Physiol Genomics*, 45(2), pp. 79–88. doi: 10.1152/physiolgenomics.00114.2012.

Mogil, L. S. *et al.* (2018) 'Genetic architecture of gene expression traits across diverse populations', *PLoS Genetics*, 14(8), pp. 1–21. doi: 10.1371/journal.pgen.1007586.

Morgan, T. M., Krumholz, H. M. and Lifton, R. P. (2007) 'Nonvalidation of Reported Genetic Risk in a Large-Scale Replication Study', *Jama*, 297(14). doi: 10.1001/jama.297.14.1551.

Morris, J. A. *et al.* (2019) 'An atlas of genetic influences on osteoporosis in humans and mice', *Nature Genetics*, 51(2), pp. 258–266. doi: 10.1038/s41588-018-0302-x.

Mota, R. *et al.* (2017) 'Atherosclerosis: Pathogenesis, Genetics and Experimental Models', *eLS*, (October), pp. 1–10. doi: 10.1002/9780470015902.a0005998.pub3.

Mozaffarian, D. (2017) 'Global Scourge of Cardiovascular Disease: Time for Health Care Systems Reform and Precision Population Health', *Journal of the American College of Cardiology*, 70(1), pp. 26–28. doi: 10.1016/j.jacc.2017.05.007.

Musunuru, K. *et al.* (2010) 'From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus', *Nature*. Nature Publishing Group, 466(7307), pp. 714–719. doi: 10.1038/nature09266.

Nazarenko, M. S. *et al.* (2015) 'A comparison of genome-wide DNA methylation patterns between different vascular tissues from patients with coronary heart disease', *PLoS ONE*, 10(4), pp. 1–15. doi: 10.1371/journal.pone.0122601.

Nazarian, A., Yashin, A. I. and Kulminski, A. M. (2019) 'Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex disparities', *Alzheimer's Research and Therapy*. Alzheimer's Research & Therapy, 11(1), pp. 1–21. doi: 10.1186/s13195-018-0458-8.

NCD Risk Factor Collaboration (2019) 'Rising rural body-mass index is the main driver of the global obesity epidemic in adults', *Nature*, 569(7755), pp. 260–264. doi: 10.1038/s41586-019-1171-x.

Nikpay, M. *et al.* (2015) 'A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease.', *Nature genetics*, 47(10). doi: 10.1038/ng.3396.

Njock, M.-S. *et al.* (2015) 'Endothelial cells suppress monocyte activation through secretion of extracellular vesicles containing antiinflammatory microRNAs', *Blood*, 125(20), pp. 3202–3212. doi: 10.1182/blood-2014-11-611046.

Nonterah, E. A. *et al.* (2019) 'Classical Cardiovascular Risk Factors and HIV are Associated With Carotid Intima-Media Thickness in Adults From Sub-Saharan Africa: Findings From H3Africa AWI-Gen Study', *Journal of the American Heart Association*. doi: 10.1161/JAHA.118.011506.

Norman, J. P. *et al.* (2008) 'HIV-1 Tat activates neuronal ryanodine receptors with rapid induction of the unfolded protein response and mitochondrial hyperpolarization', *PLoS ONE*, 3(11). doi: 10.1371/journal.pone.0003731.

Noubiap, J. J. *et al.* (2018) 'Prevalence of dyslipidaemia among adults in Africa: a systematic review and meta-analysis', *The Lancet Global Health*, 6(9), pp. e998–e1007. doi: 10.1016/S2214-109X(18)30275-4.

O'Donnell, C. J. *et al.* (2007) 'Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study.', *BMC Medical Genetics*, 8 Suppl 1, p. S4. doi: 10.1186/1471-2350-8-S1-S4.

O'Donnell, C. J. *et al.* (2011) 'Genome-Wide Association Study for Coronary Artery Calcification With Follow-Up in Myocardial Infarction', *Circulation*, 124(25), pp. 2855–2864. doi: 10.1161/circulationaha.110.974899.

O'Leary, D. H. *et al.* (1999) 'Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults', *The New England Journal of Medicine*, 340, pp. 14–22. doi: 10.1056/NEJM199901073400103.

Oduro, A. R. *et al.* (2012) 'Profile of the Navrongo health and demographic surveillance system', *International Journal of Epidemiology*, 41(4), pp. 968–976. doi: 10.1093/ije/dys111.

Okeahialam, B. N. *et al.* (2011) 'Carotid intima media thickness as a measure of cardiovascular disease Burden in Nigerian Africans with hypertension and diabetes mellitus', *International*

Journal of Vascular Medicine, 2011. doi: 10.1155/2011/327171.

Olt, S. *et al.* (2016) 'The relationship between HbA1c and carotid intima-media thickness in type 2 diabetic patients', *Pan African Medical Jounal*, 23(224), pp. 1–5. doi: 10.11604/pamj.2016.23.224.8970.

Van Den Oord, S. C. H. *et al.* (2013) 'Carotid intima-media thickness for cardiovascular risk assessment: Systematic review and meta-analysis', *Atherosclerosis*. Elsevier Ltd, 228(1), pp. 1–11. doi: 10.1016/j.atherosclerosis.2013.01.025.

Owolabi, M. O. *et al.* (2015) 'Can common carotid intima media thickness serve as an indicator of both cardiovascular phenotype and risk among black Africans?', *European Journal of Preventive Cardiology*, 22(11), pp. 1442–1451. doi: 10.1177/2047487314547656.

Paladugu, R. *et al.* (2003) 'HIV Tat protein causes endothelial dysfunction in porcine coronary arteries', *Journal of Vascular Surgery*, 38(3), pp. 549–555. doi: 10.1016/S0741-5214(03)00770-5.

Paone, S. *et al.* (2019) 'Endothelial cell apoptosis and the role of endothelial cell-derived extracellular vesicles in the progression of atherosclerosis', *Cellular and Molecular Life Sciences*. Springer International Publishing, 76(6), pp. 1093–1106. doi: 10.1007/s00018-018-2983-9.

Park, S. *et al.* (2015) 'Mercapturic Acids Derived from the Toxicants Acrolein and Crotonaldehyde in the Urine of Cigarette Smokers from Five Ethnic Groups with Differing Risks for Lung Cancer', *PloS one*, 10(6), pp. 1–17. doi: 10.1371/journal.pone.0124841.

Penney, M. E. *et al.* (2019) 'A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer', *BMC cancer*. BMC Cancer, 19(1), p. 133. doi: http://dx.doi.org/10.1186/s12885-019-5346-5.

Peprah, E., Tekola-ayele, F. and Royal, C. D. (2015) 'Genome-Wide Association Studies in Africans and African Americans : Expanding the Framework of the Genomics of Human Traits and Disease', pp. 40–51. doi: 10.1159/000367962.

Pers, T. H. *et al.* (2015) 'Biological interpretation of genome-wide association studies using predicted gene functions', *Nature Communications*, 6, p. 5890. doi: 10.1038/ncomms6890.

Polfus, L. M. *et al.* (2013) 'Genome-Wide Association Study of Gene by Smoking Interactions in Coronary Artery Calcification', *PLoS ONE*, 8(10). doi: 10.1371/journal.pone.0074642.

Popejoy, Alice B. and Fullerton, S. M. (2016) 'Genomics is failing on diversity', *Nature*. Nature Publishing Group, pp. 161–164. doi: 10.1038/538161a.

Popejoy, Alice B and Fullerton, S. M. (2016) 'Supplementary information to: Genomics is failing on diversity (Comment in Nature 538, 161–164; 2016)', *Nature*, (538), pp. 161–164.

Pott, J. *et al.* (2017) 'Genome-wide meta-analysis identifies novel loci of plaque burden in carotid artery', *Atherosclerosis*, 259, pp. 32–40. doi: 10.1016/j.atherosclerosis.2017.02.018.

Poussin, C. *et al.* (2015) 'Systems Biology Reveals Cigarette Smoke-Induced Concentration-Dependent Direct and Indirect Mechanisms That Promote Monocyte – Endothelial Cell Adhesion', *Toxicological Sciences*, 147(2), pp. 370–385. doi: 10.1093/toxsci/kfv137.

Powell, S. R. and Divald, A. (2010) 'The ubiquitin-proteasome system in myocardial ischaemia and preconditioning', *Cardiovascular Research*, 85(2), pp. 303–311. doi: 10.1093/cvr/cvp321.

Puig, O. *et al.* (2011) 'A gene expression signature that classifies human atherosclerotic plaque by relative inflammation status', *Circulation: Cardiovascular Genetics*, 4(6), pp. 595–604. doi: 10.1161/CIRCGENETICS.111.960773.

Purcell, S. *et al.* (2007) 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

Raines, E. W. and Ferri, N. (2005) 'Cytokines affecting endothelial and smooth muscle cells in vascular disease', *Journal of Lipid Research*, 46(6), pp. 1081–1092. doi: 10.1194/jlr.r500004-jlr200.

Rajavashisth, T. B. *et al.* (1990) 'Induction of endothelial cell expression of granulocyte and macrophage colony-stimulating factors by modified low-density lipoproteins', *Nature*, 344(6263), pp. 254–257. doi: 10.1038/344254a0.

Ramsay, M. *et al.* (2016) 'H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries', *Global Health, Epidemiology and Genomics*, 1(e20). doi:

10.1017/gheg.2016.17.

Ramsay, M. *et al.* (2018) 'Regional and sex-specific variation in BMI distribution in four sub-Saharan African countries : The H3Africa AWI-Gen study Regional and sex-specific variation in BMI distribution in four sub-Saharan', *Global Health Action*. Taylor & Francis, 11(2). doi: 10.1080/16549716.2018.1556561.

Ramsey, S. A., Gold, E. S. and Aderem, A. (2010) 'A systems biology approach to understanding atherosclerosis', *EMBO Molecular Medicine*, 2(3), pp. 79–89. doi: 10.1002/emmm.201000063.

Randall, J. C. *et al.* (2013) 'Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits', *PLoS Genetics*. Edited by G. Gibson, 9(6), p. e1003500. doi: 10.1371/journal.pgen.1003500.

Rawlik, K. *et al.* (2016) 'Evidence for sex-specific genetic architectures across a spectrum of human complex traits', *Genome Biology*. Genome Biology, 17(1), p. 166. doi: 10.1186/s13059-016-1025-x.

Reiger, S. *et al.* (2017) 'Health & Aging in Africa : A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI)', *PLoS ONE*, 12(10), pp. 1–12. doi: 10.7910/DVN/F5YHML.

REITSMA, P. H. (2007) 'No praise for folly: genomics will never be useful in arterial thrombosis', *Journal of Thrombosis and Haemostasis*, 5(3), pp. 454–457. doi: 10.1111/j.1538-7836.2007.02405.x.

Rheeder, P. *et al.* (2012) 'Carotid intima-media thickness and its associations with type 2 diabetes mellitus in South Africans', *Journal of Endocrinology, Metabolism and Diabetes of South Africa*, 17(3), pp. 135–140. doi: 10.1080/22201009.2012.10872292.

Richter, L. *et al.* (2007) 'Cohort profile: Mandela's children: The 1990 birth to twenty study in South Africa', *International Journal of Epidemiology*, 36(3), pp. 504–511. doi: 10.1093/ije/dym016.

Ritchie, M. D. *et al.* (2017) 'Incorporation of Biological Knowledge into the Study of Gene-Environment Interactions', *American Journal of Epidemiology*, 186(7), pp. 771–777. doi:

10.1093/aje/kwx229.

Ritz, B. R. *et al.* (2017) 'Lessons Learned from Past Gene-Environment Interaction Successes', *American Journal of Epidemiology*, 186(7), pp. 778–786. doi: 10.1093/aje/kwx230.

Rocha, D. M. *et al.* (2016) 'Saturated fatty acids trigger TLR4-mediated inflammatory response', *Atherosclerosis*. Elsevier Ltd, 244, pp. 211–215. doi: 10.1016/j.atherosclerosis.2015.11.015.

Rodford, J. L. *et al.* (2008) 'Endothelial dysfunction and reduced antioxidant protection in an animal model of the developmental origins of cardiovascular disease.', *The Journal of physiology*, 586(Pt 19), pp. 4709–20. doi: 10.1113/jphysiol.2008.156976.

Roth, G. A. *et al.* (2015) 'Global and regional patterns in cardiovascular mortality from 1990 to 2013', *Circulation*, 132(17), pp. 1667–1678. doi: 10.1161/CIRCULATIONAHA.114.008720.

Roth, G. A. *et al.* (2017) 'Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015.', *Journal of the American College of Cardiology*, 70(1), pp. 1–25. doi: 10.1016/j.jacc.2017.04.052.

Royston, P. and Wrightt, E. M. (1998) 'A Method for Estimating Age-Specific Reference Intervals (` Normal Ranges ') Based on Fractional Polynomials and Exponential Transformation', *Journal of the Royal Statistical Society*, 161(1), pp. 79–101.

Rundek, T. *et al.* (2013) 'Traditional risk factors are not major contributors to the variance in carotid intima-media thickness.', *Stroke; a journal of cerebral circulation*, 44(8), pp. 2101–8. doi: 10.1161/STROKEAHA.111.000745.

Sacco, R. L. *et al.* (2009) 'Heritability and Linkage Analysis for Carotid Intima-Media Thickness: The Family Study of Stroke Risk and Carotid Atherosclerosis', *Stroke*, 40(7), pp. 2307–2312. doi: 10.1161/STROKEAHA.109.554121.

Saleheen, D. *et al.* (2017) 'Loss of Cardioprotective Effects at the ADAMTS7 Locus as a Result of Gene-Smoking Interactions', *Circulation*, 135(24), pp. 2336–2353. doi: 10.1161/CIRCULATIONAHA.116.022069.

Salvatore, J. E. *et al.* (2018) 'Beyond genome-wide significance: Integrative approaches to the

interpretation and extension of GWAS findings for alcohol use disorder', *Addiction Biology*, 24(2), pp. 275–289. doi: 10.1111/adb.12591.

Samson, M. *et al.* (2004) 'Human endocrine gland-derived vascular endothelial growth factor: Expression early in development and in Leydig cell tumors suggests roles in normal and pathological testis angiogenesis', *Journal of Clinical Endocrinology and Metabolism*, 89(8), pp. 4078–4088. doi: 10.1210/jc.2003-032024.

Santos, I. S. *et al.* (2014) 'Carotid intima-media thickness value distributions in The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)', *Atherosclerosis*, 237(1), pp. 227–235. doi: 10.1016/j.atherosclerosis.2014.09.004.

Savage, J. E. *et al.* (2018) 'Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence', *Nature Genetics*, 50(7), pp. 912–919. doi: 10.1038/s41588-018-0152-6.

Schlebusch, C. M. *et al.* (2012) 'Genomic variation in seven Khoe-San groups reveals adaptation and complex African history', *Science*, 338(6105), pp. 374–379. doi: 10.1126/science.1227721.

Schmeisser, A. *et al.* (2006) 'Apoptosis of human macrophages by Flt-4 signaling: Implications for atherosclerotic plaque pathology', *Cardiovascular Research*, 71(4), pp. 774–784. doi: 10.1016/j.cardiores.2006.06.012.

Schoffelen, A. F. *et al.* (2015) 'Carotid Intima Media Thickness in Mainly Female HIV-Infected Subjects in Rural South Africa: Association with Cardiovascular but Not HIV-Related Factors', *Clinical Infectious Diseases*, 61(10), pp. 1606–1614. doi: 10.1093/cid/civ586.

Schork, N. J. *et al.* (2009) 'Common vs. rare allele hypotheses for complex diseases', *Current Opinion in Genetics and Development*, 19(3), pp. 212–219. doi: 10.1016/j.gde.2009.04.010.

Schroeder, S. A. (2013) 'New Evidence That Cigarette Smoking Remains the the Most Important Health Hazard', *New England Journal of Medicine*, 368(4), pp. 389–390. doi: 10.1056/NEJMe1215043.

Schunkert, H. *et al.* (2011) 'Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease', *Nature Genetics*, 43(4), pp. 333–340. doi: 10.1038/ng.784.

Schutte, R. *et al.* (2009) 'Blood Glutathione and Subclinical Atherosclerosis in African Men : The SABPA Study', *American Journal of Hypertension*. Nature Publishing Group, 22(11), pp. 1154–1159. doi: 10.1038/ajh.2009.158.

Shendre, A. *et al.* (2014) 'RYR3 gene variants in subclinical atherosclerosis among HIV-infected women in the Women's Interagency HIV Study (WIHS)', *Atherosclerosis*. Elsevier Ltd, 233(2), pp. 666–672. doi: 10.1016/j.atherosclerosis.2014.01.035.

Shendre, A., Wiener, H. W., *et al.* (2017) 'Genome-wide admixture and association study of subclinical atherosclerosis in the Women's Interagency HIV Study (WIHS)', *PLoS ONE*, 12(12), pp. 1–23. doi: 10.1371/journal.pone.0188725.

Shendre, A., Irvin, M. R., *et al.* (2017) 'Local Ancestry and Clinical Cardiovascular Events Among African Americans From the Atherosclerosis Risk in Communities Study', *Journal of the American Heart Association*, 6(4). doi: 10.1161/JAHA.116.004739.

Shrestha, S. *et al.* (2010) 'A genome-wide association study of carotid atherosclerosis in HIV-infected men', *Aids*, 24(4), pp. 583–592. doi: 10.1097/QAD.0b013e3283353c9e.

Shrestha, S. *et al.* (2012) 'Replication of RYR3 gene polymorphism association with cIMT among HIV-infected whites', *Aids*, 26(12), pp. 1571–1573. doi: 10.1097/QAD.0b013e328355359f.

Shrestha, S., Irvin, M. M. D. and Grunfeld, C. (2010) 'A Genome-wide Association Study of Carotid Atherosclerosis in HIV-infected Men', *Aids*, 24(February 20), pp. 583–592. doi: 10.1097/QAD.0b013e3283353c9e.A.

Siewert, K. M. and Voight, B. F. (2018) 'Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated With Lipid Levels and Coronary Artery Disease', *Circulation. Genomic and precision medicine*, 11(12), p. e002239. doi: 10.1161/CIRCGEN.118.002239.

Sillesen, H. *et al.* (2012) 'Carotid plaque burden as a measure of subclinical atherosclerosis: Comparison with other tests for subclinical arterial disease in the high risk plaque bioimage study', *JACC: Cardiovascular Imaging*. Elsevier Inc., 5(7), pp. 681–689. doi: 10.1016/j.jcmg.2012.03.013.

Siti, H. N., Kamisah, Y. and Kamsiah, J. (2015) 'The role of oxidative stress, antioxidants and

vascular inflammation in cardiovascular disease (a review)', *Vascular Pharmacology*. Elsevier Inc., 71, pp. 40–56. doi: 10.1016/j.vph.2015.03.005.

Skelton, R. J. P. *et al.* (2014) 'SIRPA, VCAM1 and CD34 identify discrete lineages during early human cardiovascular development', *Stem Cell Research*. Elsevier B.V., 13(1), pp. 172–179. doi: 10.1016/j.scr.2014.04.016.

Spence, J. D. and Pilote, L. (2015) 'Importance of sex and gender in atherosclerosis and cardiovascular disease', *Atherosclerosis*. Elsevier Ltd, 241(1), pp. 208–210. doi: 10.1016/j.atherosclerosis.2015.04.806.

Stamova, B. *et al.* (2012) 'The X-chromosome has a different pattern of gene expression in women compared with men with ischemic stroke', *Stroke*, 43(2), pp. 326–334. doi: 10.1161/STROKEAHA.111.629337.

Stangl, K. and Stangl, V. (2010) 'The ubiquitin-proteasome pathway and endothelial (dys)function', *Cardiovascular Research*, 85(2), pp. 281–290. doi: 10.1093/cvr/cvp315.

Stefanovic, B. *et al.* (2019) 'Discovery and evaluation of inhibitor of LARP6 as specific antifibrotic compound', *Scientific Reports*, 9(1), pp. 1–15. doi: 10.1038/s41598-018-36841-y.

Stein, J. H. *et al.* (2008) 'Use of Carotid Ultrasound to Identify Subclinical Vascular Disease and Evaluate Cardiovascular Disease Risk: A Consensus Statement from the American Society of Echocardiography Carotid Intima-Media Thickness Task Force Endorsed by the Society for Vascular', *Journal of the American Society of Echocardiography*, 21(2), pp. 93–111. doi: 10.1016/j.echo.2007.11.011.

Strawbridge, R. J. *et al.* (2011) 'Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes', *Diabetes*, 60(10), pp. 2624–2634. doi: 10.2337/db11-0415.

Strawbridge, R. J. *et al.* (2017) 'Identification of a novel proinsulin-associated SNP and demonstration that proinsulin is unlikely to be a causal factor in subclinical vascular remodelling using Mendelian randomisation', *Atherosclerosis*, 266, pp. 196–204. doi: 10.1016/j.atherosclerosis.2017.09.031.

Suh-Hang, H. J. *et al.* (2004) 'Genetic and environmental contributions to carotid intima-media

thickness and obesity phenotypes in the Northern Manhattan family study', *Stroke*, 35(10), pp. 2243–2247. doi: 10.1161/01.STR.0000142132.20442.d8.

Sukhanov, S. *et al.* (2018) 'SM22α (Smooth Muscle Protein 22-α) Promoter-Driven IGF1R (Insulin-Like Growth Factor 1 Receptor) Deficiency Promotes Atherosclerosis', *Arteriosclerosis, Thrombosis, and Vascular Biology*. American Heart Association, 38(10), pp. 2306–2317. doi: 10.1161/ATVBAHA.118.311134.

Sulovari, A. *et al.* (2017) 'Atlas of human diseases influenced by genetic variants with extreme allele frequency differences', *Human Genetics*. Springer Berlin Heidelberg, 136(1), pp. 39–54. doi: 10.1007/s00439-016-1734-y.

Sun, B. B. *et al.* (2018) 'Genomic atlas of the human plasma proteome', *Nature*. Springer US, 558(7708), pp. 73–79. doi: 10.1038/s41586-018-0175-2.

Sung, Yun J. *et al.* (2016) 'An Empirical Comparison of Joint and Stratified Frameworks for Studying GxE Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group', *Genetic Epidemiology*, 40(5), pp. 404–415. doi: 10.1002/gepi.21978.

Sung, Y. J. *et al.* (2016) 'Genome-wide association studies suggest sex-specific loci associated with abdominal and visceral fat', *International Journal of Obesity*, 40(4), pp. 662–674. doi: 10.1038/ijo.2015.217.

Sung, Y. J. *et al.* (2018) 'A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure', *American Journal of Human Genetics*, 102(3), pp. 375–400. doi: 10.1016/j.ajhg.2018.01.015.

Szilagyi, K. *et al.* (2014) 'Defective signal regulatory protein alpha (sirpa) signaling reduces atherosclerosis in mice', *Atherosclerosis*. Elsevier Ltd, 235(2), p. e91. doi: 10.1016/j.atherosclerosis.2014.05.242.

Tam, V. *et al.* (2019) 'Benefits and limitations of genome-wide association studies', *Nature Reviews Genetics*. doi: 10.1038/s41576-019-0127-1.

Tan, X. *et al.* (2017) 'Identification of Key Pathways and Genes in Advanced Coronary Atherosclerosis Using Bioinformatics Analysis', *BioMed Research International*. Hindawi,

2017, pp. 1–12. doi: 10.1155/2017/4323496.

Taylor, J. Y. *et al.* (2016) 'A Genome-wide study of blood pressure in African Americans accounting for gene-smoking interaction.', *Scientific reports*. Nature Publishing Group, 6(January 2015), p. 18812. doi: 10.1038/srep18812.

Tekola-ayele, I. F. and Rotimi, N. C. (2015) 'Translational Genomics in Low- and Middle-Income Countries : Opportunities and Challenges', *Public Health Genomics*, 18, pp. 242–247. doi: 10.1159/000433518.

The GTExArd Consortium *et al.* (2015) 'The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.', *Science*, 348(6235), pp. 648–60. doi: 10.1126/science.1262110.

Toma, I. and MacCaffrey, T. A. (2012) 'Transforming growth factor-β and atherosclerosis: interwoven atherogenic and atheroprotective aspects', *Cell Tissue Res.*, 118(24), pp. 155–175. doi: doi:10.1007/s00441-011-1189-3.

Trillhaase, A. *et al.* (2018) 'Di ff erentiation of human iPSCs into VSMCs and generation of VSMC-derived calcifying vascular cells', *Stem Cell Research*. Elsevier, 31(July), pp. 62–70. doi: 10.1016/j.scr.2018.07.008.

Tuenter, A. *et al.* (2016) 'High shear stress relates to intraplaque haemorrhage in asymptomatic carotid plaques', *Atherosclerosis*. Elsevier Ltd, 251, pp. 348–354. doi: 10.1016/j.atherosclerosis.2016.05.018.

Tzoulaki, I. *et al.* (2016) 'Worldwide Exposures to Cardiovascular Risk Factors and Associated Health Effects: Current Knowledge and Data Gaps', *Circulation*, 133(23), pp. 2314–2333. doi: 10.1161/CIRCULATIONAHA.115.008718.

United Nations (2015) *Sustainable Development Goals: 17 Goals to Transform our World*. Available at: https://www.un.org/sustainabledevelopment/ (Accessed: 22 May 2019).

Venter, J. C. *et al.* (2001) 'The Sequence of the Human Genome', *Science*, 291(February), pp. 1304–1351.

Verdugo, R. A. *et al.* (2013) 'Graphical Modeling of Gene Expression in Monocytes Suggests Molecular Mechanisms Explaining Increased Atherosclerosis in Smokers', *PLoS ONE*, 8(1).

doi: 10.1371/journal.pone.0050888.

Villines, T. C. *et al.* (2017) 'Relation of Carotid Intima-Media Thickness to Cardiovascular Events in Black Americans (From the Jackson Heart Study)', *The American Journal of Cardiology*. Elsevier Inc., 120(9), pp. 1528–1532. doi: 10.1016/j.amjcard.2017.07.046.

Viola, J. and Soehnlein, O. (2015) 'Atherosclerosis - A matter of unresolved inflammation', *Seminars in Immunology*. Elsevier Ltd, 27(3), pp. 184–193. doi: 10.1016/j.smim.2015.03.013.

Vu, K. N. *et al.* (2016) 'Causal role of alcohol consumption in an improved lipid profile: The atherosclerosis risk in communities (aric) study', *PLoS ONE*, 11(2), pp. 1–16. doi: 10.1371/journal.pone.0148765.

Vuorio, T. *et al.* (2018) 'Downregulation of VEGFR3 signaling alters cardiac lymphatic vessel organization and leads to a higher mortality after acute myocardial infarction', *Scientific Reports*, 8(1), pp. 1–13. doi: 10.1038/s41598-018-34770-4.

Wan, L. *et al.* (2018) 'Screening key genes for abdominal aortic aneurysm based on gene expression omnibus dataset'. BMC Cardiovascular Disorders, pp. 1–13.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research*, 38(16), pp. 1–7. doi: 10.1093/nar/gkq603.

Wang, L. *et al.* (2014) 'Genome-wide interaction study identifies RCBTB1 as a modifier for smoking effect on carotid intima-media thickness', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(1), pp. 219–225. doi: 10.1161/ATVBAHA.113.302706.

Ward-Caviness, C. K. *et al.* (2017) 'A genome-wide trans-ethnic interaction study links the PIGR-FCAMR locus to coronary atherosclerosis via interactions between genetic variants and residential exposure to traffic', *PLoS ONE*, 12(3), pp. 1–16. doi: 10.1371/journal.pone.0173880.

Wassel, C. L. *et al.* (2009) 'Genetic ancestry is associated with subclinical cardiovascular disease in African-Americans and Hispanics from the multi-ethnic study of atherosclerosis', *Circulation: Cardiovascular Genetics*, 2(6), pp. 629–636. doi: 10.1161/CIRCGENETICS.109.876243.

Watanabe, K. *et al.* (2017) 'Functional mapping and annotation of genetic associations with FUMA', *Nature Communications*. Springer US, (8), pp. 1–10. doi: 10.1038/s41467-017-01261-5.

Waterland, R. A. and Michels, K. B. (2007) 'Epigenetic Epidemiology of the Developmental Origins Hypothesis', *Annual Review of Nutrition*, 27(1), pp. 363–388. doi: 10.1146/annurev.nutr.27.061406.093705.

Weber, C., Zernecke, A. and Libby, P. (2008) 'The multifaceted contributions of leukocyte subsets to atherosclerosis: lessons from mouse models', *Nature Reviews Immunology*. Nature Publishing Group, 8, p. 802. Available at: https://doi.org/10.1038/nri2415.

Whincup, P. H. *et al.* (2012) 'Ethnic differences in carotid intima-media thickness between UK children of black African-Caribbean and white European origin', *Stroke*, 43(7), pp. 1747–1754. doi: 10.1161/STROKEAHA.111.644955.

Winkel, L. C. *et al.* (2015) 'Animal models of surgically manipulated flow velocities to study shear stress-induced atherosclerosis', *Atherosclerosis*. Elsevier Ltd, 241(1), pp. 100–110. doi: 10.1016/j.atherosclerosis.2015.04.796.

Winkler, T. W., Kutalik, Z., *et al.* (2015) 'EasyStrata: Evaluation and visualization of stratified genome-wide association meta-Analysis data', *Bioinformatics*, 31(2), pp. 259–261. doi: 10.1093/bioinformatics/btu621.

Winkler, T. W., Justice, A. E., *et al.* (2015) *The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study*, *PLoS genetics*. doi: 10.1371/journal.pgen.1005378.

Winkler, T. W. *et al.* (2017) 'Approaches to detect genetic effects that differ between two strata in genome-wide meta-analyses: Recommendations based on a systematic evaluation', *PLoS ONE*, 12(7). doi: 10.1371/journal.pone.0181038.

Wojczynski, M. K. *et al.* (2013) 'Genetics of coronary artery calcification among African Americans, a meta-analysis', *BMC Medical Genetics*, 14:75. Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L52698119%5Cnhttp://www.biomedcentral.com/1471-2350/14/75%5Cnhttp://dx.doi.org/10.1186/1471-2350-14-

75%5Cnhttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=14712350&id=doi:10.1186%2F1.

Wright, A. F. and Hastie, N. D. (2001) 'Complex genetic diseases: controversy over the Croesus code', *Genome biology*, 2(8), p. COMMENT2007. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11532206%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC138948.

Wu, Y. *et al.* (2013) 'Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained', *PLoS Genetics*, 9(3). doi: 10.1371/journal.pgen.1003379.

Xiang, A. H. *et al.* (2002) 'Heritability of subclinical atherosclerosis in Latino families ascertained through a hypertensive parent', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 22(5), pp. 843–848. doi: 10.1161/01.atv.0000015329.15481.e8.

Xie, G. *et al.* (2015) 'Genome-wide association study on progression of carotid artery intima media thickness over 10 years in a Chinese cohort', *Atherosclerosis*. Elsevier Ltd, 243(1), pp. 30–37. doi: 10.1016/j.atherosclerosis.2015.08.034.

Xie, M. *et al.* (2007) 'Abstract 1949: The Protein Kinase MAP4K4 Is Activated in Failing Human Hearts and Mediates Cardiomyocyte Apoptosis in Experimental Models, in vitro and in vivo', *Circulation*. American Heart Association, 116(suppl_16), p. II_420-II_420. doi: 10.1161/circ.116.suppl_16.II_420.

Xie, T. *et al.* (2014) 'A genome-wide association study combining pathway analysis for typical sporadic amyotrophic lateral sclerosis in Chinese Han populations', *Neurobiology of Aging*. Elsevier Ltd, 35(7), pp. 1778.e9-1778.e23. doi: 10.1016/j.neurobiolaging.2014.01.014.

Yang, D. *et al.* (2015) 'Cigarette smoking and carotid plaque echodensity in the Northern Manhattan study', *Cerebrovascular Diseases*, 40(3–4), pp. 136–143. doi: 10.1159/000434761.

Yang, H. *et al.* (2004) 'Retardation of atherosclerosis by overexpression of catalase or both Cu/Zn-superoxide dismutase and catalase in mice lacking apolipoprotein E', *Circulation Research*, 95(11), pp. 1075–1081. doi: 10.1161/01.RES.0000149564.49410.0d.

Yang, J. *et al.* (2014) 'Advantages and pitfalls in the application of mixed-model association

methods', *Nature Genetics*. Nature Publishing Group, 46(2), pp. 100–106. doi: 10.1038/ng.2876.

Yang, J. *et al.* (2019) 'The emerging role of sorting nexins in cardiovascular diseases', *Clinical Science*, 133(5), pp. 723–737. doi: 10.1042/cs20190034.

Yang, Q. and Wang, Y. (2012) 'Methods for analyzing multivariate phenotypes in genetic association studies', *Journal of Probability and Statistics*, 2012. doi: 10.1155/2012/652569.

Youn, Y. J. *et al.* (2011) 'Normative values and correlates of mean common carotid intima-media thickness in the Korean rural middle-aged population: The atherosclerosis RIsk of rural areas iN Korea general population (ARIRANG) study', *Journal of Korean Medical Science*, 26(3), pp. 365–371. doi: 10.3346/jkms.2011.26.3.365.

Young, A. (2019) 'Solving the missing heritability problem using UK Biobank families Sib-Regression', *PLoS genetics*, 15(6), pp. 1–7. doi: https://doi.org/10.1371/journal.pgen.1008222.

Yu, X. and Kem, D. C. (2010) 'Proteasome inhibition during myocardial infarction', *Cardiovascular Research*, 85(2), pp. 312–320. doi: 10.1093/cvr/cvp309.

Zaina, S. *et al.* (2014) 'DNA methylation map of human atherosclerosis', *Circulation: Cardiovascular Genetics*, 7(5), pp. 692–700. doi: 10.1161/CIRCGENETICS.113.000441.

Zemlin, A. E. *et al.* (2017) 'Correlation of E-selectin concentrations with carotid intima-media thickness and cardio-metabolic profile of mixed ancestry South Africans: a cross-sectional study', *Annals of Clinical Biochemistry*, 54(1), pp. 92–100. doi: 10.1177/0004563216640001.

Zhang, X. Y. *et al.* (2014) 'Cognitive function, plasma MnSOD activity, and MnSOD Ala-9Val polymorphism in patients with schizophrenia and normal controls', *Schizophrenia Bulletin*, 40(3), pp. 592–601. doi: 10.1093/schbul/sbt045.

Zhang, Y. and Stefanovic, B. (2016) 'LARP6 meets collagen mRNA: Specific regulation of type I collagen expression', *International Journal of Molecular Sciences*, 17(3). doi: 10.3390/ijms17030419.

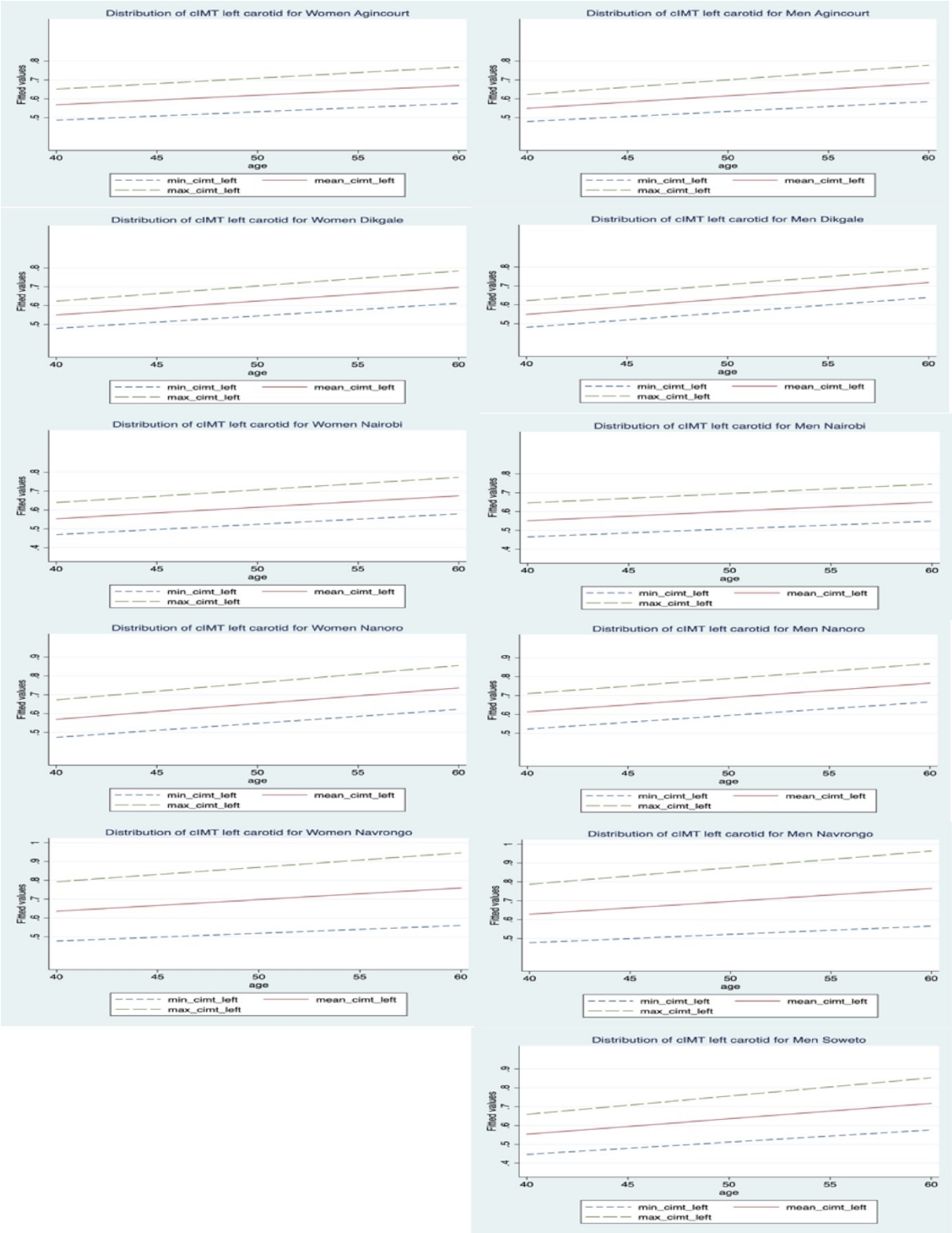
Zhou, X. and Stephens, M. (2012) 'Genome-wide efficient mixed model analysis for association studies', *Nature Genetics*, 44(7), pp. 821–824. doi: 10.1038/ng.2310.

Supplementary

SUPPLEMENTARY CHAPTER 2

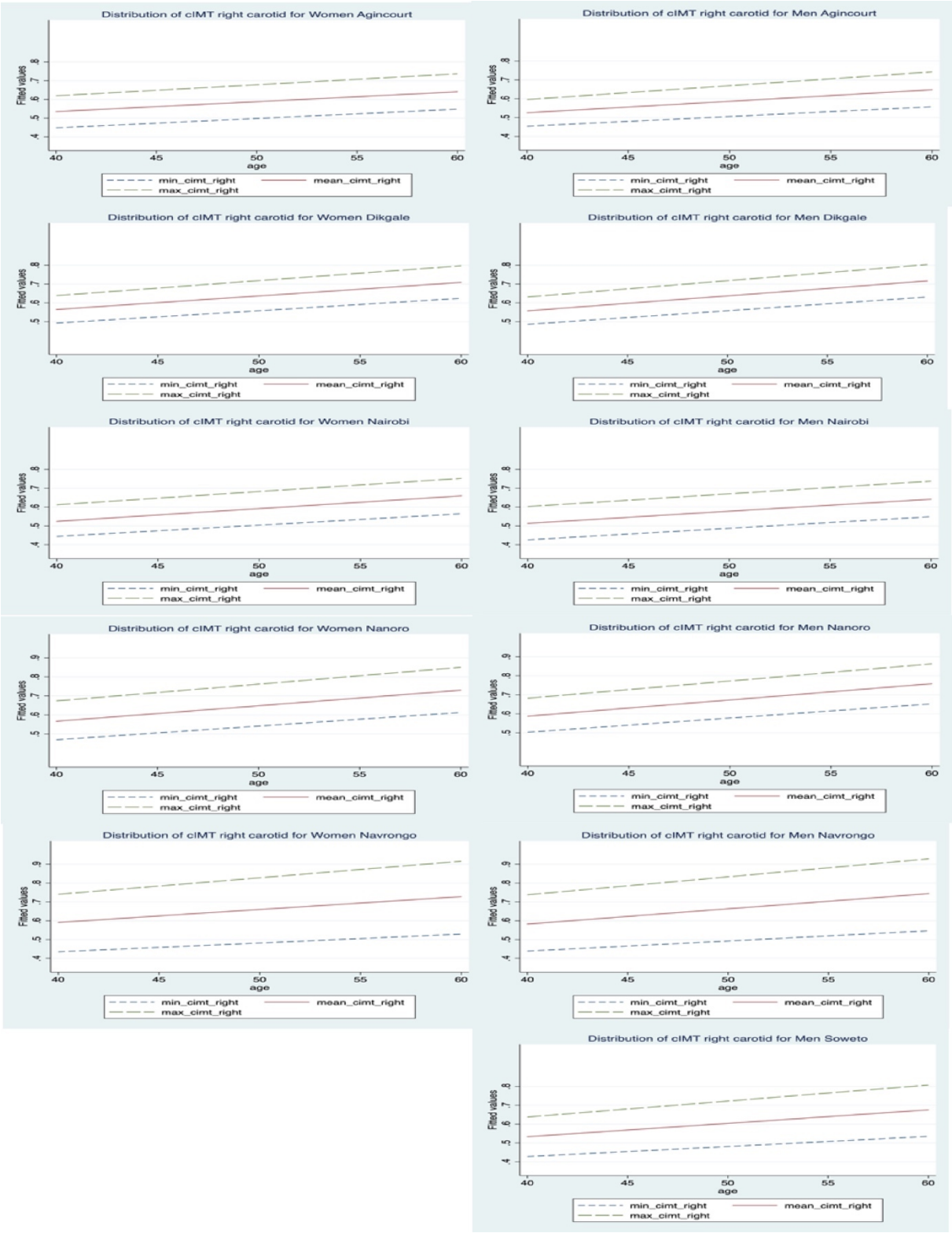
Supplementary Figure 2.1a:

Age-specific medians of cIMT measurements (minimum, mean maximum) for left carotid



Supplementary Figure 2.1b:

Age-specific medians of cIMT measurements (minimum, mean maximum) for right carotid



Supplementary Table 2.1 : Characteristics of study population by site and sex stratified by cardiovascular risk (Healthy vs non-healthy)

	Agincourt				Dikgale				Nairobi				Nanoro				Navrongo				Soweto				All males			
	Male (N=504)				Male (N=348)				Male (N=856)				Male (N=1,029)				Male (N=775)				Male (N= 892)				N=4404			
	Healthy=94	Healthy=94	non-Healthy=410	non-Healthy=411	Healthy=43		non-Healthy=305		Healthy=264		non-Healthy=592		Healthy=471		non-Healthy=558		Healthy=243		non-Healthy=532		Healthy=62		non-Healthy=830		Healthy=1177		non-Healthy=3227	
Age (years)	47	12	52	9	46,5	11	50	10	47	9	48	10	51	10	49	11	50	10	51	10	47	9	49	11	49	10	50	10
Bmi	21,6	3,12	24,37	7,49	20,09	2,86	20,64	5,98	21,9	2,9	22,71	6,5	21,02	2,74	21,13	6,72	20,93	2,69	20,27	3,55	21,94	3,12	24,25	8,03	21,25	2,88	22,01	6,74
Systolic blood pressure	118,75	15	133,5	28	112,5	17,5	123,5	25,5	114,5	17,25	122	29	115	16	120,75	28,5	117,5	16,5	124,25	31	119,5	17,5	130,5	27	115,5	16,5	126,5	29
Diastolic blood pressure	72,5	13	81	16	70,5	11	78,25	15,5	73,5	10	79,5	16,5	72,5	11,5	77	17	72,5	11,5	77,5	20	79,5	9,5	89	16	73	11,5	81,5	18
Total cholesterol	3,45	1,28	3,77	1,36	3,49	1,01	3,95	1,34	3,79	0,97	4,38	1,59	3,37	1,12	3,69	1,51	2,99	1,11	3,18	1,1	3,64	1,33	4,03	1,4	3,42	1,16	3,83	1,52
Triglycerides	0,68	0,41	0,82	0,53	0,88	0,51	0,97	0,64	0,8	0,46	0,99	0,66	0,65	0,4	0,75	0,46	0,52	0,3	0,57	0,31	0,82	0,5	0,83	0,58	0,67	0,45	0,8	0,57
Ldl (Friedwald)	1,88	0,97	2,15	1,15	2,01	0,9	2,19	1,13	2,25	0,82	2,53	1,34	1,88	0,86	2,07	1,31	1,61	0,95	1,73	0,84	2,24	0,89	2,45	1,17	1,94	0,92	2,19	1,26
Hdl	1,18	0,48	1,1	0,45	1,1	0,39	1,17	0,53	1,13	0,42	1,19	0,59	1,1	0,41	1,14	0,46	1,12	0,49	1,15	0,49	1,04	0,42	1,16	0,55	1,11	0,43	1,16	0,51
Glucose	4,67	1,04	4,78	1,19	4,42	0,63	4,7	1,05	4,96	0,71	5,07	0,9	4,87	0,82	4,95	1,01	4,38	0,97	4,41	0,85	4,89	0,68	5,07	0,83	4,78	0,89	4,91	1
Serum Creatinin	68,53	17,73	73,66	20,19	71,6	16,99	74,3	19,68	71,84	16,76	70,94	17,65	69,31	15,43	70,15	17,35	72,07	17,36	72,94	17,98	71,79	17,44	71,27	20,59	70,77	16,43	71,72	19,06
MVPA (min/week)	600	1410	720	1240	1080	2070	1210	2100	1080	2580	1350	2760	1800	2850	2160	2940	2520	1950	2340	2220	720	1680	630	1590	1545	2640	1080	2370
Visceral fat	5,68	1,84	6,46	2,83	5,37	1,98	5,94	2,65	4,67	2,05	5,14	2,52	4,23	1,48	4,38	1,82	4,03	1,45	4,13	1,48	5,04	2,07	6,31	2,83	4,46	1,82	5,27	2,66
Subcutaneous fat	0,8	0,54	1,18	0,83	0,85	0,54	0,81	0,7	0,89	0,56	1,12	0,87	0,74	0,47	0,87	0,73	0,72	0,31	0,7	0,33	1,1	0,74	1,39	1,08	0,79	0,49	0,99	0,87
	Agincourt				Dikgale				Nairobi				Nanoro				Navrongo				Soweto				All Females			
			Female (N=723)				Female (N=767)				Female (N=1,034)				Female (N=1,024)				Female (N=920)				Female (N=920)		N=4404		N=4468	
	Healthy=85		non-Healthy=638		Healthy=81		non-Healthy=686		Healthy=171		non-Healthy=863		Healthy=522		non-Healthy=502		Healthy=441		non-Healthy=479						Healthy=1300		non-Healthy=3168	
Age (years)	50	10	52	10	48	10	51	10	45	7	48	9	49	10	51	9	51	9	53	10					49	10	50	10
Bmi	22,1	3,02	29,51	8,24	21,57	2,77	30,87	9,36	22,12	2,87	27,95	7,58	20,49	2,14	18,14	4,63	21,18	2,73	22,39	7,4					21,01	2,71	26,84	10,47
Systolic blood pressure	115	19	134	27,5	112,5	15	124	28	106,5	20	115,5	27,5	106	17	111,5	27,5	113,5	20,5	128	35,5					109,5	19,5	122,5	32
Diastolic blood pressure	70	15	81	16,25	73	10,5	82,5	17	68,5	13,5	78,5	17	69	10	71,5	16	72	12	82	20					70,5	11	79	18
Total cholesterol	3,75	1,1	4,16	1,43	3,82	1,07	4,25	1,45	3,77	1,1	4,46	1,32	3,04	0,98	3,13	1,12	3,09	0,99	3,3	1,28					3,24	1,07	3,98	1,55
Triglycerides	0,65	0,31	0,8	0,47	0,8	0,43	0,98	0,61	0,78	0,45	0,95	0,55	0,65	0,36	0,68	0,37	0,52	0,24	0,57	0,34					0,62	0,35	0,82	0,55
Ldl (Friedwald)	2,22	0,76	2,58	1,22	2,13	0,95	2,57	1,14	2,14	0,9	2,68	1,15	1,66	0,76	1,72	0,88	1,79	0,76	1,94	1					1,82	0,82	2,38	1,26
Hdl	1,17	0,48	1,16	0,39	1,28	0,43	1,14	0,37	1,18	0,53	1,17	0,46	0,99	0,39	1,02	0,36	1,07	0,41	1,11	0,45					1,06	0,43	1,13	0,42
Glucose	4,51	0,77	4,73	0,91	4,43	0,8	4,86	0,94	4,95	0,81	5,09	1,01	4,88	0,71	4,92	0,91	4,46	0,75	4,61	0,84					4,71	0,83	4,86	0,97
Serum Creatinin	63,6	16,96	64,76	17,3	63,47	21,86	66,69	19,41	61,81	15,5	61,44	16,25	57,16	13,24	57,57	13,17	60,47	12,77	62,78	15,18					59,49	14,21	62,36	17
MVPA (min/week)	600	1020	615	1260	1080	1080	1305	1470	1080	2100	900	1800	2940	3060	2940	3060	1665	2340	1560	2355					1740	2820	1200	2100
Visceral fat	4,23	1,8	6,04	2,89	4,46	1,27	6,94	3,09	3,69	1,41	4,83	2,28	4,26	1,34	4,26	1,68	3,26	1,18	3,39	1,48					3,83	1,51	4,97	2,92
Subcutaneous fat	1,22	0,62	2,27	1,26	1,01	0,77	2,31	1,27	1,28	0,73	2,05	0,91	0,95	0,52	0,81	0,66	0,95	0,47	1,18	0,88					0,99	0,58	1,82	1,38

Supplementary Table 2.2:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Agincourt.

	Male					Female				
age	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
40	2	0.497	0.538	0.580	0.617	6	0.505	0.550	0.596	0.637
41	28	0.502	0.544	0.587	0.626	23	0.510	0.556	0.602	0.643
42	22	0.507	0.551	0.595	0.634	45	0.515	0.561	0.608	0.650
43	21	0.512	0.557	0.602	0.643	36	0.519	0.567	0.614	0.656
44	29	0.517	0.564	0.610	0.652	30	0.524	0.572	0.620	0.662
45	30	0.522	0.570	0.618	0.661	25	0.529	0.577	0.626	0.669
46	17	0.527	0.576	0.625	0.670	38	0.534	0.583	0.632	0.676
47	16	0.532	0.583	0.633	0.679	17	0.538	0.588	0.638	0.682
48	15	0.537	0.589	0.641	0.688	30	0.543	0.593	0.644	0.689
49	14	0.542	0.596	0.649	0.697	27	0.548	0.599	0.650	0.695
50	23	0.547	0.602	0.657	0.706	35	0.553	0.604	0.656	0.702
51	37	0.552	0.608	0.665	0.715	43	0.557	0.610	0.662	0.709
52	35	0.557	0.615	0.672	0.724	53	0.562	0.615	0.668	0.715
53	18	0.562	0.621	0.680	0.734	43	0.567	0.620	0.674	0.722
54	34	0.567	0.627	0.688	0.743	51	0.571	0.626	0.680	0.729
55	28	0.571	0.634	0.696	0.753	37	0.576	0.631	0.686	0.736
56	34	0.576	0.640	0.704	0.762	28	0.581	0.636	0.692	0.742
57	27	0.581	0.647	0.712	0.772	38	0.585	0.642	0.698	0.749
58	20	0.585	0.653	0.720	0.781	46	0.590	0.647	0.704	0.756
59	28	0.590	0.659	0.729	0.791	34	0.594	0.652	0.711	0.763
60	26	0.595	0.666	0.737	0.801	38	0.599	0.658	0.717	0.770

Supplementary Table 2.3:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Dikgale.

	Male					Female				
age	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
40	17	0.514	0.558	0.602	0.641	35	0.512	0.559	0.606	0.648
41	17	0.519	0.566	0.612	0.653	27	0.518	0.566	0.615	0.658
42	19	0.525	0.573	0.621	0.665	32	0.524	0.574	0.623	0.668
43	13	0.531	0.581	0.631	0.677	33	0.529	0.581	0.632	0.678
44	15	0.536	0.589	0.641	0.689	23	0.535	0.588	0.641	0.689
45	18	0.542	0.597	0.651	0.701	35	0.541	0.595	0.650	0.699
46	18	0.547	0.604	0.661	0.713	33	0.546	0.603	0.659	0.709
47	17	0.553	0.612	0.671	0.725	36	0.552	0.610	0.668	0.720
48	16	0.558	0.620	0.682	0.737	40	0.557	0.617	0.677	0.730
49	9	0.564	0.628	0.692	0.750	37	0.563	0.624	0.686	0.741
50	20	0.569	0.635	0.702	0.762	44	0.568	0.632	0.695	0.752
51	18	0.574	0.643	0.712	0.775	41	0.574	0.639	0.704	0.762
52	16	0.579	0.651	0.723	0.787	40	0.579	0.646	0.713	0.773
53	19	0.584	0.659	0.733	0.800	42	0.584	0.653	0.722	0.784
54	15	0.589	0.666	0.744	0.813	26	0.590	0.661	0.731	0.795
55	26	0.595	0.674	0.754	0.826	49	0.595	0.668	0.741	0.806
56	16	0.599	0.682	0.765	0.839	30	0.600	0.675	0.750	0.817
57	14	0.604	0.690	0.775	0.852	43	0.605	0.682	0.759	0.828
58	15	0.609	0.698	0.786	0.865	36	0.611	0.690	0.768	0.840
59	12	0.614	0.705	0.797	0.879	50	0.616	0.697	0.778	0.851
60	18	0.619	0.713	0.807	0.892	35	0.621	0.704	0.787	0.862

Supplementary Table 2.4:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Nairobi.

	Male					Female				
age	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
40	12	0.484	0.536	0.589	0.636	20	0.491	0.543	0.594	0.641
41	40	0.488	0.542	0.595	0.643	60	0.496	0.549	0.602	0.649
42	71	0.493	0.547	0.602	0.651	89	0.501	0.555	0.609	0.658
43	64	0.497	0.553	0.608	0.658	70	0.506	0.561	0.616	0.666
44	53	0.501	0.558	0.614	0.665	61	0.511	0.567	0.624	0.675
45	67	0.506	0.563	0.621	0.673	87	0.515	0.573	0.631	0.684
46	60	0.510	0.569	0.627	0.680	70	0.520	0.579	0.639	0.692
47	43	0.514	0.574	0.634	0.688	61	0.525	0.586	0.646	0.701
48	45	0.518	0.579	0.641	0.696	58	0.530	0.592	0.654	0.710
49	33	0.522	0.585	0.647	0.703	59	0.534	0.598	0.662	0.719
50	33	0.526	0.590	0.654	0.711	55	0.539	0.604	0.669	0.728
51	45	0.530	0.595	0.660	0.719	44	0.543	0.610	0.677	0.737
52	46	0.534	0.601	0.667	0.727	56	0.548	0.616	0.684	0.746
53	45	0.538	0.606	0.674	0.735	51	0.552	0.622	0.692	0.755
54	30	0.542	0.612	0.681	0.743	32	0.557	0.628	0.700	0.764
55	36	0.546	0.617	0.687	0.751	43	0.561	0.635	0.708	0.774
56	28	0.550	0.622	0.694	0.759	29	0.566	0.641	0.716	0.783
57	33	0.554	0.628	0.701	0.767	27	0.570	0.647	0.723	0.792
58	23	0.558	0.633	0.708	0.775	24	0.575	0.653	0.731	0.802
59	21	0.562	0.638	0.715	0.783	19	0.579	0.659	0.739	0.811
60	28	0.566	0.644	0.721	0.791	19	0.583	0.665	0.747	0.821

Supplementary Table 2.5:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Nanoro.

	Male					Female				
age	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
40	40	0.538	0.605	0.672	0.733	21	0.522	0.575	0.629	0.677
41	53	0.545	0.613	0.681	0.742	54	0.528	0.583	0.638	0.687
42	61	0.552	0.621	0.689	0.751	53	0.535	0.591	0.647	0.697
43	63	0.560	0.629	0.698	0.760	54	0.542	0.599	0.656	0.708
44	41	0.567	0.637	0.706	0.769	57	0.548	0.607	0.665	0.718
45	38	0.574	0.644	0.715	0.778	52	0.555	0.615	0.674	0.728
46	45	0.581	0.652	0.723	0.787	45	0.561	0.622	0.684	0.739
47	49	0.588	0.660	0.732	0.797	59	0.567	0.630	0.693	0.749
48	65	0.596	0.668	0.741	0.806	45	0.574	0.638	0.702	0.760
49	48	0.603	0.676	0.749	0.815	51	0.580	0.646	0.711	0.770
50	54	0.610	0.684	0.758	0.824	54	0.587	0.654	0.721	0.781
51	54	0.617	0.692	0.766	0.834	54	0.593	0.661	0.730	0.792
52	51	0.624	0.699	0.775	0.843	65	0.599	0.669	0.739	0.802
53	47	0.631	0.707	0.784	0.852	52	0.605	0.677	0.749	0.813
54	49	0.638	0.715	0.792	0.862	59	0.612	0.685	0.758	0.824
55	46	0.645	0.723	0.801	0.871	45	0.618	0.693	0.768	0.835
56	45	0.652	0.731	0.810	0.881	53	0.624	0.701	0.777	0.846
57	53	0.659	0.739	0.818	0.890	45	0.630	0.708	0.787	0.857
58	31	0.666	0.747	0.827	0.900	43	0.636	0.716	0.796	0.868
59	44	0.673	0.754	0.836	0.909	43	0.642	0.724	0.806	0.879
60	52	0.680	0.762	0.845	0.919	20	0.648	0.732	0.815	0.891

Supplementary Table 2.6:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Navrongo.

	Male					Female				
age	Sample size	25th	50th	75th	90th	Sample size	25th	50th	75th	90th
40	6	0.550	0.609	0.668	0.721	8	0.549	0.611	0.674	0.730
41	42	0.556	0.616	0.676	0.730	34	0.554	0.618	0.681	0.738
42	31	0.562	0.623	0.684	0.740	33	0.560	0.624	0.688	0.746
43	31	0.568	0.630	0.693	0.749	33	0.565	0.630	0.695	0.754
44	29	0.574	0.638	0.701	0.759	22	0.570	0.637	0.703	0.762
45	37	0.580	0.645	0.710	0.768	48	0.576	0.643	0.710	0.770
46	47	0.586	0.652	0.718	0.778	51	0.581	0.649	0.717	0.779
47	29	0.592	0.659	0.727	0.787	48	0.586	0.656	0.725	0.787
48	46	0.597	0.666	0.735	0.797	32	0.592	0.662	0.732	0.795
49	41	0.603	0.673	0.744	0.807	44	0.597	0.668	0.739	0.803
50	45	0.609	0.681	0.752	0.817	37	0.602	0.675	0.747	0.812
51	39	0.615	0.688	0.761	0.826	45	0.607	0.681	0.754	0.820
52	46	0.621	0.695	0.769	0.836	50	0.613	0.687	0.762	0.829
53	42	0.626	0.702	0.778	0.846	39	0.618	0.694	0.769	0.837
54	17	0.632	0.709	0.787	0.856	35	0.623	0.700	0.777	0.846
55	52	0.638	0.717	0.795	0.866	90	0.628	0.706	0.784	0.854
56	53	0.643	0.724	0.804	0.877	63	0.633	0.713	0.792	0.863
57	35	0.649	0.731	0.813	0.887	52	0.638	0.719	0.799	0.872
58	37	0.654	0.738	0.822	0.897	31	0.643	0.725	0.807	0.881
59	31	0.660	0.745	0.831	0.907	57	0.649	0.732	0.815	0.889
60	39	0.666	0.752	0.839	0.918	68	0.654	0.738	0.822	0.898

Supplementary Table 2.7:

Age- and sex-specific percentiles of References intervals for carotid artery intima-media thickness (in mm) in the general population for Soweto.

	Male				
age	Sample size	25th	50th	75th	90th
40	59	0.496	0.544	0.592	0.634
41	44	0.502	0.551	0.601	0.645
42	53	0.508	0.559	0.610	0.656
43	32	0.513	0.566	0.619	0.666
44	52	0.519	0.573	0.628	0.677
45	43	0.524	0.581	0.637	0.688
46	39	0.530	0.588	0.646	0.698
47	53	0.536	0.595	0.655	0.709
48	45	0.541	0.603	0.665	0.720
49	57	0.546	0.610	0.674	0.731
50	31	0.552	0.618	0.683	0.743
51	40	0.557	0.625	0.693	0.754
52	45	0.563	0.632	0.702	0.765
53	41	0.568	0.640	0.712	0.777
54	36	0.573	0.647	0.721	0.788
55	35	0.578	0.655	0.731	0.800
56	34	0.583	0.662	0.740	0.811
57	42	0.589	0.669	0.750	0.823
58	46	0.594	0.677	0.760	0.835
59	35	0.599	0.684	0.770	0.846
60	30	0.604	0.691	0.779	0.858

SUPPLEMENTARY CHAPTER 3

Supplementary Table 3.1:

Summary statistics of selected variants (p≤1E-04) for the combined set (All), female-specific (Female), male-specific (Male) and sex-difference (SexDiff) with genes and traits mapped and references from GWAS Catalog

Due to the size of the file Supplematry Table 3.1 is in digital excel file.

Link:

https://drive.google.com/file/d/14AFeTksfXqGOawsIXYee95oeDGYS_36e/view?usp=sharing

Supplementary Table 3.2:

Summary statistics of selected variants (p≤1E-04) for the combined set (All), female-specific (Female), male-specific (Male) and sex-difference (SexDiff) with genes and traits mapped and references of all GWAS Catalog within 25 kb of the variants

Due to the size of the file Supplematry Table 3.1 is in digital excel file.

Link:

https://drive.google.com/file/d/14AFeTksfXqGOawsIXYee95oeDGYS_36e/view?usp=sharing

Supplementary Table 3.3:

GWAS Catalog look-up of selected index variants (p≤1E-04) displaying loci coordinates and number of position in the locus for the combined set (All), female-specific (Female), male-specific (Male) and sex-difference (SexDiff) with genes mapped and references of all GWAS Catalog within 25 kb of the variants

Due to the size of the file Supplematry Table 3.1 is in digital excel file.

Link:

https://drive.google.com/file/d/14AFeTksfXqGOawsIXYee95oeDGYS_36e/view?usp=sharing

Supplementary Table 3.4a:

Functional annotation for selected SNPs associated (p≤1E-05) with cIMT in the combined analysis

Note: ANNOVAR = functional variant classification based on position in or outside of a gene; CADD = Combined Annotation-Dependent depletion score, a marker of deleteriousness of SNP's effect on a protein product (higher scores are more deleterious); RDB = RegulomeDB scores, which predict likelihood that a SNP has a regulatory role (lower scores are more likely); minChrState = minimum chromatin state across 127 tissue types (lower scores indicate more open chromatin); commonChrState = most common chromatin state in 127 tissue types; Not Annotated = SNP not available in database. LD r2 are based on African from 100GP.

Due to the size of this filem Supplementary Table is in digital excel file.

https://drive.google.com/file/d/1ksoTO2uGTIGscIOAj5t-Moy63VGE_21X/view?usp=sharing

Supplementary Table 3.4b:

Functional annotation for selected SNPs associated (p≤1E-05) with cIMT in the female-specific analysis

Note: ANNOVAR = functional variant classification based on position in or outside of a gene; CADD = Combined Annotation-Dependent depletion score, a marker of deleteriousness of SNP's effect on a protein product (higher scores are more deleterious); RDB = RegulomeDB scores, which predict likelihood that a SNP has a regulatory role (lower scores are more likely); minChrState = minimum chromatin state across 127 tissue types (lower scores indicate more open chromatin); commonChrState = most common chromatin state in 127 tissue types; Not Annotated = SNP not available in database. LD r2 are based on African from 100GP.

Due to the size of this filem Supplementary Table is in digital excel file.

https://drive.google.com/file/d/1hDm7nAN03surYuhEeukzzjOItgGL7H2g/view?usp=sharing

Supplementary Table 3.4c:

Functional annotation for selected SNPs associated (p≤1E-05) with cIMT in the male-specific analysis

Note: ANNOVAR = functional variant classification based on position in or outside of a gene; CADD = Combined Annotation-Dependent depletion score, a marker of deleteriousness of SNP's effect on a protein product (higher scores are more deleterious); RDB = RegulomeDB scores, which predict likelihood that a SNP has a regulatory role (lower scores are more likely); minChrState = minimum chromatin state across 127 tissue types (lower scores indicate more open chromatin); commonChrState = most common chromatin state in 127 tissue types; Not Annotated = SNP not available in database. LD r2 are based on African from 100GP.

Due to the size of this filem Supplementary Table is in digital excel file.

https://drive.google.com/file/d/1rOMh_C3TSPaiUTHAFTZqhYXAi461oskN/view?usp=sharing

Supplementary Table 3.5a:

Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with cIMT in the combined analysis

Note: SNPs were identified by a genome-wide suggestive (P<1e-5) association in a sample-size weighted meta-analysis, grouped into LD-independent genomic loci, and mapped to genes using positional mapping, eQTL mapping, and chromatin interaction strategies. Loci for genes mapped by two genomic risk loci are listed as locus1:locus2. Blue highlighted rows represent low confidence loci; see Supplementary Results 2.3.1. pLI = probability of loss of function mutation intolerance (higher scores indicate higher probability); posMap = SNPs mapped to this gene by positional mapping; CADD = Combined Annotation-Dependent depletion score (higher scores indicate more deleterious effects of a SNP); eqtlMap = SNPs mapped to this gene by expression quantitative traic locus mapping; minP = minimum P value of eQTL association for the mapped gene; minQ = minimum false discovery rate Q value of eQTL association for the mapped gene; Tissues = tissues in which the significant eQTL association or chromatin interaction was observed; Direction = direction of association of eQTL SNP with expression levels (upregulation or downregulation); ci = chromatin interaction mapping; minGwasP =

minimum GWAS P value of SNP(s) implicating the gene; IndSigSNPs = SNPs with significant GWAS p-values from which the gene was mapped.

https://drive.google.com/file/d/1HKLjYKTSm8Zw7I5imIE31xJDE-uIQ9nK/view?usp=sharing

Supplementary Table 3.5b:

Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with cIMT in the female-specific analysis

Note: SNPs were identified by a genome-wide suggestive (P<1e-5) association in a sample-size weighted meta-analysis, grouped into LD-independent genomic loci, and mapped to genes using positional mapping, eQTL mapping, and chromatin interaction strategies. Loci for genes mapped by two genomic risk loci are listed as locus1:locus2. Blue highlighted rows represent low confidence loci; see Supplementary Results 2.3.1. pLI = probability of loss of function mutation intolerance (higher scores indicate higher probability); posMap = SNPs mapped to this gene by positional mapping; CADD = Combined Annotation-Dependent depletion score (higher scores indicate more deleterious effects of a SNP); eqtlMap = SNPs mapped to this gene by expression quantitative traic locus mapping; minP = minimum P value of eQTL association for the mapped gene; minQ = minimum false discovery rate Q value of eQTL association for the mapped gene; Tissues = tissues in which the significant eQTL association or chromatin interaction was observed; Direction = direction of association of eQTL SNP with expression levels (upregulation or downregulation); ci = chromatin interaction mapping; minGwasP = minimum GWAS P value of SNP(s) implicating the gene; IndSigSNPs = SNPs with significant GWAS p-values from which the gene was mapped.

https://drive.google.com/file/d/10Ghzf7EGpFFJooXhVabuBhYq407RziA8/view?usp=sharing

Supplementary Table 3.5c:

Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with cIMT in the male-specific analysis

Note: SNPs were identified by a genome-wide suggestive (P<1e-5) association in a sample-size weighted meta-analysis, grouped into LD-independent genomic loci, and mapped to genes using positional mapping, eQTL mapping, and chromatin interaction strategies. Loci for genes

mapped by two genomic risk loci are listed as locus1:locus2. Blue highlighted rows represent low confidence loci; see Supplementary Results 2.3.1. pLI = probability of loss of function mutation intolerance (higher scores indicate higher probability); posMap = SNPs mapped to this gene by positional mapping; CADD = Combined Annotation-Dependent depletion score (higher scores indicate more deleterious effects of a SNP); eqtlMap = SNPs mapped to this gene by expression quantitative traic locus mapping; minP = minimum P value of eQTL association for the mapped gene; minQ = minimum false discovery rate Q value of eQTL association for the mapped gene; Tissues = tissues in which the significant eQTL association or chromatin interaction was observed; Direction = direction of association of eQTL SNP with expression levels (upregulation or downregulation); ci = chromatin interaction mapping; minGwasP = minimum GWAS P value of SNP(s) implicating the gene; IndSigSNPs = SNPs with significant GWAS p-values from which the gene was mapped.

https://drive.google.com/file/d/1Ti-3YRKA03Iy0EmYTrQTJsgOON5mIiRu/view?usp=sharing

Supplementary Table 3.6a:

Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive SNPs (p<1E-05) with cIMT in the combined analysis.

Note: Hypergeometric tests of enrichment of gene-sets was implemented in FUMA and two-tailed P-values are reported. These tests do not include genes implicated solely by GWGAS. All curated gene-sets were obtained from MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp). Set names link to an online description. N_set = number of genes in curated set; n_overlap = number of genes mapped by GWAS SNPs that overlap with the gene-set; genes = overlapping genes.

https://drive.google.com/file/d/1qYLXYcSmtthA98YRHBfmuqA4jjDSpgYs/view?usp=sharing

Supplementary Table 3.6a:

Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive SNPs (p<1E-05) with cIMT in the female-specific analysis.

Note: Hypergeometric tests of enrichment of gene-sets was implemented in FUMA and two-tailed P-values are reported. These tests do not include genes implicated solely by GWGAS. All curated gene-sets were obtained from MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp). Set names link to an online description. N_set = number of genes in curated set; n_overlap = number of genes mapped by GWAS SNPs that overlap with the gene-set; genes = overlapping genes.

https://drive.google.com/file/d/1kG0GvpzN7HvuMUu1dEZF1uB12hFnYQed/view?usp=sharing

Supplementary Table 3.6a:

Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive SNPs (p<1E-05) with cIMT in the female-specific analysis.

Note: Hypergeometric tests of enrichment of gene-sets was implemented in FUMA and two-tailed P-values are reported. These tests do not include genes implicated solely by GWGAS. All curated gene-sets were obtained from MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp). Set names link to an online description. N_set = number of genes in curated set; n_overlap = number of genes mapped by GWAS SNPs that overlap with the gene-set; genes = overlapping genes.

https://drive.google.com/file/d/1sG38S2lyRuFdBcGszURLXuOgDaf9AIPr/view?usp=sharing

SUPPLEMENTARY CHAPTER 4:

(All files located in the same spreadsheet)

https://drive.google.com/file/d/1z9SQaAB828jJY1QFpj3iUAk48fwUhmlM/view?usp=sharing

Supplementary Table 4.1. Summary of suggestive and significant SNPs with betas in non-smokers and smokers group.

Supplementary Table 4.2a. Functional annotation for suggestive SNPs associated with Gene-Smoking interaction for cIMT in Nanoro sample

Supplementary Table 4.2b. Functional annotation for suggestive SNPs associated with Gene-Smoking interaction for cIMT in Navrongo sample

Supplementary Table 4.2c. Functional annotation for suggestive SNPs associated with Gene-Smoking interaction for cIMT in the combined sample

Supplementary Table 4.3a. Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with Gene-Smoking interaction for cIMT in Nanoro sample

Supplementary Table 4.3b. Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with Gene-Smoking interaction for cIMT in Navrongo sample

Supplementary Table 4.3c. Genes implicated by positional, eQTL, or chromatin interaction mapping of SNPs associated with Gene-Smoking interaction for cIMT in combined sample

Supplementary Table 4.4a. Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive GWAS SNPs ($p<1E-05$) in Nanoro sample.

Supplementary Table 4.4b. Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive GWAS SNPs ($p<1E-05$) in Navrongo sample.

Supplementary Table 4.4c. Gene-sets with significant enrichment for genes associated with gene-smoking interaction for cIMT based on positional, eQTL, or chromatin interaction of suggestive GWAS SNPs (p<1E-05) in Combined sample.

Supplementary Table 4.5: Summary of GWAS Catalog lookup for selected SNPs (P-values<1E-04)

Supplementary Figure 4.1: Regional association plots of selected loci (p-values) in Nanoro (a). Regional association plots of selected loci (p-values) in Navrongo (b). Regional association plots of selected loci (p-values) in the combined sample (c)

https://drive.google.com/file/d/1NGYUCjG4tjIvPY_2Ky0km0K-PGy99q-v/view?usp=sharing

Supplementary Figure 4.2: Genotypes plots of selected SNPs (p-values<1E-05) in Nanoro showing distibutions of mean cIMT residuals in smokers and non-smokers groups. Mean_cIMT_Res are in mm (a). Genotypes plots of selected SNPs (p-values<1E-05) in Navrongo showing distibutions of mean cIMT residuals in smokers and non-smokers groups. Mean_cIMT_Res are in mm (b). Genotypes plots of selected SNPs (p-values<1E-05) in the combined sample showing distibutions of mean cIMT residuals in smokers and non-smokers groups. Mean_cIMT_Res are in mm (c).

https://drive.google.com/file/d/1m2w782SV7Q2xCzpBoFRhE5gD5CaBOejQ/view?usp=sharing

Supplementary Figure 4.3 : Circos plots showing genes on chromosomes that were linked to risk (P<1E-05) loci in the GWAS of Nanoro sample (a). Circos plots showing genes on chromosomes that were linked to risk (P<1E-05) loci in the GWAS of Navrongo sample (b). Circos plots showing genes on chromosomes that were linked to risk (P<1E-05) loci in the GWAS of the combined sample (c).

https://drive.google.com/file/d/1pyv_BCD5cA4ulO-nfuGmJLkiJp5Zw24O/view?usp=sharing

Supplementary Figure 4.4: Data for expression heatmap of selected gene expression data sets. Nanoro (a). Navrongo (b). Combined sample (c).

https://drive.google.com/file/d/1LmOeTElZmBkcKq_cVRg4C-Y5Il9K3-8i/view?usp=sharing