Epigenetic analysis of cardio-metabolic health in an African population

HT Cronjé

orcid.org/0000-0001-6855-8324

Thesis submitted in fulfilment of the requirements for the degree Doctor of Philosophy in Nutrition at the North-West University

Promoter:	Prof Marlien Pieters
Co-promoter:	Prof Cornelie Nienaber-Rousseau
Co-promoter:	Dr Hannah Rebecca Elliott

Graduation: July 2020

Student number: 23520825

Douw and Carine Cronjé, this one is for you.
Thank you.

And I said to the man who stood at the gate of the year:
“Give me a light that I may tread safely into the unknown.”
And he replied:
“Go out into the darkness and put your hand into the Hand of God.
That shall be to you better than light and safer than a known way.”
So I went forth, and finding the Hand of God, trod gladly into the night.
And He led me towards the hills and the breaking of day in the lone East.
So heart be still:
What need our little life
Our human life to know,
If God hath comprehension?
In all the dizzy strife
Of things both high and low,
God hideth His intention.

God knows. His will
Is best. The stretch of years
Which wind ahead, so dim
To our imperfect vision,
Are clear to God. Our fears
Are premature; In Him,
All time hath full provision.

Then rest: until
God moves to lift the veil
From our impatient eyes,
When, as the sweeter features
Of Life’s stern face we hail,
Fair beyond all surmise
God’s thought around His creatures
Our mind shall fill.

The Gate of the Year – Minnie Louise Haskins

ACKNOWLEDGEMENTS

The past 40 months have been the greatest of my life, so far, and will forever be remembered as a highlight of my life. These past years I had the privilege of getting to know the most incredible people from all across the globe. I had the privilege of loving them and being loved by them, making memories with them and learning from each and every one. These past years I've failed more times than I can count and I have achieved more than I ever thought I could. These 40 months. This project. This thesis. These people. Looking back, I see a myriad of people to whom I will be eternally grateful. I would like this opportunity to thank all of you in my well known 'concise' writing style.

First and foremost, **Prof Marlien Pieters**, my principal investigator, promotor and mentor. Thank you for creating an environment that allows your students to excel. The past five years you have gone above and beyond even the most generous expectations one can have of a supervisor, in a constant effort to teach me, help me, open doors for me and protect me. Thank you for the incredible opportunities you have given me. Much of the first paragraph of these acknowledgements stem from your generosity and willingness to invest in me as much as you did.

Secondly, **Prof Cornelie Nienaber-Rousseau**, my co-promotor. I will never forget the day you convinced me to pursue further study (in genetics no less) under your guidance, because you were sure there was potential in this undergraduate student. That was day one. More than five years later you are still consistently cheering me on as you see the seed you planted finally break the ground. Any career I have will be built on that seed. Thank you!

Lastly, **Dr Hannah Elliott**, my co-promotor and teacher. Hannah, thank you for joining our team – you were essential in the completion of this project. Thank you for taking the time to teach me. Thank you for patiently supporting me in the many months of bioinformatic trial and error. Thank you for hosting me in Bristol, twice, and for making me feel welcome and safe in an intimidating world.

Thank you to the research team for the ownership you gave me. I am acutely aware that I sometimes spent months on things you may have been able to do in days – thank you for patiently supporting me from the sidelines and letting me learn.

Outside of the research team there were a multitude of co-workers who made this endeavour possible. **Dr Fiona R. Green**, thank you for your valuable contribution in acquiring funding for this project and for your co-authorship on two of my PhD manuscripts. **Dr Lizelle Zandberg**,

thank you for taking the time to teach me everything I know about molecular laboratory work. Thank you for consistently challenging me to understand every step instead of simply following protocols. I owe thanks to you for instilling confidence in me. **Prof Alta Schutte and Dr Josine Min,** thank you for your valuable input as co-authors of one of my manuscripts. Sincerest thanks also to the **Centre of Excellence for Nutrition**, particularly my director, **Prof Marius Smuts**, and the brilliant **Mrs Ronel Benson. Mrs Henriëtte Claassen,** thank you for all your help with the Newton Fund bookkeeping. Thank you for patiently helping me with so many purchase orders and invoices and for all your help with making my travels possible. Special thanks also to the **National Research Foundation of South Africa** and the **North-West University** that generously enabled me to pursue my studies full-time and granted me the opportunity to travel and attend courses and conferences.

I would also like to use this opportunity to pay tribute to the behind-the-scenes tribe that carried me to the finish line. The greatest debt is undoubtedly to **my parents**. Thank you for creating the environment foundational to this PhD. This PhD was a non-starter without you. Your constant, selfless investment of immeasurable support, love and resources made this possible. **Pierre Cronje**, definitely my biggest fan! Thank you for weekly conversations and for your continued attempts to understand what I do. Thank you for printing manuscripts and showing them off to anyone who would want to see them! **Pierre, Mariska, Johan, Kirstin, Leandi, Jaaco and Sellies**, thank you for extending grace and patience to me the many times I was preoccupied, absent from events, took longer to visit than I should have and was absent-minded. Thank you for your continued love and encouragement. **Erna and Phillip Kemp**. You are the grace given to me before I knew I needed it. Thank you for being my second home and for the many times you carried me through. **Mariechen**. Thank you for being the wise, brilliant, funny and beautiful best friend that you are. Thank you for Italy and for seeing the *signs* with me. **Jonathan**. You were the toughest non-academic challenge of this whole journey. I have learnt, and keep learning, so much from you. The passion you and **Stephie** have for the research you do and the way you pour yourself into every aspect of it and meet every challenge with a level of resilience I can't really fathom, is incredible. I learnt something about confidence and determination from you and mostly that quiet, restful, 'wasted' moments are the prep-work for a wave on its way. In your own words, Jon: "*The wave always comes to an end, but what doesn't is finding myself sitting on the backline waiting for the next wave to pick me up ...The beautiful part is that during this down time, you are unconsciously preparing yourself for the next wave."* **Yannick**. Cambridge, London, Bristol, Copenhagen, the Netherlands, South Africa. What an adventure life has been with you the past two years. Thank you for choosing depth over distance and for being my PhD partner in crime. Meeting you was my ultimate highlight. Thank you for your encouragement, help and comfort.

ABSTRACT

INTRODUCTION AND AIM

Eighty-five percent of the 41 million annual non-communicable disease (NCD) mortalities worldwide occur in low- and middle-income countries (LMICs). A large proportion of these deaths are caused by cardio-metabolic diseases (CMDs). The prevalence of CMDs continues to increase in part owing to the rapid urbanisation experienced by these countries. Evidence has shown that epigenetic mechanisms, such as DNA methylation (DNAm), associate with CMDs and CMD risk factors. These mechanisms potentially mediate the relationship between genetic/environmental exposure (such as the behaviour and lifestyle changes related to urbanisation) and disease. Valuable insights have so far come from investigations of DNAm in the context of CMD through epigenome-wide association analyses (EWASs), white blood cell count (WBC) ratios and DNAm clocks. Although these investigations could be of great benefit to CMD prevention and treatment in LMICs, thus far data have largely been collected in individuals from European descent, mostly living in urbanised, high-income countries. Data on populations from different ancestries, living in LMICs, including continental Africans, are scarce. Because there are known genetic and epigenetic differences between ancestral groups, the generalisability of the current epigenetic literature, mostly resulting from European cohorts, to understudied African populations is unknown. This thesis reports the first investigation into the relationship between DNAm and cardio-metabolic health in black South Africans. First, the urban-rural divide, as is experienced in developing countries such as South Africa, was described as an epidemiological approach to investigate the role of DNAm in the association between urbanisation and NCD risk, in the form of a review. This formed part of the literature required to understand and interpret the experimental data. Empirically, DNAm was investigated using EWASs and analysis of methylation-derived WBC ratios and DNAm clocks, in relation to a range of CMD-related phenotypes including chronological and biological age, alcohol consumption, smoking status, body composition, biochemical indicators of metabolic health and inflammation, as well as markers of cardiovascular function (CVF) and risk.

METHODS

A sub-sample of 120 apparently healthy Batswana men, aged 45 to 88 years, who participated in the 2015 arm of the Prospective Urban and Rural Epidemiology study in the North West province of South Africa (PURE-SA-NW) were investigated. Genome-wide DNAm data were generated from whole-blood DNA using the Illumina® Infinium HumanMethylationEPIC bead chip (EPIC array). Multiple CMD-related EWASs were performed and compared to previously published EWASs conducted in different ethnicities, to evaluate the reproducibility of current literature and

to contribute novel findings from the PURE-SA-NW cohort. Next, methylation-derived WBC ratios were investigated and compared to protein-based inflammatory markers in their associations with CVF markers and their literature-based portrayal of CVD risk. Lastly, DNAm ages were estimated using five widely used DNAm clocks. Age estimates from the Horvath, Hannum and skin and blood clocks were compared in terms of their accuracy of chronological age estimation and those from PhenoAge and GrimAge clocks were compared according to their ability to characterise biophysiological decline.

RESULTS

Up to 86% of previously identified epigenome-wide associations overlapped with the findings from the PURE-SA-NW study, and a further 13% were directionally consistent. Only 1% of the replicated associations presented with effects opposite to findings in other ancestral groups and were largely explained by population-specific genomic variance. Nineteen novel CpG associations with alcohol consumption (11 EPIC probes and eight 450K probes also present on the EPIC array) and one with high-density lipoprotein (450K probe) were observed. The WBC ratio estimates of the PURE-SA-NW group were comparable to previously investigated ostensibly healthy ethnic groups. The CVD risk portrayed by these markers was also similar to that of conventionally used risk markers, including C-reactive protein. The methylation-derived WBC ratio indicators performed better than the protein-based inflammatory markers when disentangling variance in CVF. Optimal clarification of CVF variance was obtained when the methylation-derived and protein-based markers were used in tandem. The skin and blood clock had a stronger correlation with chronological age and less variation in age acceleration compared to the Horvath and Hannum clocks. All three of these clocks, however, tended to underestimate the chronological age of the cohort. This underestimation was increasingly pronounced with older chronological age. GrimAge provided superior characterisation of biophysiological decline compared to the PhenoAge estimate, partly because of its incorporation of smoking-related effects, which were not encapsulated by the PhenoAge estimate or any of its constituents. This was of particular importance in this study population, given that more than half of them were current smokers.

CONCLUSION

This thesis demonstrates that the methylation associations observed in this black South African population are largely in agreement with the epigenetic data published on other ethnicities, with some differences related to genomic variance, highlighting the need for population-specific data. The enhanced coverage of the EPIC array proved useful in expanding the current epigenetic literature. Methylation-derived WBC ratio markers provided additional value to conventionally

used inflammatory markers in the elucidation of the role of inflammation in CVF, even in population-based research without overt inflammatory diseases. The DNAm clocks require further optimisation for their use in older populations, as was observed in their systematic underestimation of biological age in the PURE-SA-NW data. The fact that the GrimAge incorporates, for the first time, lifestyle-related exposure, such as smoking, seemed to add to its accuracy in characterising biophysiological decline. Empirically, this thesis shows that investigations of diverse populations are valuable and can reveal new associations. The critical narrative literature review highlights the need for epidemiological studies of DNAm across urban-rural divides where suitable data sets exist. Future studies can replicate the data reported here and further investigate causal pathways and utility in disease prediction.

KEYWORDS: cardiovascular disease, cardiovascular function, DNA methylation, epigenetic clocks, EWAS, inflammation, LMICs, methylation age, PURE, urbanisation

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

450K array	Illumina® Infinium HumanMethylation450K bead chip
$5_{ca}C$	5-carboxylcytosine
$5_{f}C$	5-formylcytosine
$5_{hm}C$	5-hydroxymethylcytosine
AA	African American (Chapter 4)
AA	Age acceleration (Chapter 6)
bDBP	Brachial diastolic blood pressure
BIOS	Biobank-based Integrative Omics Studies
BMI	Body mass index
bSBP	Brachial systolic blood pressure
bPP	Brachial pulse pressure
CAD	Coronary artery disease
CMD	Cardio-metabolic disease
cfPWV	Carotid-femoral pulse wave velocity
Chr	Chromosome
CpGs	Cytosine-phosphate-guanine sites
CRP	C-reactive protein
CVD	Cardiovascular disease
CVF	Cardiovascular function
DNA	Deoxyribonucleic acid
DNAm	DNA methylation

DNAmAge	DNA methylation age
DNAmAgeAccel	DNA methylation age acceleration
DNMTs	DNA methyltransferases
EA	European American
EEAA	Extrinsic epigenetic age acceleration
EPIC array	Illumina® Infinium HumanMethylationEPIC bead chip
EU	European
EWAS	Epigenome-wide association study
eQTL	Expression quantitative trait locus
HDL-C	High-density lipoprotein cholesterol
HR	Heart rate
HT	Hypertension
IA	Indian Asian
IEAA	Intrinsic epigenetic age acceleration
IFN-γ	Interferon gamma
IL	Interleukin
IQR	Inter-quartile range
LD	Linkage disequilibrium
LDL-C	Low-density lipoprotein cholesterol
LMR	Lymphocyte-to-monocyte-ratio
MAF	Minor allele frequency
MAP	Mean arterial pressure
MBD	Methyl-CpG-binding domain proteins

md	Methylation-derived
mQTL	Methylation quantitative trait locus
NCD	Non-communicable disease
NLR	Neutrophil-to-lymphocyte-ratio
PAI1	Plasminogen activation inhibitor 1
PURE-SA-NW	South Africa, North-West arm of the Prospective Urban and Rural Epidemiology study
QC	Quality control
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SB	Skin and blood
SNP	Single nucleotide polymorphism
TC	Total cholesterol
TET	Ten-eleven translocation enzyme
TG	Triglycerides
TNF-α	Tumour necrosis factor alpha
UHRF1	Ubiquitin-like plant homeodomain and RING finger domain 1
UTR	Untranslated region
WBC	White blood cell
WHO	World Health Organization
wPAI	Weighted physical activity index

CHAPTER 1

INTRODUCTION

INTRODUCTION

1.1 Background and problem statement

Non-communicable diseases (NCDs) are responsible for 71% of all deaths globally (World Health Organization [WHO], 2018a). A large proportion of these deaths are caused by cardio-metabolic diseases (CMDs), such as cardiovascular diseases (CVDs) and diabetes, which account for 44% and 4% of all NCD deaths, respectively (WHO, 2018a). Of all mortalities registered in South Africa in 2016, 12.3% were the result of CVD and another 5.5% were due to diabetes. Overall, NCDs accounted for 57% of all deaths in that year (Stats SA, 2018). As in the rest of the developing nations, the proportion of NCD-related deaths compared to death from infectious diseases and malnutrition is increasing annually. It is estimated that 85% of the global NCD mortalities occur in low- and middle-income countries (LMICs) (WHO, 2018b). Adults in these regions face twice the risk of NCD mortality than their counterparts living in high-income countries (WHO, 2018a). Although this discrepancy is partly accounted for by the lack of adequate health care and infrastructure, the main driver of the increase in NCDs in LMICs has been ascribed to be the rapid urbanisation and globalisation of lifestyles experienced by LMIC residents (Miranda *et al.*, 2019; WHO, 2018b). In South Africa specifically, urbanisation, accompanied by nutritional transition, has been identified as the driving force of the increased NCD prevalence (Department of Health *et al.*, 2019; Nienaber-Rousseau *et al.*, 2017; Pieters & Vorster, 2008; Popkin, 2015; Vorster, 2002). The association of lifestyle and environmental changes with NCD prevalence shows, in line with global findings, that a large component of the NCD/CMD epidemic is the result of environmental influence, notwithstanding the known genetic component of these diseases (Holland, 2017; Popkin, 2015; Prioreschi et al., 2017; Rider & Carlsten, 2019).

Because a significant proportion of CMD cases are preventable, modifiable risk factors of CMDs and the ways in which these may alter disease risk and mortality have received considerable attention (Bennett *et al.*, 2018; Forouzanfar *et al.*, 2016; Reddy, 2016). It is, however, becoming increasingly evident that the epigenetic mechanisms, such as DNA methylation (DNAm), which bridge genetic predisposition and environmental exposure with complex disease, may be a valuable novel target for intervention (Jhun *et al.*, 2017; Ladd-Acosta & Fallin, 2016; Ligthart *et al.*, 2018; Richard *et al.*, 2017; Richardson *et al.*, 2017).

Although the genome is fixed at conception, the body is able to respond to short- and long-term environmental exposure through the epigenome (West-Eberhard, 1989) by altering the way in

which this genomic blueprint is expressed (Holland, 2017; Laubach *et al.*, 2018; Martin & Fry, 2018). DNAm is the most investigated epigenetic modification because of its ease of measurement and its intervention potential (Christman, 2002; Lioznova *et al.*, 2019; Singh *et al.*, 2013). DNAm, in the context of cardio-metabolic health and specifically disease, is the focus of this thesis. Multiple epigenome-wide association studies (EWASs) have highlighted the relationship between the methylation of specific cytosine-guanine-phosphate sites (CpGs) and the risk of CVD (Davis Armstrong *et al.*, 2018; Richardson *et al.*, 2017) and diabetes (Chambers *et al.*, 2015; Elliott *et al.*, 2017; Ortiz *et al.*, 2018). Similarly, differential CpG methylation has been associated with well-known risk factors of CMDs, such as smoking (Joehanes *et al.*, 2016), alcohol consumption (Liu *et al.*, 2018), physical inactivity (Hunter *et al.*, 2019), adiposity (Aslibekyan *et al.*, 2015; Mendelson *et al.*, 2017), blood lipid concentrations (Braun *et al.*, 2017; Hedman *et al.*, 2017), impaired glucose metabolism (Kriebel *et al.*, 2016) and inflammation (Ligthart *et al.*, 2016). Investigations into the potential causal influence of DNAm on CMD development are increasing (Elliott *et al.*, 2017; Mendelson *et al.*, 2017; Richardson *et al.*, 2017; Richmond *et al.*, 2016), although conclusive evidence is yet to be published.

A recognised limitation of the current epigenetic literature is that the vast majority of research only represents European populations (Popejoy & Fullerton, 2016). This limits the generalisability of current literature to different ethnic and demographic landscapes and consequently the ability to validate these findings in alternative settings. Although efforts are being made in the larger genomics research community to address this, only marginal increases in data on African Americans (Akinyemiju *et al.*, 2018; Barcelona *et al.*, 2019; Chitrala *et al.*, 2019), Asians (Zhu *et al.*, 2016) and Latin Americans (Galanter *et al.*, 2017) have been observed (Popejoy & Fullerton, 2016). In relation to CMD, thus far only one continental African EWAS cohort has represented a West African sample population (Agyemang *et al.*, 2014; Meeks *et al.*, 2018).

One of the integral features of this thesis is that it utilises the unique genetic and environmental aspects of the South African population to broaden the current perspectives on the role of DNAm in cardio-metabolic health and disease. Continental Africans are known for their vast genomic diversity and low genetic linkage (Schlebusch & Jakobsson, 2018; Schuster *et al.*, 2010). This has proven useful in genetic research when attempting to narrow down on causal variants (Marigorta *et al.*, 2018; Park, 2019; Pranavchand & Reddy, 2016). In terms of epigenetic research, investigating a population with a different ancestral origin from any referred to in the previously published literature is useful for both validation and expansion of current findings. As with all epidemiology, validation of epigenetic research is crucial to rule out any spurious findings.

In addition, as a relatively new area of research, advancements in methylation research, such as the new methylation quantification platform, Illumina® Infinium HumanMethylationEPIC bead chip (EPIC array), allows for a continual increase in the loci available for testing and consequently the acceleration in the discovery of novel associations. This array currently also provides the most cost-effective way (price per CpG) to generate genome-wide methylation data. Of the ~850 000 loci represented on the EPIC array, ~450 000 can be investigated for validation of previous findings (from data generated from the EPIC predecessor, the 450K array) and ~400 000 can be leveraged for the discovery of novel associations. Apart from EWASs (reported in Chapter 4), many of the computational methods developed for methylation data, have not been tested in continental Africans, including two of the methods applied in this thesis: methylation-derived cell ratios (Chapter 5) and epigenetic clocks (Chapter 6).

Added to the genomic novelty of continental African groups are the differences in environmental exposure and, therefore, potential DNAm modifiers these cohorts can contribute. South Africa, for example, is a developing country that contains vast open stretches of agricultural land inhabited by small farming communities, but also world-class cities and large, densely-populated informal settlements (Government Communications, 2019; Miranda *et al.*, 2019). These socio-economic distinctions are further intertwined with numerous ethno-linguistic, religious and cultural groups showcasing vast behavioural differences. The country is furthermore home to eight distinct biomes, ranging from desert to forest (Government Communications, 2019; Department of Health *et al.*, 2019; Ramsay, 2012), all of which can modulate gene expression by altering the DNAm.

1.2 Study cohort

Funding for this project was received to pilot an epigenetics investigation of cardio-metabolic health within the South African arm of the international Prospective Urban and Rural Epidemiology (PURE) study. The PURE study was established to investigate the relationship between social, behavioural, genetic and environmental factors and NCD progression in urban and rural individuals (N = 225 000) residing in 27 developing countries around the world, of which South Africa is one (Teo *et al.*, 2009). The North-West University administrates the North West province arm of the PURE study (PURE-SA-NW). At baseline, in 2005, the PURE-SA-NW cohort recruited 2 010 self-identified black Batswana participants. Participants had to be healthy adults (no diagnosed chronic or acute disease, including negative status for the human immunodeficiency virus), older than 30. These participants were followed up for 10 years. In these 10 years, three

major data collections took place, in 2005, 2010 and 2015 (De Lange *et al.*, 2012; Jacobs *et al.*, 2019; Wentzel-Viljoen *et al.*, 2018).

In line with the acknowledged strengths of African cohorts, prior evidence confirmed vast genetic variation and low linkage disequilibrium in the PURE-SA-NW group (Chikowore *et al.*, 2015; Cronjé *et al.*, 2017a; Cronjé *et al.*, 2017b), as well as environmental diversity through its inclusion of rural, urban and transitioning study sites (Nienaber-Rousseau *et al.*, 2013; Wentzel-Viljoen *et al.*, 2018). Other aspects of this cohort that make it ideal for an epigenetic epidemiology pilot study include (1) a longitudinal design spanning 10 years, (2) a large sample size (N = 2 010), (3) an extensive data set on a range of phenotypes over the 10 years, including data on demographics, biochemical markers, body composition, lifestyle (e.g. diet, stress and physical activity), cardiovascular function, injuries and non-fatal events, (4) the availability of cryo-stored samples from all three time points, and (5) a multi-disciplinary research team from multiple faculties and areas of expertise. Challenges associated with using this cohort were that (1) major rural development occurred during the 10-year follow-up in the rural study area, resulting in the rural and urban groups being largely indistinguishable in their lifestyles by 2015, (2) the sample size was reduced through attrition to 980 by the tenth year of follow-up, (3) epigenetics was only introduced as a specific aim in 2015, which means that DNA-containing samples collected previously (2005 and 2010) might be of sub-optimal quality; (4) limited genetic data were available (118 single nucleotide polymorphisms in the larger cohort and whole-genome sequences for a sub-sample of 30 individuals) and, (5) mortality data were not yet available.

1.3 Research questions, aims and objectives

Based on the above, four research questions were identified that best suited the strengths and limitations of our study population and available PURE-SA-NW data within the overarching theme of: The relationship between DNAm and cardio-metabolic health in black South Africans. These are explored in four manuscripts, presented in this thesis. The aims, objectives and context of each of these will be discussed next.

1.3.1 Leveraging the urban-rural divide for epigenetic research

Aim: To evaluate the use of an urban-rural divide as an epidemiological approach when investigating the role of DNAm in the association between urbanisation and NCD risk.

Objectives:

1. Critically review current literature on urbanisation-NCD, urbanisation-DNAm and DNAm-NCD relationships.

2. Evaluate the suitability of the migration, income-comparative and urban-rural study designs to clarify the role of DNAm in the association between urbanisation and NCDs by:

    a. Comparing their strengths and limitations;

    b. Describing the best suited research questions that can be addressed with each of these study designs; and

    c. Describing appropriate cohorts that are/can be leveraged for each study design.

These objectives were pursued in the form of a narrative review manuscript. The results of a vast number of investigations on associations between DNAm and individual components of urbanisation (such as reduced physical activity, Westernised diet, increased alcohol consumption) and urbanisation-related NCD/CMD risk factors (insulin resistance, adiposity, hypertension, unfavourable blood lipid profiles) have been published (Kriebel *et al.*, 2016; Levine *et al.*, 2018; Ligthart *et al.*, 2016; Liu *et al.*, 2018; Lu *et al.*, 2019; Martin & Fry, 2018; Richardson *et al.*, 2017; Ryan *et al.*, 2019). The urban-rural divide may provide a unique opportunity to investigate the amalgamated environmental exposure that defines the *urban*-dwelling compared to *rural*-dwelling individual, and to extend that to an investigation of the associated DNAm differences and increased NCD risk. This review evaluates the ability of three epidemiological study designs (migration, income-comparative and urban-rural designs) to investigate the role of DNAm in the association between urbanisation and the rise in NCD prevalence and mortality. The varied aims, strengths and weaknesses of each of these investigative frameworks are reviewed in this manuscript.

1.3.2 Replication and expansion of epigenome-wide association literature in a black South African population

Aim: To replicate findings from previously published NCD-related EWASs and contribute novel findings from the PURE-SA-NW cohort to the current literature.

Objectives:

1. Find and extract summary statistics from the largest reproducible EWAS of each of the traits for which data are available in the PURE-SA-NW cohort (age, alcohol consumption, smoking status, body mass index, waist circumference, blood lipids and C-reactive protein).

2. Perform EWASs as described in the respective reference manuscripts using the PURE-SA-NW data.

3. Evaluate and report the overall agreement between the summary statistics of each EWAS performed and the reference cohorts by:

    a. Determining whether any systematic differences exist between the PURE-SA-NW EWAS findings and the respective reference studies;

    b. Evaluating causes for such systematic differences;

    c. Determining whether there are any population-specific differences at an individual CpG-trait association level; and

    d. Evaluating the causes of such probe-specific population differences.

4. Identify any novel epigenome-wide associations from the PURE-SA-NW results, including:

    a. Loci previously investigated (450K probes present on the EPIC array) and found not to be associated with the trait of interest in other populations; and

    b. Associations of novel EPIC array probes, investigated for the first time.

Since this was the first epigenetic study in this South African ethnic group, the first empirical aim was to replicate the available, validated EWAS literature obtained in other ethnic groups. Replication serves the purpose of validating the sufficiency of the data quality and statistical power

of this study to replicate well-known associations. This also provides a first look at the degree of generalisability of findings between the PURE-SA-NW cohort and European and African American cohorts. This aim was extended to expand the current literature with any novel associations observed in the EWASs conducted. Novel associations can either be previously investigated probes that resulted in population-specific associations, or new probes (represented for the first time on the EPIC array) that have not been investigated before.

1.3.3 Methylation *vs* protein inflammatory biomarkers and their associations with cardiovascular function

Aim: Compare the association of methylation-derived markers of cell count and protein-based inflammatory markers with cardiovascular function (CVF) and their literature-based portrayal of CVD risk.

Objectives:

1. Compare the CVD risk of the PURE-SA-NW cohort reflected by the measured protein-based (C-reactive protein, interleukins 6 and 10, tumour-necrosis factor alpha and interferon-gamma) and methylation-derived (neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratio (mdNLR and mdLMR) and five myeloid differentiation-associated CpGs) inflammatory markers according to literature-based cut-offs.

2. Compare the PURE-SA-NW cell count ratios to ratios reported in population-based studies from other ethnic groups.

3. Investigate the relationship among and between the methylation-derived and protein-based inflammatory biomarkers in the PURE-SA-NW data.

4. Compare the associations of the methylation-derived and protein-based inflammatory biomarkers with CVF markers.

5. Investigate whether a combination of methylation-derived and protein-based inflammatory biomarkers provides added benefit in explaining CVF variance, or whether one proxies the other.

In the PURE-SA-NW cohort, apart from the whole-genome methylation data generated for this sub-study, various protein-based inflammatory markers have been measured and CVF has been

well characterised. This strength of extensive phenotype data is leveraged to pursue a unique research question regarding the methylation-derived cell count ratios (mdNLR and mdLMR) that have emerged over the past few years. When directly measured using flow cytometry, NLRs and LMRs are used to characterise systemic inflammation and inform prognosis and disease progression (Angkananard *et al.*, 2019; Corriere *et al.*, 2018; Venkatraghavan *et al.*, 2015; Wang *et al.*, 2018). Cancer and rheumatoid arthritis case-control research has reported very similar observations when using methylation-derived ratios compared to the literature on directly measured cell counts (Ambatipudi *et al.*, 2018a; Ambatipudi *et al.*, 2018b; Koestler *et al.*, 2016). Directly measured cell counts have also been used in previous investigations of CVD risk, progression and prognosis (Angkananard *et al.*, 2019; Corriere *et al.*, 2018; Haybar *et al.*, 2019; Jhuang *et al.*, 2019; Wang *et al.*, 2018). However, to date, no data on methylation-derived cell count ratios and CVD risk or CVF have been reported. No data comparing the information provided by these cell ratios and conventionally used inflammatory markers (interleukin-6, C-reactive protein etc.) have been published either. Inflammation is an integral part of CMD development and risk (Angkananard *et al.*, 2019; Bao *et al.*, 2018; Lopez-Candales *et al.*, 2017), and the use of methylation-based cell ratios in CMD remains unexplored. The second empirical aim was, therefore, to compare methylation-derived markers of cell count ratios and protein-based inflammatory markers in their associations with CVF markers and their literature-based portrayal of CVD risk in the PURE-SA-NW cohort.

1.3.4 Comparing DNAm clocks in black South African men

Aim: Characterise the behaviour of five DNAm clocks in the PURE-SA-NW cohort.

Objectives:

1. Compare the three first-generation (Horvath, Hannum and skin and blood) DNAm clocks in terms of:
    a. Their accuracy in estimating chronological age;
    b. Their relative chronological age underestimation in older adults; and
    c. Their age acceleration estimates.
2. Compare two next-generation clocks (PhenoAge and GrimAge) in terms of:

a. Their portrayal of the cohort's age-related biophysiological decline;

b. Their relative underestimation of age-related biophysiological decline in older adults; and

c. The contribution of their constituents to the accuracy of and the differences between their biological age estimates.

Similar to global trends, life expectancy in South Africa is increasing (Mathers *et al.*, 2015; Stats SA, 2018). This is largely the result of better access to medication and appropriate health care. However, with increased life expectancy, a heavier burden of chronic CMD and CMD co-morbidities is experienced (Banerjee, 2015; Miranda *et al.*, 2019). Biological clocks have been suggested as a way to investigate aging when the aim is to discern the level of biophysiological decline in same-aged (chronological age from birth) individuals (Horvath & Raj, 2018; Jylhävä *et al.*, 2017). Ultimately, this discernment would allow for timely intervention to prolong a healthier life span as opposed to only a longer one (Horvath & Raj, 2018). Currently, most of the evidence suggests DNAm-based biological age markers as the most promising biological clocks for epidemiology (Jylhävä *et al.*, 2017). To date, a number of these clocks, divided into first- and next-generation clocks, have been developed and used in epigenetic epidemiology research. First-generation DNAm clocks are used when the goal is to estimate the chronological age of unknown donors or compare the biological age of multiple tissues in the same individual (Hannum *et al.*, 2013; Horvath, 2013; Horvath *et al.*, 2018). The three first-generation clocks that have undergone the most frequent validation and are most often used are the Horvath, Hannum and skin and blood clocks. Next-generation clocks, however, have been developed as indicators of the physiological dysregulation and morbidity and mortality risk associated with age acceleration, as opposed to estimating chronological age. The two currently available next-generation clocks are the PhenoAge (Levine *et al.*, 2018) and GrimAge (Lu *et al.*, 2019) clocks. The GrimAge clock is set apart from the PhenoAge (and all the first-generation clocks) by its inclusion of a lifestyle-related risk factor, smoking pack-years, in its estimation of biological age (Lu *et al.*, 2019). DNAmAge acceleration (when the estimated biological age is older than the chronological age) has been associated with numerous CMD risk factors (adiposity, inflammation, fasting glucose concentration, blood pressure) and have been proven accurate in predicting CMD incidence and prognosis (Fransquet *et al.*, 2019; Horvath *et al.*, 2014; Levine *et al.*, 2015; Levine *et al.*, 2018; Lu *et al.*, 2019; Quach *et al.*, 2017; Ryan *et al.*, 2019). These clocks have, however, not yet been applied to any continental African cohorts. Furthermore, recent evidence shows that the Horvath

and Hannum clocks systematically underestimate the biological age of older adults (El Khoury *et al.*, 2019). Because CMD is mainly a concern in older adults, it is necessary to evaluate this underestimation in all clocks in independent cohorts and to draw comparisons between the currently available DNAm clocks to determine the best-fitting clocks for future research.

1.4 Structure of this thesis

This thesis is presented as seven chapters, written in article format. A literature review, Chapter 2, follows this introductory chapter. Chapter 3 is a narrative review article and is followed by three original research manuscripts (Chapters 4, 5 and 6). The article presented in Chapter 3 has been accepted for pulication at *Epigenomics* (impact factor (IF): 4.40). Chapter 4 presents a manuscript published in *Clinical Epigenetics* (Cronjé *et al.* (2020), IF: 5.50). Chapter 5 has been accepted for publication at *Frontiers in Immunology* (IF: 4.72) and Chapter 6 is currently under review at *Aging* (IF: 5.52). A discussion and conclusions chapter, Chapter 7, concludes the thesis, followed by a reference list for citations used in Chapters 1, 2 and 7. Addendum A presents the manuscript version of Chapter 4, as published by *Clinical Epigenetics*. Each article includes its own reference list according to the guidelines of the journal to which the manuscript was submitted or in which it was published. These chapters are also formatted in the respective journal language and layout. As such, Chapters 5 and 6 are written in US English while the rest of this thesis, including Chapters 3 and 4, makes use of UK English.

1.5 Research team

Table 1-1 Research team members and contributions

Ms H. Toinét Cronjé (PhD candidate)[1]
Conceptualised the study with the rest of the research team; Critically appraised the literature and wrote the study proposal and ethics application; Isolated and curated the DNA; Performed all bioinformatic and statistical analyses; Interpreted the results; Was the first author of all four manuscripts; Was resposible for all rebuttals (including the scientific and ethics review of the proposal and peer review of all manuscripts); Planned, wrote and compiled this thesis.
Prof. Marlien Pieters (Promotor)[1]
Acted as principal investigator and funder of the project; Co-conceptualised and critically reviewed the study proposal and ethics application; Served as primary supervisor of the statistical analysis and writing of this thesis; Critically reviewed and commented on the entire content of this thesis; Co-authored all manuscripts; Critically reviewed all rebuttals.
Prof Cornelie Nienaber-Rousseau (Co-promotor)[1]
Co-conceptualised and reviewed the study proposal and ethics application; Critically reviewed and commented on the entire content of this thesis; Co-authored all manuscripts; Critically reviewed all rebuttals.
Dr Hannah R. Elliott (Co-promotor)[2,3]
Co-conceptualised and reviewed the study proposal and ethics application; Served as primary supervisor of the bioinformatic analysis and data interpretation; Critically reviewed and commented on the entire content of this thesis; Co-authored all manuscripts; Critically reviewed all rebuttals.

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

1.6 Contributing authors

The following individuals contributed to one or more of the manuscripts presented in this thesis. By signing below, permission is granted to the doctoral candidate to include the co-authored manuscripts as chapters in this thesis for degree examination purposes.

Table 1-2 Permission from co-authors to submit manuscripts for degree purposes

Title and name	Co-authored	Signature
Prof. Marlien Pieters[1]	Chapters 3, 4, 5, 6	
Prof. Cornelie Nienaber-Rousseau[1]	Chapters 3, 4, 5, 6	
Dr Hannah R. Elliott[2,3]	Chapters 3, 4, 5, 6	
Dr Fiona R. Green[4]	Chapters 5, 6	
Prof. Aletta E. Schutte[5,6]	Chapter 5	
Dr Josine Min[2,3]	Chapter 6	

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[4] Formerly School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

[5] Hypertension in Africa Research Team, Medical Research Council Unit for Hypertension and Cardiovascular Disease, North-West University, Potchefstroom, South Africa

[6] School of Public Health and Community Medicine, University of New South Wales; George Institute for Global Health, Sydney, Australia

CHAPTER 2

LITERATURE REVIEW

LITERATURE REVIEW

2.1 Introduction

This thesis focusses on DNA methylation (DNAm) as an epigenetic mechanism in the context of cardio-metabolic health and disease. DNAm is the most frequently investigated epigenetic modification, partly owing to its ease of measurement, but also because of its ability to respond to environmental change and, therefore, possibly, intervention (Christman, 2002; Singh *et al.*, 2013; Xiao *et al.*, 2013).

This chapter presents a literature review covering the key themes of this thesis. First, a summary of the relevant concepts and potential applications of epigenetics, specifically DNAm, is given. Here, general epigenetic concepts are briefly discussed before narrowing the focus to DNAm. The biochemical framework of DNAm is explored, including the addition to and the preservation and removal of methyl groups from the human genome. This is followed by a review of the physiological utility of DNAm and some of the distinct ways in which the methylome is investigated in subsequent chapters. Then, the roles of DNAm in the context of cardio-metabolic health research are described. This entails (i) differential DNAm as a biomarker of exposures related to cardio-metabolic health; (ii) differential methylation as a biomarker of cardio-metabolic health and disease itself and (iii) differential DNAm as a mediator between the exposures and the outcomes of cardio-metabolic processes. The ways in which these roles are investigated in subsequent chapters are also highlighted. Lastly, the African, particularly the South African, context is described. This final section focusses on the advantages of conducting epigenetic research in a (South) African population, particularly the genetic and environmental diversity, and how these advantages are leveraged in this thesis.

2.2 Epigenetics: concepts and applications

The differential output of the same genetic code is the result of epigenetic mechanisms that regulate how DNA is transcribed to RNA before proteins are formed. In contrast with the static genome, it is the malleability of the epigenome that makes epigenetic research critical, as it represents a regulatory mechanism that can be altered through targeted intervention. Genetics of common complex diseases explain some but not all of the variation in disease risk (Ladd-Acosta & Fallin, 2016; Shah *et al.*, 2015; Shah *et al.*, 2014). Investigation of epigenetic variation is therefore important, as it can enhance researchers' understanding of disease aetiology and progression and also offer a route for possible intervention.

The four main interrelating systems of epigenetic control are covalent modifications, higher-order chromatin remodelling, histone variants and non-coding RNAs (Jones & Liang, 2012). Covalent

modifications occur when chemical groups change the way DNA is read by either attaching to the DNA itself (such as DNAm) or attaching to the histone proteins around which the DNA is wrapped (histone modifications). Chromatin remodelling refers to the shifting of DNA or histone proteins to alter the density of the chromatin, which affects the accessibility of the DNA for transcription (Clapier & Cairns, 2009). Histone variants include the post-translational acetylation, methylation or phosphorylation of histones three and four, which alter their interactions with DNA and other nuclear proteins (Jones & Liang, 2012). Non-coding RNAs (functional RNA not translated to protein) encompass micro-RNAs, short interfering RNAs, Piwi-interacting RNAs and long non-coding RNAs, which influence gene expression in a transcriptional or post-transcriptional manner (Lee, 2012; Mercer *et al.*, 2009).

2.2.1 DNA methylation

DNA methylation entails a post-replication addition of a methyl group to the fifth carbon of a cytosine base (5_mC) adjacent to a guanine nucleotide (cytosine-phosphate-guanine site [CpGs]) (Razin & Riggs, 1980). CpGs often gather in high densities referred to as CpG islands (more than 500 consecutive base pairs with a GC content greater than 55%) that are flanked by CpG island shores and more distally, CpG island shelves with CpG density decreasing in each region (Deaton & Bird, 2011; Irizarry *et al.*, 2009; Jones, 2012; Takai & Jones, 2002). CpG islands are often observed in the promoter regions of genes, where they are predominantly unmethylated and frequently associated with increased transcription (Bird & Wolffe, 1999; Deaton & Bird, 2011).

In humans, DNAm of somatic cells is almost exclusively present at cytosine residues in the CpG context, in a symmetrical manner. It is estimated that 20 of the 28 million (60–80%) cytosine nucleotides in the human genome are methylated; therefore, methylated rather than unmethylated CpGs are referred to as the "default state" of the genome (Borgel *et al.*, 2010; Suzuki & Bird, 2008; Zemach *et al.*, 2010). Unmethylated regions (often observed in CpG islands) are primarily co-localised with regions overlapping active regulatory elements occupied with transcription factors such as gene promotors and enhancers. The unmethylated state of these regulatory elements is associated with their active transcription (Krebs *et al.*, 2014; Ziller *et al.*, 2013).

The DNAm process is regulated by three interrelated pathways (discussed separately in the following sub-sections) that direct the generation, maintenance and loss of methyl groups at cytosine nucleotides. These mechanisms are co-regulated and allow methylomic response to environmental stimuli by tightly regulating the transcriptional competence of the genome (Ziller *et al.*, 2013).

2.2.1.1 The methylation process

2.2.1.1.1 *De novo* cytosine methylation

The selective acquisition of methyl groups by cytosine is catalysed by a set of DNA methyltransferases (DNMTs), namely DNMT3A and 3B (Okano *et al.*, 1999; Okano *et al.*, 1998), which are modulated by DNMT3L (Aapola *et al.*, 2000; Bourc'his *et al.*, 2001; Goll & Bestor, 2005). S-adenosylmethionine acts as the methyl donor to cytosine and is converted to S-adenosylhomocysteine in a reaction incorporating zinc and the above-mentioned DNMTs. S-adenosylhomocysteine is then hydrolysed to form homocysteine, a bio-synthesiser of the amino acid methionine that can itself be remethylated back to methionine. Cytosine methylation occurs in a symmetric manner where methyl groups can be added to both sides of the double-stranded DNA (Jones, 2012). Figure 2-1 depicts the *de novo* methylation of cytosine nucleotides in the context of CpGs.

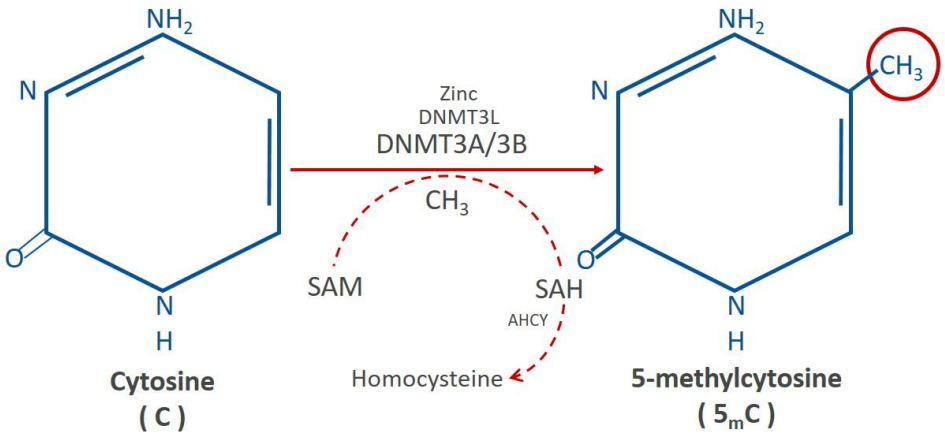


Figure 2-1 *De novo* cytosine methylation

Adapted from Jones and Liang (2012), Wu and Zhang (2014), Schübeler (2015), Ambrosi et al. (2017), Iurlaro et al. (2017) and Laubach et al. (2018)

A = adenosine; AHCY = adenosylhomocysteinase; C = cytosine; CH_3 = methyl; DNMT = deoxyribonucleic acid methyltransferases; G = guanine; H = hydrogen; N = nitrogen; O = oxygen; SAH = s-adenosylhomocysteine; SAM = s-adenosylmethionine; T = thymine

2.2.1.1.2 Preservation of existing methylation marks

Each cell division has the potential for partial or full methylation loss. Newly synthesised CpGs require active re-methylation to maintain methylation status faithfully (Chen *et al.*, 2003; Wu & Zhang, 2014). This process is depicted in Figure 2-2 and is discussed below.

Upon symmetric CpG *de novo* methylation, DNA undergoes semi-conservative replication, resulting in hemi-methylated CpGs. The daughter strand retains the methylation pattern of the

mother strand, and it is the role of DNMT1 to replicate the mother strand's methylation to the newly formed strand (Bostick *et al.*, 2007). Through its association with the replication complex and the co-operation of ubiquitin-like plant homeodomain and RING finger domain 1 (UHRF1), DNMT1 is able to restore symmetrical CpG methylation (Bostick *et al.*, 2007; Chen *et al.*, 2007; Hermann *et al.*, 2004; Li *et al.*, 1992; Sharif *et al.*, 2007).

Although DNMT1 is referred to as the "maintenance DNMT", it has become clear that the DNMT3s are also critical for the maintenance process (Chen *et al.*, 2003). DNMT3A and 3B are responsible for the restoration of methylation at sites left untouched by DNMT1s. This was observed by Chen *et al.* (2003) when the deactivation of DNMT3s resulted in progressive dilution of methyl groups during cell differentiation. DNMT3s are referred to as the "proof-readers" of DNMT1's work (Chen et al., 2003). The DNMT3s associate with the replication complex and replicate methylation using the *de novo* methylation process discussed above (Chen *et al.*, 2003; Jackson *et al.*, 2004; Liang *et al.*, 2002). In the event of failed methylation maintenance, methyl groups dilute and finally disappear throughout multiple cell divisions. This phenomenon is discussed in the following section.

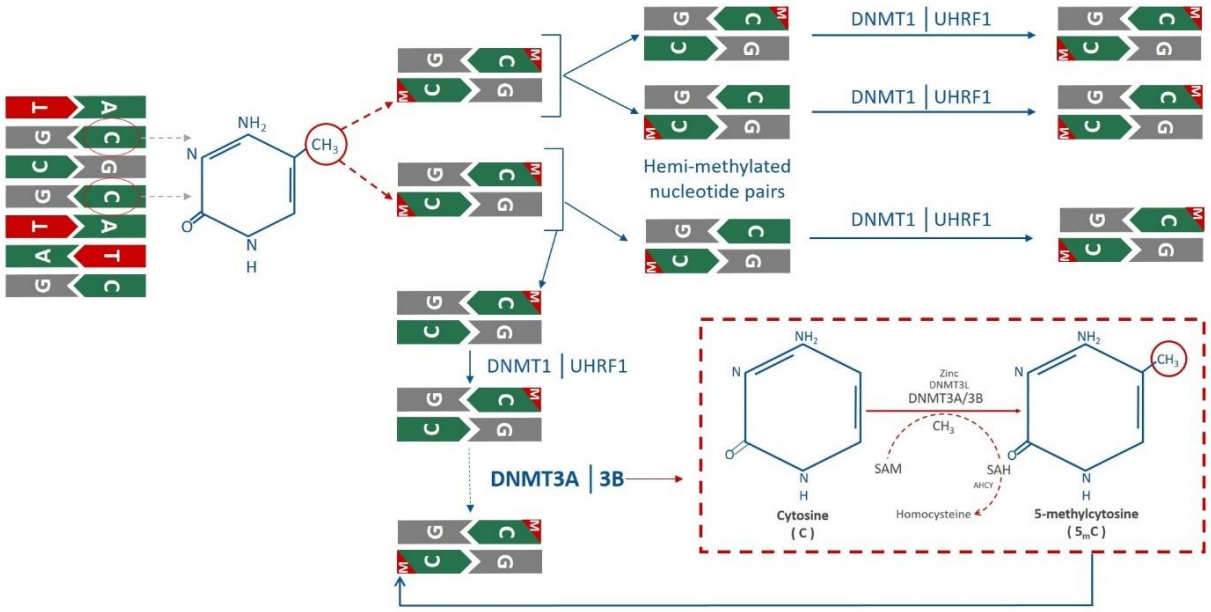


Figure 2-2 Maintenance of cytosine methylation

Adapted from Chen et al. (2003), Moore et al. (2013) and Wu and Zhang (2014)

A = adenosine; AHCY = adenosylhomocysteinase; C = cytosine; CH_3 = methyl; DNMT = deoxyribonucleic acid methyltransferases; G = guanine; H = hydrogen; N = nitrogen; O = oxygen; SAH = s-adenosylhomocysteine; SAM = s-adenosylmethionine; T = thymine; UHRF = ubiquitin-like plant homeodomain and RING finger domain 1

2.2.1.1.3 Removal and dilution of methyl groups

During embryonic development the methylome undergoes two rapid erasures of almost all its methyl marks in an effort to reset the methylome and potency of the inherited DNA (reviewed by Jones (2012); Wu and Zhang (2014) and Lee *et al.* (2014)). As cells begin to differentiate upon exiting pluripotency, re-methylation takes place (Section 2.2.1.1.1) as cells create their individual epigenetic identity and memory (Lee *et al.*, 2014). After the methylome has been reset, methyl groups are removed in a locus-specific manner through either passive or active removal processes throughout the rest of the life course (Figure 2-3) (Sadakierska-Chudy *et al.*, 2015).

The faithful inheritance of methylation marks through cell division and replication is performed with 95% accuracy through the preservation of existing methylation marks (Section 2.2.1.1.2). In the absence of proper methylation maintenance, replication-dependent or passive demethylation occurs, resulting in stochastic methylomic variation (Chen *et al.*, 2003). The active loss of methylation is replication-independent and is achieved by a group of ten-eleven translocation enzymes (TETs) that oxidise 5_mCs to remove specific methyl groups (Ito *et al.*, 2010; Tahiliani *et al.*, 2009). The intermediate products of TET-mediated demethylation are 5-hydroxymethylcytosine (5_{hm}C), 5-formylcytosine (5_fC) and 5-carboxylcytosine (5_{ca}C) and are produced through stepwise oxidation in the presence of α-ketoglutarate and water (Ito *et al.*, 2011; Kriaucionis & Heintz, 2009; Tahiliani *et al.*, 2009).

The demethylation derivatives each plays a unique biological role such as recruitment of specific binders or DNA reparation machinery (reviewed by Breiling and Lyko (2015) and Fong *et al.* (2013)). Only 0.1–0.2% of the cytosines in mammalian tissues are observed as 5_{hm}C, where the maximum concentration (0.6% of cytosines) occurs in the brain and spinal cord (Globisch *et al.*, 2010; Kriaucionis & Heintz, 2009). The relative scarcity of these intermediates in comparison to 5_mC makes their quantification and physiological effects difficult to investigate. These intermediate products are removed through human thymine DNA glycosylase (5_{ca}C specifically, Zhang *et al*., 2012), DNA reparation mechanisms or through further cell division, as these are poor substrates for DNMT1 and UHRF1 (Hashimoto *et al.*, 2012; He *et al.*, 2011).

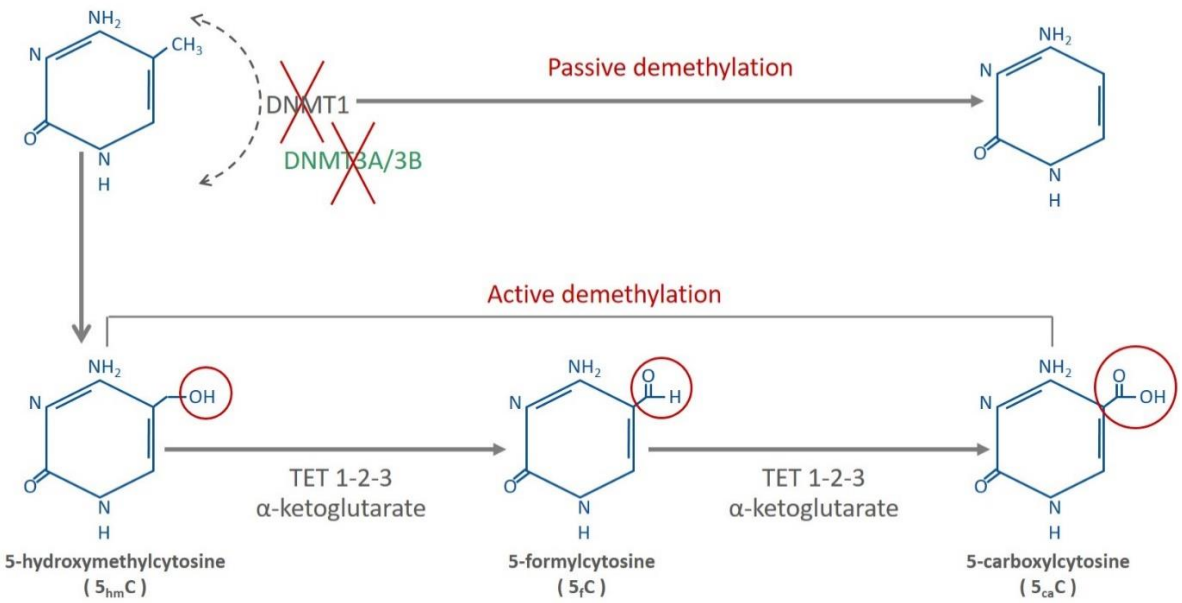


Figure 2-3 DNA demethylation and methylation dilution

Adapted from Jones and Liang (2012), Wu and Zhang (2014), Breiling and Lyko (2015) and Schübeler (2015)

CH_3 = methyl; DNMT = deoxyribonucleic acid methyltransferases; H = hydrogen; N = nitrogen; O = oxygen; OH = hydroxide; TET = ten-eleven translocation enzymes

2.2.1.2 The physiological importance of DNA methylation

Practically, the identical genomic blueprint is not expressed in all cells simultaneously, and genes are not expressed in the same manner or at the same intensity in different cell types. This is controlled by the transcriptional competence of cells, partly through methylomic variation. Although more complex and context-specific, hypermethylation at CpGs situated in gene promoter regions are typically associated with transcriptional silencing, whereas the opposite is true for CpGs situated within the gene body (Dirks *et al.*, 2016; Razin & Cedar, 1991; Siegfried & Simon, 2010). Methylated cytosine nucleotides are specifically recognised by methyl-CpG-binding domain proteins (MBDs). These MDBs are significantly enriched in promoter regions and are known to boost methylation-mediated transcriptional repression (Baubec *et al.*, 2013; Meehan *et al.*, 1989). The two primary ways in which DNAm alters transcriptional competency is through physically altering the accessibility of transcription factors (Campanero *et al.*, 2000; Watt & Molloy, 1988) or selectively recruiting protein complexes that influence chromatin structure and in turn, also transcription (Buck-Koehntop & Defossez, 2013).

Transcriptional repression is often critical, as in the case of the inactivation of one of the female X-chromosomes during embryonic development (Beard *et al.*, 1995; Borgel *et al.*, 2010; Panning & Jaenisch, 1996; Riggs, 1975). Animal studies have revealed the importance of

hypermethylation during embryonic development (Chen *et al.*, 2003; Dodge *et al.*, 2005). Removal of DNMTs during the embryonic development of mice resulted in substantially reduced methylation, which was associated with premature death (Li *et al.*, 1992). Removing only DNMT3A and 3B completely inhibited *de novo* methylation in mice, also resulting in premature death (Okano *et al.,* 1999). Hypotheses as to why disruption of DNMTs has these severe effects are all centred on the loss of transcriptional control, such as inability to inactivate the X-chromosome, inefficient genomic imprinting and unwanted mobilisation of transposons (Bird & Wolffe, 1999).

Appropriate transcriptional activation or silencing as a result of DNAm is also crucial to prevent the expression of potentially harmful proteins. An example of this is the hypermethylation of the promoter regions of cancer- or virus-inducing proteins, which effectively reduces their expression (Egger *et al.*, 2004; Jones & Baylin, 2002; Walsh *et al.*, 1998). Similarly, hypermethylation of selected retrotransposons is necessary to silence their transcription to reduce viral sequences effectively (Jähner *et al.*, 1982; Walsh *et al.*, 1998). Conversely, hypomethylation of tumour suppressor gene promoters actively challenges tumour growth, thus promoting health (Jones & Baylin, 2002; Jones & Laird, 1999). Furthermore, hypomethylation of selected retrosponsons (ALU elements) allows them to fulfil critical physiological roles in response to stress and infections (Schulz *et al.*, 2006).

In the context of epidemiology and cardio-metabolic health, however, it is the differential methylation in response to disease-related genetic variance or environmental exposure that is of interest. While methylation at many sites either does not change over the life course, or changes in a predictable manner to aid a specific developmental process, it is the subset of loci that reveals environmental plasticity or relationships with disease outcomes when behaviour deviates from the norm that is the interest of the research community. It is these sites that can potentially be targeted through intervention strategies, or the sites that can provide insight into the biological mechanisms of common complex diseases that remain to be understood (Egger *et al.*, 2004; Jones, 2012; Moore *et al.*, 2013; Schübeler, 2015).

2.2.1.3 Information content of DNA methylation analysis

The methylation state of each CpG site is binary: methylated (denoted by a value of 1), or unmethylated (0). The reported CpG methylation represents the fraction of methylated cytosines of the specific locus in the total repeated measures. A beta CpG methylation value of 1, therefore, represents a methyl group on both alleles in all the quantified cells (Laubach *et al.*, 2018). Determining the methylated *vs* unmethylated state of a cytosine can be done using various methods, all based on one of three primary strategies: bisulphite-, restriction enzyme- or affinity-

based quantification. Extensive reviews of these strategies have been published (Dedeurwaerder *et al.*, 2011; Michels *et al.*, 2013; Olkhov-Mitsel & Bapat, 2012; Sun *et al.*, 2015). For the project presented in this thesis, a bisulphite-based strategy was used. The principle of bisulphite conversion is that a methyl group, when attached to a cytosine nucleotide, protects the nucleotide from deamination. When DNA is treated with sodium bisulphite, unmethylated cytosines are deaminated, yielding uracil that is read as thymine when sequenced. Conversely, 5_mC is protected from deamination and yields a cytosine when sequenced. Sequencing prior to and after bisulphite treatment is then used to determine whether a site is unmethylated or methylated (Kurdyukov & Bullock, 2016).

The following sub-sections provide more detail on the approaches followed in this thesis to use the data generated from whole-genome methylation profiling (discussed below) in the context of cardio-metabolic health research. Genome-wide methylation (Chapter 4), methylation-derived cell count estimations (Chapter 5) and methylation age and age acceleration (Chapter 6) approaches are discussed.

2.2.1.3.1 Genome-wide methylation analysis

The method that provides the most extensive genome-wide DNAm coverage is whole genome bisulphite sequencing, but it is extremely costly and labour-intensive. An alternative to sequencing the entire genome prior to, and upon bisulphite treatment, is to sequence regions that have been identified as 5_mC-enriched. Genome-wide DNAm analysis makes use of the percentage methylation quantified at some, but not all, CpGs throughout the genome. Illumina® is currently the largest supplier of genome-wide methylation assays. Their latest release, the Infinium HumanMethylationEPIC bead chip (also referred to as the 850K or EPIC array), quantifies the methylation of more than 850 000 CpGs and is used as the 'next generation' replacement of the widely used Illumina® precursor, the 450K array (Dedeurwaerder *et al.*, 2011; Fortin *et al.*, 2016). Although this is currently the most comprehensive alternative to whole genome bisulphite sequencing, only approximately 5% of the total amount of genomic CpGs is included.

An effort was made by Illumina®, however, to select a diverse set of CpGs that were likely to be the most useful to future research endeavours. The Illumina® arrays, therefore, include CpGs in the RefSeq genes, known differentially methylated regions, and CpG-enriched regions (including gene promotor regions, CpG islands, shores and shelves) (Fernandez-Jimenez *et al.*, 2019; Mansell *et al.*, 2019; Nakabayashi, 2017; Zhou *et al.*, 2017). The EPIC array is set apart by not only its increased genomic coverage, but also by its increased inclusion of CpGs at regulatory enhancers (Pidsley *et al.*, 2016). Data from the encyclopaedia of DNA elements and the

functional annotation of the mammalian genome projects were used to identify CpGs that could, through methylation, alter transcriptional competency by modulating transcription factor binding (Pidsley *et al.* (2016), https://emea.support.illumina.com/). The inclusion of these elements significantly increases the probability that transcriptionally relevant sites are sequenced. Other genome-wide approaches include reduced-representation bisulphite sequencing, target enrichment bisulphite sequencing and methylated DNA immunoprecipitation sequencing. These methods have been reviewed and compared by Michels *et al.* (2013) and Sun *et al.* (2015). For the current study, we generated genome-wide methylation data using the Illumina® EPIC platform. Because it contains more than 90% of the loci present on the 450K array, data from the EPIC array can be integrated with previously generated data when replicating previously published associations (one of the aims of the first empirical manuscript of this thesis presented in Chapter 4). In addition, because the EPIC array contains twice the amount of probes on the 450K array, it provides ample opportunity (~400 000) for novel findings. Lastly, validated and widely used protocols and platforms for data pre-processing (Min *et al.*, 2018), epigenome-wide association studies (EWASs) (Mansell *et al.*, 2019; Min *et al.*, 2018) and cell count (Houseman *et al.*, 2014; Salas *et al.*, 2018) and methylation age (https://DNAmAge.genetics.ucla.edu/home; Horvath *et al.*, 2013; McEwen *et al.*, 2018) estimation, are easily accessible and available. Although the EPIC array data are used and reported throughout this thesis, Chapter 4, in particular, incorporates the single base resolution data by reporting the results of ten EWASs on traits related to cardio-metabolic health.

2.2.1.3.2 Methylation-derived cell distributions

Methylation-derived cell distribution estimates are possible because each leukocyte sub-type (B-cells, T-cells, neutrophils, natural killer cells, granulocytes and monocytes) represented in a peripheral whole blood sample (origin of the DNA reported on in this thesis), has a unique methylomic signature (Reinius *et al.*, 2012). As such, methylation profiles quantified in whole blood represent a mixture of cell sub-types and therefore require adjustment for cellular composition between individuals to ensure that methylation-disease associations are not, instead, cell count-disease associations (Jaffe & Irizarry, 2014). Adjusting for cell proportions measured using flow cytometry is the gold standard for minimising cell sub-type confounding. However, because flow cytometry requires fresh blood samples and most cohorts analyse DNA-containing samples retrospectively, their inability to perform flow cytometry necessitates the use of methylation-derived cell count estimates. Statistical models inferring cell composition from methylation data have been developed, are widely used and are continually improved (Houseman *et al.*, 2012; Houseman *et al.*, 2014; Salas *et al.*, 2018; Titus *et al.*, 2017).

The role of methylation-derived cell distribution in the context of cardio-metabolic epidemiology is primarily fulfilled through its association with inflammation. Inflammation alters the leukocyte composition in peripheral blood (Kelsey & Wiencke, 2018). This characteristic of the inflammatory response has led to the development of neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios (NLRs and LMRs) as acknowledged biomarkers for the degree of inflammation and the prognosis of inflammation-related diseases, including cardiovascular diseases and cancer (Angkananard *et al.*, 2019; Guthrie *et al.*, 2013; Suárez-Cuenca *et al.*, 2019; Turkmen *et al.*, 2012). These ratios are useful because of their ability to reflect the relative contribution of the adaptive and innate immune response, thereby providing a more integrated view of systemic inflammation than protein-based inflammatory markers such as interleukin-6 and C-reactive protein (Balta *et al.*, 2014). Through the ability of methylation data to determine cell composition accurately, methylation-derived NLRs and LMRs (mdNLRs and mdLMRs) are used as surrogates of those directly measured (Koestler *et al.*, 2016; Koestler *et al.*, 2017; Salas *et al.*, 2018). Validation analysis reported R^2 >0.95 when comparing directly measured and methylation-derived cell count and NLR estimates with each other (Koestler *et al.*, 2017; Salas *et al.*, 2018). Recently, five CpGs, robustly associated with neutrophil and monocyte differentiation (adjusted R^2 ≥0.80), have been suggested as proxies for the mdNLR (Wiencke *et al.*, 2017) in relation to cancer. One subsequent case-control study confirmed their possible proxy status by reporting hypermethylation of these five loci (associated with lower NLR) in cancer patients who had survived, during follow-up (Ambatipudi *et al.*, 2018a). The behaviour of these CpGs in relation to cardio-metabolic disease has not been investigated.

In addition to conventional markers of inflammation (acute-phase proteins and cytokine concentrations) the mdNLR, mdLMR and the percentage methylation of the five myeloid CpGs were quantified in a black South African study population, for this thesis. Chapter 5 reports the usefulness of these methylation-derived cell count estimates (compared with and in relation to inflammatory proteins) in reflecting systemic inflammation and cardiovascular function.

2.2.1.3.3 DNA methylation age

A group of same-aged peers can represent a spectrum of biophysiological health and age-related deterioration. Some show disease-induced rapid aging (such as Alzheimer's disease), whereas others follow a lineage of centenarians and seem to age slowly. It is in these chronological *vs* biological age discrepancies that the science of gerontology aims to find the cellular mechanisms of aging that can be targeted to delay the functional decline and prolong the health span of the global population. Within epigenetic epidemiology, various methods of estimation of biological as opposed to chronological ages have been developed. A recent review of these methods provided evidence that epigenetic clocks outperform telomere length, proteomic-, metabolomic- and

transcriptomic-based methods and composite markers in their ability to estimate biological age (Horvath & Raj, 2018). Epigenetic clocks make use of targeted differentially methylated CpGs to estimate a biological/methylation age, hereafter referred to as DNAmAge (Horvath, 2013; Horvath *et al.*, 2018; Horvath & Raj, 2018; Levine *et al.*, 2018; Lu *et al.*, 2019). The two most frequently cited epigenetic clocks are those from Steve Horvath (Horvath, 2013) and Gregory Hannum (Hannum *et al.*, 2013). More recently, the skin and blood (SB) (Horvath *et al.*, 2018), PhenoAge (Levine *et al.*, 2018) and GrimAge (Lu *et al.*, 2019) clocks have been developed and are becoming increasingly popular (Horvath & Raj, 2018; Nelson *et al.*, 2019; Ryan *et al.*, 2019; Zhao *et al.*, 2019). These clocks all ultimately aim to identify instances where discrepancies exist between chronological age and DNAmAge, referred to as DNAm age acceleration (DNAmAgeAccel). A brief description of these clocks follows.

Horvath's epigenetic clock (Horvath, 2013) comprises a weighted average of the methylation of 353 CpGs, selected using an elastic net regression of genome-wide methylation on chronological age. Of these CpGs, 193 correlate positively and 160 negatively with chronological age. Gene enrichment analysis revealed significant enrichment for pathways related to cellular survival, growth, proliferation and death, as well as tissue development and cancer. Horvath's clock was developed as a multi-tissue, robust age estimator. His manuscript reports correlation estimates >0.96 with a median absolute difference (error) of <3.6 years ($p < 1 \times 10^{-200}$) across 8 000 samples from 51 healthy tissues and cell types (Horvath, 2013).

Hannum's clock, on the other hand, was developed as a blood-based (single tissue) age estimator (Hannum *et al.*, 2013). This clock incorporates 71 CpGs selected using a similar elastic net regression model described above. Again, these statistically selected probes were all within or near genes for which clear associations with age-related functions had been published. The model predicted chronological age with accuracy in DNA from peripheral blood samples, reporting a correlation of 0.96 and a median error of 3.88 years in 656 individuals. Validation of the model in a separate cohort of 174 individuals revealed slightly lower predictive power (r = 0.91, error = 4.89 years) than the training data, also in whole blood. When tested on other tissue types (breast, kidney, lung and skin), a further loss of predictive accuracy was observed, emphasising that the model should be used as developed, in DNA obtained from whole blood (Hannum *et al.*, 2013).

Building on the foundation of these two, in 2018, Steve Horvath and his colleagues developed the SB clock. The motivation for this clock was the fact that the multi-tissue (Horvath, 2013) and blood-based estimators (Hannum *et al.*, 2013) both performed poorly in cases where the age of skin cells was estimated. Because skin cells are often incorporated in clinical investigations, Horvath and his team argued that research on, for example, anti-aging therapeutics will suffer in the absence of a clock that is accurate both *in* and *ex vivo*. In addition to skin cells, DNA from

other easily accessible tissues (and therefore feasible candidates for clinical research), including saliva samples and blood, was also incorporated in the training of this algorithm. A weighted average of 391 CpGs, selected using an elastic net regression of genome-wide methylation on chronological age, comprises the SB clock (Horvath *et al.*, 2018). Validation of the SB clock revealed that it outperforms both the Hannum and Horvath algorithms in its accuracy of chronological age prediction, when applied to blood, skin, fibroblasts, keratinocytes and endothelial cells (Horvath *et al.*, 2018).

After the success of the SB clock, and the preceding multi-tissue and blood clocks, Morgan Levine and his colleagues, including Steve Horvath, developed PhenoAge. PhenoAge is an epigenetic clock that provides information about the biological process of aging instead of the chronological age itself. The premise was to replace chronological age with a composite marker that captures the functional state of organ systems as a proxy of organismal aging between same-aged individuals (Levine *et al.*, 2018). The authors regressed the hazard of aging-related mortality against chronological age and 42 clinical biomarkers in 9 926 adults with follow-up data over 23 years. A weighted average of the ten best performing biomarkers (albumin, alkaline phosphatase, creatinine, C-reactive protein, serum glucose, mean cell volume, lymphocyte percentage, red cell distribution width, white blood cell counts and chronological age) was calculated and regressed on genome-wide methylation levels using an elastic net regression model. This resulted in the statistical selection of 513 CpGs that, when expressed as a weighted average, constituted PhenoAge. In contrast to the Horvath and Hannum clocks that sought the most accurate prediction of chronological age, the PhenoAge algorithm sought to optimise the prediction of mortality due to age-associated physiological dysregulation. The CpGs incorporated in the PhenoAge model did not strongly correlate with age, but rather with the relative age acceleration. As hypothesised, the PhenoAge predictor outperforms both the Horvath and Hannum clocks in its ability to predict age-related pathological outcomes (Horvath & Raj, 2018).

Finally, in an attempt to improve on all of the above-described age predictors, Lu *et al.* (2019), under the guidance of Steve Horvath, developed the GrimAge estimate. As with the PhenoAge, the GrimAge incorporates surrogate markers of disease risk, rather than chronological age, although this model uses DNAm surrogates instead of clinical biomarkers per se. The GrimAge estimator comprises DNAm surrogate markers for smoking pack-years and seven plasma proteins (adrenomedullin, beta-2 microglobulin, cystatin C, growth differentiation factor 15, leptin, plasminogen activation inhibitor type 1 and tissue inhibitor metalloproteinase 1). These markers where chosen in a similar fashion as Levine's method, although this time regressing on time-to-death and taking in to account sex and chronological age. It is worth noting that this is the only epigenetic clock that directly incorporates a modifiable lifestyle-related exposure such as smoking

habits. Cigarette smoking is known to affect most chronic diseases and DNAm (Bollepalli *et al.*, 2019; Gakidou *et al.*, 2017; Joehanes *et al.*, 2016). The GrimAge model was validated in five independent cohorts and outperformed the three preceding epigenetic clocks in its ability to predict time-to-death (Lu *et al.*, 2019).

Apart from the ability of these clocks to estimate the risk of all-cause mortality (with increased accuracy) (Chen *et al.*, 2016; Fransquet *et al.*, 2019; Levine *et al.*, 2018; Lu *et al.*, 2019; 2016; 2015; Perna *et al.*, 2016), numerous investigations on the associations between DNAmAge and DNAmAgeAccel and cardio-metabolic diseases, risk factors and mortality have been published. Lifestyle factors known to reduce cardio-metabolic disease risk, such as smoking abstinence (Levine *et al.*, 2018; Zhao *et al.*, 2019), high fruit and vegetable consumption (Lu *et al.*, 2019), omega-3 supplementation (Lu *et al.*, 2019), level of education (Fiorito *et al.*, 2019; Liu *et al.*, 2019b; Zhao *et al.*, 2019) and physical activity (Levine *et al.*, 2018; Quach *et al.*, 2017), have consistently been associated with a lower DNAmAge (most recent publications cited). In contrast, obesity (Fiorito *et al.*, 2019; Ryan *et al.*, 2019), alcohol consumption (Rosen *et al.*, 2018; Zhao *et al.*, 2019), increased blood lipid (Lu *et al.*, 2019), glucose and insulin (Lu *et al.*, 2019) and C-reactive protein (Levine *et al.*, 2018; Lu *et al.*, 2019) concentrations are associated with older DNAmAges. Cardio-metabolic diseases, including ischemic stroke (Soriano-Tárraga *et al.*, 2016; Soriano-Tárraga *et al.*, 2017), hypertension (Lu *et al.*, 2019), congestive heart failure (Lu *et al.*, 2019), type II diabetes (Lu *et al.*, 2019) and incident CVD (Lind *et al.*, 2018) are associated with DNAmAgeAccel.

All the age estimators discussed in this thesis are publicly available on a platform where the respective algorithms are applied to user-uploaded data (Horvath *et al*. (2013), https://DNAmAge.genetics.ucla.edu/home). This platform was used to determine and compare DNAmAge and DNAmAgeAccel, using the five clocks described in this section (Chapter 6).

2.3 Epigenetic epidemiology

The investigation of methylomic variability, in the context of cardio-metabolic health, can be approached in three ways. Differential methylation can, firstly, be investigated as a biomarker of the cardio-metabolic health-related exposures to which an individual is subject. Secondly, methylation differences could be investigated as a non-causal biomarker of the risk, presence or progression of cardio-metabolic disease. Finally, methylation differences could act as a mediator in the causal pathway that links these exposures with cardio-metabolic disease risk/outcomes. Figure 2-4 depicts these three approaches (numerically indicated) and is followed by a brief review of each concept. It should be acknowledged that genetic architecture contributes to this

framework by influencing individual translation of exposures, determining baseline methylomic variability and competency and informing heritable disease risk.

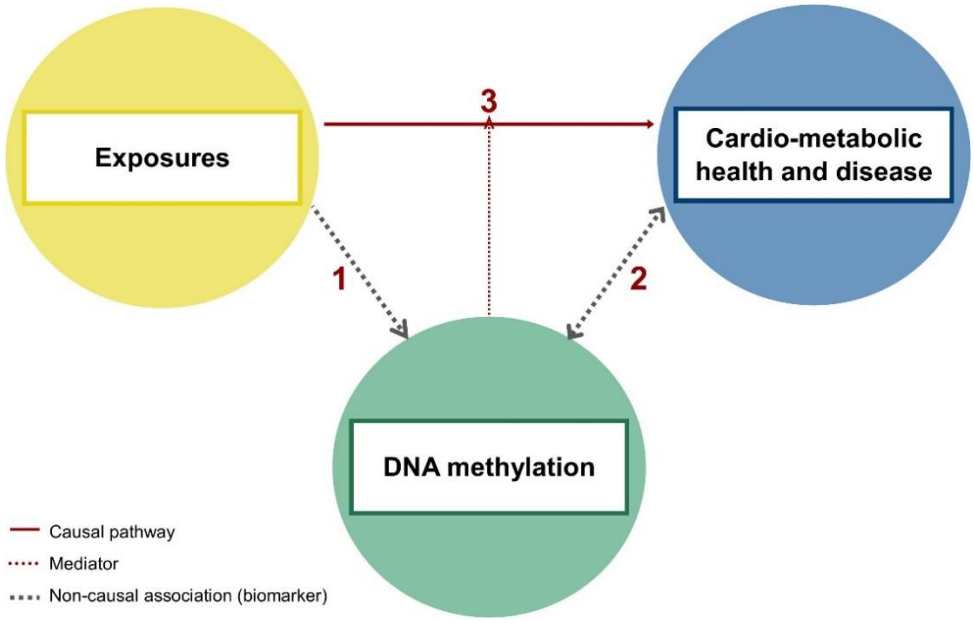


Figure 2-4 Investigation framework of epigenetics in epidemiology

2.3.1 Methylation as a biomarker of exposure

Methylation differences as biomarkers of exposures are particularly useful in cases where reporting bias is frequent, and a more reliable surrogate marker is needed. An example is the classification of smoking status and behaviour, which is often under- or over-reported in epidemiology (Klesges *et al.*, 1995; Stanton *et al.*, 1996). DNAm markers offer the ability to indicate true present and past smoking behaviour accurately (Bojesen *et al.*, 2017; Elliott *et al.*, 2014; Joehanes *et al.*, 2016; Zhang *et al.*, 2016). Similarly, accurate discrimination between heavy, light and non-drinkers using a methylation-based biomarker of alcohol consumption has been reported (Liu *et al.*, 2018). Evidence for the associations between global or targeted methylation or methylation age and specific environmental exposures (related to cardio-metabolic health) has accumulated in the last five years, including findings related to diet (Quach *et al.*, 2017), cigarette smoking (Bollepalli *et al.*, 2019; Joehanes *et al.*, 2016), alcohol consumption (Liu *et al.*, 2018), air pollution (Ding *et al.*, 2017), socio-economic status (Fiorito *et al.*, 2019; Needham *et al.*, 2015), physical activity (Dimauro *et al.*, 2016) and stress (Wolf *et al.*, 2017; Zannas *et al.*, 2015).

2.3.2 Methylation as a biomarker of cardio-metabolic disease

A great deal of literature has been published on methylation differences as biomarkers of cardio-metabolic disease or disease progression. Many EWASs have, for example, reported and validated associations between CpG methylation and diabetes (Meeks *et al.*, 2018), metabolic syndrome (Chitrala *et al.*, 2019), stroke (Davis Armstrong *et al.*, 2018), adiposity (Aslibekyan *et al.*, 2015; Meeks *et al.*, 2017) and inflammation (Braun *et al.*, 2017; Hedman *et al.*, 2017; Ligthart *et al.*, 2016). Similarly, DNAmAge and DNAmAgeAccel have been associated with obesity (Horvath *et al.*, 2014), CVD development (Lind *et al.*, 2018) and cardiovascular mortality (Perna *et al.*, 2016; Soriano-Tárraga *et al.*, 2017). Methylation-derived cell count ratios have been used as biomarkers for several cancers (Ambatipudi *et al.*, 2018a; Koestler *et al.*, 2017; Wiencke *et al.*, 2017) and rheumatoid arthritis (Ambatipudi *et al.*, 2018b), although this has not been investigated in the context of cardio-metabolic disease.

2.3.3 Methylation as a mediator

Apart from its role as a biomarker of either exposure or disease, methylation differences could be the mediator between genetic susceptibility, environmental exposures and cardio-metabolic disease risk/outcomes. The majority of the genome-wide significant associations reported to date are of single nucleotide polymorphisms (SNPs) residing in non-coding genomic regions, suggesting that the basis of these associations is the regulatory role of SNPs. The identification of numerous methylation and expression quantitative trait loci (mQTLs and eQTLs; SNPs associating with methylation levels or gene expression competency at a separate genomic location) has solidified the notion that DNAm could be a potential causal role player in the SNP-disease pathway (Bonder *et al.*, 2016; Zhernakova *et al.*, 2016). Mendelian randomisation is the most frequently used causal inference method and involves identification of robust associations between genetic variants and modifiable phenotypes as the anchor to evaluating the role of DNAm in the causal pathway (Haycock *et al.*, 2016; Relton & Davey Smith, 2010; Richardson *et al.*, 2017). The past years have brought about an accumulation of evidence for the potential mediatory role of DNAm in inflammation (Jhun *et al.*, 2017), diabetes (Elliott *et al.*, 2017), CVD (Richardson *et al.*, 2017) and adiposity (Mendelson *et al.*, 2017; Richmond *et al.*, 2016) – although the possibility of horizontal pleiotropy, genetic linkage, reverse causality and unmeasured confounding indicates the need for replicated, well-powered, well-phenotyped and genotyped, analytically rigorous studies. Mediation analyses investigating methylation as the mediator between specific exposure and disease without a genetic anchor have shown promise, particularly for the role of smoking-related methylation changes and cancer (Bojesen *et al.*, 2017; Jordahl *et al.*, 2019).

Figure 2-5 depicts the layout of this thesis in terms of the three approaches discussed above. Chapter 3 presents literature encapsulating all three these approaches in the form of a review. This review explores *urbanisation* as an amalgamated exposure and *non-communicable diseases* (NCDs) as the urbanisation-related outcome. Chapter 4, then, reports on epigenome-wide methylation associations with age, alcohol consumption, smoking status, body composition, blood lipid levels and C-reactive protein. Chapter 5 discusses the ability of DNAm indicators of cell count distribution to reflect inflammation, cardiovascular function and CVD risk. Finally, Chapter 6 reports the association of alcohol consumption and smoking habits with DNAmAge and DNAmAgeAccel.

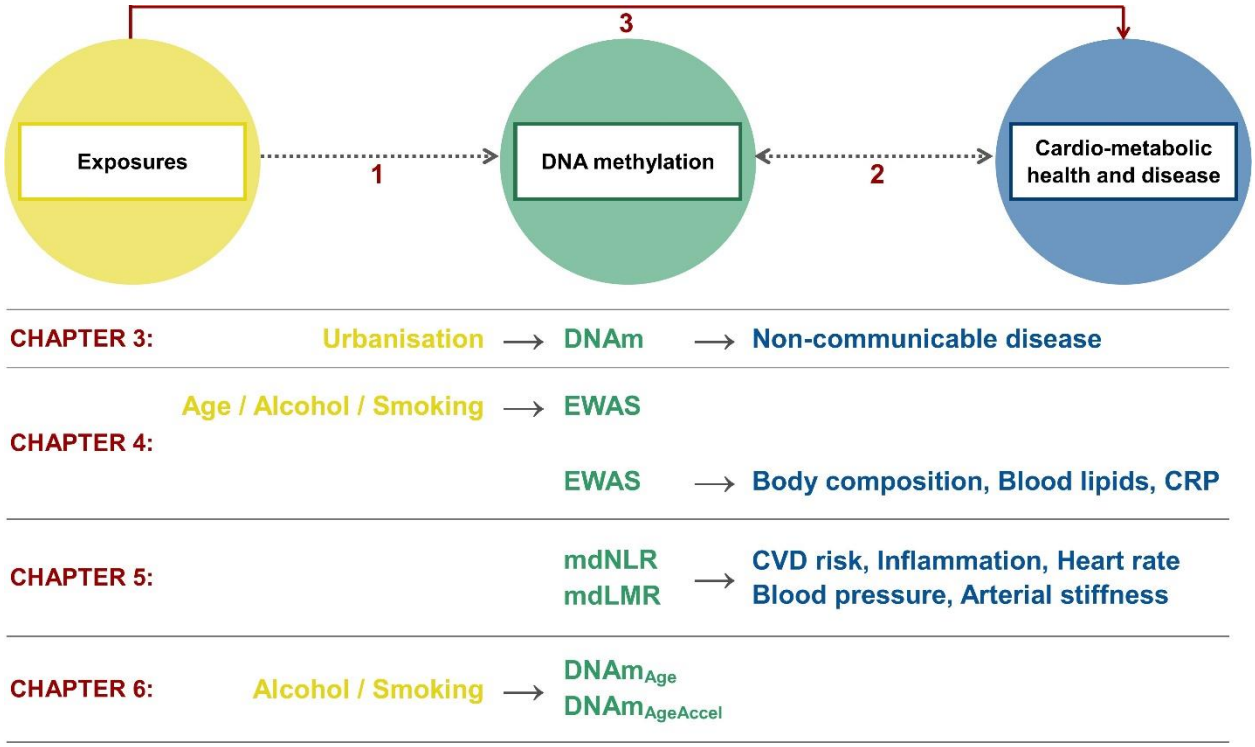


Figure 2-5 Thesis outline according to the three investigative approaches and three methylation quantification methods discussed

DNAm: DNA methylation; DNAmAgeAccel: DNA methylation age acceleration; CRP: C-reactive protein CVD: cardiovascular disease; EWAS: epigenome-wide association study; mdLMR: methylation-derived lymphocyte-to-monocyte ratio; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio.

2.4 Epigenetic epidemiology in the (South) African framework

The major informants of the epigenome are the genome, the environment and stochastic variability (Grundberg *et al.*, 2013; McRae *et al.*, 2014; Van Dongen *et al.*, 2016). The population-specific nature of genetics and environmental influencers requires epigenetic epidemiology research to incorporate multiple ancestral groups and environmental landscapes in order to fully

understand and contextualise global epigenetic variability. The South African framework is particularly beneficial in this regard.

Despite the recognised value of genomic research in Africans, most of the ethno-linguistic groups in Africa remain uncharacterised and research investigating genetic, particularly epigenetic, associations in African populations is scarce. This underrepresentation is partly because of financial limitations, scarcity in training and skilled researchers and a shortage of adequate health-service infrastructure, that result in a lack of proper record-keeping (Ramsay, 2012). The diversity within Africa also poses some language and cultural barriers between researchers and participants, which make recruiting, truly informing participants prior to consent being requested and getting to remote locations more challenging (Ramsay, 2012; Ramsay *et al.*, 2014). This section discusses the value of the South African context in terms of the benefit to be gained when its genomic and environmental characteristics are leveraged.

2.4.1 Genomic diversity

The methylome is regulated by the genome through: (i) polymorphic variation that determines the presence or absence of cytosine nucleotides with methylation potential (Galanter *et al.*, 2017), (ii) the efficacy of the proteins responsible for establishing, maintaining and removing methylation marks (Bjornsson et al., 2004; Lienert et al., 2011; Gutierrez-Arcelus et al., 2013; Krebs et al., 2014) and (iii) genetic variation that influences the methylation of CpG sites elsewhere in the genome (mQTLs) (Gaunt *et al.*, 2016; McRae *et al.*, 2014).

A vast body of literature describes the value of the continental African genome in biomedical research (Gurdasani *et al.*, 2015; Ramsay, 2012; Retshabile *et al.*, 2018). As Africa is the ancestral home of *Homo Sapiens* (Tishkoff *et al.*, 2009; Tishkoff & Williams, 2002), more genetic variation is observed among Africans than among their non-African counterparts (International HapMap 3 Consortium, 2010). A breakthrough in this regard was made when Schuster *et al.*, in 2010, reported more genetic variation between three San exomes than one would expect between an Asian and a European individual. In addition, low linkage disequilibrium (LD) has consistently been reported in African populations (Park, 2019). This enables researchers to narrow down on potential causal variants in African *vs* non-African populations (Cronjé *et al.*, 2017b; Jallow *et al.*, 2009; Teo *et al.*, 2010). The inherent depth of genetic diversity is coupled with a complex population structure comprising various ethnic and linguistic affiliations (Gurdasani *et al.*, 2015; Schlebusch & Jakobsson, 2018).

More consistent methylation patterns are observed within than between families (Bjornsson et al., 2008) and in monozygotic twin pairs *vs* dizygotic twin pairs or non-twin siblings (Grundberg *et al.*,

2013; McRae *et al.*, 2014). A recent investigation on disease-related mQTLs reported that common SNPs accounted for approximately 20% of methylation variance in mothers investigated during pregnancy and at middle age and their children at birth, during childhood and as adolescents, and that the effects of these mQTLs probably remain stable throughout life (Gaunt *et al.*, 2016). In terms of population-based investigations, a genome-wide study comparing 107 self-declared African American new-born babies with 94 Caucasian American new-born babies reported differential DNAm in whole blood in approximately 14% of the investigated CpGs (Adkins *et al.*, 2011). Teh *et al.* (2014), investigating new-born infants (Chinese, Indian and Malayan), reported that 25% of the most variable methylomic regions (quantified from the umbilical cord) were accounted for by genetics. Ethnic-specific patterns of DNAm, DNAmAge and DNAm-associations have since been reported in numerous population-based studies (Akinyemiju *et al.*, 2018; Barcelona *et al.*, 2019; Chitrala *et al.*, 2019; Liu *et al.*, 2018; Liu *et al.*, 2019a; Liu *et al.*, 2019b; Park *et al.*, 2018; Tajuddin *et al.*, 2019).

In terms of genomic diversity, investigating DNAm in Africans offers three main advantages. Firstly, it opens up the opportunity for the identification of novel, population-specific mQTLs. Secondly, it allows for rigorous validation of previous findings regarding environmental influence, because of its ability to contribute alternative genetic structures and in so doing, elucidate unknown genetic confounding. Lastly, the low LD in African populations makes them particularly helpful in statistical frameworks such as Mendelian randomisation. This characteristic allows better ability to narrow down on potential causality of identified mQTLs.

Chapter 4 of this thesis incorporates two of these research advantages by replicating EWAS findings from European and African American cohorts and exploring novel associations in our cohort. The degree to which current literature can be extrapolated to a South African cohort, such as the one investigated in this thesis, is not yet known. The implication is that, for this and many other underrepresented populations that have limited resources to conduct their own large-scale analysis, the ability to assume the applicability of current research findings is uncertain. Chapter 4 addresses this gap and contribute to the epigenetic epidemiology field by investigating novel associations.

2.4.2 Environment

The influence of environmental exposure commences *in utero*. The best known research in this regard centres around the Dutch hunger winter cohort (Heijmans *et al.*, 2008; Tobi *et al.*, 2014; Tobi *et al.*, 2009; Tobi *et al.*, 2018). Six decades after the Dutch lived through famine in the German-occupied Netherlands, same-sex siblings (with less than a five-year age difference between them), one of whom was exposed to famine during early gestation and the other not,

were investigated. Landmark findings from these investigations include vast methylation differences in regulatory genomic regions enriched for pathways related to growth and metabolism between sibling pairs (Tobi *et al.*, 2014) and the observation that some methylomic differences were specific to periconceptional deprivation, but not to deprivation in late gestation (Heijmans *et al.*, 2008). A recent publication provides evidence that periconceptional deprivation is linked to metabolic disease risk in adulthood (Tobi *et al.*, 2018). These findings have added to the molecular framework of the well-known Developmental Origins of Health and Disease hypothesis, which states that prenatal malnutrition is associated with metabolic diseases later in life, because of the offspring's inability to process nutritional affluence metabolically (Mandy & Nyirenda, 2018; Nyirenda & Byass, 2019). Differential methylation in new-born infants as a result of maternal smoking (Cardenas *et al.*, 2019; Joubert *et al.*, 2016a), plasma folate concentrations (Joubert *et al.*, 2016b) and level of education (Alfano *et al.*, 2018) has been observed, although, surprisingly, findings on maternal alcohol consumption have been unconvincing (Sharp *et al.*, 2018).

Apart from the extremes investigated in maternal and offspring health, evidence for environmental influence on methylation is also observed in the discrepancy in methylation between monozygotic twins as they grow older, even though they started their lives practically indistinguishably methylomically (Martin, 2005). More similar methylation profiles have also been observed in unrelated spouse-pairs that cohabit and, therefore, share an environment, than in those who live apart (Li *et al.*, 2018).

The African continent is known for its stark contrasts in environmental exposure, ranging from tropical forests and affluence to drought and immense food scarcity and poverty (Campbell & Tishkoff, 2008; Ramsay, 2012; Teo *et al.*, 2010). South Africa is no different. South Africa is home to eight distinct biomes, ranging from dessert to forest. The country is divided into nine provinces, including highly populated urbanised provinces (e.g. Gauteng province accommodating ~15 million residents in 18 178 km^2, which provides ~15% of the country's domestic product) to deeply rural less densely populated provinces (e.g. Northern Cape province accommodating ~1.2 million residents in 372 889 km^2, who are largely dependent on sheep farming for economic sustenance). Food insecurity estimates per province varied from 13% to 28% in 2016 (Stats SA, 2016). The percentage of households with access to a formal dwelling place and safe water ranged from 65% to 89% and 73% to 93%, respectively. The proportions of households sustained by agriculture varied between 4% and 28% (Government Communications, 2019; Stats SA, 2016).

These provincial discrepancies are factual indicators of the socio-economic divide still present in South Africa. Amidst the socio-economic extremities, the entire country is undergoing continual

and rapid urbanisation. The urban proportion of South African residents has risen by 10% in the past 20 years, with just over 66% of the population currently classified as urban (World Bank, 2018). The consequence of this urban transition has been a rapidly accelerating NCD epidemic added to the high burden of infectious disease still dominating rural regions. Of the reported natural deaths in the year 2016, 57% resulted from non-communicable and 31% from communicable diseases. Ten years prior, in 2006, the distribution was reversed, with 44% and 48% attributed to NCD and infectious diseases, respectively (Stats SA, 2018).

The increased NCD prevalence is largely the result of shifts in lifestyles related to urbanisation (Department of Health *et al.*, 2019; Nienaber-Rousseau *et al.*, 2017; Pieters & Vorster, 2008; Popkin, 2015; Vorster, 2002). For instance, urban populations in South Africa and around the world generally smoke more and consume alcohol more regularly (Dixon & Chartier, 2016; Department of Health *et al.*, 2019; Roberts *et al.*, 2016; WHO, 2017; WHO, 2018c). Urbanicity, furthermore, is positively associated with education, sedentary lifestyles and the consumption of fat and processed foods (Brathwaite *et al.*, 2017; Department of Health *et al.*, 2019; Wentzel-Viljoen *et al.*, 2018; WHO, 2018d). Differential methylation is associated with and provides biomarkers for most of these urban-rural discrepancies, from systemic (including population density, pollution, access to education and sanitation), to lifestyle-related (including diet, substance use and physical activity) characteristics (Bollepalli *et al.*, 2019; Dimauro *et al.*, 2016; Ding *et al.*, 2017; Fiorito *et al.*, 2019; Joehanes *et al.*, 2016; Liu *et al.*, 2018; Needham *et al.*, 2015; Olden *et al.*, 2015; Quach *et al.*, 2017; Smith *et al.*, 2017). Similarly, methylation is associated with the NCDs related to urbanisation (Aslibekyan *et al.*, 2015; Braun *et al.*, 2017; Chitrala *et al.*, 2019; Davis Armstrong *et al.*, 2018; Hedman *et al.*, 2017; Ligthart *et al.*, 2016; Meeks *et al.*, 2017).

Chapter 3 showcases a narrative review summarising evidence of the potential role of DNAm in the association between urbanisation (and the shifting environment it implies) and NCDs. It proposes the urban-rural divide present in many developing countries, such as South Africa, as a platform that can, and should, be leveraged for epigenetic epidemiology.

2.5 Conclusion

This chapter provided an overview of the current literature pertaining to the overarching theme of this thesis: “Epigenetic analysis of cardio-metabolic health in an African population”. A vast body of research supports the fact that epigenetic, particularly DNAm, modifications may be useful biomarkers and/or mediators in the framework of cardio-metabolic health and disease. This includes markers such as epigenetic age, genome-wide CpG associations and methylation-derived cell count estimators. Currently, epigenetic epidemiology is, however, deficient in its representation of genetic and environmental diversity. This limits the usefulness of current

knowledge in other ethnicities, but also limits the research field in its ability to replicate findings and narrow down on potentially causal markers to identify novel environmental or genetic targets. The African continent is characterised by incredible genetic and environmental diversity, although lack of adequate resources has restrained research in most of its regions. Validating current findings will allow for better understanding of the generalisability of what is largely European-based knowledge, favouring both the continent and the research environment. The following chapters present the manuscripts reporting on the research done in this thesis, which focuses on aspects pertaining to the South African framework as set out in this chapter.

CHAPTER 3

MANUSCRIPT ONE – NARRATIVE REVIEW

This manuscript has been accepted for publication in Epigenomics (EPI-2020-0049).

Publisher: Future Medicine

Impact factor: 4.40

Journal aims and scope:

Epigenomics provides the forum to address the rapidly progressing research developments in this ever-expanding field; to report on the major challenges ahead and critical advances that are propelling the science forward. The journal delivers this information in concise, at-a-glance article formats – invaluable to a time constrained community. Substantial developments in our current knowledge and understanding of genomics and epigenetics are constantly being made, yet this field is still in its infancy. Epigenomics provides a critical overview of the latest and most significant advances as they unfold and explores their potential application in the clinical setting.

Author's guidelines:

https://epigeneticsandchromatin.biomedcentral.com/submission-guidelines

LEVERAGING THE URBAN-RURAL DIVIDE FOR EPIGENETIC RESEARCH

H. Toinét Cronjé[1*], Hannah R. Elliott[2,3], Cornelie Nienaber-Rousseau[1], Marlien Pieters[1]

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

***Corresponding author:**

H. Toinét Cronjé

23520825@nwu.ac.za

Keywords: DNA methylation, non-communicable diseases, epigenetic epidemiology, urbanisation, LMICs

3.1 Abstract

Urbanisation coincides with a complex change in environmental exposure and a rapid increase in non-communicable diseases (NCDs). Epigenetics, including DNA methylation (DNAm), is thought to mediate part of the association between genetic/environmental exposure and NCDs. The urban-rural divide provides a unique opportunity to investigate the effect of the combined presence of multiple forms of environmental exposure on DNAm and the related increase in disease risk. This review evaluates the ability of three epidemiological study designs (migration, income-comparative and urban-rural designs) to investigate the role of DNAm in the association between urbanisation and the rise in NCD prevalence. We also discuss the ability of each study design to address the gaps in current literature, including the complex methylation-mediated risk attributable to the cluster of forms of exposure characterising urban and rural living, while providing a platform for developing countries to leverage their demographic discrepancies in future research ventures.

3.2 Introduction

Cancer, type 2 diabetes, respiratory disorders and cardiovascular diseases (CVDs) are collectively responsible for 80% of the 41 million deaths caused by non-communicable diseases (NCDs) each year [1]. These diseases result from the interplay between fixed genetic [2] and modifiable environmental and behavioural factors [3]. It is the modifiable factors, such as pollution and industrial toxin exposure, physical activity, diet, smoking and alcohol consumption, that are the current focus of NCD prevention and education worldwide [4].

Low- and middle-income countries (LMICs) currently carry 85% of the global NCD death toll. Adults in these regions face twice the risk of NCD mortality than their counterparts living in high-income countries (HIC) [1]. Although this discrepancy is partly accounted for by the lack of adequate health care and infrastructure, the main driver of the disproportionate burden of NCDs has been ascribed to the globalisation of lifestyles and environmental changes resulting from the rapid rate of urbanisation experienced by LMICs. At the same time, these regions remain vulnerable to malnutrition and infectious diseases and are therefore considered to carry the 'double burden of disease' [4, 5]. Urban areas are broadly defined by their dense population, commercial activity, non-agricultural employment, level of available education and infrastructure [6]. The strong link between the degree of urbanisation (urbanicity) and NCDs supports the notion of environmental factors being instrumental in the aetiology of NCDs, rather than purely inherited risk [5]. The possibility exists that differences in epigenetic regulation between urban and rural-dwelling individuals could be the basis by which the differences in environmental exposure shift disease prevalence with urbanicity.

The epigenome, unlike the genome, is environmentally modifiable and involves the alteration of gene expression without altering the genes themselves [7]. It is also influenced by genetic variation [8]. Epigenetic mechanisms include DNA methylation (DNAm), miRNAs, histone and chromatin modifications [7]. DNA methylation is currently the most studied epigenetic modification and has been implicated in the aetiology and progression of several NCDs over and above genetic predisposition [9]. Methylation differences can be (i) biomarkers of exposure that do not affect disease, (ii) part of the causal pathway between exposures and disease, or (iii) a biomarker of current disease [9]. Apart from its potential role as a disease mediator, the plasticity of DNAm also makes it a valuable topic of investigation owing to its intervention potential in preventative care [10].

To date, most methylation studies have investigated only single exposures or disease outcomes. This does not take into account that an individual, in any given environment, experiences a combination of exposures simultaneously such as those clustered together in rural and urban

landscapes. In this review we evaluate the ability of three epidemiological study designs to investigate the role of DNAm in the association between urbanisation and the NCDs. Investigating urbanisation, as a well-defined cluster of exposures, could allow for a better understanding of methylation's role in the global urbanisation–NCD trend. First, we review the current evidence for the urbanisation-NCD, urbanisation-DNAm and DNAm-NCD relationships as the theoretical backdrop to the research models discussed. We then discuss and compare the migration, income-comparative and urban-rural study designs in terms of the questions they are best suited to answer, suitable cohorts and their respective strengths and weaknesses. Figure 3-1 provides an illustrated summary of this review.

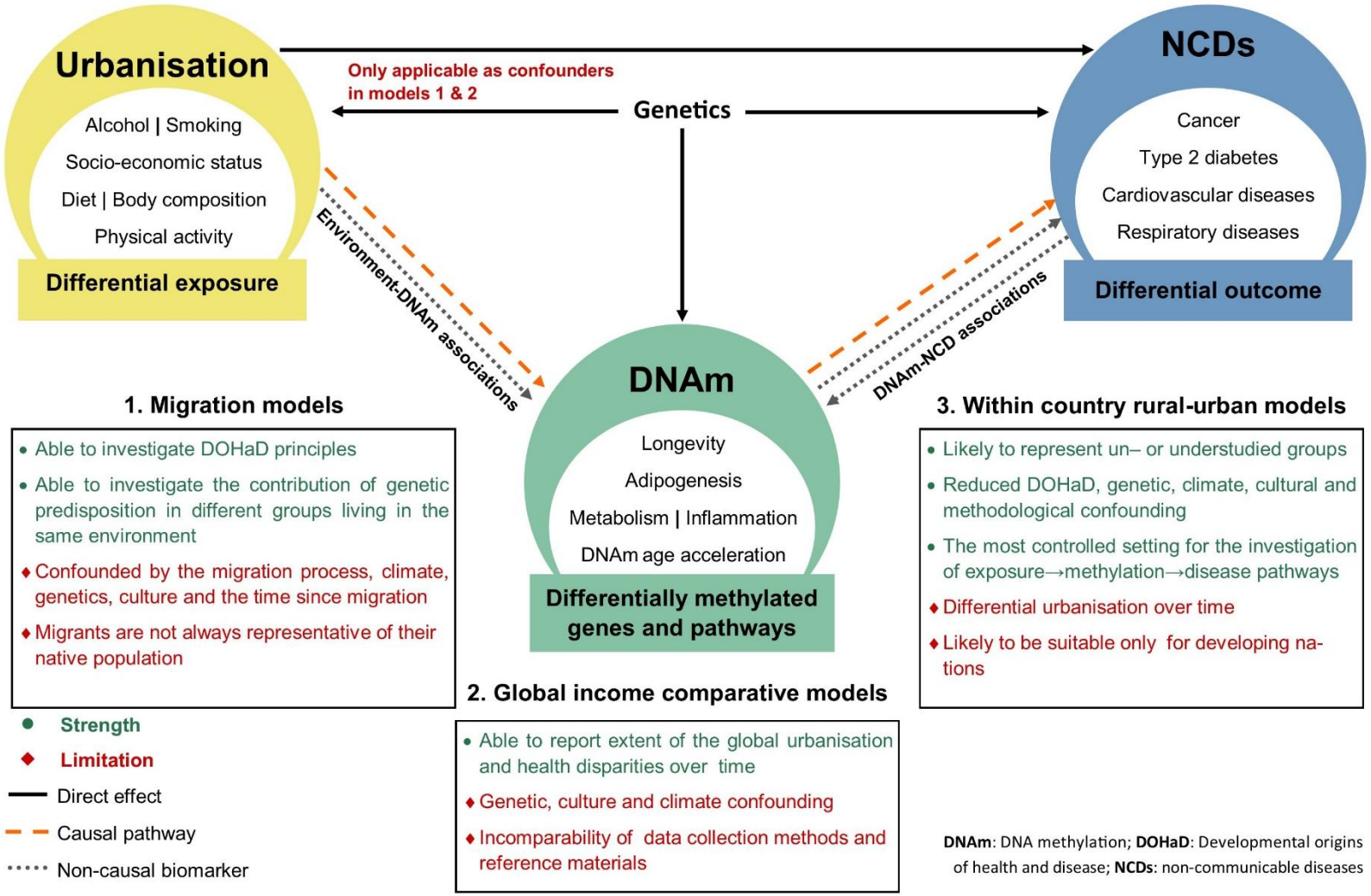


Figure 3-1 **The role of DNAm in mediating the association between urbanisation and NCDs and the strengths and limitations of different study designs aimed at investigating these associations**

3.3 Contextualising methylation

Genetics and environmental exposures both influence phenotype, partly through epigenetic modifications. Genetic architecture affects DNAm in terms of the availability of cytosine-phosphate-guanine (CpG) sites, the efficacy of methylation-related enzyme expression [8, 11], how DNAm responds to exposure [12] and inherent NCD risk [2]. Environmental exposures can be either external (behavioural or lifestyle-related factors such as diet, exercise and air purity) or internal (metabolic and biochemical factors such as inflammation and adiposity). The external environment's role in methylomic variance is best observed in monozygotic twins that start their lives methylomically indistinguishable (genetically determined), but grow ever more discordant throughout the life course [13]. Additionally, it has been observed that unrelated spouse-pairs that share an environment have more similar methylation profiles than those who live apart [14]. Methylomic variance attributed to the internal environment, on the other hand, is most prominent in the presence of disease, as a result of disease-related metabolic and biochemical changes [15, 16]. Three key aspects of the exposure-methylation-disease framework will be discussed in the sections to follow. These are also depicted in Figure 3-1.

3.3.1 Urbanisation is associated with NCD risk

The association between urbanisation and NCDs is predominantly driven by characteristics of the urban environment and behavioural factors. Urbanisation-associated environmental factors known to increase NCD risk include increased exposure to pollution and occupational toxins [3, 4]. Urbanisation-associated behavioural factors contributing to increased NCD risk include increased food availability [17], decreased non-recreational physical activity [18], a higher portion of energy intake from fat [19], protein [20] and processed foods [21], a reduction in relative energy from carbohydrates [19] and increased adiposity [21]. While smoking and alcohol consumption are known contributors to NCD risk, their relationship with urbanisation is more complex [3]. Urbanisation coincides with increased purchase of commercial tobacco products and exposure to tobacco-encouraging advertising, while second-hand smoke inhalation tends to be reduced [20, 22, 23]. Although urban individuals are less likely to be subject to alcohol abuse, they are more likely to consume alcohol during their lifetime [20, 24]. Urbanisation is also associated with increased psychosocial stress as a result of social inequity and exclusion, job insecurity and growing concerns of violence and crime [25]. These are particularly prevalent when urbanisation coincides with the growth of informal settlements within the urban landscapes [5]. Access to education, on the other hand, decreases NCD risk [26].

3.3.2 Urbanisation is associated with DNAm

The same types of exposure related to NCD risk, described above, have also been independently associated with altered DNAm. In this context mostly non-causal associations between exposures and DNAm have been investigated, although evidence for causal associations are accumulating (Figure 3-1). Evidence that DNAm is causally affected by urbanisation-related exposures (i.e. the exposure alters DNAm not vice versa) have been published for body mass index [27] and smoking [28]. Genome-wide [29] and gene-specific [27] investigations have reported adiposity-related methylation changes. These methylation signatures have been used successfully to predict the efficacy of weight-loss interventions [30]. Smoking status, cumulative smoking exposure and smoking intensity can also be determined using DNAm as biomarker [31, 32]. Smoking-associated differential DNAm is only partially reversible upon cessation [31, 32]. Similarly, heavy drinking can be identified using a methylation-based biomarker [33]. Methylation differences associated with alcohol consumption seem to be completely reversible, indicating possible causality [33].

Regarding non-casual associations, dietary patterns, such as high fat [34] and Western diets [35], have been associated with methylation differences in genes involved in lipid metabolism, adipogenesis, inflammation and glucose regulation. Physical activity-based intervention studies reported beneficial methylation and transcription changes in genes related to longevity, inflammation and metabolism in blood, skeletal muscle and adipose tissue [36-38].

Methylation levels at specific CpGs have also been incorporated into methylation-derived biological age predictors [39, 40]. The discrepancy between DNAmAge and chronological age is referred to as biological/methylation age acceleration (DNAmAgeAccel) and, when positive, is used as a biomarker of accelerated cellular ageing [39, 40]. Urbanicity-related factors such as adiposity [41], meat consumption [41] and cigarette smoking [42] are associated with DNAmAgeAccel. Alcohol associates with DNAmAge in a non-linear manner, where light and heavy drinkers experience DNAmAgeAccel and moderate alcohol consumers have a relative deceleration of DNAmAge [43, 44]. Education, aligned with its negative association with NCD risk [26], also protects against DNAmAgeAccel [42, 43].

In terms of the urban environment, a vast amount of literature has reported on the genome-wide, global and gene-specific DNAm associations with exposure to general pollution and distinct pollutants [45]. Increased exposure to traffic-related air pollution, for example, has been associated with altered methylation at the Ten-eleven translocation 1 gene, which encodes a key enzyme involved in DNA demethylation [46]. A dose-response association between traffic-related pollution and DNAm changes has also been observed [47]. Accelerated DNAmAge has also

been observed in groups exposed to pollution and pesticides [48]. From a social environment point of view, neighbourhood unity, aesthetics and safety have been associated with favourable DNAm and downstream expression changes in particularly stress- and inflammation-related genes [49, 50]. Such neighbourhood characteristics also enable outdoor recreational physical activity that in itself has proven beneficial [5], although these characteristics are likely to be largely present in urban-dwellers of high socio-economic status [5, 51].

3.3.3 DNAm is associated with NCDs

Associations between DNAm and NCDs have been reported in both directions (Figure 3-1), as exposure-related differential DNAm may precede disease (DNAm being on the causal pathway between exposure and disease), and, conversely, disease-related metabolic changes can affect DNAm (DNAm as a non-causal biomarker of disease). Investigations into the potential causal influence of DNAm on type 2 diabetes [52] and CVD development [53] are increasing, although conclusive evidence is yet to be published. As non-causal biomarkers, both DNAm and DNAmAge have been used to identify several sub-types of cancer [54, 55] and CVD [56, 57]. Tumorous tissues are epigenetically older than their non-cancerous counterparts [58]. As a prognostic marker, DNAm, particularly DNAmAge, has been useful in predicting cancer incidence [59, 60] and survival [61], cardiac events [62], premature CVD [40, 63] and all-cause mortality [60] independent of traditional risk factors. Lastly, as an intervention strategy, methylation-altering drugs are proving to be increasingly successful in the treatment of CVD [64], cancer [65, 66] and type 2 diabetes [67].

3.3.4 The missing link

Collectively, the evidence summarised in the previous sections highlight the potential role of DNAm as a mediator between urbanisation and NCDs. The only robust evidence that has been able to link the environment to DNAm, and then DNAm to disease concerns the relationship between smoking and bladder cancer in postmenopausal women [68]. Preliminary findings have, however, associated BMI-related changes in DNAm with cardio-metabolic disease development [29]. In addition, a randomised controlled trial has provided some evidence of DNAm mediating the association between exposure to air pollution and adverse cardiovascular profiles [69].

The main gaps remaining in the literature, and the best way to address them, are the topic of interest of this review. Key questions include: i) Have we identified all the key risk factors in the urbanicity-NCD relationship? ii) How does the research currently address the amalgamated risk posed by the entirety of the urban vs rural context? iii) How can we investigate and understand the role of DNAm in this lifestyle-disease model better? Thus far, the role of DNAm in NCDs has

been investigated typically by focusing on one form of exposure at a time. Investigating DNAm in the context of urbanisation provides the opportunity to aggregate NCD-related exposures to provide not only a more accurate reflection of the amalgamated disease risk associated with urbanicity, but also to start identifying currently unknown contributing factors that explain the variance in risk after accounting for all the known single forms of exposure. Such an investigation could also provide insight regarding the extent of potential additive risk compared to a wash-out effect when numerous methylation-altering exposures are clustered together. By identifying DNAm mediators involved in the relationship between urbanisation and NCD we might find modifiable targets for improving population health. The sections to follow evaluate the ability, strengths and weaknesses of different approaches to best answer key questions and elaborate on our current understanding of the role of DNAm in urbanisation-related population health (Figure 3-1).

3.4 Contextualising urbanisation

Urbanisation can be driven by the net movement of individuals from rural to urban residency in search of, among others, better education or health care, economic success, safety or food security. In this context, urbanisation can be the result of individuals moving from a rural to an urban community within their own countries or to another country/culture entirely (migration). Alternatively, urbanisation can occur as a specific region progressively urbanises. There are three main epidemiological approaches that can be used to investigate the health-related consequences of urbanisation: the migration, income-comparative and urban-rural divide approach.

3.4.1 Migration models

Migration studies are able to investigate the effects of environmental shifts in two ways. The first studies groups of similar ancestral and geographical origin, living in different countries, such as those who remained in the home country compared to those who moved to different locations [70]. These studies are useful in that they allow investigation of an altered environment while controlling for early life exposure and ethnicity. A study of Japanese migrants, for example, reported a dramatic increase in CVD risk in the migrant compared to the non-migrant group, providing evidence that the environmental shift increased CVD risk in this population [71]. A second approach is the comparison of the migrant population with their host. In these cases, ethnic inequalities in NCD risk and outcomes can be assessed, for example the major differences in the rate of specific CVD incidence and mortality between migrants from different countries and the same HIC host population [70].

The complexity of using migration models to investigate the association of urbanisation with health stems from the numerous migration-associated confounders that are not part of the rural-urban shift. First, the migrant group themselves are not necessarily reflective of the group they originated from in that migrants are often better resourced to enable their migration than those staying behind [72]. Second, the circumstances of migration complicate the separation of the health effects related to the rural-urban shift, through the stress and impact of the migration itself [72, 73]. Third, migrants often experience a vast shift in culture and climate. In addition, cultural differences in health-seeking behaviour may lead to a lack of timely disease diagnosis or non-compliance in treatment, particularly in new migrants [74]. Lastly, migration timing may profoundly influence the relative severity of the effect of urbanisation on health outcomes. According to the Developmental Origins of Health and Disease hypothesis, individuals born in an environment with limited nutritional resources (often rural settlements) are prenatally programmed to survive in these conditions. If they are then subsequently exposed to nutritional abundance, they are not metabolically equipped to manage this affluence and are pre-disposed to develop NCDs [75, 76].

3.4.2 Income-comparative models

An alternative model that can be used to investigate urbanisation is a global income-comparative research model where groups of differing demographic backgrounds from across the world are compared with respect to disease prevalence. The largest investigation currently implementing such a study design is the international Prospective Urban and Rural Epidemiology (PURE) study. The study includes 225 000 individuals residing in 27 low-, middle- and high-income countries [77]. The PURE study has provided many insights on global NCD risk progression and contributors. Meta-analyses by the PURE cohort include topics such as carbohydrate and fat intake [78] fruit, vegetable and legume intake [79], dietary nutrients [78], education [26], physical activity [80] and alcohol consumption [81] in relation to CVD and its related health outcomes such as blood lipid concentrations and blood pressure. The primary focus of these meta-analyses is better understanding of the relationship between country-income classification, subsequent exposure and the influence of this on CVD incidence and mortality.

Although these investigations provide vital information on the extent of the NCD crisis, particularly in LMICs, this approach is limited in a few ways. First, it is often unable to account for the genetic diversity risk of specific groups when comparing and combining multiple ethnicities. It is widely known that the risk models, ranges and cut-off created for one population are not always indicative of the same risk variance in other populations [82]. For this reason, continual attempts are being made by the World Health Organisation to recalibrate NCD risk models that are currently used in HICs to be used in LMIC population groups [83].

The numerous genome-wide association analyses that have indicated ethnic differences in genotype frequencies and their associations with intermediate phenotypes and ultimate risk indicate that although phenotypic risk assessment might be the most feasible, the gap in risk variability might only be fully addressed when also considering genetic contributors [2]. In the epigenetic context, methylation differences have also been reported among ethnic groups [11].

Second, as with migrant studies, differing geographical locations also introduce confounding by climate and diet [21, 84]. Cross-cultural adaption of data collection methods is critical in these cases, as reference material developed for one population might leave many factors unstudied in the population to which it is applied, purely because of their absence in the reference group [85]. Many developing nations remain severely underrepresented in genetic and epigenetic research, suggesting that the driver of observed DNAm associations might not have been identified or studied previously, resulting in potentially unquantifiable confounding when comparing these population groups [86].

Lastly, although socio-economic status is associated with urbanicity, national economic status (such as the World Bank status used in most income-comparative models) does not reflect urbanicity. Developing nations, for example, are often LMICs, but generally have urban capitals, informal urban and rural settlements and rural agricultural landscapes [87].

3.4.3 Within-country rural-urban models

A third approach that can be used to investigate the process of progressive urbanisation is to consider those who do not undergo urbanisation during their lifetime, but are, instead, subject to the urban-rural divide still common in developing countries [51]. This research design can be used under the condition of having a cohort that represents communities of a single genetic origin, part of which resides in an urban, and the other in a rural area, and does so for its entire lifespan. These individuals should have been born, and remained, on either the rural or urban side of the socio-demographic divide throughout their lifetime. A cohort of this nature will allow for the investigation of discrepant environmental exposure and health outcomes while limiting many of the confounding factors discussed in the previous sections.

Two examples of large-scale studies that can leverage this approach are the PURE [77] and the Research on Obesity and Diabetes among African Migrants (RODAM, [88]) cohorts. Although spanning continents, all the countries participating in the PURE study contribute a variety of both rural and urban sub-cohorts [87]. The RODAM study, on the other hand, includes a rural and an urban site in Ghana (Africa), in addition to the Ghanaian migrants residing in Europe [88].

Only one of the PURE sub-cohorts has published epigenetic data [89], but many of the other sub-cohorts have access to previously collected peripheral blood samples in cryo-storage facilities [77]. No epigenetic data from the PURE cohort have been used to investigate urban-rural disparities. The advantage of a cohort such as PURE is the availability of longitudinal data on the disparity between large well-defined urban and rural communities in at least 27 countries [87]. The RODAM cohort has published genome-wide DNAm data in relation to obesity [90] and type 2 diabetes [91], although no urban-rural epigenetic comparisons have been made to date. Although the RODAM study is currently of cross-sectional design, there are plans to transform it into a longitudinal cohort [92]. The PURE cohort was established in 2003, and the RODAM cohort in 2012. These cohorts are, therefore, able to capture urbanisation at the pace it is currently experienced [77].

The country-specific urban-rural research platform has the benefit of being able to investigate the clusters of types of exposure that represent rural or urban living, while factors such as genetics, climate and geographical influencers remain constant and similar between groups. Developing nations, such as those included in the PURE and RODAM studies, are particularly likely to benefit from this approach, as the urban-rural divide is most severe in these countries. Furthermore, particularly in the context of epigenetic epidemiology, these countries often contain many under- or unstudied ethnic groups. Currently, most of the available evidence on NCDs, NCD risk factors and the role of epigenetics originates from study populations in developed countries [83, 93]. As there are vast genomic and socio-economic differences between these countries and the ethnic groups they contain [2, 11], the feasibility of simply extrapolating findings from what is largely HIC European literature is unknown. Inclusion of more LMICs in large-scale research efforts will, therefore, not only provide an opportunity to generate population-relevant information to inform prevention, detection and treatment of NCDs in these countries, but will also contribute to closing the knowledge gaps in the global literature. Findings from such investigations will provide external validation of generalisable findings, while highlighting the circumstances where population-specific research is needed.

3.5 Current challenges

One of the limitations of most epigenetic investigations is the unavailability of disease-relevant tissues. It has been well established that DNAm signatures differ among tissues, although the available evidence on environment-methylation-disease patterns is almost exclusively derived from blood-based methylation investigations [94]. Urbanicity-associated DNAm changes are, therefore, more likely to be leveraged as biomarkers of exposure or disease indicators, as the unavailability of target tissues for specific disease or outcomes limits causal inference. Mediation analyses such as Mendelian randomisation could be employed to help with this, although multiple

causal inference methods might be needed for triangulation of evidence [52, 95]. Because, to our knowledge, population-specific genomic data are not available for many LMICs currently investigated, the addition of genetic data will be a valuable contribution.

As progressive urbanisation is likely to affect both the rural and urban groups in LMICs, longitudinal measurements of DNAm will be a beneficial and informative resource. Research has shown that altered environments affect health at different rates. Adiposity, for example, seems to increase rapidly once individuals relocate to urban areas, whereas fasting insulin increases at a much more gradual pace [96]. Cross-sectional representations of urbanicity and health are therefore limited, as they capture only the factors that have an impact at the specific point in time. Should longitudinal data collection be performed, not only will there be better control of the epidemiological transition over time, but this will also allow researchers to address the gap in longitudinal epigenetic research in terms of causality, reversibility and/or stability of DNAm. Standardised protocols for blood collection, handling and storage will be critical in avoiding the limitation of time-point-related batch variance.

Leveraging richly phenotyped, genetically similar, rural and urban communities with genome-wide epigenetic data and the ability to track NCD risk progression and mortality prospectively provides a unique opportunity to investigate the full environment → DNAm → NCD framework where such pathways exist, and where they do not, the value of DNAm as a biomarker for either environmental exposure or existing disease risk can be evaluated.

3.6 Future perspectives

As the 21st century continues to be marked by urbanisation, it is essential to improve our understanding of the molecular mechanisms driving the effect of the environmental shift on NCD prevalence and incidence. The current global landscape allows numerous approaches to be taken to investigate these mechanisms, each with its own strengths, limitations and answerable questions. In the era of big data and the continual pressure of the scientific community to promote open access and increase data availability, we expect the use of these models to significantly add to the genetic and environmental diversity captured in global epigenetic epidemiology data. Integrating the knowledge gained from the different perspectives of each of these three models will allow for a more holistic view of the different genetic and environmental origins of disease and the epigenetic mechanisms that bridge them. Ultimately, it is within the rounded understanding of methylation's role in the urbanisation-NCD relationship that modifiable targets can be identified to translate research to population-based NCD prevention strategies.

Executive summary

The non-communicable disease (NCD) death toll is rising globally

- In LMIC, this is thought to be the result of urbanisation

DNA methylation (DNAm) could mediate the urbanisation-disease relationship

- Urbanisation-related exposures associates with DNAm
- DNAm associates with NCD risk factors and outcomes
- Urbanisation associates with NCD risk factors and outcomes

The following models can be used to explore this hypothesis

- *Migration model*

This model is particularly useful when investigating genetic predisposition and the developmental origins of health and disease. The migration process itself may, however, confound associations. Keep in mind that migrants are not always representative of their native population.

- *Global income comparative model*

This model can be used to report the extent of global urbanisation and health disparities over time, but is susceptible to genetic, cultural and climate confounding. Comparable data collection methods and reference material is a necessity when this model is used.

- *Within country urban-rural models*

This model significantly reduces abovementioned confounding and is, therefore, the most controlled setting to explore the hypothesis of DNAm mediating the effect of urbanisation on NCD risk. This model is likely to be suitable only for developing nations, but can, therefore, be used when investigating un- or understudied populations.

Integrating knowledge gained from these three models is the key

Integrating the knowledge gained from the different perspectives of each of these three models will allow for a more holistic view of the different genetic and environmental origins of disease and the epigenetic mechanisms that bridge them. Ultimately, it is within the rounded understanding of methylation's role in the urbanisation-NCD relationship that modifiable targets can be identified to translate research to population-based NCD prevention strategies.

3.7 Declarations

Funding

This work was supported by the South African National Research Foundation [SFH106264 to H. Toinét Cronjé]; and the Newton Fund Foundation [AMS-NAF1-Pieters to Marlien Pieters]. Hannah R. Elliott works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol [MC_UU_00011/5].

Authors' contributions

All authors contributed to the study concept and design. H. Toinét Cronjé performed the literature search and wrote the first draft. All authors critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

3.8 References

1. World Health Organization. Noncommunicable diseases. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. (2018).

2. Pranavchand R, Reddy B. Genomics era and complex disorders: Implications of GWAS with special reference to coronary artery disease, type 2 diabetes mellitus, and cancers. *Journal of postgraduate medicine* 62(3), 188-98 (2016).

3. Gakidou E, Afshin A, Abajobir AA *et al*. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet* 390(10100), 1345-1422 (2017).

4. World Health Organization. World health statistics 2018: monitoring health for the SDGs, sustainable development goals. https://www.who.int/gho/publications/world_health_statistics/2018/en/ (2018).

5. Miranda JJ, Barrientos-Gutiérrez T, Corvalan C *et al*. Understanding the rise of cardiometabolic diseases in low- and middle-income countries. *Nature Medicine* 25(11), 1667-1679 (2019).

****An excellent review on the broader context of NCDs rising in LMICs including environmental and macro drivers of NCDs and valuable thoughts on how LMICs (as part of a global community) can respond to the mounting threat.**

6. Mcgranahan G, Satterthwaite D. *Urbanisation: concepts and trends*. IIED working paper; IIED London, (2014).

7. Kim J, Samaranayake M, Pradhan S. Epigenetic mechanisms in mammals. *Cellular and molecular life sciences* 66(4), 596-612 (2009).

8. Gaunt TR, Shihab HA, Hemani G *et al*. Systematic identification of genetic influences on methylation across the human life course. *Genome biology* 17(1), 61 (2016).

9. Sharp GC, Relton CL. Epigenetics and noncommunicable diseases. *Epigenomics* 9(6), 789-791 (2017).

10. Jin Z, Liu Y. DNA methylation in human diseases. *Genes & diseases* 5(1), 1-8 (2018).

11. Galanter JM, Gignoux CR, Oh SS *et al*. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *eLife* 6,e20532 (2017).

12. Chitrala KN, Hernandez DG, Nalls MA *et al*. Race-specific alterations in DNA methylation among middle-aged African Americans and Whites with metabolic syndrome. *Epigenetics* DOI: 10.1080/15592294.2019.1695340 (2019).

13. Talens RP, Christensen K, Putter H *et al*. Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging cell* 11(4), 694-703 (2012).

14. Li S, Wong EM, Dugué P-A *et al*. Genome-wide average DNA methylation is determined in utero. *International journal of epidemiology* 47(3), 908-916 (2018).

15. Ligthart S, Marzi C, Aslibekyan S *et al*. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome biology* 17(1), 255 (2016).

16. Wahl S, Drong A, Lehne B *et al*. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 541(7635), 81-6 (2017).

17. NCD Risk Factor Collaboration. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *Lancet* 390(10113), 2627-2642 (2017).

18. World Health Organization. Prevalence of insufficient physical activity. https://www.who.int/gho/ncd/risk_factors/physical_activity_text/en/ (2018).

19. Dehghan M, Mente A, Zhang X *et al*. Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (PURE): a prospective cohort study. *Lancet* 390(10107), 2050-2062 (2017).

20. Department of Health (South Africa), Statistics South Africa, South African Medical Research Council, & International Coach Federation. South Africa demographic and health survey 2016. Pretoria, South Africa and Rockville, Maryland, USA (2019)

21. Swinburn BA, Kraak VI, Allender S *et al*. The Global Syndemic of Obesity, Undernutrition, and Climate Change: The Lancet Commission report. *Lancet* 393(10173), 791-846 (2019).

22. Chow CK, Corsi DJ, Gilmore AB *et al*. Tobacco control environment: cross-sectional survey of policy implementation, social unacceptability, knowledge of tobacco health harms and relationship to quit ratio in 17 low-income, middle-income and high-income countries. *BMJ open* 7(3), e013817 (2017).

23. Brathwaite R, Addo J, Kunst AE *et al*. Smoking prevalence differs by location of residence among Ghanaians in Africa and Europe: The RODAM study. *PloS one* 12(5), e0177291 (2017).

24. Dixon MA, Chartier KG. Alcohol use patterns among urban and rural residents: demographic and social influences. *Alcohol research: current reviews* 38(1), 69-77 (2016).

25. Pikhart H, Pikhartova J. The relationship between psychosocial risk factors and health outcomes of chronic diseases: a review of the evidence for cancer and cardiovascular diseases. WHO regional office for Europe, Copenhagen (2015)

26. Jacobs DR, Kromhout D. Education, diet, and incident cardiovascular disease: ecological interactions and conclusions. *Lancet global health* 7(6), e684-e5 (2019).

27. Richmond RC, Sharp GC, Ward ME *et al*. DNA methylation and body mass index: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* 65(5), 1231-44 (2016).

28. Li S, Wong EM, Bui M *et al*. Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study. *Clinical epigenetics* 10(1), 18 (2018).

29. Mendelson MM, Marioni RE, Joehanes R *et al*. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: A Mendelian randomization approach. *PLoS medicine* 14(1), e1002215 (2017).

30. Perez-Cornago A, Mansego ML, Zulet MA, Martinez JA. DNA hypermethylation of the serotonin receptor type-2A gene is associated with a worse response to a weight loss intervention in subjects with metabolic syndrome. *Nutrients* 6(6), 2387-2403 (2014).

31. Joehanes R, Just A, Marioni R *et al*. Epigenetic signatures of cigarette smoking. *Circulation. Cardiovascular genetics* 9(5), 436-447 (2016).

32. Dugué P-A, Jung C-H, Joo JE *et al*. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. *Epigenetics* doi:10.1080/15592294.2019.1668739 1-11 (2019).

33. Liu C, Marioni RE, Hedman ÅK *et al*. A DNA methylation biomarker of alcohol consumption. *Molecular psychiatry* 23(2), 422 (2018).

34. Zhang P, Chu T, Dedousis N *et al*. DNA methylation alters transcriptional rates of differentially expressed genes and contributes to pathophysiology in mice fed a high fat diet. *Molecular metabolism* 6(4), 327-339 (2017).

35. Yokoyama AS, Dunaway K, Rutkowsky J, Rutledge JC, Milenkovic D. Chronic consumption of a western diet modifies the DNA methylation profile in the frontal cortex of mice. *Food & function* 9(2), 1187-1198 (2018).

36. Hibler E, Huang L, Andrade J, Spring B. Impact of a diet and activity health promotion intervention on regional patterns of DNA methylation. *Clinical Epigenetics* 11(1), 133 (2019).

37. Dimauro I, Scalabrin M, Fantini C *et al*. Resistance training and redox homeostasis: Correlation with age-associated genomic changes. *Redox biology* 10 34-44 (2016).

38. Grazioli E, Dimauro I, Mercatelli N *et al*. Physical activity in the prevention of human diseases: role of epigenetic modifications. *BMC genomics* 18(8), 802 (2017).

39. Horvath S. DNA methylation age of human tissues and cell types. *Genome biology* 14(10), 3156 (2013).

40. Lu AT, Quach A, Wilson JG *et al*. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* 11(2), 303 (2019).

41. Dugué P-A, Bassett JK, Joo JE *et al*. Association of DNA methylation-based biological age with health risk factors and overall and cause-specific mortality*. American journal of epidemiology* 187(3), 529-538 (2017).

42. Zhao W, Ammous F, Ratliff S *et al*. Education and lifestyle factors are associated with DNA methylation clocks in older African Americans*. International journal of environmental research and public health* 16(17), 3141 (2019).

43. Quach A, Levine ME, Tanaka T *et al*. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors*. Aging* 9(2), 419-446 (2017).

44. Rosen AD, Robertson KD, Hlady RA *et al*. DNA methylation age is accelerated in alcohol dependence*. Translational psychiatry* 8(1), 182 (2018).

45. Martin EM, Fry RC. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations*. Annual review of public health* 39 309-333 (2018).

46. Somineni HK, Zhang X, Myers JMB *et al*. Ten-eleven translocation 1 (TET1) methylation is associated with childhood asthma and traffic-related air pollution*. Journal of allergy and clinical immunology* 137(3), 797-805 (2016).

47. Ding R, Jin Y, Liu X *et al*. Dose-and time-effect responses of DNA methylation and histone H3K9 acetylation changes induced by traffic-related air pollution*. Scientific reports* 7, 43737 (2017).

48. Ryan J, Wrigglesworth J, Loong J, Fransquet PD, Woods RL. A systematic review and meta-analysis of environmental, lifestyle, and health factors associated With DNA methylation age*. The journals of gerontology: series A* 75(3), 481-94 (2019).

49. Needham BL, Smith JA, Zhao W *et al*. Life course socioeconomic status and DNA methylation in genes related to stress reactivity and inflammation: The multi-ethnic study of atherosclerosis*. Epigenetics* 10(10), 958-969 (2015).

50. Smith JA, Zhao W, Wang X *et al*. Neighborhood characteristics influence DNA methylation of genes involved in stress response and inflammation: The multi-ethnic study of atherosclerosis*. Epigenetics* 12(8), 662-73 (2017).

51. Zhang XQ. The trends, promises and challenges of urbanisation in the world*. Habitat International* 54, 241-252 (2016).

***This review contextualises the economic, social and environmental impacts of urbanisation which need to be well understood before focussing on any one discipline (e.g. NCDs).**

52. Elliott HR, Shihab HA, Lockett GA *et al*. The role of DNA methylation in Type 2 diabetes aetiology–using genotype as a causal anchor*. Diabetes* 66(6):1713-22 (2017).

53. Richardson TG, Zheng J, Smith GD *et al*. Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk*. The American journal of human genetics* 101(4), 590-602 (2017).

54. Vrba L, Futscher BW. A suite of DNA methylation markers that can detect most common human cancers*. Epigenetics* 13(1), 61-72 (2018).

55. Patel PG, Wessel T, Kawashima A *et al*. A three-gene DNA methylation biomarker accurately classifies early stage prostate cancer*. The prostate* 79(14), 1705-1714 (2019).

56. Muka T, Koromani F, Portilla E *et al*. The role of epigenetic modifications in cardiovascular disease: a systematic review*. International journal of cardiology* 212 174-183 (2016).

57. Ward-Caviness CK, Agha G, Chen BH *et al*. Analysis of repeated leukocyte DNA methylation assessments reveals persistent epigenetic alterations after an incident myocardial infarction*. Clinical Epigenetics* 10(1), 161 (2018).

58. Hannum G, Guinney J, Zhao L *et al*. Genome-wide methylation profiles reveal quantitative views of human aging rates*. Molecular cell* 49(2), 359-367 (2013).

59. Zheng Y, Joyce BT, Colicino E *et al*. Blood epigenetic age may predict cancer incidence and mortality*. EBioMedicine* 5 68-73 (2016).

60. Levine ME, Lu AT, Quach A *et al*. An epigenetic biomarker of aging for lifespan and healthspan*. Aging* 10(4), 573 (2018).

61. Dugué PA, Bassett JK, Joo JE *et al*. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies*. International journal of cancer* 142(8), 1611-9 (2018).

62. Lind L, Ingelsson E, Sundström J, Siegbahn A, Lampa E. Methylation-based estimated biological age and cardiovascular disease*. European journal of clinical investigation* 48(2), (2018).

63. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clinical epigenetics* 8(1), 64 (2016).

64. Watson CJ, Horgan S, Neary R *et al*. Epigenetic therapy for the treatment of hypertension-induced cardiac hypertrophy and fibrosis. *Journal of cardiovascular pharmacology and therapeutics* 21(1), 127-137 (2016).

65. Liang G, Weisenberger DJ. DNA methylation aberrancies as a guide for surveillance and treatment of human cancers. *Epigenetics* 12(6), 416-432 (2017).

66. Miranda Furtado CL, Dos Santos Luciano MC, Silva Santos RD, Furtado GP, Moraes MO, Pessoa C. Epidrugs: targeting epigenetic marks in cancer treatment. *Epigenetics* 14(12), 1164-76 (2019).

67. Arguelles AO, Meruvu S, Bowman JD, Choudhury M. Are epigenetic drugs for diabetes and obesity at our door step? *Drug discovery today* 21(3), 499-509 (2016).

68. Jordahl KM, Phipps AI, Randolph TW *et al*. Differential DNA methylation in blood as a mediator of the association between cigarette smoking and bladder cancer risk among postmenopausal women. *Epigenetics* 14(11), 1065-1073 (2019).

69. Chen R, Meng X, Zhao A *et al*. DNA hypomethylation and its mediation in the effects of fine particulate air pollution on cardiovascular biomarkers: a randomized crossover trial. *Environment international* 94, 614-619 (2016).

70. Agyemang C, Van Den Born B-J. Non-communicable diseases in migrants: an expert review. *Journal of travel medicine* 26(2), (2018).

71. Marmot M, Syme SL, Kagan A, Kato H, Cohen J, Belsky J. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: prevalence of coronary and hypertensive heart disease and associated risk factors. *American journal of epidemiology* 102(6), 514-525 (1975).

72. Castelli F. Drivers of migration: why do people move? *Journal of travel medicine* 25(1), (2018).

73. Pavli A, Maltezou H. Health problems of newly arrived migrants and refugees in Europe. *Journal of travel medicine* 24(4), (2017).

**** Critical review regarding the wide range of health-related complications in the rapidly growing migrant and refugee populations, while providing thoughtful insights on appropriate interventions that may benefit both the host and the migrant populations.**

74. Maneze D, Digiacomo M, Salamonson Y, Descallar J, Davidson PM. Facilitators and barriers to health-seeking behaviours among Filipino migrants: Inductive analysis to inform health promotion*. BioMed research international* 2015(506269) (2015).

75. Mandy M, Nyirenda M. Developmental origins of health and disease: the relevance to developing nations. *International health* 10(2), 66-70 (2018).

76. Nyirenda MJ, Byass P. Pregnancy, programming, and predisposition*. Lancet global health* 7(4), e404-e405 (2019).

77. Teo K, Chow CK, Vaz M, Rangarajan S, Yusuf S. The Prospective Urban Rural Epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. American Heart Journal 158(1), 1-7 (2009).

78. Mente A, Dehghan M, Rangarajan S *et al*. Association of dietary nutrients with blood lipids and blood pressure in 18 countries: a cross-sectional analysis from the PURE study*. Lancet diabetes & endocrinology* 5(10), 774-787 (2017).

79. Miller V, Mente A, Dehghan M *et al*. Fruit, vegetable, and legume intake, and cardiovascular disease and deaths in 18 countries (PURE): a prospective cohort study*. Lancet* 390(10107), 2037-2049 (2017).

80. Attaei MW, Khatib R, Mckee M *et al*. Availability and affordability of blood pressure-lowering medicines and the effect on blood pressure control in high-income, middle-income, and low-income countries: an analysis of the PURE study data*. Lancet public health* 2(9), e411-e419 (2017).

81. Smyth A, Teo KK, Rangarajan S *et al*. Alcohol consumption and cardiovascular disease, cancer, injury, admission to hospital, and mortality: a prospective cohort study*. Lancet* 386(10007), 1945-1954 (2015).

82. Kruger HS, Schutte AE, Walsh CM, Kruger A, Rennie KL. Body mass index cut-points to identify cardiometabolic risk in black South Africans*. European journal of nutrition* 56(1), 193-202 (2017).

83. Kaptoge S, Pennells L, De Bacquer D *et al*. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet global health* 7(10), e1332-e1345 (2019).

***This study highlights the importance of population-specific risk assessment and provides first steps in recalabrating risk models, in this case for CVD, to be equally accurate in high, middle and low income countries.**

84. Dwyer JT, Wiemer KL, Dary O *et al*. Fortification and health: challenges and opportunities. *Advances in nutrition* 6(1), 124-131 (2015).

85. Macintyre UE, Venter CS, Vorster HH. A culture-sensitive quantitative food frequency questionnaire used in an African population: 1. development and reproducibility. *Public health nutrition* 4(1), 53-62 (2001).

86. Hobbs A, Ramsay M. Epigenetics and the burden of noncommunicable disease: a paucity of research in Africa. *Epigenomics* 7(4), 627-639 (2015).

87. Corsi DJ, Subramanian SV, Chow CK *et al*. Prospective Urban Rural Epidemiology (PURE) study: Baseline characteristics of the household sample and comparative analyses with national data in 17 countries. *American heart journal* 166(4), 636-646 (2013).

88. Agyemang C, Beune E, Meeks K *et al*. Rationale and cross-sectional study design of the research on obesity and type 2 diabetes among African migrants: the RODAM study. *BMJ open* 4(3), e004877 (2014).

89. Cronjé HT, Elliott HR, Nienaber-Rousseau C, Pieters M. Replication and expansion of epigenome-wide association literature in a black South African population. *Clinical epigenetics* 12(1), 6 (2020).

90. Meeks KA, Henneman P, Venema A *et al*. An epigenome-wide association study in whole blood of measures of adiposity among Ghanaians: the RODAM study. *Clinical epigenetics* 9(1), 103 (2017).

91. Meeks KA, Henneman P, Venema A *et al*. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. *International journal of epidemiology* 48(1), 58-70 (2018).

92. Agyemang C, Beune E, Meeks K *et al*. Innovative ways of studying the effect of migration on obesity and diabetes beyond the common designs: lessons from the RODAM study. *Annals of the New York academy of sciences* 1391(1), 54-70 (2017).

93. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature* 538(7624), 161-164 (2016).

94. Zhong J, Agha G, Baccarelli AA. The role of DNA methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. *Circulation research* 118(1), 119-131 (2016).

95. Battram T, Richmond RC, Baglietto L *et al*. Appraising the causal relevance of DNA methylation for risk of lung cancer. *International journal of epidemiology* 48(5), 1493-1504 (2019).

96. Kinra S, Andersen E, Ben-Shlomo Y *et al*. Association between urban life-years and cardiometabolic risk: the Indian migration study. *American journal of epidemiology* 174(2), 154-164 (2011).

CHAPTER 4

MANUSCRIPT TWO – ORIGINAL RESEARCH

This manuscript has been published by Clinical Epigenetics and can be viewed in its final format in Annexure A.

Publisher:	Springer Nature
Impact factor:	5.50
Section:	Endocrinology and metabolic epigenetics

Journal aims and scope:

Encompassing the broad spectrum of epigenetics research from basic research to innovations in therapeutic treatments, Clinical Epigenetics is a top tier, open access journal devoted to the study of epigenetic principles and mechanisms as applied to human development, disease, diagnosis and treatment. The journal particularly welcomes submissions involving clinical trials, translational research, new and innovative methodologies and model organisms providing mechanistic insights.

Author's guidelines:

https://clinicalepigeneticsjournal.biomedcentral.com/submission-guidelines

Supplementary material:

https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-019-0805-z#Sec18

REPLICATION AND EXPANSION OF EPIGENOME-WIDE ASSOCIATION LITERATURE IN A BLACK SOUTH AFRICAN POPULATION

H. Toinét Cronjé[1*], Hannah R. Elliott[2,3], Cornelie Nienaber-Rousseau[1], Marlien Pieters[1]

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

***Corresponding author:**

H. Toinét Cronjé

23520825@nwu.ac.za

Keywords: ancestry, DNAm, EPIC, epigenetic epidemiology, EWAS, methylation, NCD, PURE

4.1 Abstract

Background: DNA methylation is associated with non-communicable diseases (NCDs) and related traits. Methylation data on continental African ancestries are currently scarce, even though there are known genetic and epigenetic differences between ancestral groups and a high burden of NCDs in Africans. Furthermore, the degree to which current literature can be extrapolated to the understudied African populations, who have limited resources to conduct independent large-scale analysis, is not yet known. To this end, this study examines the reproducibility of previously published epigenome-wide association studies of DNA methylation conducted in different ethinicities, on factors related to NCDs, by replicating findings in 120 South African Batswana men aged 45 to 88 years. In addition, novel associations between methylation and NCD-related factors are investigated using the Illumina® EPIC BeadChip.

Results: Up to 86% of previously identified epigenome-wide associations with NCD-related traits (alcohol consumption, smoking, body mass index, waist circumference, C-reactive protein, blood lipids and age) overlapped with those observed here and a further 13% were directionally consistent. Only 1% of the replicated associations presented with effects opposite to findings in other ancestral groups. The majority of these inconcistencies were associated with population-specific genomic variance. In addition, we identified eight new 450K array CpG associations not previously reported in other ancestries, and 11 novel EPIC CpG associations with alcohol consumption.

Conclusions: The successful replication of existing EWAS findings in this African population demonstrates that blood-based 450K EWAS findings from commonly investigated ancestries can largely be extrapolated to ethnicities for which epigenetic data are not yet available. Possible population-specific differences in 14% of the tested associations do, however, motivate the need to include a diversity of ethnic groups in future epigenetic research. The novel associations found with the enhanced coverage of the Illumina® EPIC array support its usefulness to expand epigenetic literature.

4.2 Background

The role of epigenetics in the aetiology of non-communicable diseases (NCDs) is of interest owing to its valuable addition to the limited variance of disease risk explained by genetics alone [1]. The modifiable nature of the epigenome also offers opportunities to predict, detect and prevent lifestyle-related diseases [2]. DNA methylation (DNAm) is the most intensively researched epigenetic modification, partly because of its ease of measurement from stored samples commonly collected in epidemiological studies. A number of robust associations between differentially methylated cytosine-guanine dinucleotides (CpGs) and NCD-related traits or exposures have been reported [3-7]. Epigenetic research has allowed for richer insight into the origin and progression of complex diseases, and is expected to continue doing so, thereby enhancing our ability to combat the continued rise in NCD prevalence [2, 8].

Despite its importance in the global context of NCDs, current epigenetic literature remains limited by the lack of ethnic diversity, with most investigating associations between DNAm and health outcomes/traits within European (EU) populations. Although several large-scale epigenome-wide association studies (EWASs) have used data collected from African American (AA) individuals [4, 5, 9], information on continental African populations remain particularly limited. Sub-Saharan Africans are known to be genetically different from AA individuals, who typically stem from West African ancestors, with varying levels of admixture [10]. Because DNAm differences have been reported among ethnic groups [11-13], the degree to which current EWAS results can be extrapolated to other populations, including Sub-Saharan Africans, remains to be established. Understanding the degree of generalisability of EWAS results to different ethnicities informs one whether existing knowledge can be extrapolated to understudied ethnic groups or whether additional research is needed in these populations, where resources are often limited [14].

To this end, we replicated data extracted from the EWAS Catalog (http://www.ewascatalog.org) on traits related to NCDs (alcohol consumption, smoking status, body mass index (BMI), waist circumference (WC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglycerides (TG), C-reactive protein (CRP) and age), in a subset of Batswana men from the North-West (NW) province of South Africa, who participated in the international Prospective Urban and Rural Epidemiology study (PURE-SA-NW). In doing so, we evaluated the reproducibility of previous EWAS findings in a Sub-Saharan African population that has never been investigated before. In addition, because the majority of EWASs to date have been conducted using the older Illumina® 450K BeadChip, our secondary aim was to report novel DNAm associations by using the new Illumina® MethylationEPIC platform, to extend existing knowledge on methylation and traits related to NCDs in this population [15].

4.3 Results

For each trait, we report the degree of replication between EWAS findings in the PURE-SA-NW cohort and the reference studies identified, using the EWAS catalog (complete test statistics in Additional file 1). In cases where the reference study included cohorts of different ancestries, the PURE-SA-NW cohort was compared to these ancestries separately. We first report the agreement between the effect sizes obtained in the PURE-SA-NW data and the reference studies for all the tested CpGs per trait, to evaluate the overall consensus between the studies (PURE-SA-NW *vs* reference study). We then examine the similarity between studies at the individual CpG level by determining whether or not the individual PURE-SA-NW association's confidence intervals (CIs) overlap with those of the reference study. This allows us to identify systematic differences (e.g. attributable to exposure variation) between cohorts before investigating differences at an individual CpG level (e.g. attributable to site-specific genetic variation). To permit further investigation of individual CpG association differences, we inspect probes previously identified to measure methylation at polymorphic sites of which either the global minor allele frequency (MAF) is higher than 1% [16], or variation has been documented in Africans, specifically [17] (Additional file 1). Probes identified to hybridise to multiple genomic regions or to be cross-reactive are also noted [18]. Replication analyses are followed by a report of any methylation associations of newly investigated EPIC probes and novel 450K associations (of 450K probes present on the EPIC array used here), where applicable (Additional file 2). Table 4-1 provides the descriptive statistics for (i) the PURE-SA-NW cohort for traits used as covariates in the models, (ii) the trait of interest as reported by the EWAS catalog reference study, and (iii) the PURE-SA-NW cohort trait of interest reported in the same unit as in the specific reference study. For the different traits, the sample size here differs because we applied the specific inclusion criteria of the respective reference studies to our population to permit comparison (Additional file 1).

Comparatively, our study population had a more favourable body composition and blood lipid profile, but a much higher CRP concentration than those included in the reference studies [3, 9, 22]. The proportion of current smokers in our study population was twice as high as the reference cohort [4], and they consumed larger volumes of alcohol than the EU, but less than the AA reference cohorts [5]. The remaining traits were similar between our population and that of the reference studies.

Table 4-1 Descriptive characteristics of the study and reference cohorts

Trait	PURE-SA-NW	Reference study	Comparative PURE-SA-NW	Reference study citation
N	120	See Additional file 1		
Age (y)	64 [55–70]	62 [58–67]	64 [55–70]	[19]
BMI (kg/m^2)	22.5 ± 4.9	27.6 ± 4.4[1]* 27.7 ± 4.5[2]*	22.4 ± 5.0	[20]
WC (cm)	83.8 ± 12.8	101 ± 15.1[3]* 97 ± 16[4]*	83.6 ± 12.7	[9] [21]
Physical activity (index)	2.41 ± 0.94			
Smoking status [N(%)]				
Never smoker	56 (47)	6 956 (74)[2,4]*	56 (48)	[4]
Current smoker	61 (51)	2 433 (26)[2,4]*	61 (52)	
Ever smoker	64 (53)			
Alcohol use [N(%)]				
Never user	56 (47)			
Ever user	64 (53)			
Alcohol consumption (g/d)	16.7 ± 36.6	1.3 (0, 301)[3] 5.6 (0, 181)[4]	0 (0, 240)	[5]
CRP (mg/L)	9.7 ± 27.2	6.2 ± 8.8[3]* 3.3 ± 5.6[4]*	9.9 ± 27.5	[22]
TC (mg/dL)	171 ± 41.6	207 ± 37.1[4]*	171 ± 41.6	[3]
LDL-C (mg/dL)	96.5 ± 35.7	125 ± 30.9[4]*	96.5 ± 35.7	
HLD-C (mg/dL)	54.1 ± 22.7	57.0 ± 16.8[4]	54.1 ± 22.7	
TG (mg/dL)	48.5 ± 30.5	126 ± 69.0[4]*	48.5 ± 30.5	
Education [N(%)]				
None	26 (22)			
1–7 years of schooling	66 (55)			
8–12 years of schooling	28 (23)			
Blood cell type proportions (%)				
B-cells	0.04 ± 0.02			
CD4 T-cells	0.11 ± 0.04			
CD8 T-cells	0.11 ± 0.06			
Granulocytes	0.47 ± 0.11			
Monocytes	0.09 ± 0.02			
Natural killer cells	0.11 ± 0.03			

[1]Indian Asian ancestry, [2]European American ancestry, [3]African American ancestry, [4]European ancestry. *Population means differ between the reference study and comparative PURE-SA-NW study population at p <0.05 following Bonferroni adjustment. BMI: body mass index; CRP: C-reactive protein; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; TC: total cholesterol; TG: triglycerides; WC: waist circumference. Values are presented as median [IQR], mean ± standard deviation, *N* (%) or median (minimum, maximum). Blood cell proportions were determined using methylation-based estimates [23].

4.3.1 Alcohol consumption

Ancestry-stratified (European American (EA) and AA) findings from the meta-analysis by Liu *et al*. [5] on the association of alcohol consumption (g/d) with differential methylation at individual CpGs were compared with those from the PURE-SA-NW (Figure 4-1). In the study of Liu *et al*. [5], alcohol consumption was more strongly associated with DNAm in AA than EA individuals (regression slope = 3.2, p = 8.6 x 10^{-70}). Effect sizes in the PURE-SA-NW cohort were larger than in either of the reference groups (regression slope = 0.12, p = 3.2 x 10^{-16} and 0.47, p = 3.2 x 10^{-17} for the AA and EA comparisons, respectively).

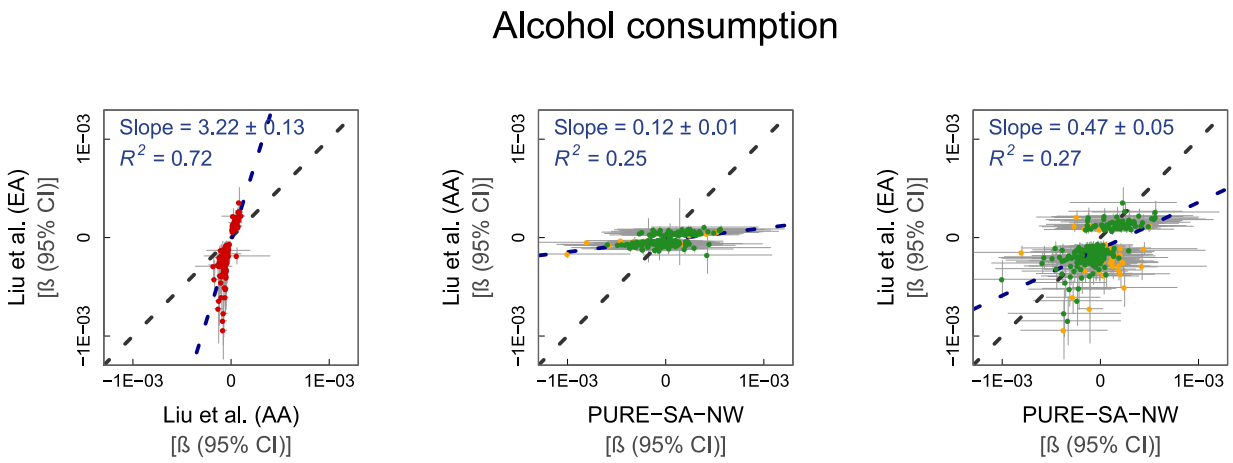


Figure 4-1 % Methylation change per gram of alcohol intake

From left to right: (i) reference AA *vs* EA data (247 CpGs), (ii) PURE-SA-NW *vs* AA data (228 CpGs), and (iii) PURE-SA-NW *vs* EA data (228 CpGs). Model used: methylation ~ alcohol consumption+ age + BMI + cell counts + surrogate variables. Reference data: Liu *et al*. (2018). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line.

Individual association results showed stronger similarity between the PURE-SA-NW and the AA than with the EA findings. Overall, 361 CpGs were investigated (two unique AA, 131 unique EA and 228 associations reported for both ethnicities). Out of the 230 association tests to compare the AA reference cohort to the PURE-SA-NW data, 93% (213) of the regression CIs overlapped, compared to 80% (287) of the 359 comparisons between the EA and PURE-SA-NW. Where CIs did not overlap, directional consistency was nevertheless observed with the exception of the associations for cg15636519 (EA and AA comparisons), cg08471846 (EA comparison only) and cg21227253 (EA comparison only) with alcohol consumption (Additional file 1a). Data from the

Biobank-based Integrative Omics Studies (BIOS) Consortium indicated that, apart from cg08471846, methylation quantitative trait loci (mQTLs) have been identified for each of these CpGs with absolute reported Z-scores ranging from 4.15 to 12.9 [24, 25]. Data from the 1000 Genomes project support that the differences observed here could be partly influenced by ancestry-specific genetic variance; for example, the MAF of rs7153432 (*cis* mQTL for cg21227253) is 18% in Africans and 40% in Europeans [26].

The EWAS conducted on alcohol consumption in the PURE-SA-NW cohort resulted in 19 genome-wide significant findings ($p < 9.4 \times 10^{-8}$), 11 of which were newly investigated EPIC probes and eight were part of those previously investigated by 450K probes, that were present on the EPIC array, but failed to reach association thresholds in other cohorts (Additional file 2a). Table 4-2 provides the test statistics for these CpGs.

The proportion of methylation variance of these CpGs explained by including alcohol consumption in the model methylation ~ age + BMI + cell counts + smoking status, ranged from 10.3 to 43.8%. When alcohol consumption was used as the outcome variable, the addition of these 19 probes to the regression model increased the percentage of alcohol consumption variance explained by 57% (adjusted $R^2 = 0.05$ before and 0.62 after including the CpGs, $p = 5.5 \times 10^{-26}$).

Table 4-2 EWAS CpG-alcohol consumption associations $p < 9.4 \times 10^{-8}$

ProbeID	Location	Gene	Region	β	SE	*p*	% Variance explained	X^2 *p-value*
cg13153796*	14:101405628	*SNORD*113-6	TSS1500	–6.78E–04	8.45E–05	2.2E–11	29.4 (38.4)	3.8E–12
cg00712390*	17:79373624	*BAHCC*1	1stExon	8.08E–04	1.14E–04	9.7E–10	37.5 (47.0)	5.6E–18
cg05706661	7:36134301	*LOC*101928618	TSS1500	–1.05E–03	1.51E–04	2.0E–09	17.6 (57.1)	6.7E–12
cg24252287*	17:40250379			1.48E–04	2.21E–05	4.8E–09	36.5 (41.8)	7.7E–15
cg12177743*	11:113185079	*TTC*12	TSS200	1.59E–04	2.41E–05	7.5E–09	13.4 (23.4)	2.8E–05
cg19323439	17:9136232	*NTN*1	Body	5.06E–04	7.93E–05	1.9E–08	14.1 (59.0)	2.1E–08
cg19683675*	5:142077712	*FGF*1	TSS200	–1.13E–03	1.78E–04	2.0E–08	35.1 (43.6)	2.7E–15
cg08333974	12:1956337	*CACNA2D*4	Body	–1.24E–03	1.95E–04	2.2E–08	25.8 (38.3)	8.3E–12
cg12325997	15:59280148	*RNF*111	1stExon	9.84E–05	1.57E–05	3.2E–08	10.5 (58.4)	9.4E–08
cg19642811	13:95453039	*LOC*101927284	Body	–6.02E–04	9.64E–05	3.4E–08	19.3 (37.3)	2.3E–08
cg06943216	8:102683096			–1.33E–03	2.13E–04	3.5E–08	17.5 (33.9)	4.3E–08
cg26187237*	2:217498574	*IGFBP*2	1stExon	4.19E–04	6.72E–05	3.6E–08	15.5 (53.0)	2.2E–09
cg16358446*	1:1534984			8.10E–05	1.31E–05	4.4E–08	43.8 (52.4)	1.9E–21
cg08724692	6:133646558	*EYA*4	Body	–6.26E–04	1.03E–04	6.4E–08	10.3 (43.4)	1.2E–06
cg08035774	9:136600662	*SARDH*	5'UTR	–1.12E–03	1.85E–04	7.5E–08	23.6 (32.8)	6.3E–10
cg18780412*	3:179755086	*PEX*5*L*	TSS1500	6.36E–04	1.06E–04	8.6E–08	27.0 (33.5)	6.4E–11
cg15942324	1:38482118	*UTP*11*L*	Body	–6.63E–04	1.10E–04	8.8E–08	23.3 (33.2)	3.4E–09
cg25278025	2:103378026	*TMEM*182	TSS1500	5.99E–04	9.98E–05	8.8E–08	15.4 (26.0)	5.0E–06
cg22572934[&]	5:173171061	*LINC*01484	Body	–1.21E–03	2.02E–04	9.3E–08	13.3 (24.3)	5.5E–05

Model: methylation ~ alcohol consumption (g/d) + age + BMI + smoking + cell counts + surrogate variables. * 450K probes. [&] Probe that should be interpreted with caution owing to the presence of genomic variance at probe measurement site [17]. The percentage variance explained reflects the added value of alcohol consumption to the variance in CpG methylation, reported as percentage explained by alcohol as an added exposure (percentage variance explained by the total model). X^2 *p* value = Chi-square *p* value when the regression models with and without alcohol consumption are compared.

4.3.2 Smoking status

The association of smoking status with the DNAm of 3 618 CpGs in the PURE-SA-NW cohort was compared to a multi-ethnic (EA and AA) EWAS conducted by Joehanes *et al.* [4]. 'Current' users in the PURE-SA-NW cohort included individuals regularly smoking any bought or self-made tobacco product (commercial cigarettes, bidis, pipes and cigars). Joehanes *et al*. [4], however, restricted the definition of 'current' smokers to those specifically reporting cigarette use. Regardless of the discrepancy in the product smoked, results from the respective EWASs were fairly similar. No ancestral comparisons were made by Joehanes *et al.* [4], who combined data from a number of different ethnic groups in a meta-analysis.

Effect sizes were generally larger in the PURE-SA-NW than in the reference data (regression slope = 0.34, $p = 1.7 \times 10^{-206}$). Of the 3 618 CpGs tested for their independent association with smoking status, 3 315 (92%) of the regression β 95% CIs overlapped and 269 were directionally consistent between cohorts (Figure 4-2). Only 34 CpGs showed a difference in the direction of effect between the findings of Joehanes *et al*. [4] and the PURE-SA-NW cohort (Additional file 1b). Thirteen of these probes measure methylation at polymorphic sites, 20 had *cis*-mQTLs and five had *trans*-mQTLs, all of which had differing AA and EU ancestry MAFs, suggesting that genetic variation between cohorts could drive some of the dissimilarities observed [24-26]. No novel associations with smoking were identified and the only genome-wide significant CpG association was for a previously identified CpG (cg05575921) that was associated with a 17% ($p = 4.2 \times 10^{-10}$) reduction in DNAm in *current* smokers compared to participants who had *never* smoked (Additional file 2b).

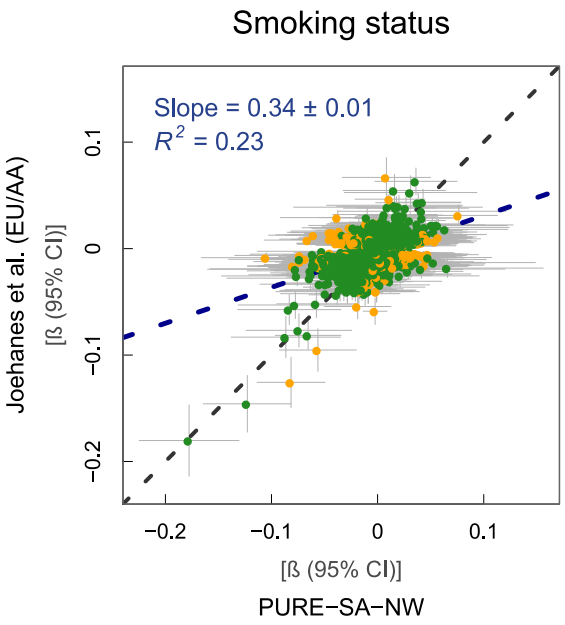


Figure 4-2 % Methylation difference between current and never smokers in reference *vs* PURE-SA-NW data.

Model used for PURE-SA-NW EWAS: methylation ~ smoking + age + cell counts + surrogate variables. Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Black dashed line: line of equality. Blue dashed line: regression line.

4.3.3 Body mass index

We replicated findings from the largest EWAS on BMI conducted to date, that of Wahl *et al.* [20]. These authors investigated the relationship of methylation with BMI in individuals of Indian Asian (IA) and EU descent. Wahl *et al.* [20] observed larger effect sizes among the IA than the EU group (regression slope = 0.48, $p = 4.9 \times 10^{-72}$). PURE-SA-NW data reflected the IA data better than the EU data, but in both instances, PURE-SA-NW data showed larger effect sizes than either reference group (regression slope = 0.57, $p = 6.0 \times 10^{-7}$ and 0.37, $p = 1.8 \times 10^{-8}$ for IA and EU groups, respectively). However, when comparing the overlap between individual effect estimates, PURE-SA-NW mirrored findings from the EU group better. The 95% CIs of the 265 regression estimates between the cohorts overlapped 55% (147) and 77% (203) of the time when compared with IA data and EU data, respectively (Figure 4-3). All regression CIs that did not overlap were directionally consistent between the PURE-SA-NW and reference cohorts. No genome-wide significant associations with BMI were identified (Additional file 2c).

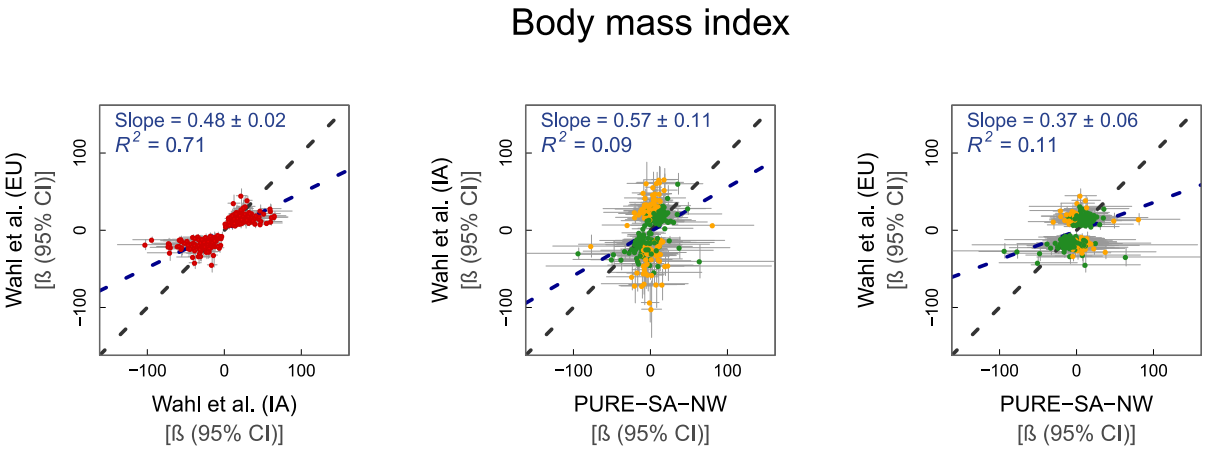


Figure 4-3 Change in BMI (kg/m^2) per % methylation change

From left to right: (i) reference EU *vs* IA data, (ii) IA *vs* PURE-SA-NW data, and (iii) EU *vs* PURE-SA-NW data. Model used for PURE-SA-NW EWAS: BMI ~ methylation + age + smoking status + alcohol consumption + physical activity + cell counts + surrogate variables. Reference data: Wahl *et al*. (2017). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line.

4.3.4 Waist circumference

Eight previously reported associations of WC with DNAm in cohorts of AA and EA descent [9] were replicated in the PURE-SA-NW cohort (Figure 4-4). The regression model used to quantify the relationship between WC and DNAm differed between the reference cohort subgroups. In addition to the covariates adjusted for in the EA regression model (age, smoking and white blood cell counts), the AA model also included alcohol consumption status, physical activity, education and household income. The use of the two different models was justified, as it resulted in highly comparable findings between the reference study's AA and EA groups (r = 0.96), with a slightly larger average effect size observed in the EA than in the AA data (regression slope = 0.56, p = 0.0001). Applying the fully adjusted (AA) model to the PURE-SA-NW data resulted in a 10.4% increase in average effect size compared to the model used for the EA group, justifying the use of the fully adjusted model in our cohort.

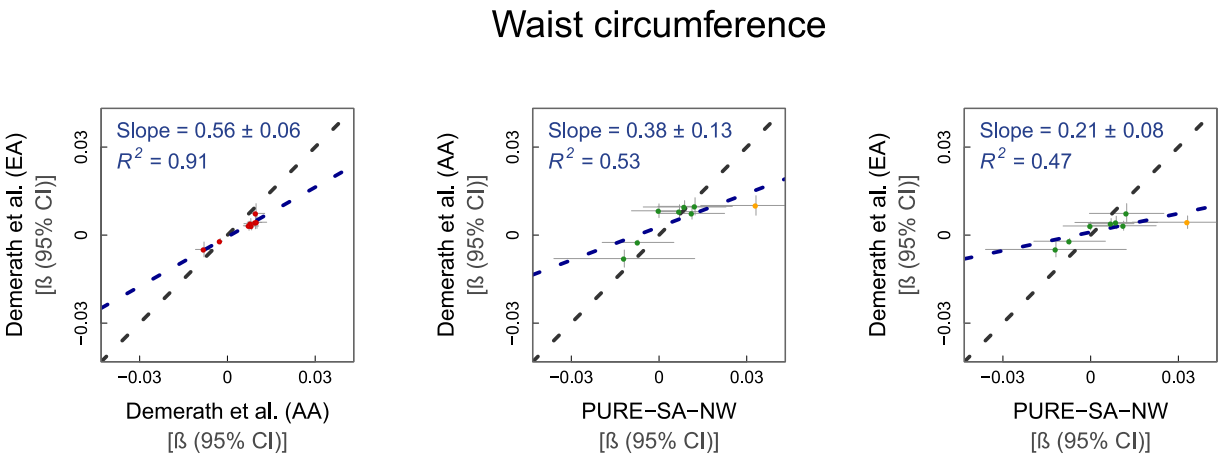


Figure 4-4 % Methylation change per centimetre change in WC

From left to right: (i) reference EA *vs* AA data, (ii) AA *vs* PURE-SA-NW data, and (iii) EA *vs* PURE-SA-NW data. WC was normalised to have a mean of 0 and a standard deviation of 1. Model used for PURE-SA-NW EWAS: methylation ~ WC + age + alcohol consumption + smoking + physical activity + education + cell counts + surrogate variables. Reference data: Demerath *et al*. (2015). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line.

As for the previous traits, larger effect sizes were observed in the PURE-SA-NW than both the AA (regression slope = 0.38, p = 0.03) and EA (regression slope = 0.21, p = 0.04) cohorts with a closer resemblance to the AA than the EA data (R^2 = 0.53 *vs* 0.47). When comparing individual effect estimates between the PURE-SA-NW and reference data, the 95% CIs overlapped in seven

of the eight assessed associations in both groups (Additional file 1d). The non-overlapping associations were directionally consistent between studies, overall indicating strong comparability between WC's association with DNAm across the investigated ancestral groups. The single non-overlapping locus was the same in both ethnic groups compared. This site, cg26403843, is associated with five *cis*-mQTLs and one *trans*-mQTL with absolute Z-scores ranging from 4.9 to 39.8. Population differences between the mQTL-associated SNPs were observed; rs6556405, for example, has an MAF of 26% in Europeans compared to a frequency of 66% in Africans [24-26].

4.3.5 Blood lipids

Findings from the largest TC, LDL-C, HDL-C and TG EWASs to date, reported by Hedman *et al*. [3], were compared to those of the PURE-SA-NW cohort. For each of the four lipids, larger effect sizes were observed in the PURE-SA-NW than in the EU reference cohort. The regression slopes when modelling the PURE-SA-NW effect sizes against those of the reference cohorts were 0.12 (p = 0.18), 0.13 (p = 0.27), 0.19 ($p = 9.9 \times 10^{-06}$) and 0.30 (p = 0.01) for TC, LDL-C, HDL-C and TG, respectively (Figure 4-5). Effect estimates and 95% CIs overlapped for 38/40 (95%) for TC, 18/21 (86%) for LDL-C, 96/102 (94%) for HDL-C and 15/16 (94%) for TG, of the associations tested (Additional file 1e). Ten of the 12 non-overlapping associations were directionally consistent, leaving only two associations divergent in the direction of effect: cg24939194-HDL-C and cg15878619-TC. Two mQTLs have been identified for cg24939194 (rs748097 and rs2969017), the strongest of which has an MAF of 6% in Africans and 37% in Europeans, indicating that genetic ancestry may be important for the association of cg24939194 with HDL-C [26].

Despite the consistency in the effect sizes between the PURE-SA-NW and the reference data, the large CIs observed in our data do not allow for further interpretation of these findings. There was one genome-wide significant lipid-DNAm association in our cohort (Additional file 2e). High-density lipoprotein cholesterol associated with cg23636606 at a regression β of $2.6 \times 10^{-04} \pm 4.4 \times 10^{-05}$ ($p = 4.8 \times 10^{-08}$).

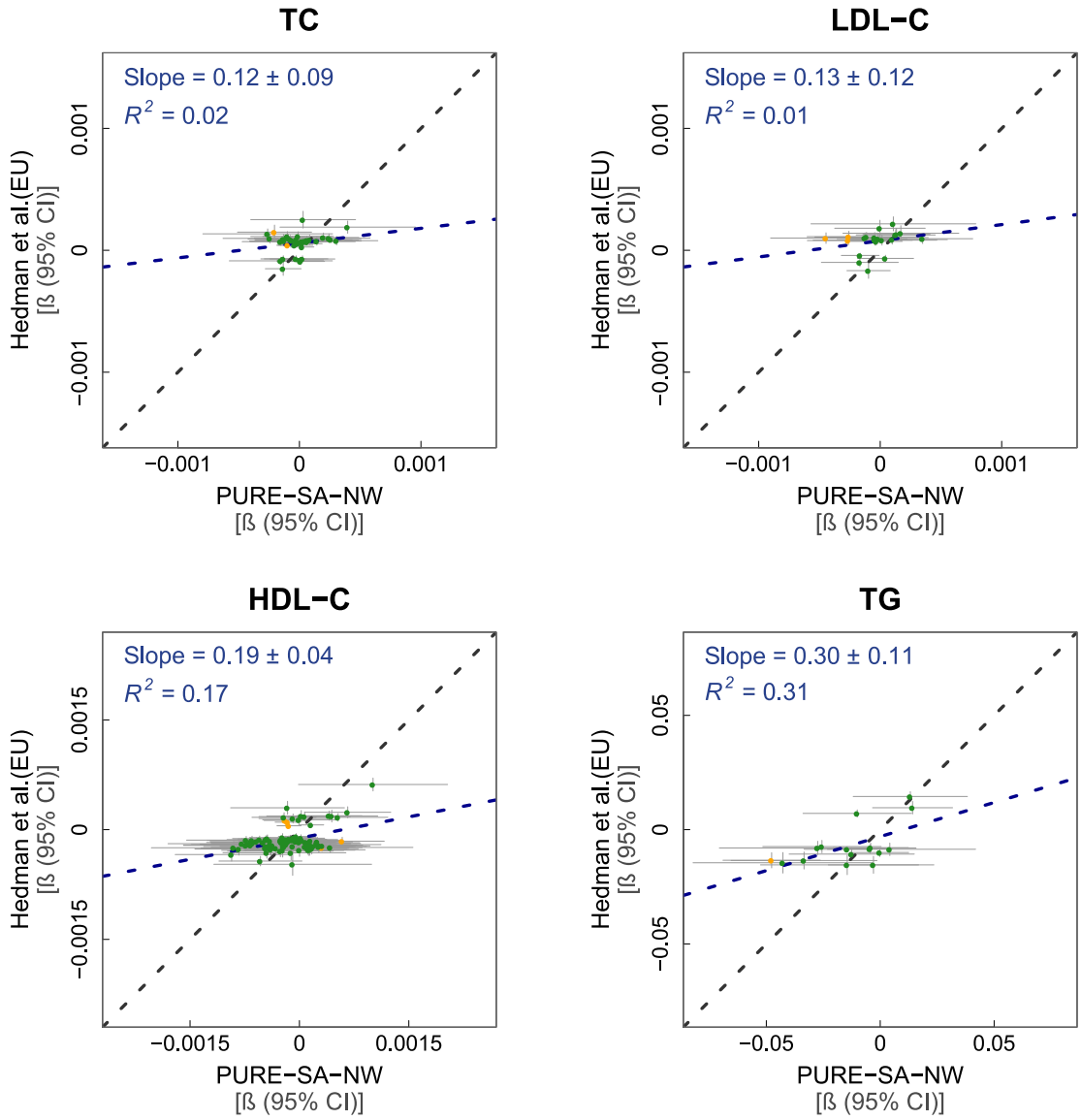


Figure 4-5 % Methylation change per mg/dL change in lipid concentration in reference *vs* PURE-SA-NW data

Models used: methylation ~ lipid (TC, LDL-C, HDL-C or TG) + age + cells + surrogate variables. Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line.

4.3.6 CRP

Ancestry-stratified (AA and EU) data on the effect of CRP on the DNAm of 207 loci, by Ligthart *et al.* [22] were compared to PURE-SA-NW. The reference study reported highly comparable effect sizes between the AA and EU ancestral groups (regression slope = 0.82, p = 1.25 x 10^{-107}), with

slightly larger effects observed in the AA group. The comparison of the regression slope of effect sizes between the reference data and our own showed moderately larger effect sizes in the PURE-SA-NW findings than in the reference data, more so for the EU (regression slope = 0.25, $p = 2.5 \times 10^{-10}$) than the AA (regression slope = 0.22, $p = 1.3 \times 10^{-10}$) comparison (Additional file 1f). Confidence intervals of the individual effect estimates between the reference and PURE-SA-NW data overlapped for 192 out of the 207 tests (93%) in each ethnicity (Figure 4-6).

Twenty-two of the 30 non-overlapping associations were directionally consistent. Two CpGs had associations in opposing directions of effects compared to EU (cg01588592 and cg23740758) and three compared to the EU and AA (cg24174557, cg26846781, cg27184903) reference datasets. All the non-overlapping CpGs have *cis*-mQTLs with absolute reported Z-scores ranging from 4.06 to 22.95 [24, 25]. Data from the 1000 Genomes project support the notion that the differences observed here could be partly influenced by ancestry-specific genetic variance: for example, the MAF of rs9791189 (*cis*-mQTL for cg23740758) is 12% in Africans and 23% in Europeans [26]. There were no genome-wide significant or novel CRP-DNAm associations in our cohort (Additional file 2f).

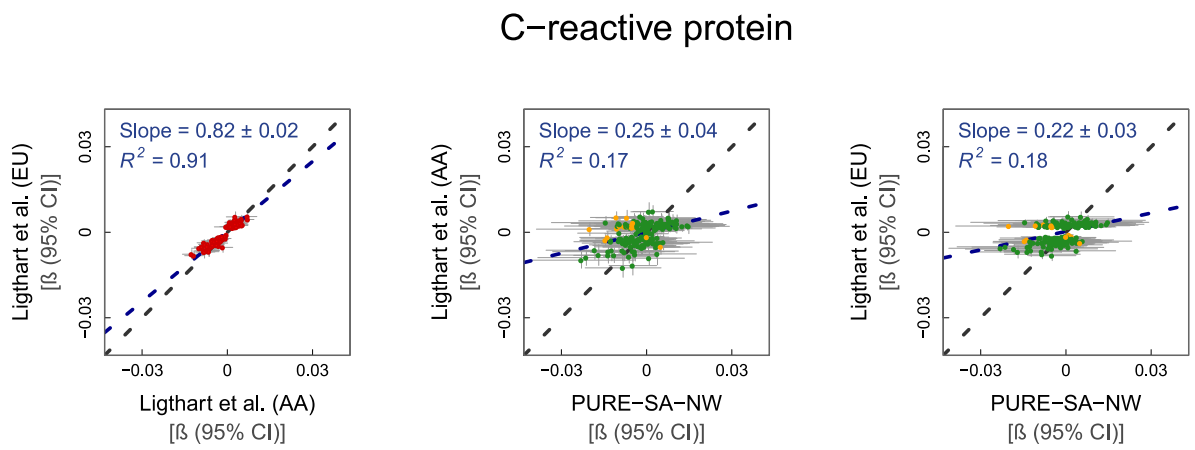


Figure 4-6 Change in logarithmic CRP (mg/L) per % methylation change

From left to right: (i) reference EU *vs* AA data, (ii) AA *vs* PURE-SA-NW data, and (iii) EU *vs* PURE-SA-NW data. Model used for PURE-SA-NW EWAS: methylation ~ CRP + age + smoking + BMI + cells + surrogate variables. Reference data: Ligthart *et al*. (2016). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line.

4.3.7 Age

Previous findings from EU-based research on the association of age with DNAm of 152 CpGs [19] were compared to those from the PURE-SA-NW cohort (Figure 4-7). In contrast to all other traits, a much weaker association between age and DNAm was observed in our data than in the reference data (regression slope = 12.9, p = 4.2 x 10^{-31}). Although the direction of effects was consistently similar between the two studies, none of the regression CIs overlapped when comparing the individual associations (Additional file 1g).

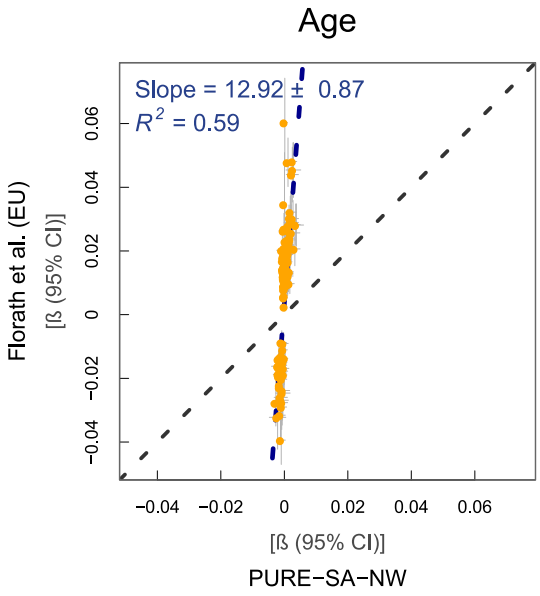


Figure 4-7 % Methylation change per year of age in reference *vs* PURE-SA-NW data

Model used for PURE-SA-NW EWAS: methylation ~ age + smoking + cell counts + surrogate variables. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Black dashed line: line of equality. Blue dashed line: regression line.

Formal data on disease diagnosis were not available for the PURE-SA-NW cohort and were, therefore, not included in regression models, as done by Florath *et al*. [19]. Furthermore, cell counts were not adjusted for in the reference study, but were included in our models, since cell counts are recognised confounder in our data. Sensitivity analyses were, however, performed by including data on chronic medication use (as a proxy for disease) as well as excluding cell count adjustments. These analyses did not result in any discernible differences in findings (inclusion of medication use: regression slope = 13.0, p = 4.5 x 10^{-30}, exclusion of cells: regression slope = 12.4, p = 1.7 x 10^{-32}). There were no genome-wide significant or novel age-DNAm associations in our cohort (Additional file 2g).

4.4 Discussion

Our primary analysis focussed on the replication of relevant EWAS literature in 120 Batswana men from the PURE-SA-NW cohort. Secondary analysis included the discovery of novel findings, either investigated for the first time on the EPIC array, or with the 450K probes incorporated in the EPIC array that had not previously been associated with these traits.

Overall the 95% CI of effect estimates for 86% (4 730 out of the 5 498 CpG-trait association tests) of the PURE-SA-NW associations overlapped with previously reported findings, and a further 13% (720 out of the 5 498 CpG-trait association tests) were directionally uniform. Generally, larger effect sizes were observed in the PURE-SA-NW data than those of the reference studies. Although the reason for differing effect sizes cannot be answered definitively, given the small sample size, the degree of association seems to be related to population-specific differences. Only ~1% of our findings (48 out of the 5 498 CpG-trait association tests, including 44 unique CpGs) were directionally inconsistent with its compared association reported in the reference study. No data quality concerns were observed for any of these directionally contradicting findings. Of the 44 CpGs, 36 have mQTLs [24, 25] for which population differences in MAFs have been observed by the 1000 genomes project [26].

Overall, these results indicate general consistency in epigenome-wide associations among ethnicities, but ancestry may be important in up to 14% of the tested associations. This is supported by the fact that regardless of the similarity in traits measured among groups, the associations observed in PURE-SA-NW data consistently reflected those reported in AA better than in EU/EA cohorts and better in EU than IA in the case of methylation-BMI associations. Furthermore, eight novel associations between the methylation of 450K array probes, present on the EPIC platform, and alcohol consumption are reported in the Batswana South Africans that were not previously observed in populations of different ancestral origins. These population distinctions indicate the value of ethnic diversity in epigenetic research.

The only trait for which we were unable to replicate any associations was age. Apart from the reference study for age being the smallest of the reference studies included (N = 498), there were also clear differences in the pre-processing, data normalisation and EWAS approach followed between PURE-SA-NW and Florath *et al.* [20]. The reference cohort's analyses were restricted to a pre-selected set of 200 CpGs, the methylation levels of which were normalised using Box–Cox transformations. A mixed regression model with plate and BeadChip as random effects was used. For the PURE-SA-NW data, however, we employed a functional normalisation strategy on the raw methylation data of all the EPIC BeadChip probes, followed by linear regression where surrogate variables were adjusted for as fixed effects to control for possible unaccounted

variance. Our findings remained directionally consistent with the reference study's, with the average difference in effect size amounting to 0.87% methylation change per year increase in age (calculated as the percentage difference between the average of the 152 tests' absolute regression βs of the PURE-SA-NW *vs* Florath *et al.* [19] results).

In terms of findings related to the EPIC array, 11 genome-wide significant alcohol associations are reported here. An additional eight genome-wide significant alcohol associations were observed for 450K probes present on the EPIC array. Alcohol consumption contributed to a large portion of the variance in the methylation of these probes, as well as, when reversed, the probes to the variance in alcohol consumption. Previous 450K CpG-alcohol associations have been used successfully to identify risky and heavy drinkers [5]. Our sample size did not allow stratification of alcohol intake, but we expect the addition of the alcohol-associated EPIC probes reported here to enhance the discriminatory potential of the current methylation-based biomarker of alcohol consumption [5]. The variance explained by these findings and their usefulness as potential biomarkers warrant replication in large and ethnically diverse cohorts. Larger sample sizes and ethnic diversity will also permit further exploration of the biological basis of these findings and their potential application in NCD-related epigenetic research.

The strengths of this study are the expansion of current data, both by using the EPIC array and investigating a novel study population, after first being able to observe similar findings to those from independent, highly powered, previously replicated literature. The overall consistency between effect sizes is reassuring, in terms of not only the comparability of the PURE-SA-NW data with previous findings, but also the consistency in the effect size and explained variability of novel associations compared to previous EWASs on similar traits [5, 9, 20, 21]. The efficacy of the enhanced coverage of the EPIC array, to uncover new associations with a range of traits, is shown in our study, even with our limited sample size. We motivate the use of this array in future large-scale analyses, as it is likely to add to the variance that can be explained using methylation markers and also to identify novel sites that may be important in prediction, risk stratification or understanding causal disease pathways.

In this study, however, the corresponding limitation to doubling the coverage of the 450K array was the relative loss of statistical power, given our sample size. The lack of power resulted in wide regression CIs for most association estimates that limited our capacity for the fine scale inference of findings. We were able to comment on general patterns and large differences, but we do not know whether more subtle differences between population subgroups exist. Furthermore, the unavailability of genomic data in our cohort and the absence of data on Southern African populations in the 1000 genomes' database restricted our ability to evaluate MAF differences between the reference and Batswana South African groups. We are, therefore,

unable to quantify the relative contribution of genetic compared to environmental factors in the associations and association differences observed. The overall congruence in replication results between cohorts—even when large differences in phenotypes are demonstrated—does, however, suggest that these associations might be the result of genetic architecture rather than environmental differences, which we expect to affect the investigated traits as well.

The inclusion of only one sex also limits this study in that no assumptions can be made regarding the generalisability of these results to black South African women. However, because all the reference studies we replicated contained mixed-gender data, there are likely not major differences in these associations between the sexes.

4.5 Conclusions

This study reports that up to 86% of the previously reported epigenome-wide associations observed in other ethnicities are present in this black male South African population. While acknowledging the value of ethnic-specific genomic data, our results support the notion that current blood-based 450K EWAS findings can largely be extrapolated to under-represented ethnicities for whom epigenetic data are not yet available. However, the population-specific differences in up to 14% of the CpGs, tested together with the unique associations reported here, do motivate the inclusion of a diversity of ethnic groups in epigenetic association studies. Investigating multi-ethnic data in epigenome-wide studies should be considered the golden standard.

4.6 Methods

4.6.1 Study design

This study was performed on a sub-sample of individuals participating in the international PURE study [27]. The PURE study includes sub-cohorts across the world, including one comprising individuals residing in the NW province of South Africa. This sub-cohort represents a single, self-reported ethnicity, Batswana South Africans, who were born and still reside in the NW province of South Africa. Detailed descriptions of the international and PURE-SA-NW cohorts have been published previously [27, 28].

PURE-SA-NW data were collected in 2005, 2010 and 2015. A total of 126 participants were randomly selected for the current investigation, from a group of 990 individuals who took part in the 2015 PURE-SA-NW data collection. Eligibility depended on the following inclusion criteria: availability of bio-samples, testing negative for the human immunodeficiency virus at the time of data collection and male sex. These criteria were incorporated to eliminate confounding by sex

and CD4 cell counts in a study with already limited power. The participants included in this study are referred to as the PURE-SA-NW cohort in this manuscript.

4.6.2 Data collection

Height and weight were quantified using a stadiometer and an electronic scale. BMI was calculated as weight per unit height squared (kg/m^2). WC was measured at the appropriate landmarks, by qualified anthropometrists using a steel tape.

An adapted physical activity index questionnaire was used to gather data to calculate a physical activity index [29]. Alcohol intake (g/d) was determined from a quantitative food frequency questionnaire adapted and validated for use in this population [30]. Participants reported the amount, frequency and any relevant description of the alcoholic drinks they had consumed in the preceding month. Data were processed to an amount in g/d, based on the South African food composition tables using FoodFinder3® software (available from http://foodfinder.mrc.ac.za). Smoking status (current, former or never) was self-reported, using a standardised questionnaire. When used as a covariate, smoking and drinking status were dichotomised into *never* and *ever* groups, with former smokers/drinkers included in the *ever* category. When investigated as the EWAS exposure, smoking status and alcohol consumption were classified according to the classification used in the reference studies.

Fasting blood samples were collected and handled as described previously [31]. High-sensitivity CRP and fasting blood lipids (TC, LDL-C, HDL-C, TG) were quantified using the Cobas® Integra 400 (Roche® Clinical System, Roche Diagnostics, Indianapolis, IN, USA).

4.6.3 DNAm data generation and processing

Whole blood intended for the isolation of genomic DNA was collected in 9 mL Tempus tubes (Applied Biosystems™, Foster city, CA, USA) at the same time as blood used for the quantification of all other phenotypes. Tubes were vortexed for 10 seconds prior to storage in a -20°C freezer for up to 5 days, after which samples were transferred to cryostorage (–80°C) until analysis. DNA was isolated using QIAGEN Flexigene DNA extraction kits (QIAGEN® Valencia, CA, USA). The manufacturer's protocol was followed with minor modifications.

Upon extraction, the picoGreen® dsDNA quantitation assay (Invitrogen™, Carlsbad, CA, USA) was used to quantify DNA. Five hundred nanograms DNA from each participant was bisulphite-converted using the Zymo EZ DNAm™ kit (Zymo Research, Irvine, CA, USA), followed by genome-wide DNAm profiling on the Illumina® Infinium MethylationEPIC BeadChip according to the manufacturer's protocol (Illumina®, San Diego, CA, USA).

Samples were randomised across slides to minimise the possibility of confounding by batch. Raw signal intensity data were processed from .idat files using functional normalisation as described by the R package *meffil* [32]. The quality threshold for samples and probes was set at 95%. All probes or samples with a detection *p* value > 0.01 for more than 5% of the evaluated measures were excluded. Six samples were removed on account of low quality: four samples because of a proportion of undetected probes above the quality control (QC) threshold and two with outlying control probes. Probes failing QC were removed prior to data normalisation (*N*= 8343). Eventually 857 516 probes and 120 individuals were included in subsequent data normalisation and analysis. Principal component analysis of the control probes identified 12 principal components to be included in the functional normalisation. In addition, *slide* was specified as a random effect to be included to address batch variance. Sample cell fractions (B-cells, CD4 and CD8 T-cells, neutrophils, monocytes and natural killer cells) were estimated using the IDOL optimised L-DMR library for whole blood samples [23].

4.6.4 Identification of reference data using the EWAS catalog

Data we sought to replicate were extracted from the EWAS catalog (http://www.ewascatalog.org, date of access: 27 April 2019). The EWAS catalog indexes EWAS studies performed in a study sample of at least 100 individuals for whom at least 100 000 CpGs were available genome-wide. Only associations with $p < 1 \times 10^{-4}$ are included in the catalog.

Data from the catalog were pruned according to the following criteria: (i) the EWAS catalog trait had to be available in the PURE-SA-NW study cohort in a comparable unit, (ii) methylation-trait associations had to be replicated (below a *p* value threshold of 1×10^{-4}) in at least one independent cohort, regardless of tissue, to reduce the possibility of including false positive findings from among the reference studies, (iii) the DNA had to have been extracted from a blood-based sample, (iv) DNAm had to be reported in Beta units; and (v) an effect estimate (β) and standard error had to be available for each association. Traits that fitted these criteria were age, alcohol consumption, smoking, BMI, WC, CRP, HDL-C, LDL-C, TG and TC. To simplify data analysis, we attempted to replicate results from the largest study indexed by the EWAS catalog for each investigated trait only. The results reported in each replication sub-section make reference to the particular study used for comparison, which would have been the largest EWAS included in the catalog at the time of writing.

4.6.5 Statistical analysis

Statistical analysis was conducted using R 3.4.3 [33]. The normality of trait data was assessed using Shapiro-Wilks tests. Linear regression models were used to identify epigenome-wide

associations using the *meffil* [32] and *ewaff* (https://github.com/perishky/ewaff) packages. DNAm was modelled as a β value between 0 and 1, representing the ratio of methylated to unmethylated probes. The relative contribution of exposures to the variance of outcome variables was determined using the *relaimpo* package's *lmg* metric from the *calc.relimp* function applied to linear models.

For the replication analysis, because of the small sample size of the PURE-SA-NW study population and, therefore, limited power, replication of previously published results focusses on the size and direction of effect sizes rather than comparison of *p* values. Associations were considered replicable when the 95% CI of the regression β of the reference and the PURE-SA-NW cohort overlapped. Most reference studies extracted from the EWAS catalog adjust regression models for "technical variation". In PURE-SA-NW, surrogate variables were added to all models to reduce any unknown or unmeasured confounding [34]. The *sva* and *generate.confounders* functions within the *meffil* and *ewaff* packages estimated the surrogate variables that were included in each model based on the method described by Leek and Storey [34]. Annotation data were obtained from *meffil*.

For the investigation of novel findings, only associations with $p < 9.4 \times 10^{-8}$ were considered genome-wide significant [15]. Within our cohort, we estimated 80% power to detect a 5% difference in methylation at this threshold for 69% of the EPIC probes, assuming an alpha level of 0.05 and 530 639 independent tests [15]. Packages used in analyses, in addition to those already specified, include *BaseR, dplyr, FlowSorted.Blood.EPIC, ggplot, IlluminaHumanMethylationEPICanno.ilm10b2.hg19, minfi, readxl* and *xlsx.*

4.7 Additional files

Additional file 1: EWAS test statistics (PURE-SA-NW *vs* reference study) for: (a) alcohol consumption; (b) smoking status; (c) BMI; (d) WC; (e) lipids; (f) CRP and (g) age. (XLSX 659kb)

Additional file 2: EWAS associations with $p < 1 \times 10^{-4}$ in the PURE-SA-NW study for: (a) alcohol consumption; (b) smoking status; (c) BMI; (d) WC; (e) lipids; (f) CRP and (g) age. (XLSX 162kb)

4.8 Declarations

Ethics approval and consent to participate

Ethical approval for the 2015 data collection of the PURE-SA-NW study was granted by the Health Research Ethics Committee of the North-West University (NWU-00016-10-A1, NWU-00119-17-A1). All participants provided written informed consent, including consent for genetic/epigenetic

analysis. All procedures described were performed in accordance with the revised version of the Helsinki Declaration of 1975 [35].

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study are available upon reasonable request and with the permission of the Health Research Ethics Committee of the North-West University and the principal investigator of the PURE-SA-NW study, Prof. I.M. Kruger (lanthe.kruger@nwu.ac.za) at the North-West University's Africa Unit for Transdisciplinary Health Research.

Competing interest

The authors declare that they have no competing interests.

Funding

Financial support for the PURE-SA-NW study was provided by the North-West University, South African National Research Foundation (SANRF), Population Health Research Institute, South African Medical Research Council (SAMRC), the North West Province Health Department, and the South African Netherlands Partnerships in Development. Grants from the SANRF, Academy of Medical Sciences UK (Newton Fund Advanced Fellowship Grant) and the SAMRC supported the additional epigenetic work reported herein. HTC is supported by a PhD scholarship from the SANRF (SFH106264); HRE works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/5). None of the funding bodies were involved in the design of the study, collection, analysis or interpretation of the data or the writing of this manuscript. Opinions expressed and conclusions arrived at are those of the authors and are not to be attributed to the funding sources.

Authors' contributions

HTC isolated and curated the DNA, performed all statistical analysis and wrote the original draft. HRE oversaw the EPIC analysis, conceptualised the manuscript, supervised statistical analysis and critically reviewed and edited the manuscript. CN-R critically reviewed and edited the manuscript. MP is the principal investigator, acquired the funding for this project and critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank all those that participated in the PURE-SA-NW study and those that made the PURE-SA-NW study possible, including the fieldworkers, researchers and staff of both the PURE-SA-NW (Africa Unit for Transdisciplinary Health Research (AUTHeR), Faculty of Health Sciences, NWU, Potchefstroom, South Africa) and PURE International (S Yusuf and the PURE project office staff at the Population Health Research Institute (PHRI), Hamilton Health Sciences and McMaster University, Ontario, Canada) teams. We thank the staff of the Bristol bioresource laboratories (Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK) who generated the DNAm data.

Abbreviations

AA: African American; BIOS: Biobank-based Integrative Omics Studies; BMI: Body mass index; Chr: chromosome; CpGs: cytosine-phosphate-guanine sites; CRP: C-reactive protein; EA: European American; DNAm: DNA methylation; ENCODE: Encyclopedia of DNA Elements; EU: European; EWAS: epigenome-wide association study; HDL-C: high-density lipoprotein cholesterol; IA: Indian Asian; IQR: inter-quartile range; LDL-C: low-density lipoprotein cholesterol; MAF: minor allele frequency; mQTLs: methylation quantitative trait loci; NCD: non-communicable diseases; PURE-SA-NW: South Africa, North-West arm of the Prospective Urban and Rural Epidemiology study; QC: Quality control; TC: total cholesterol; TG: triglycerides; UTR: untranslated region.

4.9 References

1. Rodger EJ, Chatterjee A. The epigenomic basis of common diseases. Clin Epigenetics. 2017;9:5.

2. Sharp GC, Relton CL. Epigenetics and noncommunicable diseases. Epigenomics. 2017;9:789-91.

3. Hedman AK, Mendelson MM, Marioni RE, Gustafsson S, Joehanes R, Irvin MR, Zhi D, Sandling JK, Yao C, Liu C, Liang L, Huan T, McRae AF, Demissie S, Shah S, Starr JM, Cupples LA, Deloukas P, Spector TD, Sundstrom J, Krauss RM, Arnett DK, Deary IJ, Lind L, Levy D, Ingelsson E. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. Circ Cardiovasc Genet. 2017;10:e001487.

4. Joehanes R, Just A, Marioni R, Pilling L, Reynolds L, Mandaviya P, Guan W, Xu T, Elks C, Aslibekyan S. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436-47.

5. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, Just AC, Duan Q, Boer CG, Tanaka T, Elks CE, Aslibekyan S, Brody JA, Kuhnel B, Herder C, Almli LM, Zhi D, Wang Y, Huan T, Yao C, Mendelson MM, Joehanes R, Liang L, Love SA, Guan W, Shah S, AF MR, Kretschmer A, Prokisch H, Strauch K, Peters A, Visscher PM, Wray NR, Guo X, Wiggins KL, Smith AK, Binder EB, Ressler KJ, Irvin MR, Absher DM, Hernandez D, Ferrucci L, Bandinelli S, Lohman K, Ding J, Trevisi L, Gustafsson S, Sandling JH, Stolk L, Uitterlinden AG, Yet I, Castillo-Fernandez JE, Spector TD, Schwartz JD, Vokonas P, Lind L, Li Y, Fornage M, Arnett DK, Wareham NJ, Sotoodehnia N, Ong KK, van Meurs JBJ, Conneely KN, Baccarelli AA, Deary IJ, Bell JT, North KE, Liu Y, Waldenberger M, London SJ, Ingelsson E, Levy D. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23:422-33.

6. Mendelson MM, Marioni RE, Joehanes R, Liu C, Hedman AK, Aslibekyan S, Demerath EW, Guan W, Zhi D, Yao C, Huan T, Willinger C, Chen B, Courchesne P, Multhaup M, Irvin MR, Cohain A, Schadt EE, Grove ML, Bressler J, North K, Sundstrom J, Gustafsson S, Shah S, AF MR, Harris SE, Gibson J, Redmond P, Corley J, Murphy L, Starr JM, Kleinbrink E, Lipovich L, Visscher PM, Wray NR, Krauss RM, Fallin D, Feinberg A, Absher DM, FornageM, Pankow JS, Lind L, Fox C, Ingelsson E, Arnett DK, Boerwinkle E, Liang L, Levy D, Deary IJ. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a mendelian randomization approach. PLoS Med. 2017;14:e1002215.

7. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, Volkova N, Schramm K, Carstensen-Kirberg M, Waldenberger M, Gieger C, Peters A, Illig T, Prokisch H, Roden M, Grallert H. Association between DNA methylation in whole blood and measures of glucose metabolism: Kora f4 study. PLoS One. 2016;11:e0152314.

8. World Health Organization. Noncommunicable diseases. 2018. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed 9 Aug 2019.

9. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, Hedman AK, Sandling JK, Li LA, Irvin MR, Zhi D, Deloukas P, Liang L, Liu C, Bressler J, Spector TD, North K, Li Y, Absher DM, Levy D, Arnett DK, Fornage M, Pankow JS, Boerwinkle E. Epigenome-wide association study (ewas) of bmi, bmi change and waist circumference in african american adults identifies multiple replicated loci. Hum Mol Genet. 2015;24:4464-79.

10. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O. The genetic structure and history of africans and african americans. Science. 2009;324:1035-44.11.

11. Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, Wahl S, Elliott HR, Rota F, Scott WR. Epigenome-wide association of DNA methylation markers in peripheral blood from indian asians and europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol. 2015;3:526-34.

12. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, Davey Smith G, Hughes AD, Chaturvedi N, Relton CL. Differences in smoking associated DNA methylation patterns in south asians and europeans. Clin Epigenetics. 2014;6:4.

13. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP. Accounting for population stratification in DNA methylation studies. Genet Epidemiol. 2014;38:231-41.

14. Teo Y, Small KS, Kwiatkowski DP. Methodological challenges of genomewide association analysis in africa. Nat Rev Genet. 2010;11:149-60.

15. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, Hannon E. Guidance for DNA methylation studies: statistical insights from the illumine epic array. BMC Genomics. 2019;20:366.

16. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45:e22.

17. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationEPIC beadchip. Genomics Data. 2016;9:22-4.

18. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.

19. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H. Crosssectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated cpg sites. Hum Mol Genet. 2014;23:1186-201.

20. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC, Ried JS, Zhang W, Yang Y. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81-6.

21. Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, Sha J, Pankow JS, Liu C, Irvin MR, Fornage M, Hidalgo B, Lin LA, Thibeault KS, Bressler J, Tsai MY, Grove ML, Hopkins PN, Boerwinkle E, Borecki IB, Ordovas JM, Levy D, Tiwari HK, Absher DM, Arnett DK. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity. 2015;23:1493-501.

22. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, Colicino E, Waite LL, Joehanes R, Guan W, Brody JA, Elks C, Marioni R, Jhun MA, Agha G, Bressler J, Ward-Caviness CK, Chen BH, Huan T, Bakulski K, Salfati EL, Fiorito G, Wahl S, Schramm K, Sha J, Hernandez DG, Just AC, Smith JA, Sotoodehnia N, Pilling LC, Pankow JS, Tsao PS, Liu C, Zhao W, Guarrera S, Michopoulos VJ, Smith AK, Peters MJ, Melzer D, Vokonas P, Fornage M, Prokisch H, Bis JC, Chu AY, Herder C, Grallert H, Yao C, Shah S, AF MR, Lin H, Horvath S, Fallin D, Hofman A, Wareham NJ, Wiggins KL, Feinberg AP, Starr JM, Visscher PM, Murabito JM, Kardia SL, Absher DM, Binder EB, Singleton AB, Bandinelli S, Peters A, Waldenberger M, Matullo G, Schwartz JD, Demerath EW, Uitterlinden AG, van Meurs JB, Franco OH, Chen YI, Levy D, Turner ST, Deary IJ, Ressler KJ, Dupuis J, Ferrucci L, Ong KK, Assimes TL, Boerwinkle E, Koenig W, Arnett DK, Baccarelli AA, Benjamin EJ, Dehghan A. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016;17:255.

23. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.

24. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, Slieker RC, Jhamai PM, Verbiest M, HED S, Verkerk M, van der Breggen R, van Rooij J, Lakenberg N, Arindrarto W, Kielbasa SM, Jonkers I, van 't Hof P, Nooren I, Beekman M, Deelen J, van Heemst D, Zhernakova A, Tigchelaar EF, Swertz MA, Hofman A, Uitterlinden AG, Pool R, van Dongen J, Hottenga JJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, van den Berg LH, van Zwet EW, Mei H, Li Y, Lemire M, Hudson TJ, the BC, Slagboom PE, Wijmenga C, Veldink JH, van Greevenbroek MMJ, van Duijn CM, Boomsma DI, Isaacs A, Jansen R, van Meurs JBJ, t Hoen PAC, Franke L, Heijmans BT. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics. 2016;49:131.

25. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van 't Hof P, Mei H, van Dijk F, Westra H-J, Bonder MJ, van Rooij J, Verkerk M, Jhamai PM, Moed M, Kielbasa SM, Bot J, Nooren I, Pool R, van Dongen J, Hottenga JJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, Zhernakova A, Li Y, Tigchelaar EF, de Klein N, Beekman M, Deelen J, van Heemst D, van den Berg LH, Hofman A, Uitterlinden AG, van Greevenbroek MMJ, Veldink JH, Boomsma DI, van Duijn CM, Wijmenga C, Slagboom PE, Swertz MA, Isaacs A, van Meurs JBJ, Jansen R, Heijmans BT, t Hoen PAC, Franke L. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics. 2016;49:139.

26. Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68.

27. Teo K, Chow CK, Vaz M, Rangarajan S, Yusuf S. The prospective urban rural epidemiology (pure) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. Am. Heart J. 2009;158:1-7. e1

28. De Lange Z, Pieters M, Jerling JC, Kruger A, Rijken DC. Plasma clot lysis time and its association with cardiovascular risk factors in black africans. PLoS One. 2012;7:e48881.

29. Nienaber-Rousseau C, Sotunde OF, Ukegbu PO, Myburgh PH, Wright HH, Havemann-Nel L, Moss SJ, Kruger IM, Kruger HS. Socio-demographic and lifestyle factors predict 5-year changes in adiposity among a group of black south african adults. Int J Environ Res Public Health. 2017;14:1089.

30. MacIntyre UE, Venter CS, Vorster HH. A culture-sensitive quantitative food frequency questionnaire used in an african population: 1. Development and reproducibility. Public Health Nutr. 2001;4:53-62.

31. Pieters M, Kotze RC, Jerling JC, Kruger A, Ariens RA. Evidence that fibrinogen gamma' regulates plasma clot structure and lysis and relationship to cardiovascular risk factors in black africans. Blood. 2013;121:3254-60.

32. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;34:3983-9.

33. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017. URL https://www.Rproject.org/

34. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:1724-35.

35. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310:2191-4.

CHAPTER 5

MANUSCRIPT THREE – ORIGINAL RESEARCH

This manuscript has been accepted for publication in Frontiers in Immunology.

Publisher: Frontiers Media
Impact factor: 4.72
Section: Inflammation

Journal aims and scope:

Inflammation plays a fundamental role in nearly all chronic degenerative diseases and the contribution of pro-inflammatory cytokines in neurodegenerative, cardiovascular and bone diseases has become a major area of investigation. Lessons from cytokine biology are at the core of understanding inflammation. Blocking a cytokine with a “biological” outside the cell continues to be highly effective in treating several diseases and without organ toxicity. Several post-receptor signaling kinases are located downstream from cytokine-activated cells, which have become new targets to tame inflammation. Orally active small molecule inhibitors of intracellular signaling kinases will likely be the new frontier of anti-inflammatory agents. The scope of the section Inflammation will be broad and include clinically relevant areas of investigation which apply to the following areas: Pro-inflammatory cytokines; The effects of innate inflammation; How inflammation affects the immune system and Cell signaling during inflammation.

Author’s guidelines:

https://www.frontiersin.org/journals/immunology/sections/inflammation#author-guidelines

Notes:

This manuscript is presented in US English

METHYLATION *VS* PROTEIN INFLAMMATORY BIOMARKERS AND THEIR ASSOCIATIONS WITH CARDIOVASCULAR FUNCTION

H. Toinét Cronjé[1*], Hannah R. Elliott[2,3], Cornelie Nienaber-Rousseau[1], Fiona R. Green[4], Aletta E. Schutte[5,6], Marlien Pieters[1]

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[4] Formerly School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

[5] Hypertension in Africa Research Team, Medical Research Council Unit for Hypertension and Cardiovascular Disease, North-West University, Potchefstroom, South Africa

[6] School of Public Health and Community Medicine, University of New South Wales; George Institute for Global Health, Sydney, Australia

***Corresponding author:**

H. Toinét Cronjé

23520825@nwu.ac.za

Keywords: cell-counts, epigenetics, epigenetic epidemiology, inflammation, neutrophil-to-lymphocyte, lymphocyte-to-monocyte

Running head: Methylation *vs* protein inflammatory markers

5.1 Abstract

DNA methylation data can be used to estimate proportions of leukocyte subsets retrospectively, when directly measured cell counts are unavailable. The methylation-derived neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios (mdNLRs and mdLMRs) have proven to be particularly useful as indicators of systemic inflammation. As with directly measured NLRs and LMRs, these methylation-derived ratios have been used as prognostic markers for cancer, although little is known about them in relation to other disorders with inflammatory components, such as cardiovascular disease (CVD). Recently, methylation of five genomic cytosine-phosphate-guanine sites (CpGs) was suggested as proxies for mdNLRs, potentially providing a cost-effective alternative when whole-genome methylation data are not available. This study compares seven methylation-derived inflammatory markers (mdNLR, mdLMR and individual CpG sites) with five conventionally used protein-based inflammatory markers (C-reactive protein, interleukins 6 and 10, tumor-necrosis factor alpha and interferon-gamma) and a protein-based inflammation score, in their associations with cardiovascular function (CVF) and risk. Markers of CVF were more strongly associated with methylation-derived than protein-based markers. We found that the protein-based and methylation-derived inflammatory markers complemented rather than proxied one another in their contribution to the variance in CVF. There were no strong correlations between the methylation and protein markers either. Therefore, the methylation markers could offer unique information on the inflammatory process and are not just surrogate markers for inflammatory proteins. Although the five CpGs mirrored the mdNLR well in their capacity as proxies, they contributed to CVF above and beyond the mdNLR, suggesting possible added functional relevance. We conclude that methylation-derived indicators of inflammation enable individuals with increased CVD risk to be identified without measurement of protein-based inflammatory markers. In addition, the five CpGs investigated here could be useful surrogates for the NLR when the cost of array data cannot be met. Used in tandem, methylation-derived and protein-based inflammatory markers explain more variance than protein-based inflammatory markers alone.

5.2 Contribution to the field

Neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios (NLRs and LMRs) are valuable prognostic markers in cancer and cardiovascular disease patients. Whole-genome DNA methylation data can be used to estimate the methylation-derived neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios (mdNLRs and mdLMRs) of a blood sample. Five methylation sites were recently suggested as surrogate markers for the mdNLR in the absence of whole-genome methylation data. This can be very useful because, methylation data, unlike directly performed cytometry, can be quantified retrospectively. Because many large cohorts possess DNA methylation data from suitable quantification platforms, it is important to understand the utility of these data in epidemiology. This study evaluated these methylation-derived cell ratios in the context of CVD risk by comparing their usefulness in explaining variance in cardiovascular function factors with frequently used inflammatory proteins (interleukin-6, interleukin-10, tumor-necrosis factor alpha, interferon gamma and C-reactive protein). We report that the methylation sites not only reflect the mdNLR, but provide additional information in explaining cardiovascular function variance. We also report that methylation-derived inflammatory markers can be useful above and beyond information that can be gained from protein inflammation biomarkers. We encourage cohorts to explore their methylation data in this regard and use already gathered information to expand current epidemiological research.

5.3 Introduction

Methylation-derived cell count ratios, particularly methylation-derived neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios (mdNLRs and mdLMRs), are increasingly being used as robust alternatives to flow cytometry-based cell count ratios as indicators of systemic inflammation (1-3). One key advantage is that they can be derived from archived blood in cohorts where cytometric measurements have not been performed (1, 4). Through unique methylation signatures, leukocyte subtypes can be separated and quantified with comparative accuracy. Validation analyses have reported an R2 estimate of at least 0.95 when methylation-derived estimates of leukocyte sub-types and NLRs are regressed on those measured directly (1, 5).

Similar to cytometry-based ratios, mdNLRs and mdLMRs are considered prognostic markers of overt inflammatory diseases such as rheumatoid arthritis (3) and cancer (2, 6). However, little is known about these methylation-derived ratios in relation to less pronounced inflammatory diseases such as cardiovascular disease (CVD). While directly measured cell counts have been established as indicators of CVD severity, recurrence and prognosis (7-12), the use of cell counts, methylation-derived or directly measured, in CVD risk prediction and disease progression in the epidemiological setting is less well known. There is also no consensus on the reference ranges

or thresholds to be used when characterizing the NLR or MLR as healthy, at risk or pathological (13), with large ethnic diversity also being reported (14-16). More recently, methylation levels of five cytosine-phosphate-guanine sites (CpGs), namely cg25938803, cg10456459, cg01591037, cg03621504 and cg00901982, have been suggested as proxy markers for the mdNLR, because of their robust associations with myeloid cell (neutrophil and monocyte) differentiation (2, 4). If true, measurement of this small panel of CpGs could render whole genome methylation measurement unnecessary in cohorts with limited financial resources.

Blood-based protein inflammatory markers, such as C-reactive protein (CRP), interleukin-6 (IL-6) and tumor necrosis factor alpha (TNF-α) are useful epidemiological tools for quantifying inflammatory state and disease risk (17). However, cell count ratios are considered superior to circulating inflammatory markers in their ability to quantify systemic inflammation (7, 18, 19). Cell-count ratios provide a more integrated view of systemic inflammation by reflecting the relative contribution of the innate (neutrophils/monocytes as indicators of general inflammation) and adaptive (lymphocytes as an indicator of physiological stress) immune responses (19), supporting their use in population-based research. Because few cohorts have access to data on both cell counts and protein-based inflammatory markers, we set out to determine whether cell count ratios provide added benefit in characterizing inflammatory status and CVD risk independent from protein-based inflammatory markers in our cohort where both are measured. The rapid advancement of epigenetic investigations in CVD research, together with the increasing number of samples with epigenetic data available, including increasing ethnic diversity among available samples (20, 21), also motivate our interest in exploring novel ways to mine for valuable additional information from previously analyzed samples.

To this end, we investigated methylation-derived and protein-based biomarkers of inflammation in relation to cardiovascular function (CVF) in a cohort of black South African men. We included seven methylation-derived (mdNLR, mdLMR, and the five myeloid CpGs) and five high-sensitivity protein-based (CRP, IL-6, IL-10, TNF-α, interferon (IFN)-γ) markers of inflammation. In addition, we used a protein-based inflammation score (22) to provide the combined effect of inflammatory biomarkers. First, we compared how well the protein-based and methylation-derived inflammatory markers reflected CVD risk according to literature-based cut-offs. We also compared the mdNLRs and mdLMRs reported in this sample population to ratios reported in studies on healthy individuals from different ethnicities. This is followed by an investigation of the relationship between the methylation-derived and protein-based inflammatory biomarkers in our study population, and a comparison of their relative associations with CVF markers. Lastly, we investigated whether a combination of methylation and protein inflammatory biomarkers provided added benefit in explaining CVF variance, or whether one proxies the other. Cardiovascular

function is represented by blood pressure (BP), heart rate (HR), and arterial stiffness. In contrast to previous work, we evaluated the methylation-derived biomarkers in a population-based cohort as opposed to a case-control design, to yield better understanding of the value these markers may have in the general population.

5.4 Methods

5.4.1 Study population

This is a cross-sectional investigation of 120 self-identified Batswana men who were enrolled in the North West province, South African arm of the international Prospective Urban and Rural Epidemiology study (PURE-SA-NW) in 2015 (23). Individuals were randomly selected from 926 participants residing in selected rural and urban regions, based on the following criteria: male sex, testing negative for the human-immunodeficiency virus at the time of data collection and bio-sample availability. These criteria were incorporated to minimize confounding by sex and CD4+ cell counts. The PURE-SA-NW study received ethical approval from the Health Research Ethics Committee of the North-West University, South Africa (NWU-00016–10-A1). Written informed consent was obtained from participants prior to data collection.

5.4.2 DNA methylation, cell counts and cell count ratios

Genomic DNA isolated from peripheral whole blood was bisulfite-converted prior to genome-wide methylation quantification using the Illumina Infinium MethylationEPIC BeadChip according to the manufacturer's protocol (Illumina®, San Diego, CA, USA). Details regarding DNA extraction, quality control, methylation quantification, data processing and data normalization have been reported previously (24). Sample cell fractions were estimated using the IDOL optimized L-DMR library for whole blood samples in the FlowSorted.Blood.EPIC R software package (5). Neutrophil counts were divided by lymphocytes (calculated as the sum of B-, CD4T, CD8T and natural killer cell counts), and lymphocytes by monocytes, to obtain the respective mdNLRs and mdLMRs (1, 2).

5.4.3 Inflammatory markers

Fasting blood samples were collected in ethylenediamine tetra acetic acid tubes for the analysis of cytokines and in anti-coagulant-free tubes for the quantification of CRP. Samples were centrifuged within 30 minutes of collection at 2000 x g for 15 minutes. The Cobas© Integra 400 (Roche© Clinical System, Roche Diagnostics, Indianapolis, IN) was used to quantify high-sensitivity CRP concentrations. High-sensitivity Q-Plex™ planar-based multiplexed enzyme-

linked immunosorbent assays (Quansys Biosciences, Logan, UT) were performed to measure IL-6, IL-10, TNF-α and IFN-γ. An inflammation summary score was calculated to amalgamate related inflammatory proteins as suggested previously (22, 25, 26). Data on CRP, IL-6, IL-10, TNF-α, and IFN-γ were $\log_e$-transformed to improve distribution. Thereafter data were converted to Z-scores to account for the difference in measurement units. The average of the Z-scores is reported here as the inflammatory score.

5.4.4 Measures of cardiovascular function

Systolic and diastolic BP (SBP and DBP) and HR were measured using the OMRON M6 device (Omron Healthcare, Kyoto, Japan). Participants were seated in an upright position with legs uncrossed. After participants had rested for 10 minutes, the correct cuff size was fitted on their right arms, whereafter two measurements were recorded with a five-minute interval. Data from the second measurement were used for analysis. Pulse pressure (PP) was then calculated as the difference between SBP and DBP. Large artery stiffness was investigated using the current gold standard measurement, carotid-femoral pulse wave velocity (cfPWV, (26)) using the SphygmoCor XCEL device (AtCorMedical Pty. Ltd., Sydney, Australia). The transit-distance method was used to measure PWV along the descending thoracoabdominal aorta. Two readings were taken from each participant while supine. Data from the second reading were used.

5.4.5 Cardiovascular risk factors (covariates)

Socio-demographic information and data on medicine use and smoking habits, alcohol consumption and physical activity were collected by interview, using a standardized and validated questionnaire (23). Current smoking and alcohol consumption status were reported as *current*, *former* or *never*, but has been dichotomized here to *never* and *ever* (where *ever* denotes *formerly* and *currently*). Participants also reported the frequency and quantity of usage, age at the start of use and previous attempts at abstinence. Participants were asked by interviewers to provide information on any prescribed or over-the-counter medication they regularly make use of. Body mass index (BMI) was calculated as weight (measured using an electronic scale) per unit height (measured using a stadiometer) squared (kg/m^2). Waist circumference was measured using a steel tape (Lufkin, Cooper Tools, Apex NC, USA), according to standard anthropometric procedures. Physical activity is reported as a continuous physical activity index measure determined using data from a modified Baecke's questionnaire validated for use in South Africa adults (27).

Blood samples for the subsequent analyses were collected and processed in the same manner as described above. Fasting glucose, total cholesterol, triglycerides and low- and high-density

lipoprotein cholesterol (LDL-C and HDL-C), were quantified with the Cobas© Integra 400 (Roche© Clinical System, Roche Diagnostics, Indianapolis, IN). The Cobas© Integra 400 plus (Roche©, Basel, Switzerland) was used to determine serum gamma-glutamyl transferase concentrations. Glycated hemoglobin was quantified using the D-10 Hemoglobin testing system from Bio-Rad© (#220-0101).

5.4.6 Statistical analysis

Analyses were performed using version 3.5.0 of R statistical software (28). Data distribution was evaluated using Shapiro-Wilks tests and visual inspection of histograms and quantile-quantile plots. As most of the data were not normally distributed, we proceeded with non-parametric testing where appropriate. Prior to linear modeling, all skewed data were $\log_e$-transformed. A Bonferroni adjustment, based on the number of independent comparisons, was used to account for multiple testing. Variables were considered dependent when the coefficient of determination (R^2) between them was greater than 0.2. Based on these criteria, we regarded the number of independent inflammatory markers tested here as four: ((1) methylation-derived markers, (2) CRP, (3) IL-6 and IL-10 and (4) TNF-α and IFN-γ) and the CVF markers as three: ((1) BP markers, (2) HR and (3) cfPWV).

Relationships among the biomarkers of inflammation were assessed with partial Spearman correlations, controlling for age and smoking status (*Hmisc* R package for quantification and *corrplot* for visualization, Figure 5-1), because of the well-described association of age and smoking status with methylation (29, 30). The Bonferroni threshold for these correlations was $p < 0.003$ ($\alpha = 0.05/16$ tests, calculated as 4 x 4 independent inflammatory marker comparisons).

Next, the associations between individual inflammatory biomarkers (protein-based and methylation-derived) and CVF, and the variance in CVF explained by these inflammatory markers, were investigated using linear multivariate regression models adjusted for known cardiovascular risk markers (Table 5-2 and Supplementary Table 1). The Bonferroni threshold for these models was set at $p < 0.004$ ($\alpha = 0.05/12$ tests, calculated as 4 independent inflammatory x 3 independent CVF markers).

Thereafter, a combination of methylation-derived inflammatory markers (selected using backwards-stepwise regression models) were investigated in similar linear multivariate regression models (using the *car* and *relaimpo* packages), this time adjusting for known cardiovascular risk markers including inflammation, represented by the protein-based inflammatory score (Table 5-3 and Supplementary Table 3). The Bonferroni threshold was 0.02 (0.05/3, calculated as 3 independent CVF markers x 1 independent inflammatory marker). The relative contribution of

inflammatory biomarkers to the CVF variance was determined using the *relaimpo* package's *lmg* metric from the *calc.relimp* function. Chi-square tests were used to compare linear models, before and after adding methylation-derived biomarkers (Table 5-3).

To identify covariates, known cardiovascular risk markers were entered in backward stepwise linear regression models with CVF markers as outcome, to identify risk markers strongly associated with CVF in this study population. Age, dwelling place (rural/urban), body composition (BMI and waist circumference), level of education, physical activity, smoking status, alcohol consumption, medicine use, blood lipid levels (total cholesterol, LDL-C, HDL-C and triglycerides), markers of glucose metabolism (fasting glucose and glycated hemoglobin) and gamma-glutamyl transferase were tested. With the exception of dwelling place, smoking and alcohol consumption status and medicine use, all variables were investigated as continuous variables. Only risk markers retained by the stepwise regression models were adjusted for in subsequent models to avoid overfitting, given our limited sample size. Based on these results, we made use of two main covariate clusters in all regression analyses. First (hereafter referred to as Model 1), we adjusted for age only. Second (hereafter referred to as Model 2), we adjusted for age, smoking status, dwelling place (rural/urban), BMI, LDL-C, HDL-C, and medicine use (as a binary variable, yes or no). When cfPWV was the outcome, mean arterial pressure was additionally adjusted for in both models. In Table 5-3, inflammation, quantified using the inflammatory score, was added to Model 2 and is referred to as Model 3.

5.5 Results

The clinical characteristics of our cohort are provided in Table 5-1 and Supplementary Table 1. We report on 120 ostensibly healthy men, aged between 45 and 88 years ($\bar{x}$ = 63). Sixty-nine of these men resided in rural areas, 79 reported regular medication use, and 64 classified themselves as *ever* smokers.

Table 5-1 also indicates, where available, literature-based cut-off values for increased CVD risk, in the same unit as reported in our cohort. Based on the CVF and cardiovascular risk markers reported in Table 5-1, 25–50% of our study population was at increased CVD risk. Regarding the protein-based inflammatory markers, CRP reflected a similar risk (50%), while IL-6 cut-offs categorized almost 90% of the study population as suffering from low-grade inflammation (25, 31) and increased CVD risk. No reference ranges for IL-10, TNA-α or IFN-γ in terms of chronic low-grade inflammation or CVD risk are established.

The methylation-derived cell ratios were in agreement with the CVD risk portrayed by CRP and the CVF and CVD risk markers (Table 5-1). The mdNLR and mdLMR, respectively, classified

21% and 50% of the PURE-SA-NW participants as having increased CVD risk. Nineteen participants (16%) were classified as at higher CVD risk by both ratios. Supplementary Figures 1 and 2 depict the methylation-derived cell count ratios observed in the PURE-SA-NW men (blue) in relation to directly measured reference ranges published for healthy individuals from several ethnic groups (green) and cut-offs previously used to predict the odds of specific CVD outcomes, or ranges from patients in case-control studies (orange). On average, the PURE-SA-NW cohort had comparatively lower mdNLRs than the NLR ranges reported in other population-based cohorts. The PURE-SA-NW cohort also exhibited only slight overlap with the patient groups reported. In terms of the mdLMR (where a higher ratio is more favorable), comparatively lower ratios were observed than those reported in other population-based cohorts. The PURE-SA-NW's mdLMR range also spanned the LMRs of the three CVD patient cohorts.

Table 5-1 Descriptive characteristics of the study population according to their CVD risk

Clinical characteristics	Median (25% ; 75%)	Increased CVD risk cut-off	Reference value citation	Individuals at increased risk N* (%)
Protein-based inflammatory markers				
CRP (mg/L)	3.00 (1.52 ; 7.90)	>3.0	(17)	60/119 (50.4)
IFN-γ (pg/mL)	1.51 (0.76 ; 2.79)			
IL-6 (pg/mL)	3.97 (2.10 ; 7.47)	>1.5	(32)	104/118 (89.7)
IL-10 (pg/mL)	3.53 (2.88 ; 4.86)			
TNF-α (pg/mL)	10.1 (7.68 ; 13.1)			
Methylation-derived cell ratio markers				
MdNLR	1.34 (0.90 ; 1.71)	>1.8[&]	(33)	26/120 (21.7)
MdLMR	4.30 (3.39 ; 4.88)	<4.3[&]	(34)	60/120 (50)
cg25938803 (β)	0.32 (0.26 ; 0.38)			
cg10456459 (β)	0.38 (0.31 ; 0.47)			
cg01591037 (β)	0.38 (0.32 ; 0.45)			
cg03621504 (β)	0.25 (0.19 ; 0.34)			
cg00901982 (β)	0.30 (0.21 ; 0.35)			
Markers of cardiovascular function				
bSBP (mmHg)	137 (122 ; 147)	>140		50/120 (41.6)
bDBP (mmHg)	83.0 (77.0 ; 94.0)	>90		40/120 (33.3)
bPP (mmHg)	49.0 (41.8 ; 60.3)	≥60	(35)	36/120 (30.0)
HR (bpm)	68.0 (58.0 ; 82.0)	>80		31/120 (25.8)
cfPWV (m/s)	9.35 (8.30 ; 10.5)	>10		47/111 (42.3)
Cardiovascular risk markers				
BMI (kg/m^2)	21.2 (18.7 ; 25.3)	>25	(36)	37/117 (31.6)
LDL-C (mmol/L)	2.47 (1.77 ; 3.15)	≥2.60	(37)	51/120 (42.5)
HDL-C (mmol/L)	1.29 (0.99 ; 1.65)	<1.00		31/120 (25.8)

*Expressed as number of participants at increased risk out of the number of participants with data for the specific variable [&]Directly measured cell count ratio cut-off. bDBP, brachial diastolic blood pressure; bSBP, brachial systolic blood pressure; bPP, brachial pulse pressure; CRP, C-reactive protein; cfPWV, carotid-femoral pulse wave velocity; IFN-γ, interferon-gamma; IL-6, interleukin-6; IL-10, interleukin-10; IQR, interquartile range; mdLMR, methylation-derived lymphocyte-to-monocyte ratio; mdNLR, methylation-derived neutrophil-to-lymphocyte ratio; HR, heart rate; TNF-α, tumor necrosis factor-alpha.

5.5.1 Relationship between biomarkers of inflammation

Figure 5-1 depicts the relationship between the 13 investigated indicators of inflammation, adjusted for age and smoking status. Positive correlations were observed among the protein biomarkers. The strongest correlations were between IL-6 and IL-10 ($r = 0.44$, $p = 9.9 \times 10^{-7}$) and between TNF-α and IFN-γ ($r = 0.47$, $p = 1.1 \times 10^{-7}$). All inflammatory proteins also associated strongly with the inflammatory score ($r > 0.55$, $p < 1.6 \times 10^{-10}$ for all). Comparatively stronger correlations were observed among the methylation-derived biomarkers. The negative CpG-mdNLR and positive CpG-mdLMR correlation coefficients reflect the positive associations of these CpGs with monocytes ($r > 0.27$, $p < 0.004$ in all instances) and lymphocytes ($r > 0.70$, $p < 1.0 \times 10^{-18}$ for all), and the negative association with neutrophils ($r < -0.72$, $p < 2 \times 10^{-20}$ in all instances). No convincing evidence for protein-methylation correlations was observed.

5.5.2 Association of biomarkers of inflammation with markers of cardiovascular function

Supplementary Table 2 reports the partial Spearman correlation coefficients among the protein-based and methylation-derived biomarkers of inflammation and markers of CVF, adjusted for age. No evidence was observed for a linear relationship between BP (SBP, DBP and PP) and inflammatory markers (protein-based and methylation-derived). Comparatively stronger correlations were observed where HR and cfPWV were concerned, particularly in relation to the methylation-derived inflammatory biomarkers. The only Bonferroni significant correlations observed ($p \leq 0.004$) were between cg25938803 and HR ($r = -0.27$, $p = 0.003$), cg25938803 and cfPWV ($r = -0.29$, $p = 0.002$) and cg10456459 and HR ($r = -0.30$, $p = 0.001$).

Summary statistics of the linear regression models quantifying the relative contribution of the investigated inflammatory biomarkers to the variance in CVF markers are shown in Table 5-2. Only inflammatory biomarkers that contributed to these models (either Model 1 or 2) at a Bonferroni-adjusted significance threshold of $p < 0.004$ are reported in Table 5-2. Test statistics for all 13 investigated inflammatory markers and the five CVF markers (SBP, DBP, PP, HR and cfPWV) are reported in Supplementary Table 3.

Markers of CVF appeared to be more strongly associated with methylation-derived inflammatory markers than with protein-based biomarkers (Table 5-3, Supplementary Table 3). Regarding relationships with HR, four methylation-derived biomarkers reached the Bonferroni cut-off, although the associations were attenuated when additional covariates were added to the model. The association between HR and CRP, on the other hand, strengthened upon full adjustment (Model 2). Two CpGs, cg25938803 and cg03621504, associated negatively with cfPWV. Both associations attenuated on full adjustment with evidence remaining for the association between

cg25938803 and cfPWV. To specifically determine the impact of medications impacting the cardiovascular system we repeated our analysis and replaced total medication use with CVD medication use (detailed in Supplementary Table 1). No attenuation of associations was observed upon doing so.

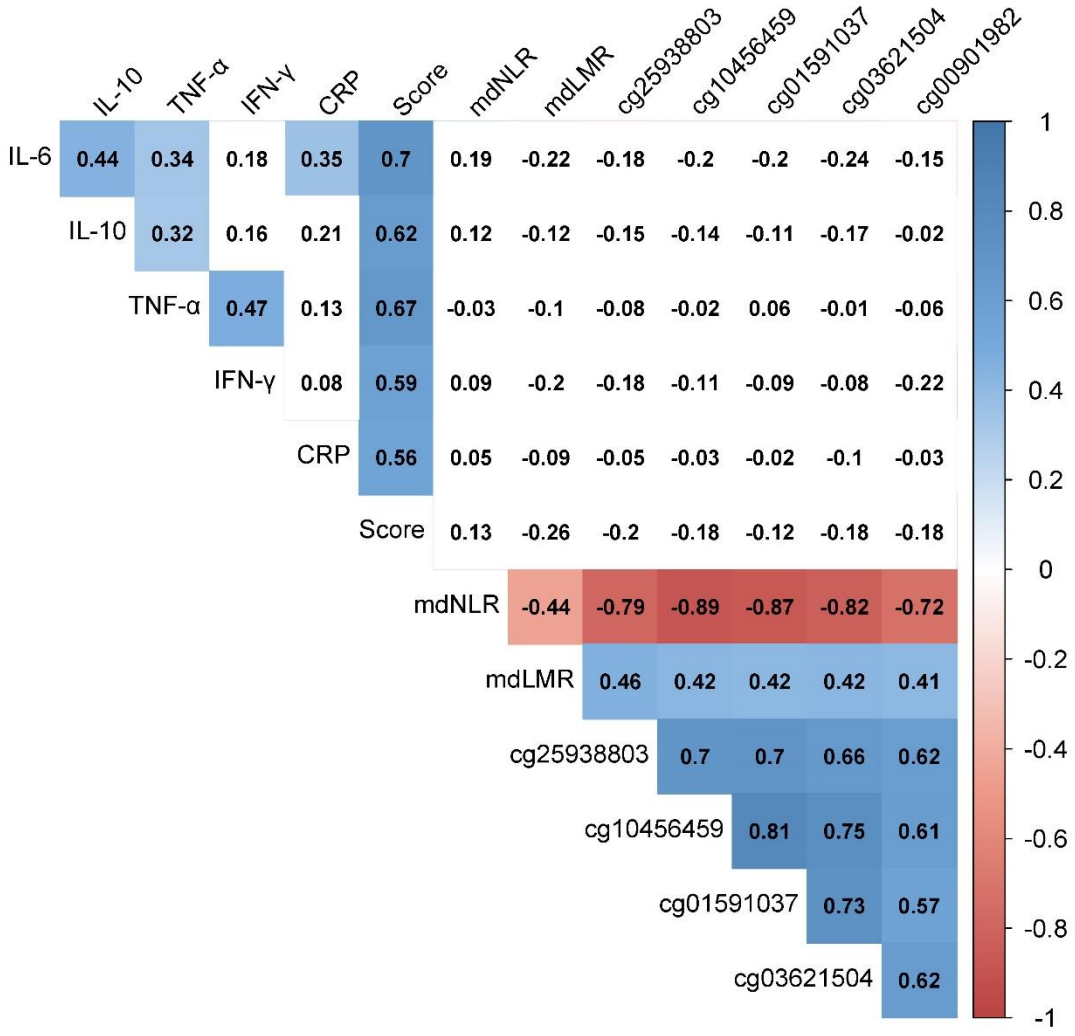


Figure 5-1 Heat map of the partial Spearman correlations among protein-based and methylation-derived biomarkers of inflammation

Numeric values indicate Spearman's rho values while controlling for age and smoking status. The presence of color indicates p <0.003 (α = 0.05/16, calculated as 4 x 4 independent inflammatory marker comparisons). The shades of color represent the strength and direction of the correlation. 'Score' represents the average of the IL-6, IL-10, TNA-α, IFN-γ and CRP Z-scores. CRP: C-reactive protein; IFN-γ: interferon-gamma; IL-6: interleukin-6; IL-10: interleukin-10; mdLMR: methylation-derived lymphocyte-to-monocyte ratio; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio; TNF-α: tumor necrosis factor-alpha.

Table 5-2 Variance in cardiovascular function explained by individual inflammatory biomarkers

Inflammatory biomarker*	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model (including the inflammatory biomarker)
			β (25% ; 75%)	P	
HR (log bpm)					
CRP	1	0%	0.04 (0.01 ; 0.07)	0.006	7%
	2	9%	0.05 (0.02 ; 0.07)	**0.001**	19%
mdNLR	1	0%	0.11 (0.04 ; 0.18)	**0.003**	7%
	2	9%	0.10 (0.02 ; 0.18)	0.01	15%
mdLMR	1	0%	–0.19 (–0.32 ; –0.07)	**0.003**	8%
	2	9%	–0.18 (–0.30 ; –0.07)	0.005	16%
cg10456459	1	0%	–0.20 (–0.32 ; –0.08)	**0.002**	8%
	2	9%	–0.17 (–0.29 ; –0.08)	0.008	15%
cg03621504	1	0%	–0.12 (–0.21 ; –0.04)	**0.004**	7%
	2	9%	–0.11 (–0.19 ; –0.04)	0.02	14%
cfPWV (log m/s)					
cg25938803	1	25%	–0.20 (–0.31 ; –0.09)	**3.8E–04**	33%
	2	36%	–0.18 (–0.29 ; –0.09)	**0.002**	42%
cg03621504	1	25%	–0.12 (–0.19 ; –0.04)	**0.002**	31%
	2	36%	–0.09 (–0.16 ; –0.04)	0.01	40%

All inflammatory and CVF biomarkers reported were $\log_e$-transformed prior to analysis. Regression coefficient interpretation should be that one per cent change in x (inflammatory marker) will induce a regression coefficient (β) per cent change in y (CVF marker). p ≤ 0.004 highlighted in bold. CRP: C-reactive protein; cfPWV: carotid-femoral pulse wave velocity; mdLMR: methylation-derived lymphocyte-to-monocyte ratio; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio; HR: heart rate. **Model 1:** CVF marker ~ (inflammatory biomarker) + age; **Model 2:** CVF marker ~ (inflammatory biomarker) + age + smoking status + dwelling place + smoking status + BMI + LDL-C + HDL-C + medication use. When cfPWV was the outcome, mean arterial pressure was additionally adjusted for.

5.5.3 Additive value of methylation-derived inflammatory markers when investigating cardiovascular function

Next, we investigated whether the methylation-derived inflammatory markers can increase the variance explained in CVF when added to a model containing known CVD risk markers, including inflammation. To this end, we included the protein-based inflammatory score, as an amalgamated biomarker of inflammation in the covariate list of Model 2 (referred to below as Model 3). To identify which methylation-derived biomarkers to investigate, we performed a backward stepwise regression analysis for all CVF phenotypes. Age (and mean arterial pressure when cfPWV was

the outcome) and the seven methylation-derived inflammatory biomarkers were added to these models as independent variables. Only the markers retained by the backward stepwise regression were included in further analyses (Table 5-3). For all three BP-related markers, cg03621504 and mdNLR were retained. For HR, cg25938803, cg10456459 and cg01591037 were retained. For cfPWV, cg25938803, cg03621504 and mdNLR were retained. The additive variance explained was determined when the retained methylation markers were added to a fully adjusted (Model 3) multivariate regression analysis. Table 5-3 reports only the models for which the addition of methylation biomarkers increased the explained variance at a Bonferroni-adjusted threshold of p <0.02. Full summary statistics are provided in Supplementary Table 4.

Table 5-3 The additive value of methylation-derived inflammatory biomarkers to known cardiovascular risk markers in relation to cardiovascular function

Regression model*	Inflammatory biomarker			Variance explained	X^2 p-value
	β (25% ; 75%)	**P**	**Contribution to CVF variance**[&]		
SBP (log mmHg)					
Model 3				14%	
+mdNLR	0.11 (0 ; 0.23)	0.05	2.2%	**22%**	**0.005**
+cg03621504	0.21 (0.08 ; 0.35)	**0.003**	7.3%		
cfPWV (log m/s)					
Model 3				41%	
+mdNLR	–0.13 (–0.26 ; –0.003)	0.05	1.5%		
+ cg25938803	–0.24 (–0.43 ; –0.06)	**0.01**	4.7%	**48%**	**0.008**
+ cg03621504	–0.11 (–0.24 ; 0.02)	0.10	1.6%		

All inflammatory and CVF biomarkers reported were $\log_e$-transformed prior to analysis. Regression coefficient interpretation should be that one per cent change in x (inflammatory marker) will induce a regression coefficient (β) per cent change in y (CVF marker). [&]*lmg* metric providing a decomposition of the model explained variance into non-negative contributions (38). X^2 p value = Chi-square p value when the regression models with and without methylation-derived inflammatory biomarkers are compared. bSBP: brachial systolic blood pressure; cfPWV: carotid-femoral pulse wave velocity; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio. **Model 3:** CVF marker ~ age + smoking status + dwelling area + BMI + LDL-C + HDL-C + medicine use + score (the average of the IL-6, IL-10, TNA-α, IFN-γ and CRP Z-scores). When cfPWV was the outcome, mean arterial pressure was additionally adjusted for.

Myeloid CpGs, cg03621504 and cg25938803 were the only methylation-derived markers that had strong evidence of individual contribution to the variance in SBP and cfPWV, respectively. The retained methylation-derived inflammatory markers contributed an additional ~7% to the variance explained in SBP and cfPWV, after age, smoking status, body composition, blood lipids, socio-economic status (represented by urban/rural status), medication use and inflammation (protein-

based) had been accounted for. For every ~2 percent methylation increase in cg03621504 a 10 mmHg change in the geometric mean of SBP was observed. For cfPWV, an increase of one m/s in the geometric mean resulted from a ~3 percent methylation increase in cg25938803. Again, findings remained robust upon replacing total medication use with CVD medication use.

5.6 Discussion

In this study we report the value of methylation-derived indicators of inflammation in relation to CVF, including BP and large artery stiffness. Although the methylation-derived and protein-based inflammatory markers did not demonstrate strong associations with each other, both reflected a similar degree of increased CVD risk. Methylation-derived markers appeared to be more strongly associated with CVF than the protein-based inflammatory markers tested. Furthermore, when exploring models performed to explain variance in CVF, we found that methylation biomarkers, particularly the myeloid CpGs, explained variance in addition to variance already explained by known CVD risk markers, including inflammation reflected by a protein-based inflammatory score. This suggests that methylation-derived inflammatory markers may complement protein-based inflammatory markers in explaining CVF marker variance, rather than simply being a proxy thereof.

5.6.1 mdNLR and mdLMR in the PURE-SA-NW cohort

Although the evidence for appropriate cut-off values for increased CVD rather than overt disease risk remains unclear, a meta-analysis of 38 studies, investigating NLRs in relation to stroke, acute coronary syndrome, coronary artery disease and a composite of these events (33), provides some guidelines. According to these guidelines, 21% of the PURE-NW cohort can be classified as at increased CVD risk. This is lower than the risk percentage indicated by CRP and IL-6 concentrations, but in agreement with the other CVF and CVD risk markers investigated, albeit in the lowest range of risk prediction. When comparing our CVD risk and mdNLRs with the individual cohorts depicted in Supplementary Figure 1, some discrepancies are, however, noted. Almost 60% of the PURE-SA-NW study population can be classified as suffering from hypertension (according to BPs and anti-hypertensive medication use), yet more than 75% of the sample population had NLRs (methylation-derived) lower than directly measured NLRs reported in a cohort with hypertension (39). It should be noted that, although validated and widely used (5), the methylation-derived cell deconvolution has been shown to underestimate neutrophil and overestimate lymphocyte proportions, by –1.66% and 0.4–1.0%, respectively (5). An average underestimation of the mdNLR *vs* directly measured NLR of 0.6 units has also been reported (1). Consequently, such a 0.6-unit increase in mdNLR will shift our study population's median to just above the CVD risk cut-off (33), resulting in a reclassification of ~50% of the cohort being at

increased risk, thereby aligning with the protein-based estimation (Supplementary Figure 1). Regardless of a possible adjustment, our data agrees with prior reports of more favorable NLRs in black populations (14), than Hispanic and white American ethnic groups. For the first time a comparison has been drawn between an African study population and European, Asian and Chinese cohorts (13, 15, 16). No notable differences were observed between our black African study population and data from European, Asian or Chinese cohorts.

In agreement with the adjusted mdNLR, the mdLMR also classified half of the PURE-SA-NW population as being at increased risk of CVD. Contrary to the mdNLR, no adjustment is required when comparing the mdLMR with directly measured ratios (5). On average, the PURE-SA-NW study population had lower mdLMRs than the LMRs previously reported for Asian, Chinese and Western Indian cohorts (15, 16, 40). An overlap in the reference ranges of these four ethnic groups that were compared was, however, clearly visible (Supplementary Figure 2). These ranges furthermore, spanned the LMRs reported in patients with coronary artery disease, coronary lesions and chronic stable angina (34, 40, 41). It is noteworthy that, altogether, the studies that investigate LMR in the context of CVD represent fewer than 1 000 individuals of whom only 162 are controls (34). The limited evidence, together with the observation that the LMRs reported in four ostensibly healthy cohorts spanned, by a wide margin, the LMRs reported for CVD patients, highlights the need for more research on the LMR and its accuracy in risk prediction, because recent evidence suggested that the LMR might, in some instances, be more useful in CVD risk prediction than the NLR (34).

5.6.2 Inflammation as a contributor to cardiovascular risk

The central finding of this study is that, although current research relies heavily on protein-based inflammatory markers for CVD risk estimation (17, 25), methylation-derived inflammatory markers appear not only to be more strongly associated with cardiovascular function than protein-based markers, but also independently so. Collectively, more variance was explained using a combination of methylation and protein markers, than the conventional protein markers on their own, suggesting that the methylation biomarkers could offer unique information on the inflammatory process and are not just surrogate markers for inflammatory proteins.

Prior evidence has shown a multitude of possible mechanisms through which individual leukocyte sub-types, proxied for in this study by methylation-derived cell count markers, may directly contribute to CVD (reviewed by 7, 34, 45). Neutrophils, for example, secrete inflammatory mediators and proteolytic enzymes related to vascular wall degeneration (7, 33). Macrophages (matured monocytes) contribute to cardiovascular risk mainly through its secretion of cytokines and reactive oxidative species once infiltrated to atherosclerotic plaque (34). On the other hand,

regulatory T-cells (a lymphocyte sub-type) is a known role player in the development of hypertension through its role in the renin–angiotensin system (39).

The implication of these findings is that using a single protein marker (typically CRP, IL-6 or TNF-α in CVD-related investigations) to adjust for 'inflammation' might not capture all the variance of the true inflammatory effect. The superiority of cell counts lies in their ability to reflect information about both the innate and complementary inflammatory processes. Adding to this advantage is the ability to provide methylation-derived cell count estimates retrospectively, from any well-preserved DNA-containing whole blood or leukocyte samples, thus enabling the reinvestigation of large sample sets already collected.

5.6.3 CpGs as an mdNLR proxy

One previous study identified (4) and another replicated (2) the use of the five investigated CpGs as surrogates for the mdNLR in cancer case-control studies in American populations. Here, we evaluated their potential use in black South African men with low-grade inflammation associated with CVD risk. All five CpGs strongly associated with the mdNLR, confirming their robust associations with myeloid cell differentiation in a population-based study of a different ethnic group than previously reported. For arterial stiffness, cg25938803 was the most important contributor, even after the mdNLR and cg03621504 had been accounted for. Similarly, for SBP, cg03621504 contributed considerably (7.3%) to the explained variance, also after protein-based inflammatory markers and the mdNLR were added to the model. Apart from its strong associations with HR and arterial stiffness, cg03621504 was the only inflammatory biomarker with evidence of a possible association with BP, contributing 3–4% to the variance in all instances, at $p < 0.05$ (Supplementary Tables 3 and 4). It is, therefore, possible that this CpG represents a novel marker of CVD risk.

Our observation that the myeloid CpGs contributed to CVF variance regardless of the prior inclusion of the mdNLR suggests that these CpGs may associate with CVF independently of cell counts. This argument is strengthened by the fact that although the five CpGs were equally reflective of the mdNLR, associations with CVF markers differed in strength. Population-specific variance may contribute to the different patterns of association. Indeed, we found that two of the investigated CpGs, cg01591037 and cg00901982, are associated with *cis* single nucleotide polymorphisms (SNPs, rs76297553 and rs6546566, respectively) that are both methylation quantitative trait loci (mQTLs) for these CpGs and are expression QTLs for local genes (42, 43). Population differences in the minor allele frequencies of these SNPs (0.3% and 26% in African Americans compared to 6% and 30% in Europeans for rs76297553 and rs6546566, respectively) have been reported by the 1000 Genomes project (44).

5.6.4 Strengths and limitations

A limitation of this study is that we only investigated men, so we are unable to generalize our findings to women from the same population. Secondly, although methylation-derived cell ratios are accurate reflections of directly measured ratios, the lack of available literature on methylation-derived cut-offs for CVD risk and outcomes hindered direct comparison with directly measured ratios, particularly in the case of NLR. As a result of our limited sample size and the stringent statistical approach followed, there may be some true positive findings not emphasized here. For this reason, we reported all our results in the supplementary material so that our findings may be replicated in larger cohorts. Lastly, although we were able to provide evidence for the independent contribution of the myeloid CpGs to CVF variance and CVD risk, we were not able to further investigate these mechanisms.

These limitations are met with various strengths. We were able to layer evidence from five inflammatory proteins frequently investigated in the context of CVD, with cytological data, whereas previous studies mostly had access to one or the other (3, 33, 45). This is particularly rare in an ostensibly healthy cohort. This is also the first time methylation-derived cell ratio estimates and the myeloid CpGs has been investigated in a Sub-Saharan African study population, which addresses the need for more ethnic representation and reference sets, particularly in epigenetic epidemiology (14, 15, 46). The lack of data on covariates known to affect cell count ratios, some even in a population-specific manner (14), such as adiposity, smoking and medication use, has hindered previous investigations of population-specific reference ranges. We addressed this by investigating a richly phenotyped study population in whom we were able to identify and evaluate many, previously unstudied, potential confounding factors.

5.7 Conclusion

The methylation-derived cell ratio estimates observed in this South African study population were comparable to previously investigated ostensibly healthy ethnic groups. The CVD risk reflected by these ratios was in accordance with that of CRP and several CVF and CVD risk markers. Five CpGs previously suggested as surrogates for the mdNLR in cancer patients were similarly highly associated with mdNLR in our cohort, regardless of the absence of overt inflammation. However, the contribution of these CpGs to CVF was independent from their effect on myeloid differentiation and robust to adjustment for known CVD risk factors, illustrating their potential functional relevance, apart from their role in myeloid differentiation. We demonstrate that population-specific genetic variance may contribute to these CpG-CVF associations even when comparable CpG-mdNLR relationships are observed. Methylation-derived and protein-based inflammatory biomarkers explain independent portions of CVF variance; the best characterization of CVF

variance is obtained when methylation biomarkers, particularly the myeloid CpGs, are included in models containing known CVD risk markers, including protein-based inflammatory markers. Cell count data are highly valuable when the aim is to characterize inflammation, particularly when DNA is available, allowing the reinvestigation of existing cohort data and potentially circumventing the need for new data collection. The widely used Infinium HumanMethylation 450K and EPIC arrays include the myeloid CpGs reported here and methods to derive cell counts have been rigorously validated for these assays. Consequently, many large epigenetic epidemiology cohorts are already in possession of the necessary data to investigate cell count-related immunomodulation.

5.8 Declarations

Authors' contributions

HTC, MP, HRE and CNR conceptualized the study. Funding was acquired by MP and FRG. AES acquired the cardiovascular function data. HTC performed the data analysis and wrote the manuscript. HRE and MP supervised the data analysis and interpreted the results with HTC. All authors contributed to the critical review and editing of the manuscript.

Funding

The PURE-SA-NW study was funded by the North-West University, South African National Research Foundation (SANRF), Population Health Research Institute, South African Medical Research Council (SAMRC), the North West Province Health Department, and the South Africa-Netherlands Research Program on Alternatives in Development. Grants from the SANRF, Academy of Medical Sciences UK (Newton Fund Advanced Fellowship Grant) and the SAMRC supported the additional epigenetic work and cytokine analysis reported herein [AMS-NAF1-Pieters to MP]. The SANRF supported HTC [SFH106264] in the time this manuscript was written. HRE works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/5). None of the funding bodies was involved in the design of the study, collection, analysis or interpretation of the data or in writing of this manuscript. Opinions expressed and conclusions arrived at are those of the authors and are not to be attributed to the funding sources.

Acknowledgments

The authors would like to thank all those that participated in the PURE-SA-NW study and those that made the PURE-SA-NW study possible, including the fieldworkers, researchers and staff of both the PURE-SA-NW (Africa Unit for Transdisciplinary Health Research (AUTHeR), Faculty of

Health Sciences, NWU, Potchefstroom, South Africa) and PURE International (S Yusuf and the PURE project office staff at the Population Health Research Institute (PHRI), Hamilton Health Sciences and McMaster University, Ontario, Canada) teams. We thank the staff of the Bristol bioresource laboratories (Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK) who generated the DNA methylation data, and Mrs Cecile Cooke for quantifying the cytokines reported here.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement

The data that support the findings of this study are available upon reasonable request and with the permission of the Health Research Ethics Committee of the North-West University and the principal investigator of the PURE-SA-NW study, Prof. I.M. Kruger (lanthe.kruger@nwu.ac.za) at the North-West University's Africa Unit for Transdisciplinary Health Research.

5.9 References

1. Koestler DC, Usset J, Christensen BC, Marsit CJ, Karagas MR, Kelsey KT, et al. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiology and Prevention Biomarkers. 2016;26(3):328-38.

2. Ambatipudi S, Langdon R, Richmond RC, Suderman M, Koestler DC, Kelsey KT, et al. DNA methylation derived systemic inflammation indices are associated with head and neck cancer development and survival. Oral Oncology. 2018;85:87-94.

3. Ambatipudi S, Sharp GC, Clarke SL, Plant D, Tobias JH, Evans DM, et al. Assessing the Role of DNA Methylation-Derived Neutrophil-to-Lymphocyte Ratio in Rheumatoid Arthritis. Journal of Immunology Research. 2018;2018:2624981.

4. Wiencke JK, Koestler DC, Salas LA, Wiemels JL, Roy RP, Hansen HM, et al. Immunomethylomic approach to explore the blood neutrophil lymphocyte ratio (NLR) in glioma survival. Clinical Epigenetics. 2017;9(1):10.

5. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biology. 2018;19(1):64.

6. Kelsey KT, Wiencke JK. Immunomethylomics: A Novel Cancer Risk Prediction Tool. Annals of the American Thoracic Society. 2018;15(Suppl 2):S76-s80.

7. Horne BD, Anderson JL, John JM, Weaver A, Bair TL, Jensen KR, et al. Which white blood cell subtypes predict increased cardiovascular risk? Journal of the American College of Cardiology. 2005;45(10):1638-43.

8. Kounis NG, Soufras GD, Tsigkas G, Hahalis G. White blood cell counts, leukocyte ratios, and eosinophils as inflammatory markers in patients with coronary artery disease. Clinical and Applied Thrombosis/Hemostasis. 2015;21(2):139-43.

9. Jhuang Y-H, Kao T-W, Peng T-C, Chen W-L, Li Y-W, Chang P-K, et al. Neutrophil to lymphocyte ratio as predictor for incident hypertension: a 9-year cohort study in Taiwan. Hypertension Research. 2019;42(8):1209-14.

10. Rudiger A, Burckhardt OA, Harpes P, Müller SA, Follath F. The relative lymphocyte count on hospital admission is a risk factor for long-term mortality in patients with acute heart failure. The American Journal of Emergency Medicine. 2006;24(4):451-4.

11. Tamhane UU, Aneja S, Montgomery D, Rogers E-K, Eagle KA, Gurm HS. Association between admission neutrophil to lymphocyte ratio and outcomes in patients with acute coronary syndrome. The American Journal of Cardiology. 2008;102(6):653-7.

12. Sun X, Luo L, Zhao X, Ye P, Du R. The neutrophil-to-lymphocyte ratio on admission is a good predictor for all-cause mortality in hypertensive patients over 80 years of age. BMC Cardiovascular Disorders. 2017;17(1):167.

13. Forget P, Khalifa C, Defour J-P, Latinne D, Van Pel M-C, De Kock M. What is the normal value of the neutrophil-to-lymphocyte ratio? BMC Research Notes. 2017;10(1):12.

14. Azab B, Camacho-Rivera M, Taioli E. Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects. PloS ONE. 2014;9(11):e112361.

15. Lee JS, Kim NY, Na SH, Youn YH, Shin CS. Reference values of neutrophil-lymphocyte ratio, lymphocyte-monocyte ratio, platelet-lymphocyte ratio, and mean platelet volume in healthy adults in South Korea. Medicine. 2018;97(26):e11138.

16. Meng X, Chang Q, Liu Y, Chen L, Wei G, Yang J, et al. Determinant roles of gender and age on SII, PLR, NLR, LMR and MLR and their reference intervals defining in Henan, China: A posteriori and big-data-based. Journal of Clinical Laboratory Analysis. 2018;32(2):e22228.

17. Pearson TA, Mensah GA, Alexander RW, Anderson JL, Cannon III RO, Criqui M, et al. Markers of inflammation and cardiovascular disease: application to clinical and public health practice: a statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. Circulation. 2003;107(3):499-511.

18. Liu X, Zhang Q, Wu H, Du H, Liu L, Shi H, et al. Blood neutrophil to lymphocyte ratio as a predictor of hypertension. American Journal of Hypertension. 2015;28(11):1339-46.

19. Balta S, Kurtoglu E, Kucuk U, Demirkol S, Ozturk C. Neutrophil–lymphocyte ratio as an important assessment tool. Expert Review of Cardiovascular Therapy. 2014;12(5):537-8.

20. Richardson TG, Zheng J, Smith GD, Timpson NJ, Gaunt TR, Relton CL, et al. Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. American Journal of Human Genetics. 2017;101(4):590-602.

21. Fernández-Sanlés A, Sayols-Baixeras S, Curcio S, Subirana I, Marrugat J, Elosua R. DNA methylation and age-independent cardiovascular risk, an epigenome-wide approach: The REGICOR (REgistre GIroní del COR) study. Arteriosclerosis, Thrombosis, and Vascular Biology. 2018;38(3):645-52.

22. van Woudenbergh GJ, Theofylaktopoulou D, Kuijsten A, Ferreira I, van Greevenbroek MM, van der Kallen CJ, et al. Adapted dietary inflammatory index and its association with a summary score for low-grade inflammation and markers of glucose metabolism: The Cohort study on Diabetes and Atherosclerosis Maastricht (CODAM) and the Hoorn study. The American Journal of Clinical Nutrition. 2013;98(6):1533-42.

23. Teo K, Chow CK, Vaz M, Rangarajan S, Yusuf S. The Prospective Urban Rural Epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. American Heart Journal. 2009;158(1):1-7.

24. Cronjé HT, Elliott HR, Nienaber-Rousseau C, Pieters M. Replication and expansion of epigenome-wide association literature in a black South African population. Clinical Epigenetics. 2020;12(1):6.

25. van Bussel BC, Henry RM, Schalkwijk CG, Dekker JM, Nijpels G, Stehouwer CD. Low-grade inflammation, but not endothelial dysfunction, is associated with greater carotid stiffness in the elderly: The Hoorn study. Journal of Hypertension. 2012;30(4):744-52.

26. Laurent S, Cockcroft J, Van Bortel L, Boutouyrie P, Giannattasio C, Hayoz D, et al. Expert consensus document on arterial stiffness: methodological issues and clinical applications. European Heart Journal. 2006;27(21):2588-605.

27. Nienaber-Rousseau C, Sotunde OF, Ukegbu PO, Myburgh PH, Wright HH, Havemann-Nel L, Moss SJ, Kruger IM, Kruger HS. Socio-Demographic and lifestyle factors predict 5-year changes in adiposity among a group of black South African adults. International Journal of Environmental Research and Public Health. 2017;14(9):1089

28. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. URL https://www.Rproject.org/

29. Joehanes R, Just A, Marioni R, Pilling L, Reynolds L, Mandaviya P, et al. Epigenetic signatures of cigarette smoking. Circulation: Cardiovascular Genetics. 2016;9(5):436-47.

30. Marttila S, Kananen L, Häyrynen S, Jylhävä J, Nevalainen T, Hervonen A, et al. Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics. 2015;16(1):179.

31. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biology. 2016;17(1):255.

32. Hoebeeck L, Rietzschel E, Langlois M, De Buyzere M, De Bacquer D, De Backer G, et al. The relationship between diet and subclinical atherosclerosis: results from the Asklepios study. European Journal of Clinical Nutrition. 2011;65(5):606.

33. Angkananard T, Anothaisintawee T, McEvoy M, Attia J, Thakkinstian A. Neutrophil lymphocyte ratio and cardiovascular disease risk: a systematic review and meta-analysis. BioMed Research International. 2018;2018: 2703518.

34. Ji H, Li Y, Fan Z, Zuo B, Jian X, Li L, et al. Monocyte/lymphocyte ratio predicts the severity of coronary artery disease: a syntax score assessment. BMC Cardiovascular Disorders. 2017;17(1):90.

35. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. European Heart Journal. 2018;39(33):3021-104.

36. NCD Risk Factor Collaboration. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants. Lancet. 2017;389(10064):37-55.

37. Mach F, Baigent C, Catapano AL, Koskina KC, Casula M, Badimon L, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. European Heart Journal. 2019;41(1):111–88.

38. Grömping U. Relative importance for linear regression in R: the package relaimpo. Journal of Statistical Software. 2006;17(1):1-27.

39. Belen E, Sungur A, Sungur MA, Erdoğan G. Increased neutrophil to lymphocyte ratio in patients with resistant hypertension. The Journal of Clinical Hypertension. 2015;17(7):532-7.

40. Sharma K, Patel AK, Shah KH, Konat A. Is neutrophil-to-lymphocyte ratio a predictor of coronary artery disease in Western Indians? International Journal of Inflammation. 2017;2017: 4136126.

41. Zouridakis EG, Garcia-Moll X, Kaski JC. Usefulness of the blood lymphocyte count in predicting recurrent instability and death in patients with unstable angina pectoris. The American Journal of Cardiology. 2000;86(4):449.

42. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics. 2016;49(1):131-38.

43. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics. 2016;49(1):139-45.

44. Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68.

45. Angkananard T, Anothaisintawee T, Ingsathit A, McEvoy M, Silapat K, Attia J, et al. Mediation Effect of Neutrophil Lymphocyte Ratio on Cardiometabolic Risk Factors and Cardiovascular Events. Scientific Reports. 2019;9(1):2618.

46. Ramsay M. Africa: continent of genome contrasts with implications for biomedical research and health. FEBS letters. 2012;586(18):2813-9.

5.10 Supplementary material

Supplementary Table 1 Descriptive characteristics of the study population

Clinical characteristics	Median (25% ; 75%)
Blood cell type proportions	
B-cells (%)	3.33 (2.51 ; 4.79)
CD4+ T-cells (%)	10.7 (7.92 ; 14.3)
CD8+ T-cells (%)	11.1 (7.18 ; 14.4)
Neutrophils (%)	46.9 (39.7 ; 53.7)
Monocytes (%)	8.94 (7.41 ; 10.8)
Natural killer cells (%)	10.6 (8.90 ; 12.6)
Education [N(%)]	
None	26 (21.7)
1–7 years of schooling	66 (55.0)
8–12 years of schooling	28 (23.3)
Alcohol consumption status [N(%)]	
Never user	56 (46.7)
Ever user	64 (53.3)
Alcohol consumption* (g/d)	20.8 (5.61 ; 47.93)
Duration of smoking [N(%)[&]]	
>10 years	45 (70.3)
5-10 years	15 (23.4)
<5 years	4 (6.3)
CVD medication use [N(%)]	
Aspirin	20 (16.7)
Statins	4 (3.33)
Anti-hypertensive medication incl.[#]	40 (33.3)
Diuretics	34 (28.3)
Beta-blockers	3 (2.5)
ACE inhibitors	13 (10.8)
Calcium channel blockers	13 (10.8)
Gamma-glutamyl transferase (U/L)	40.6 (25.8 ; 78.4)
Total cholesterol (mmol/L)	4.37 (3.64 ; 4.97)
Triglycerides (mmol/L)	1.05 (0.80 ; 1.41)
Markers of glucose metabolism	
Fasting glucose (mmol/L)	5.04 (4.65 ; 5.55)
Glycated haemoglobin (mmol/mol)	5.40 (5.10 ; 5.70)
Waist circumference (cm)	81.6 (74.4 ; 93.5)

Physical activity (index)	2.42 (1.94 ; 2.90)

CVD: Cardiovascular disease. *Current users with available consumption data only, n = 50. [&] Limited to current users, n=64. [#]Detailed summary of types of antihypertensive medication reported; participants may use more than one of these at a time.

Supplementary Table 2 **Partial Spearman correlations among protein-based and methylation-derived biomarkers of inflammation and markers of cardiovascular function**

Inflammatory biomarker	SBP	DBP	PP	HR	cfPWV
	Spearman Rho				
IL-6	–0.09	–0.09	–0.09	0.14	0.09
IL-10	–0.08	–0.01	–0.13	0.22	0.03
TNF-α	–0.05	–0.04	–0.08	0.11	0.13
IFN-γ	0.05	–0.04	0.08	–0.06	0.18
CRP	–0.07	–0.01	–0.09	0.23	0.03
Score	–0.06	–0.04	–0.09	0.24	0.15
mdNLR	–0.07	–0.04	–0.07	0.25	0.21
mdLMR	0.11	0.07	0.09	–0.23	–0.09
cg25938803	0.11	0.04	0.13	**–0.27**	**–0.29**
cg10456459	0.07	0.02	0.11	**–0.30**	–0.15
cg01591037	0.07	0.04	0.07	–0.13	–0.16
cg03621504	0.24	0.18	0.22	–0.25	–0.18
cg00901982	0.14	0.11	0.12	–0.22	–0.18

Data reported as Spearman's rho values while controlling for age. Significant correlations are highlighted in bold using the Bonferroni threshold of $p \leq 0.004$ ($\alpha = 0.05/12$ tests, calculated as 4 independent inflammatory x 3 independent CVF markers). 'Score' represents the average of the IL-6, IL-10, TNA-α, IFN-γ and CRP z-scores. CRP: C-reactive protein; IFN-γ: interferon-gamma; IL-6: interleukin-6; IL-10: interleukin-10; mdLMR: methylation-derived lymphocyte-to-monocyte ratio; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio; TNF-α: tumor necrosis factor-alpha.

Supplementary Table 3 **Variance in cardiovascular function explained by individual inflammatory biomarkers**

Inflammatory biomarker	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model
			β (25% ; 75%)	p	
SBP (log mmHg)					
IL-6	1	5%	–0.01 (–0.04 ; 0.02)	0.49	7%
	2	12%	–0.01 (–0.04 ; 0.02)	0.60	13%
IL-10	1	5%	0.02 (–0.06 ; 0.09)	0.68	5%
	2	12%	0.01 (–0.07 ; 0.09)	0.86	12%
TNF-α	1	5%	0.05 (–0.03 ; 0.13)	0.21	7%
	2	12%	0.05 (–0.04 ; 0.13)	0.27	13%
IFN-γ	1	5%	0.02 (–0.01 ; 0.05)	0.27	7%
	2	12%	0.02 (–0.02 ; 0.05)	0.30	13%
CRP	1	5%	–0.01 (–0.03 ; 0.02)	0.59	7%
	2	12%	–0.01 (–0.03 ; 0.02)	0.66	13%
Score	1	5%	–1.70 (–7.39 ; 4.34)	0.71	7%
	2	12%	–1.07 (–6.53 ; 4.70)	0.57	14%
mdNLR	1	5%	–0.03 (–0.1 ; 0.03)	0.34	6%
	2	12%	–0.04 (–0.1 ; 0.03)	0.26	13%
mdLMR	1	5%	0.08 (–0.03 ; 0.19)	0.14	7%
	2	12%	0.08 (–0.03 ; 0.19)	0.17	13%
cg25938803	1	5%	0.07 (–0.05 ; 0.2)	0.26	6%
	2	12%	0.09 (–0.04 ; 0.2)	0.17	13%
cg10456459	1	5%	0.08 (–0.03 ; 0.18)	0.16	7%
	2	12%	0.08 (–0.03 ; 0.18)	0.13	14%
cg01591037	1	5%	14.3 (–16.7 ; 56.9)	0.41	6%
	2	12%	17.7 (–14.8 ; 62.6)	0.32	12%
cg03621504	1	5%	0.10 (0.02 ; 0.17)	0.01	11%
	2	12%	0.10 (0.02 ; 0.17)	0.01	17%
cg00901982	1	5%	23.7 (–8.77 ; 67.7)	0.17	7%
	2	12%	26.0 (–7.77 ; 72.0)	0.14	13%
DBP (mmHg)					
IL-6	1	0%	–0.01 (–0.03 ; 0.01)	0.38	1%
	2	12%	–0.01 (–0.03 ; 0.01)	0.44	13%

Inflammatory biomarker	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model
			β (25% ; 75%)	p	
IL-10	1	0%	0.01 (–0.04 ; 0.06)	0.73	1%
	2	12%	0.001 (–0.05 ; 0.06)	0.97	12%
TNF-α	1	0%	0.02 (–0.04 ; 0.08)	0.52	1%
	2	12%	0.01 (–0.05 ; 0.07)	0.75	12%
IFN-γ	1	0%	–0.002 (–0.03 ; 0.02)	0.86	1%
	2	12%	–0.01 (–0.03 ; 0.02)	0.61	12%
CRP	1	0%	0.004 (–0.02 ; 0.02)	0.68	1%
	2	12%	0.004 (–0.02 ; 0.02)	0.71	12%
Score	1	0%	–0.63 (–4.77 ; 3.52)	0.77	1%
	2	12%	–1.68 (–5.89 ; 3.52)	0.43	13%
mdNLR	1	0%	–0.01 (–0.06 ; 0.03)	0.55	1%
	2	12%	–0.02 (–0.07 ; 0.02)	0.34	12%
mdLMR	1	0%	0.03 (–0.04 ; 0.11)	0.39	1%
	2	12%	0.03 (–0.05 ; 0.10)	0.47	12%
cg25938803	1	0%	0.02 (–0.07 ; 0.11)	0.62	1%
	2	12%	0.05 (–0.04 ; 0.14)	0.25	13%
cg10456459	1	0%	0.04 (–0.04 ; 0.11)	0.35	1%
	2	12%	0.05 (–0.03 ; 0.12)	0.20	13%
cg01591037	1	0%	9.62 (–12.8 ; 32.0)	0.40	1%
	2	12%	13.6 (–8.54 ; 32.0)	0.23	13%
cg03621504	1	0%	0.06 (0.01 ; 0.11)	0.03	4%
	2	12%	0.06 (0.01 ; 0.11)	0.03	15%
cg00901982	1	0%	15.3 (–6.20 ; 36.8)	0.16	2%
	2	12%	17.1 (–4.23 ; 36.8)	0.12	14%
PP (log mmHg)					
IL-6	1	9%	–0.01 (–0.07 ; 0.04)	0.60	12%
	2	12%	–0.01 (–0.06 ; 0.04)	0.78	14%
IL-10	1	9%	0.01 (–0.12 ; 0.14)	0.87	8%
	2	12%	0.001 (–0.14 ; 0.14)	0.98	11%
TNF-α	1	9%	0.08 (–0.06 ; 0.22)	0.28	10%
	2	12%	0.08 (–0.07 ; 0.22)	0.28	13%
IFN-γ	1	9%	0.06 (0.00 ; 0.11)	0.06	12%
	2	12%	0.06 (0.00 ; 0.11)	0.04	16%

Inflammatory biomarker	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model
			β (25% ; 75%)	p	
CRP	1	9%	–0.03 (–0.07 ; 0.02)	0.25	12%
	2	12%	–0.02 (–0.07 ; 0.02)	0.34	15%
Score	1	9%	–1.98 (–10.89 ; 7.82)	0.68	12%
	2	12%	–1.54 (–11.1 ; 9.06)	0.76	14%
mdNLR	1	9%	–0.06 (–0.17 ; 0.05)	0.29	10%
	2	12%	–0.06 (–0.17 ; 0.05)	0.32	13%
mdLMR	1	9%	0.14 (–0.04 ; 0.33)	0.13	10%
	2	12%	0.14 (–0.06 ; 0.33)	0.16	14%
cg25938803	1	9%	0.15 (–0.07 ; 0.36)	0.17	10%
	2	12%	0.14 (–0.09 ; 0.36)	0.23	13%
cg10456459	1	9%	0.14 (–0.05 ; 0.32)	0.15	10%
	2	12%	0.13 (–0.06 ; 0.32)	0.18	13%
cg01591037	1	9%	23.2 (–29.8 ; 116)	0.42	9%
	2	12%	24.8 (–27.4 ; 115)	0.46	12%
cg03621504	1	9%	0.14 (0.02 ; 0.27)	0.03	12%
	2	12%	0.15 (0.01 ; 0.27)	0.03	16%
cg00901982	1	9%	26.6 (–25.0 ; 114)	0.37	9%
	2	12%	29.1 (–25.0 ; 122)	0.35	13%
HR (log bpm)					
IL-6	1	0%	0.02 (–0.02 ; 0.05)	0.34	2%
	2	9%	0.02 (–0.02 ; 0.05)	0.30	12%
IL-10	1	0%	0.09 (0.00 ; 0.17)	0.04	4%
	2	9%	0.11 (0.02 ; 0.17)	0.02	14%
TNF-α	1	0%	0.04 (–0.06 ; 0.13)	0.44	1%
	2	9%	0.05 (–0.05 ; 0.13)	0.34	11%
IFN-γ	1	0%	–0.02 (–0.06 ; 0.02)	0.25	1%
	2	9%	–0.03 (–0.07 ; 0.02)	0.17	12%
CRP	**1**	**0%**	**0.04 (0.01 ; 0.07)**	**0.006**	**7%**
	2	**9%**	**0.05 (0.02 ; 0.07)**	**0.001**	**19%**
Score	1	0%	5.96 (–0.79 ; 13.2)	0.08	3%
	2	9%	7.20 (0.05 ; 14.9)	0.05	13%
mdNLR	**1**	**0%**	**0.11 (0.04 ; 0.18)**	**0.003**	**7%**
	2	**9%**	**0.10 (0.02 ; 0.18)**	**0.01**	**15%**

Inflammatory biomarker	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model
			β (25% ; 75%)	p	
mdLMR	**1**	**0%**	**–0.19 (–0.32 ; –0.07)**	**0.003**	**8%**
	2	**9%**	**–0.18 (–0.30 ; –0.07)**	**0.006**	**16%**
cg25938803	1	0%	–0.20 (–0.35 ; –0.06)	0.005	7%
	2	9%	–0.15 (–0.30 ; –0.06)	0.05	13%
cg10456459	**1**	**0%**	**–0.20 (–0.32 ; –0.08)**	**0.002**	**8%**
	2	**9%**	**–0.17 (–0.29 ; –0.08)**	**0.01**	**15%**
cg01591037	1	0%	–25.8 (–48.6 ; 7.24)	0.11	2%
	2	9%	–18.7 (–44.2 ; 18.6)	0.28	10%
cg03621504	**1**	**0%**	**–0.12 (–0.21 ; –0.04)**	**0.004**	**7%**
	2	**9%**	**–0.11 (–0.20 ; –0.04)**	**0.02**	**14%**
cg00901982	1	0%	–35.2 (–54.3 ; –7.95)	0.02	5%
	2	9%	–31.2 (–52.0 ; –1.38)	0.04	13%
cfPWV (log bpm)					
IL-6	1	25%	0.02 (–0.01 ; 0.05)	0.16	26%
	2	36%	0.02 (–0.01 ; 0.05)	0.19	38%
IL-10	1	25%	0.03 (–0.04 ; 0.09)	0.48	26%
	2	36%	0.04 (–0.02 ; 0.09)	0.20	38%
TNF-α	1	25%	0.04 (–0.03 ; 0.12)	0.28	27%
	2	36%	0.06 (–0.02 ; 0.12)	0.13	40%
IFN-γ	1	25%	0.03 (–0.01 ; 0.06)	0.12	27%
	2	36%	0.02 (–0.01 ; 0.06)	0.23	40%
CRP	1	25%	0.01 (–0.01 ; 0.03)	0.41	25%
	2	36%	0.02 (–0.01 ; 0.03)	0.13	38%
Score	1	25%	4.33 (–1.26 ; 10.2)	0.13	26%
	2	36%	5.97 (0.41 ; 11.85)	0.04	41%
mdNLR	1	25%	0.07 (0.01 ; 0.14)	0.02	29%
	2	36%	0.06 (0.00 ; 0.14)	0.05	39%
mdLMR	1	25%	–0.12 (–0.22 ; –0.02)	0.02	28%
	2	36%	–0.08 (–0.18 ; –0.02)	0.11	38%
cg25938803	**1**	**25%**	**–0.20 (–0.31 ; –0.09)**	**3.8E–04**	**33%**
	2	**36%**	**–0.18 (–0.29 ; –0.09)**	**0.002**	**42%**
cg10456459	1	25%	–0.10 (–0.21 ; 0.00)	0.05	27%
	2	36%	–0.08 (–0.18 ; 0.00)	0.12	38%

Inflammatory biomarker	Model	CVF variance explained by covariates	Inflammatory biomarker's contribution to the model		Variance explained by full model
			β (25% ; 75%)	p	
cg01591037	1	25%	–24.5 (–44.4 ; 2.40)	0.07	27%
	2	36%	–16.9 (–38.5 ; 12.2)	0.22	37%
cg03621504	**1**	**25%**	**–0.12 (–0.19 ; –0.04)**	**0.002**	**31%**
	2	**36%**	**–0.09 (–0.16 ; –0.04)**	**0.01**	**40%**
cg00901982	1	25%	–32.1 (–49.0 ; –9.58)	0.009	30%
	2	36%	–25.0 (–43.4 ; –0.55)	0.05	39%

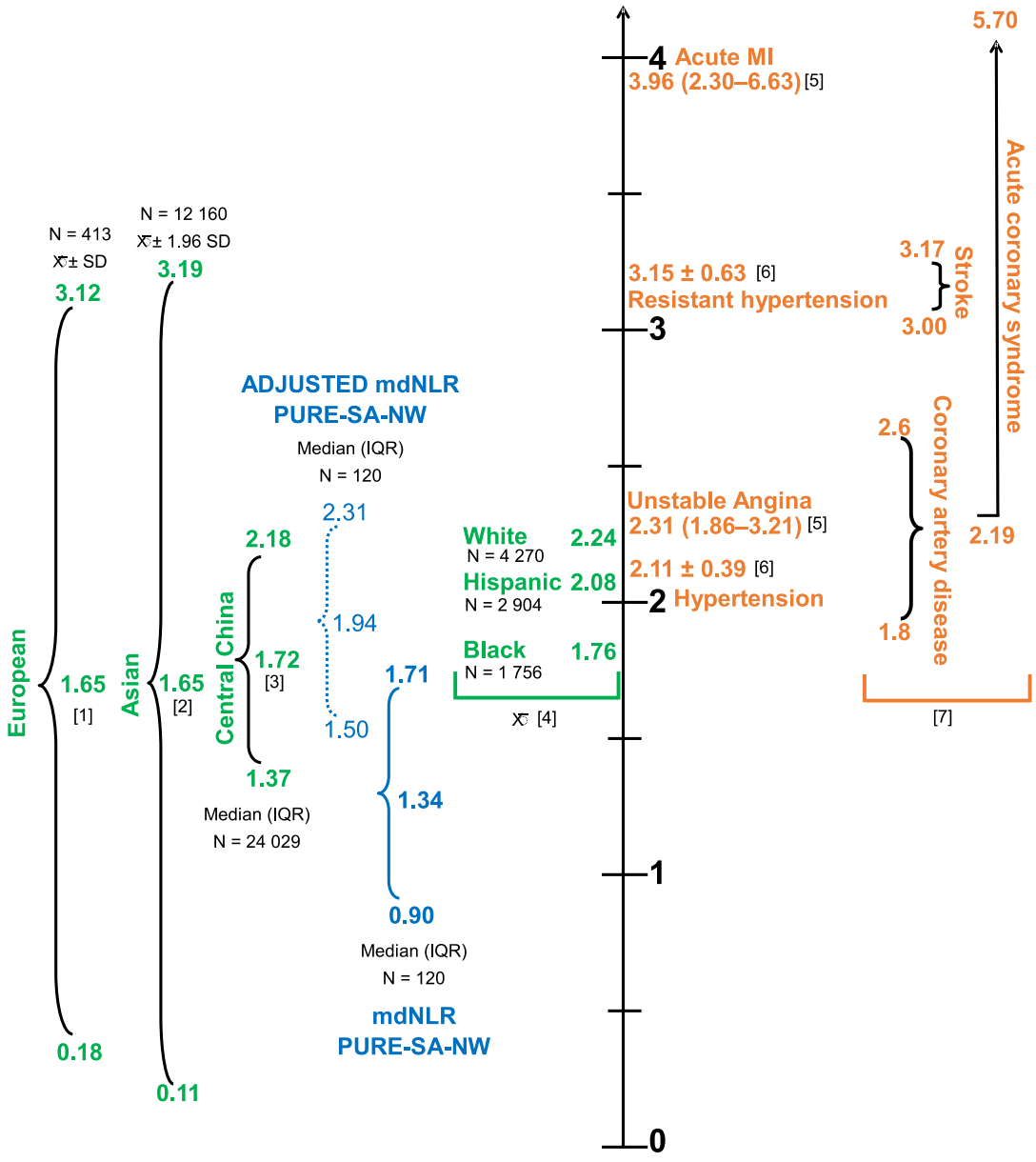
*With the exception of cg00901982, cg01591037 and DBP all inflammatory and cardiovascular biomarkers were loge-transformed prior to analysis. Linear-log and log-linear regression coefficients are presented as marginal effects. For log(CVF marker/'y') ~ log(inflammatory marker/'x'), interpretation should be that one per cent change in x will induce a regression coefficient (β) per cent change in y. For log(y)~x, interpretation should be that one unit change in x will induce a β per cent change in y. For y~log(x), interpretation should be one per cent change in x results in a β unit change in y. 'Score' represents the average of the IL-6, IL-10, TNA-α, IFN-γ and CRP Z-scores (derived from $\log_e$-transformed data for all). **Model 1:** CVF marker ~ (inflammatory biomarker) + age; **Model 2:** CVF marker ~ (inflammatory biomarker) + age + smoking status + dwelling place + smoking status + BMI + LDL-C + HDL-C + medication use. When cfPWV was the outcome, mean arterial pressure was additionally adjusted for. DBP: diastolic blood pressure; SBP: systolic blood pressure; PP: pulse pressure; CRP: C-reactive protein; cfPWV: carotid-femoral pulse wave velocity; IFN-γ: interferon-gamma; IL-6: interleukin-6; IL-10: interleukin-10; IQR: interquartile range; mdLMR: methylation-derived lymphocyte-to-monocyte ratio; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio; HR: heart rate; TNF-α: tumour necrosis factor-alpha.

Supplementary Table 4 **The additive value of methylation-derived inflammatory biomarkers to known cardiovascular risk markers in relation to cardiovascular function**

Regression model*	Inflammatory biomarker			Total variance explained	X^2 p-value
	β (25% ; 75%)	**p**	**Contribution to CVF variance&**		
SBP (log mmHg)					
Model 3				14%	**0.005**
+mdNLR	0.11 (0 ; 0.23)	0.05	2.2%	**22%**	
+cg03621504	0.21 (0.08 ; 0.35)	**0.003**	7.3%		
DBP (mmHg)					
Model 3				13%	0.03
+mdNLR	0.06 (–0.02 ; 0.01)	0.15	1.5%	19%	
+ cg03621504	0.01 (0.02 ; 0.02)	0.02	5.4%		
PP (log mmHg)					
Model 3				14%	0.03
+mdNLR	0.17 (–0.03; 0.37)	0.10	1.4%	20%	
+ cg03621504	0.30 (0.06 ; 0.54)	0.01	4.5%		
HR (log bpm)					
Model 3				13%	0.04
+ cg25938803	–0.09 (–0.31 ; 0.14)	0.44	2.0%	**20%**	
+ cg10456459	–0.27 (–0.50 ; –0.05)	0.02	4.8%		
+ cg01591037	111.7 (7.25 ; 322)	0.03	2.2%		
cfPWV (log m/s)					
Model 3				41%	**0.008**
+mdNLR	–0.13 (–0.26 ; –0.003)	0.05	1.5%	**48%**	
+ cg25938803	–0.24 (–0.43 ; –0.06)	**0.01**	4.7%		
+ cg03621504	–0.11 (–0.24 ; 0.02)	0.10	1.6%		

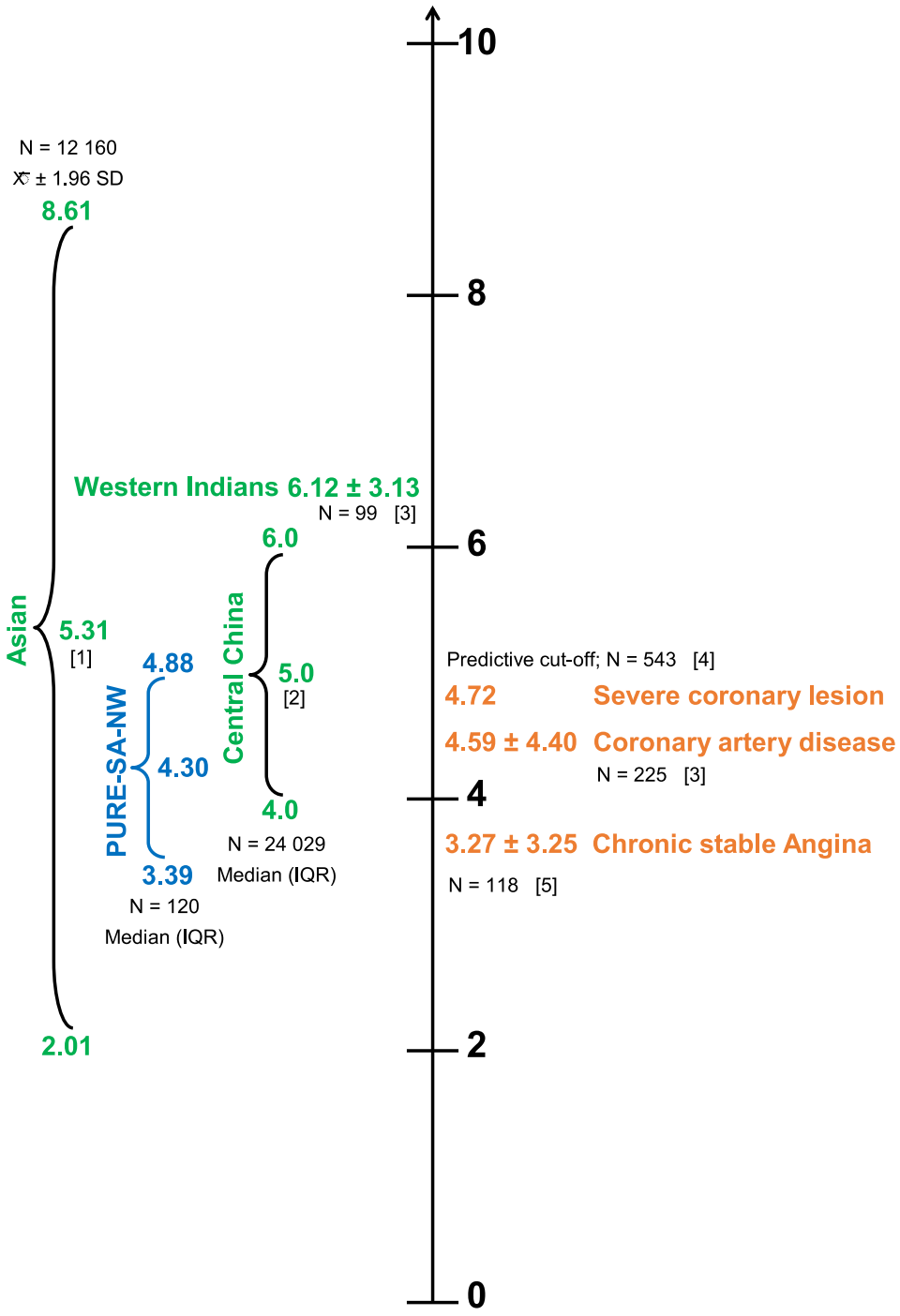
*With the exception of cg00901982, cg01591037 and DBP all inflammatory and cardiovascular biomarkers were loge-transformed prior to analysis. Linear-log and log-linear regression coefficients are presented as marginal effects. For log(CVF marker/'y') ~ log(inflammatory marker/'x'), interpretation should be that one per cent change in x will induce a regression coefficient (β) per cent change in y. For log(y)~x, interpretation should be that one unit change in x will induce a β per cent change in y. For y~log(x), interpretation should be one per cent change in x results in a β unit change in y. . The &*lmg* metric providing a decomposition of the model explained variance into non-negative contributions (Grömping, 2006). X^2 p value = Chi-square p value when the regression models with and without methylation-derived inflammatory biomarkers are compared DBP: diastolic blood pressure; SBP: systolic blood pressure; PP: pulse pressure; cfPWV: carotid-femoral pulse wave velocity; mdNLR: methylation-derived neutrophil-to-lymphocyte ratio. **Model 3:** CVF marker ~ age + smoking status + dwelling area + BMI +

LDL-C + HDL-C + medicine use + score (the average of the IL-6. IL-10. TNA-α. IFN-γ and CRP Z-scores). When cfPWV was the outcome. mean arterial pressure was additionally adjusted for.



Supplementary Figure 1 **PURE-SA-NW methylation-derived NLR compared to previously published directly measured NLR reference ranges for healthy and at-risk groups**

IQR: interquartile range; md: methylation derived; NLR: neutrophil-to-lymphocyte ratio; PURE-SA-NW: Prospective urban and rural epidemiology study cohort in South Africa's North West province. Adjusted mdNLR: mdNLR + 0.6 – accounting for reported differences in methylation-derived and directly measured NLR (Koestler et al. (2016), see Discussion). Only evidence relating to cardiovascular disease risk is reported. Higher NLRs are associated with poor survival. References: [1] Forget et al. (2017); [2] Lee et al. (2018), [3] Meng et al. (2018); [4] Azab et al. (2014); [5] Tahto et al. (2017); [6] Belen et al. (2015); [7] Angkananard et al. (2018). Reference 4 reports on cohorts from the United States. Black and White refer to self-identified non-Hispanic black and non-Hispanic white groups. References 5 and 6 report groups (n = 50) defined by specific diseases. Reference 7 is a meta-analysis reporting on the cut-offs used when evaluating the odds of CVD outcomes; multiple studies of different group sizes are reported.



Supplementary Figure 2 **PURE-SA-NW methylation-derived LMR compared to previously published directly measured LMR ranges for healthy and at-risk groups**

IQR: interquartile range; LMR: lymphocyte-to-monocyte ratio; Prospective urban and rural epidemiology study cohort in South Africa's North West province. Only evidence relating to cardiovascular disease risk is reported. Lower LMRs are associated with poor survival. References: [1] Lee et al. (2018); [2] Meng et al. (2018); [3] Sharma et al. (2017); [4] Ji et al. (2017); [5] Zouridakis et al. (2000).

References

Angkananard T, Anothaisintawee T, McEvoy M, Attia J, Thakkinstian A. Neutrophil lymphocyte ratio and cardiovascular disease risk: a systematic review and meta-analysis. BioMed Research International. 2018;2018: 2703518.

Azab B, Camacho-Rivera M, Taioli E. Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects. PloS One. 2014;9(11):e112361.

Belen E, Sungur A, Sungur MA, Erdoğan G. Increased neutrophil to lymphocyte ratio in patients with resistant hypertension. The Journal of Clinical Hypertension. 2015;17(7):532-7.

Forget P, Khalifa C, Defour J-P, Latinne D, Van Pel M-C, De Kock M. What is the normal value of the neutrophil-to-lymphocyte ratio? BMC Research Notes. 2017;10(1):12.

Grömping U. Relative importance for linear regression in R: the package relaimpo. Journal of Statistical Software. 2006;17(1):1-27.

Ji H, Li Y, Fan Z, Zuo B, Jian X, Li L, et al. Monocyte/lymphocyte ratio predicts the severity of coronary artery disease: a syntax score assessment. BMC Cardiovascular Disorders. 2017;17(1):90.

Koestler DC, Usset J, Christensen BC, Marsit CJ, Karagas MR, Kelsey KT, et al. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. Cancer Epidemiology and Prevention Biomarkers. 2016;26(3):328-38.

Lee JS, Kim NY, Na SH, Youn YH, Shin CS. Reference values of neutrophil-lymphocyte ratio, lymphocyte-monocyte ratio, platelet-lymphocyte ratio, and mean platelet volume in healthy adults in South Korea. Medicine. 2018;97(26):e11138.

Meng X, Chang Q, Liu Y, Chen L, Wei G, Yang J, et al. Determinant roles of gender and age on SII, PLR, NLR, LMR and MLR and their reference intervals defining in Henan, China: A posteriori and big-data-based. Journal of Clinical Laboratory Analysis. 2018;32(2):e22228.

Sharma K, Patel AK, Shah KH, Konat A. Is neutrophil-to-lymphocyte ratio a predictor of coronary artery disease in Western Indians? International Journal of Inflammation. 2017;2017: 4136126.

Tahto, E., Jadric, R., Pojskic, L. & Kicic, E. Neutrophil-to-lymphocyte ratio and its relation with markers of inflammation and myocardial necrosis in patients with acute coronary syndrome. Medical Archives. 2017;71(5):312-15.

Zouridakis EG, Garcia-Moll X, Kaski JC. Usefulness of the blood lymphocyte count in predicting recurrent instability and death in patients with unstable angina pectoris. The American Journal of Cardiology. 2000;86(4):449.

CHAPTER 6

MANUSCRIPT FOUR – ORIGINAL RESEARCH

This manuscript has been submitted for publication in Aging (006868).

Publisher: Impact Journals

Impact factor: 5.52

Journal aims and scope:

Aging publishes high-impact research papers of general interest and biological significance in all fields of aging research including but not limited to cellular senescence, organismal aging, age-related diseases, DNA damage response and repair, genetic control of aging from yeast to mammals, regulation of longevity, evolution of aging, anti-aging strategies and drug development and especially the role of signal transduction pathways in aging and potential approaches to modulate these signalling pathways to extend lifespan. The basic criterion for considering papers is their impact on aging research.

Author's guidelines:

https://www.aging-us.com/for-authors

Notes:

This manuscript is presented in US English

COMPARING DNA METHYLATION CLOCKS IN BLACK SOUTH AFRICAN MEN

H. Toinét Cronjé[1*], Cornelie Nienaber-Rousseau[1], Josine L. Min[2,3], Fiona R. Green[4], Hannah R. Elliott[2,3], Marlien Pieters[1]

[1] Centre of Excellence for Nutrition, North-West University, Potchefstroom, South Africa

[2] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[3] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[4] Formerly School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

***Corresponding author:**

H. Toinét Cronjé

23520825@nwu.ac.za

Keywords: GrimAge, PhenoAge, phenotypic age, biological age, smoking

6.1 Abstract

Accelerated biological *vs* chronological aging is associated with a variety of chronic diseases and mortality. DNA methylation (DNAm) clocks are widely used to estimate biological age, although only limited data are available on their behavior in non-European ethnicities. The online methylation age clock provides aging estimates from three first-generation (Horvath, Hannum and skin and blood (SB)) and two next-generation (PhenoAge and GrimAge) DNAm clocks. This study characterizes the behavior of these five clocks in a healthy cohort of older black South African men, aged 45 to 88. Of the first-generation clocks, SBAge had the strongest correlation with chronological age and the least variation in age acceleration. Our data confirm the tendency of all the tested clocks to underestimate the biological age of older individuals. The Horvath, Hannum and SB clocks best estimated chronological age at 55, 29 and 44 years, respectively. GrimAge provided superior characterization of age-related biophysiological decline compared to PhenoAge, which estimated an average biological age 16 years younger than the cohort's average chronological age. The superior performance of the GrimAge *vs* PhenoAge could be because of its incorporation of smoking-related biophysiological decline, particularly because more than half of the study sample were current smokers.

6.2 Introduction

As global life expectancy continues to increase, the chronic disease burden expands and the need for better understanding of how to promote healthy aging is emphasized [1]. Chronological age is an integral component of frailty, non-communicable disease risk and mortality [2]. Although easily accessible and standardized, chronological age as a biomarker is limited by its inability to portray changes in biological functionality accurately over the life span, especially in later life [3]. A group of peers, for example, may be the same chronological age, while exhibiting a spectrum of age-related deterioration [3]. For this reason, extensive efforts have been made to develop markers that are able to reflect biophysiological aging better than years since birth do [3, 4]. Ultimately, the availability of such markers could allow for improved targeted intervention through the identification of high-risk, functionally declining individuals before clinical symptoms appear [4].

DNA methylation (DNAm) refers to attachment of a methyl group to a DNA base. DNAm changes accumulate with age [5, 6] and are thought to mediate the effects of environmental risk factors on disease [7]. DNAm levels at specific cytosine-phosphate-guanine sites (referred to as clock CpGs), can also be used to predict chronological age [4, 8-10]. These predictors are termed “epigenetic clocks” and quantify, in years, a chronological-age-independent, biological age estimate [4]. Residuals from the regression of epigenetic age on chronological age are defined as DNAmAge acceleration (DNAmAgeAccel). When biological age is estimated to be older than chronological age, it is considered as positive age acceleration. Age acceleration is used as a biomarker of aging and has been associated with diabetes [11], cancer [12], cardiovascular disease [13] and all-cause mortality [10, 12, 14], although causality still has to be established [4, 15].

To date, multiple DNAmAge clocks have been developed, with some variation in composition and outcome. The Horvath [8], Hannum [16] and skin and blood (SB, [10]) clocks are three of the first-generation clocks most often used and readily calculated [8]. These DNAmAge clocks continue to be widely used because of their ability to robustly predict either the chronological age of unknown donors or biological age discrepancies in specific tissues in a single individual. They differ mainly in terms of the number of CpGs included in the clock model and the target tissues they were developed for: the Horvath clock is a multi-tissue clock [8], whereas the Hannum clock is blood-based [16] and the SB clock is a peripheral tissue and *ex vivo* trained age-estimator [10]. The clocks were developed using similar penalized regression protocols, where chronological age is regressed against genome-wide DNAm and consist of 353 (Horvath), 71 (Hannum) and 391 (SB) CpGs, respectively [8, 10, 16].

Because the first-generation clocks were developed with chronological age as the sole outcome of interest, they often fail to capture the inter-individual methylation differences that predict biophysiological decline above that of advancing age itself [3, 4]. The next-generation clocks; PhenoAge [17] and GrimAge [18], address this limitation by capturing the physiological dysregulation and morbidity and mortality risk associated with age acceleration, as opposed to estimating chronological age. Instead of relying solely on chronological age, these models incorporate a composite outcome of aging-related clinical measurements that differentiate between healthy and unhealthy aging. For PhenoAge, ten clinical biomarkers were chosen based on their ability to predict age-related mortality and then aggregated by means of a weighted average to create a single aging biomarker, referred to as phenotypic age. A penalized regression of DNAm on this phenotypic age resulted in the selection of the 513 CpGs that comprise the PhenoAge clock. Because these CpGs were selected based on their associations with a biomarker of age-related mortality risk, PhenoAge is said to represent the biological age-associated physiological dysregulation that relates to mortality risk [17]. For GrimAge, seven protein markers and smoking pack-years were chosen as variables of interest based on their ability to predict time to death [18]. For each of these GrimAge components, a methylation surrogate was developed, consisting of a variable number of CpGs that, when viewed collectively (1 030 CpGs in total), represent the GrimAge clock CpGs. A distinguishing characteristic of the GrimAge is its ability to encapsulate the accelerated aging brought on by a modifiable lifestyle behavior, namely smoking [18].

While these clocks are widely used and studied, a number of limitations in their use have been identified. Firstly, a systematic underestimation of Horvath and Hannum DNAmAge in older adults has been reported [19], necessitating further investigation into their use in older populations. Secondly, although ethnic differences in the behavior of the epigenetic clocks have been reported [17, 20, 21], most of the current literature represent data obtained from individuals of European ancestry only, with very limited information available on other ethnic groups. A recent editorial [22], furthermore, proposed testing multiple measures of DNAmAge and DNAmAgeAccel simultaneously, to determine the relative utility of these markers and to steer future research. We, therefore, compared the behavior of five DNAm clocks available on the new DNAm age calculator (https://dnamage.genetics.ucla.edu/new, [8]) in a group of black South African men between the ages of 45 and 88. For the Horvath, Hannum and SB clocks we investigated their accuracy in age estimation (and the potential issue of underestimation in older adults) and for the PhenoAge and GrimAge estimators, we evaluated the relative age-related mortality risk and predicted time to death, respectively, in a population with a particularly high prevalence of smoking.

6.3 Results and discussion

This study evaluates the behavior of five epigenetic clocks in a sub-cohort of 120 apparently healthy black men, aged 45 to 88, from the North-West arm of the Prospective Urban and Rural Epidemiology study in South Africa (PURE-SA-NW). Approximately half (61 and 58) of the 120 participants were current smokers and alcohol consumers, respectively, and 48 participants (40% of the cohort) were using both substances at the time of data collection. The majority of the study population were of normal weight, with only 21% and 8% of the participants, respectively, classified as overweight (body mass index (BMI) ≥25 – 29.9 kg/m^2) or obese (BMI ≥30 kg/m^2). Fifty-five percent of the study population had undergone 1–7 years, and 23% had undergone 8–12 years of schooling. The remainder of the population had received no formal education.

6.3.1 Comparison of biological age and age acceleration estimates with chronological age

Table 6-1 reports the means and standard deviations of three first-generation and two next-generation DNAmAge and DNAmAgeAccel estimates. The association of each DNAmAge estimate with chronological age is also reported. The first and next-generation clocks are discussed separately in the following sections, because of the integral differences in their aims and outcomes. While the first-generation clocks were developed with the aim of accurately estimating chronological age [8, 10, 16], the PhenoAge and GrimAge clocks aimed to provide estimators of age-related mortality risk (PhenoAge, [17]) and time to death (GrimAge, [18]).

Table 6-1 Descriptive characteristics of age, DNAmAge and DNAmAgeAccel measurements in the PURE-SA-NW study population

Age measure (years)	Mean ± SD	Correlation with chronological age	
		r	P
Chronological Age	63 ± 10		
First-generation clocks			
HorvathAge	59 ± 8	0.58	2.3E–12
HannumAge	47 ± 8	0.64	2.2E–15
SBAge	54 ± 8	0.70	7.4E–19
IEAA	0 ± 6.4		
EEAA	0 ± 7.6		
SBAA	0 ± 5.5		
Next-generation clocks			
PhenoAge	47 ± 9	0.51	2.7E–09
GrimAge	64 ± 9	0.80	1.9E–28
PhenoAA	0 ± 7.5		
GrimAA	0 ± 5.3		

AA: Age acceleration; EEAA: Extrinsic epigenetic age acceleration; IEAA: Intrinsic epigenetic age acceleration; SBAge: Skin and blood age; SBAA: Skin and blood age acceleration; SD: Standard deviation.

6.3.1.1 First-generation clocks

SBAge demonstrated the highest correlation with chronological age, although HorvathAge gave the closest estimate of mean age within our cohort (Table 6-1). These correlations were comparatively weaker than those obtained in the SBAge validation analysis, where the correlation coefficients between chronological age and predicted age were 0.96 for HorvathAge, 0.97 for HannumAge and 0.98 for SBAge [10]. The biological ages predicted by these clocks are known to differ within the same dataset, in part because of the limited overlap in represented CpGs and the likelihood that each clock (and the clock CpGs it contains) captures varying degrees of cell count, environmental and ethnic influences or confounding [4, 23, 24]. The SB clock shares 45 loci with the Hannum and 60 with the Horvath clocks [10], while the last-named two share only five [8, 16]. In our data HorvathAge correlated with

HannumAge and SBAge with a correlation coefficient of 0.62 (p = 1.5E–13) and 0.64 (p = 6.0E-15) respectively. HannumAge and SBAge associated with each other more strongly, with r = 0.71 (p = 2.8E–19), probably owing to their overlap in CpGs (63% of the Hannum clock CpGs are in the SB clock model) and the similarity of the target tissue on which they were trained.

The weaker correlations observed in our data compared to the validation study are potentially due to the differences in age ranges between the studies. The validation study compared the three first-generation clocks using samples from individuals aged 19 to 82 years [10] compared to the age range of 45-88 in this study population. El Khoury and colleagues [19] recently highlighted the tendency of the Horvath and Hannum clocks to systematically underestimate the chronological age in older individuals. However, because older individuals fall victim to age-related diseases at a much higher rate, accurate age estimation in these groups might be more important than for their younger counterparts. In addition, it has been reported that the size of associations between lifestyle factors and biological age tend to be bigger as the study population gets older [25].

Based on theories of methylation saturation with age and age-related decay, El Khoury hypothesized that the underestimation could be ameliorated by adding more loci [19]. Figure 6-1 depicts scatterplots comparing the relative underestimation of chronological age by the Horvath (top), Hannum (middle) and SB clocks (bottom), respectively. On the left, chronological age is plotted against the difference between each first-generation estimated age and chronological age (linear regression line of best fit depicted in blue). On the right-hand side of Figure 6-1, Bland Altman plots, comparable to the reporting of El Khoury *et al.* [19], depict the relationship between the mean of each estimated and chronological age against the difference between each estimated and chronological age.

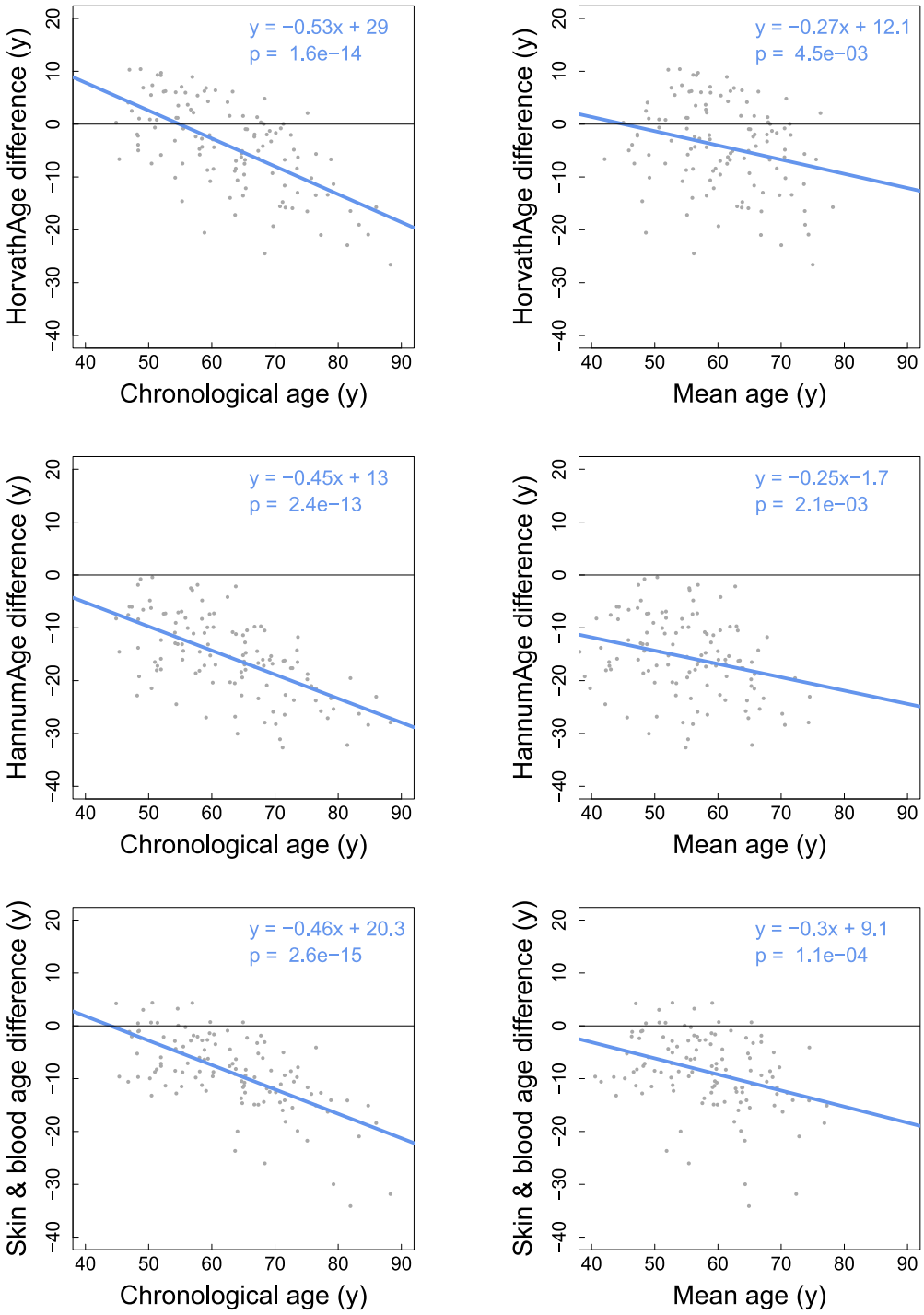


Figure 6-1 Scatterplots illustrating the relative difference in biological *vs* chronological age by three first-generation DNAmAge estimates

Left: Scatterplots depicting the relationship between chronological age and the age difference when subtracting chronological age from the DNAmAge estimates. Right: Bland Altman plots depicting the mean of each DNAmAge and chronological age against the difference between each DNAmAge and chronological age. The line of best fit from a linear regression model is formulated as y = mx + c in the top right corner and depicted in blue on the plot. The p-value represents the statistical significance of the linear regression model.

We replicated El Khoury's findings [19] by observing a similar degree of underestimation of chronological age by the Horvath and Hannum clocks, as well as a similar systematic trend where this deviation increases with age. The trend of underestimation was stronger for HorvathAge than HannumAge, although the HannumAges were consistently lower than the HorvathAges. When applied to the PURE-SA-NW group, the Horvath clock was most accurate at a chronological age of 55 years (evaluated using the y = 0 intercept of the linear regression line of best fit) and the HannumAge estimates seemed to be most accurate for individuals younger than those represented in our cohort, at the chronological age of 29.

Contrary to El Khoury's [16] hypothesis that the addition of more loci would correct the underestimation, the SB clock did not perform better than the HannumAge or HorvathAge. Evidence to reject or accept the hypothesis is, however, not yet sufficient, given that the CpGs are still relatively few, compared to the 513 and 1 030 loci of the PhenoAge and GrimAge clocks discussed in the next section. Optimal age estimation for the SB clock in this population was obtained for a chronological age of 44 years, which is older than the optimal age of estimation for the Hannum, but younger than for the Horvath clocks. As a sensitivity analysis, additional adjustment for white blood cell (WBC) counts, known to vary with chronological age, did not significantly alter the degree of underestimation.

In terms of age acceleration (AA) measures, in agreement with the SBAge having the strongest correlation with chronological age, SBAA also displayed the least amount of AA variance, followed by intrinsic epigenetic AA (IEAA) and then extrinsic epigenetic AA (EEAA) measures. These differences in AA variation probably occurred because of the varying role of WBC counts within the different DNAmAge algorithms from which each AA was derived. Based on the fact that WBC composition changes with age, the EEAA incorporates WBC changes in a weighted manner by aggregating the HannumAge estimate with plasmablasts, naïve cytotoxic T-cells and exhausted cytotoxic T-cells [16, 26]. This can, however, introduce inter-individual variation, since WBC composition is influenced by factors apart from aging itself, such as sex, medication use, disease and ethnicity [26-29]. For the IEAA, however, inter-individual variation is reduced by adjustment for cell counts. The adjustment is specifically made to optimize the performance of the multi-tissue predictor (HorvathAge) when it is applied to blood samples [8, 26]. The SBAA consequently outperforms both these estimates, as it was initially developed for blood samples and therefore needs no additional WBC count adjustments, nor does it introduce potential additional variance by incorporating differential WBC composition [10].

The SB clock was developed to improve the accuracy of age estimation of the Horvath clock (by limiting training data to a smaller number of cell types) and Hannum clock (by using bigger

training datasets and increasing represented loci). In our cohort, the superiority of the SB clock is confirmed in its higher correlation with chronological age and its improvement in AA variation, in samples derived from whole blood.

6.3.1.2 Next-generation clocks

A clear discrepancy between the biophysiological decline reflected by the PhenoAge and the GrimAge is visible (Table 6-1). The mean PhenoAge of the PURE-SA-NW study sample is much lower than its mean chronological age. The mean GrimAge, on the other hand, is strikingly similar to mean chronological age. The PhenoAge clock, therefore, estimates the PURE-SA-NW cohort at a much lower age-related mortality risk than suggested by both their chronological age and by the time to death reflected by their GrimAges. Although the GrimAge and PhenoAge estimates cannot be directly compared in terms of age estimation, the ideal would be that these clocks reveal a comparable estimation of biophysiological decline. In our cohort, however, PhenoAge correlated with GrimAge with a correlation coefficient of only 0.51 ($p = 4.2E–09$).

The role of chronological age in the development of the PhenoAge and GrimAge clocks is critical in untangling their behavior in subsequent investigations. For the PhenoAge algorithm, chronological age was incorporated as one of ten clinical markers associated with the risk of mortality, which were aggregated to represent phenotypic age. Chronological age is, therefore, used in a similar manner as used by the first-generation clocks, the difference being its aggregation with other clinical markers, rather than being the single outcome measure. The GrimAge clock incorporated age in a different manner. Instead of being a constituent of the outcome variable CpGs are tested against, chronological age was used as an adjustment variable in the CpG selection models. Age was selected together with gender, smoking pack-years and seven protein markers to best predict time to death. First a penalized regression, adjusted for gender and chronological age, was applied to select CpGs associated with each of the GrimAge proteins and smoking pack-years. After all the selected CpGs had been combined to form the GrimAge marker, linear transformation based on forcing the GrimAge mean and variance to match chronological age was applied.

Based on these protocols, a weaker correlation between chronological age and PhenoAge compared to HorvathAge, HannumAge and SBAge, was expected, because of the diluted contribution of chronological age in the prediction models. Although, even compared to external data, the PhenoAge clock performed particularly poorly in the PURE-SA-NW cohort [17, 23]. For example, the correlations between PhenoAge and chronological age reported for four independent validation cohorts were much stronger than what was observed for the

PURE-SA-NW participants (r = 0.51 in PURE-SA-NW *vs* 0.66, 0.69, 0.78 and 0. 89 in the respective validation cohorts [17]). For the GrimAge, however, the relationship we observed with chronological age (r = 0.81) was strikingly similar to those reported for the same validation cohorts mentioned above (reported here in the same order, r = 0.79, 0.80, 0.82 and 0.89, [18]). The reduced accuracy of the PhenoAge compared with the GrimAge-chronological age correlation in this population probably reflects environmental or ethnic confounding in the PURE-SA-NW group that is not captured by the design of the PhenoAge clock, but is captured by the GrimAge clock. In agreement with this hypothesis, when testing ethnic variability, Levine *et al*., [17] reported significant differences in PhenoAge between Hispanics, non-Hispanic blacks and non-Hispanic whites (p = 5.1E–05). When investigating the standard deviations of the PhenAA compared to GrimAA, Lu *et al.* [18] found an overall larger variance for the PhenoAA, as well as larger inter-ethnic differences, with the largest variation reported for African American cohorts, followed by white and then Hispanic groups. Our observation of larger variation in the PhenoAA than the GrimAA is in agreement with that of both Lu *et al.* [18] and a more recent comparison by Zhao *et al.* [23] in an independent African American study cohort. Availability of genome-wide genetic data will facilitate further investigation into potential ethnic confounding in future research.

Similar to Figure 6-1, Figure 6-2 depicts scatterplots comparing the relative underestimation of chronological age by the PhenoAge (top) and GrimAge (bottom) algorithms, respectively. Despite a two- and 15-fold increase in the number of clock CpGs included in the estimators compared to the Horvath and Hannum clocks, a stronger trend in chronological age underestimation was observed for PhenoAge. Similar to the HannumAge estimate, the PhenoAge clock likely performs best at ages outside of the PURE-SA-NW age range, as shown in the simulation on PURE-SA-NW data where the y-axis intercept is 34.3 years. For the GrimAge clock, however, optimal estimation is at age 67.4, which is the highest optimal age of all the tested clocks and is also closest to the mean age of our cohort. Moreover, for the next-generation clocks, additional adjustment for WBC count variation did not significantly alter the degree of underestimation.

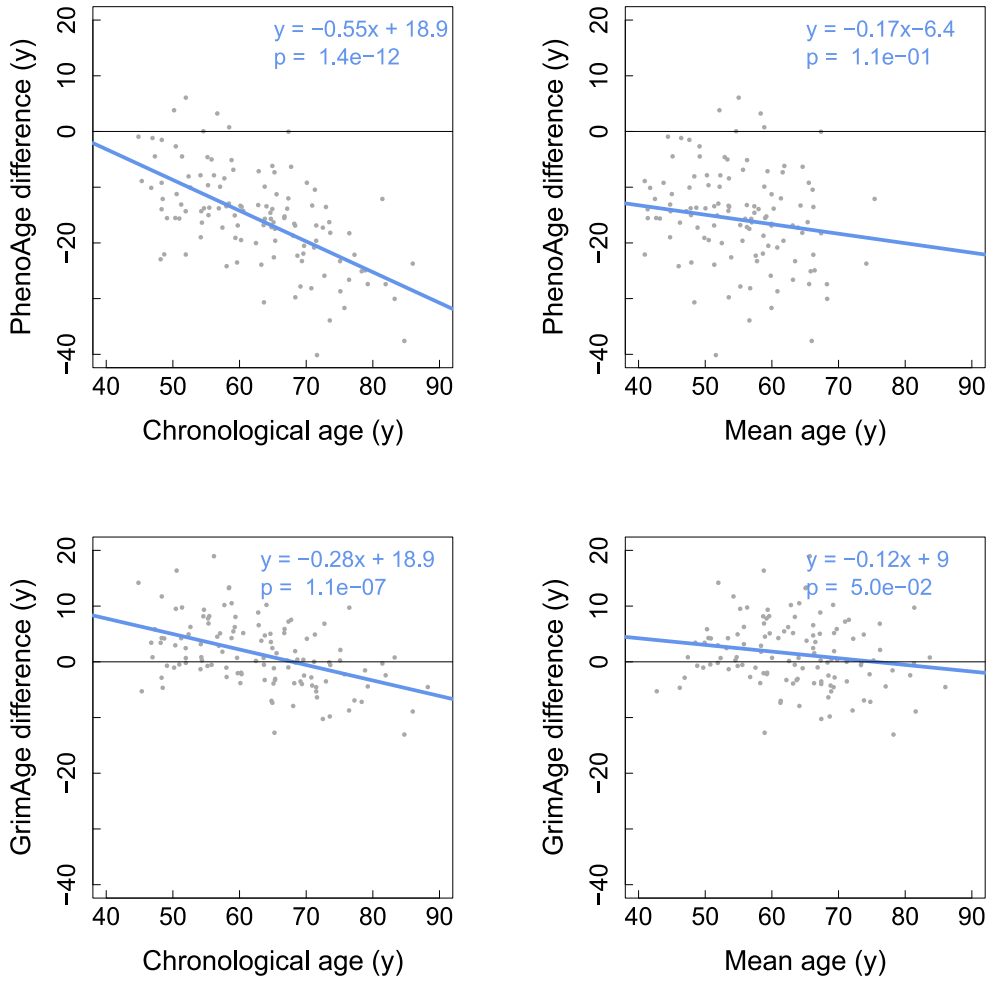


Figure 6-2 Scatterplots illustrating the relative difference in biological *vs* chronological age by two next-generation DNAmAge estimates

Left: Scatterplots depicting the relationship between chronological age and the age difference when subtracting chronological age from the DNAmAge estimates. Right: Bland Altman plots depicting the mean of each DNAmAge and chronological age against the difference between each DNAmAge and chronological age. The line of best fit from a linear regression model is formulated as y = mx + c in the top right corner and depicted in blue on the plot. The p-value represents the statistical significance of the linear regression model.

Apart from the above-mentioned validation studies, only one independent investigation compared the next-generation clocks [23]. Mean PhenoAge estimates in an African American cohort were found to be 13 years younger than chronological age (44 *vs* 57 years) compared to much more congruent behavior observed in the GrimAge estimate (54 years), analogous to our findings. Because of the limited literature, it remains difficult to confirm whether the comparatively weaker performance of the PhenoAge clock is the result of population-specific ethnic or environmental confounding, or a general limitation of the PhenoAge clock. Population-specific confounding could result from either different associations between the

clinical markers used in the design of the clock and mortality risk in the study population investigated or differing CpG-phenotypic age associations. In addition, a major strength of the GrimAge clock, advocated by its developers, is the fact that it incorporates smoking, a known influencer of advanced aging and mortality risk. Because more than half of the PURE-SA-NW study population were current smokers, we hypothesized that the difference in the contribution of smoking to the PhenoAge and GrimAge estimates might explain the large discrepancy between the two clocks.

6.3.2 Role of smoking in the next-generation clocks

Although the PhenoAge algorithm was not developed to encapsulate smoking habits specifically, previous research, including the PhenoAge validation analysis, reported that this age estimate was able to discriminate between smokers and non-smokers [17, 23]. When tested in the PURE-SA-NW data, however, mean PhenoAges of 47.4 ± 9 and 46.8 ± 9 respectively were observed when never and current smokers were compared ($p = 0.71$). No differences between the PhenoAA of smokers (–0.35 ± 7.96) and non-smokers (0.49 ± 7.23) were observed ($p = 0.55$) either. The fact that our study sample only included men should not affect this association [15, 23]. The PhenoAge biomarker, according to its developers, should be able to capture smoking-related differences through the association smoking has with the protein biomarkers that comprise phenotypic age [17]. To explore this further, we evaluated the association of smoking with seven of the ten clinical components of phenotypic age that were available for the PURE-SA-NW study population. We also evaluated the associations of each of these clinical components with PhenoAge and PhenoAA (Table 6-2).

In the PURE-SA-NW data, none of the available clinical components of phenotypic age differed significantly between smokers and non-smokers except for creatinine, where only a borderline difference ($p=0.04$) was evident. However, creatinine did not correlate with PhenoAge or PhenoAA. PhenoAge correlated with chronological age, C-reactive protein and lymphocyte percentage, while PhenoAA, correlated only with the last-named two. These results confirm that neither PhenoAge nor the clinical components of phenotypic age seem to encapsulate the biophysiological effects of smoking in this study population.

Table 6-2 Comparison of seven clinical components of phenotypic age between current and never smokers and each component's association with PhenoAge and PhenoAA

Clinical components of phenotypic age	Never smoker (n = 56)	Current smoker (n = 61)	t-test p	PhenoAge correlation		PhenoAA correlation	
	Geometric mean ± SD			r	p	r	p
Age (years)	63 ± 10	63 ± 10	0.78	0.51	**2.7E–09**	0.00	0.96
Albumin (g/L)	42.1 ± 1.19	44.2 ± 1.17	0.10	–0.15	0.10	–0.11	0.23
ALP (U/L)	83.9 ± 1.45	85.1 ± 1.39	0.82	0.04	0.63	0.04	0.69
Creatinine (umol/L)	0.08 ± 0.03	0.07 ± 0.02	**0.04**	0.10	0.30	–0.06	0.54
CRP (mg/dL)	3.62 ± 4.23	3.00 ± 3.37	0.45	0.39	**1.1E–05**	**0.36**	**9.0E–05**
Glucose (mmol/L)	5.21 ± 1.20	4.99 ± 1.16	0.19	–0.03	0.72	–0.07	0.45
Lymphocyte (%)	37.2 ± 9.30	36.6 ± 10.5	0.75	–0.31	**4.8E–04**	**–0.33**	**2.9E–04**

All clinical components were log transformed apart from age and lymphocyte %. Lymphocyte estimates are methylation-derived. Phenotypic age components for which no data are available are mean cell volume, WBC counts, and red cell distribution width. p <0.05 highlighted in bold.

Compared to PhenoAge, both GrimAge and GrimAA differed significantly between current and never smokers, even after adjusting for chronological age, education, BMI and WBC count proportions, based on the convention of current literature, to facilitate useful comparison of results (Model 1 in Table 6-3). Next, we investigated the association of smoking with each of the eight methylation surrogate markers that comprise the GrimAge estimate. Apart from the smoking pack-years methylation component (DNAmPackY), DNAmPAI1 and DNAmLeptin also differed between current and never smokers. Since alcohol consumption often coincides with smoking habits, we performed a sensitivity analysis additionally adjusting for alcohol (Model 2 in Table 6-3). This did not significantly alter the associations observed with smoking status, apart from an attenuating effect on DNAmPAI. All results can be found in the supplementary material (Supplementary Table 1) and all results at p <0.05 are included in Table 6-3, below.

Table 6-3 Adjusted group means of aging-related phenotypes for current vs never smokers

Outcome	Model	Smoking status Group mean ± SE		
		Never (n = 56)	**Current** (n = 61)	**p**
GrimAge	1	62.2 ± 0.68	66.6 ± 0.65	2.0E–05
	2	62.0 ± 0.79	66.7 ± 0.76	3.5E–04
GrimAA	1	–2.27 ± 0.68	2.19 ± 0.65	2.0E–05
	2	–2.4 ± 0.79	2.31 ± 0.76	3.5E–04
DNAmPackY	1	19.9 ± 1.6	35.1 ± 1.6	4.2E–09
	2	19.5 ± 1.9	35.5 ± 1.8	6.5E–07
DNAmLeptin	1	6448 ± 541	8514 ± 520	1.1E–02
	2	6450 ± 634	8512 ± 604	4.5E–02
DNAmPAI	1	16657 ± 503	18757 ± 483	5.5E–03
	2	16956 ± 586	18479 ± 559	1.1E–01

Group N: Never = 56, Current = 61. AA: Age acceleration; PackY: Smoking pack years. Variables prefaced by DNAm are the methylation-derived surrogates of the following component as used for the GrimAge estimate. Model 1: Outcome ~ smoking status + chronological age + BMI + education + WBC counts; Model 2: Outcome ~ smoking status + chronological age + BMI + education + WBC counts **+ alcohol consumption status**.

In the study by Zhao *et al*. [23], smoking status associated not only with DNAmPackY, but also with other methylation components such as the adrenomedullin, beta-2 microglobulin, growth differentiation factor 15, Cystatin C, plasminogen activation inhibitor 1 (PAI1) and tissue inhibitor metalloproteinase methylation components, confirming the integral role of smoking in GrimAge. Our data furthermore demonstrated that it is primarily smoking rather than combined smoking and alcohol use that is encapsulated by GrimAge. The only methylation component that additional adjustment for alcohol consumption affected was DNAmPAI1. This is in line with Zhao's findings of alcohol consumption having the strongest influence on DNAmPAI1 [23].

6.3.3 Strengths and limitations

These clocks were tested in an older population in apparently good health, which allowed us to investigate the largely disease-independent underestimation of biological age in older individuals. Literature in this area is currently lacking. We were also able to contribute to the

limited ethnic diversity represented in the epigenetic aging literature by investigating a group of continental Africans. We were, however, limited by our sample size and related to this the age range in the selected study sample. We are unable to extrapolate our findings to women or to younger individuals, or those with specific disease diagnoses. Furthermore, we were only able to investigate seven of the ten clinical phenotypic age components, which may have presented individual differences in smoking status groups. The fact that smoking-related differences were not observed in the aggregate PhenoAge marker does, however, suggest that it may be unlikely. Lastly, the lack of longitudinal and mortality data limits our ability to test the PhenoAge and GrimAge clocks' accuracy in predicting age-related mortality risk and time to death, respectively.

6.4 Conclusions

This study compared five epigenetic clocks previously proven to be highly effective in reaching their respective aims of chronological age prediction (first-generation clocks) and prediction of mortality-related functional decline (next-generation clocks). The next-generation clocks are particularly important in the context of health research because they have been developed using longitudinal data and well-defined mortality outcomes and are therefore able to generate very useful biological age biomarkers. This gives cohorts without such data, such as the PURE-SA-NW study, an opportunity to use methylation-derived biomarkers supported by causal models to indicate the health risks of study populations. Because many existing cohorts possess DNAm data, mostly generated using the Illumina platforms, understanding the multitude of uses of DNAm data is essential.

Our data echoes previous findings of discrepancies between the predicted biological ages generated by different clocks, indicating that each clock provides a different fraction of information regarding the aging body. We also confirm the tendency of all the tested clocks to underestimate the biological age of older individuals. We demonstrate that the GrimAge best fits the aging of a study sample of continental African men with high smoking and alcohol use prevalence. This could potentially be because of the larger number of methylation loci included in this algorithm, the incorporation of smoking and/or because it is built on CpGs that are more functionally relevant in this population. Epigenetic clocks provide unique possibilities to understand healthy *vs* accelerated aging better and consequently improve the quality of life of a global population with an ever-increasing lifespan.

6.5 Methods

6.5.1 Study population

The international PURE study comprises cohorts from 27 countries tracked over a period of 20 years. The data reported in this manuscript are from the PURE-SA-NW study. A subset of 120 apparently healthy black South African men with available peripheral blood samples were randomly selected for this study, provided they tested negative for the human immunodeficiency virus at the time of data collection (2015). Eligibility was restricted to reduce confounding by sex and CD4-T cell count. Ethical approval for this study was granted by the Human Research Ethics committee of the North-West University (NWU-00119-17-S1). Additional information on the international cohort [30] and this sub-study [31] has been published.

6.5.2 Data collection

Smoking and alcohol status data are interview-based. Participants reported their current status and if applicable, the frequency and quantity of use, age at the start of use and previous attempts at abstinence. Peripheral blood samples were used for DNA extraction. Genome-wide methylation data were generated using the standard protocol of the Illumina Infinium MethylationEPIC BeadChip (Illumina®, San Diego, CA, USA) platform. Quality control, sample filtering and functional normalization were done using the *meffil* [32] package in R 3.4.3 [33]. A detailed description of the DNA extraction, quality control, methylation quantification and data processing protocols has been published previously [31].

6.5.3 Cell counts and DNAmAge

The IDOL optimised L-DMR library for whole blood samples [34] was used to estimate the distribution of B-, CD4-T, CD8-T, neutrophil, monocyte and natural killer cells. For the estimation of DNAmAge and DNAmAgeAccel, the new online DNA methylation age calculator from Steve Horvath's group was used (https://dnamage.genetics.ucla.edu/new, [8]). The following age estimates were used as provided in the calculator output: *DNAmAge*, *DNAmAgeHannum*, *EEAA*, *DNAmAgeSkinBloodClock*, *BloodskinAA*, *DNAmPhenoAge*, *PhenoAA*, *DNAmGrimAge* and *GrimAA*. IEAA was self-determined and corresponded to the residuals from a linear model where *DNAmAge* was used as the outcome and chronological age, the calculator's *plasmaBlast*, *CD8pCD28nCD45Ran* and *CD8.naive* and IDOL-estimated CD4-T, natural killer cells, monocytes and neutrophils as predictors. Note that the Illumina Infinium MethylationEPIC array used to quantify methylation data for the current study

excludes 19 of the 353 Horvath and six of the 71 Hannum clock CpGs [8, 16]. The absence of these CpGs has previously been reported not to compromise the accuracy of these clocks [35]. All further cell-count adjustments were performed using cell estimates from the IDOL package [34].

6.5.4 Statistical analysis

Data normality was evaluated using Shapiro-Wilks tests. Prior to any parametric or linear modeling, skewed data were logarithmically transformed. We report the mean and standard deviation of the chronological age, DNAmAge and DNAmAA estimates in this study population. We also report Spearman correlation coefficients between chronological age and each of the DNAmAge estimates. Figures 6-1 and 6-2 were compiled with the *ggplot2* and *BlandAltmanLeh* packages. In Table 6-2 we compared the means of the clinical components of phenotypic age between current and never smokers using a t-test. Spearman tests were used to evaluate the association between the components of phenotypic age and PhenoAge and PhenoAA. Type III analysis of variance models in the *car* package were applied to linear regression objects to quantify the differences in outcome means between smokers and non-smokers, reported in Table 6-3. Adjusted group means and standard errors were extracted using the *effects* package. Two models were run for each outcome. First, a model adjusting for chronological age, BMI, education and WBC counts was developed, and then, in a second model, an additional adjustment for alcohol consumption was made. Covariates were chosen based on current literature to ease comparability [36]. R version 3.5.0 was used for all analyses [37].

6.6 Declarations

Authors' contributions

HTC, MP, CNR and HRE conceptualized the study. Funding was acquired by MP and FRG. HTC performed the data analysis and wrote the manuscript. MP supervised the data analysis and interpreted the results with JM, HRE and HTC. All authors contributed to the critical review and editing of the manuscript.

Acknowledgements

The authors would like to thank all those who participated in the PURE-SA-NW study and those who made the PURE-SA-NW study possible, including the fieldworkers, researchers and staff of both the PURE-SA-NW (Africa Unit for Transdisciplinary Health Research, Faculty of Health Sciences, NWU, Potchefstroom, South Africa) and PURE International (S Yusuf and

the PURE project office staff at the Population Health Research Institute, Hamilton Health Sciences and McMaster University, Ontario, Canada) teams. We also thank the staff of the Bristol bioresource laboratories (Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK) who generated the DNA methylation data.

Conflicts of interest

The authors have no conflicts of interest to declare.

Funding

The PURE-SA-NW study was funded by the North-West University, South African National Research Foundation (SANRF), Population Health Research Institute, South African Medical Research Council (SAMRC), the North West Province Health Department, and the South African Netherlands Partnerships in Development. Grants from the SANRF, Academy of Medical Sciences UK (Newton Fund Advanced Fellowship Grant [AMS-NAF1-Pieters to MP and FRG]) and the SAMRC funded the DNAm analysis. HTC is supported by the SANRF [SFH106264, MND 121094]. HRE works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/5). Funding bodies were not involved in the design of the study, collection, analysis or interpretation of the data or in writing of this manuscript. Opinions expressed and conclusions arrived at are those of the authors and are not to be attributed to the funding sources.

Abbreviations

AA: Age acceleration; BMI: body mass index; CpGs: cytosine-phosphate-guanine sites; DNAm: DNA methylation; DNAmAge: DNA methylation age; DNAmAgeAccel: DNA methylation age acceleration; EEAA: Extrinsic epigenetic age acceleration; IEAA: Intrinsic epigenetic age acceleration; PAI1: Plasminogen activation inhibitor 1; PURE-SA-NW: South Africa, North-West arm of the Prospective Urban and Rural Epidemiology study; SB: Skin and blood; WBC: white blood cell.

6.7 References

1. Wang H, Naghavi M, Allen C, Barber RM, Bhutta ZA, Carter A, Casey DC, Charlson FJ, Chen AZ, Coates MM. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016; 388: 1459-544.

2. Beard HPJR, Bloom DE. Towards a comprehensive public health response to population ageing. Lancet. 2015; 385: 658-61.

3. Jylhävä J, Pedersen NL, Hägg S. Biological age predictors. EBioMedicine. 2017; 21: 29-36.

4. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature reviews genetics. 2018; 19: 371-84.

5. Martin GM. Epigenetic drift in aging identical twins. Proceedings of the national academy of sciences of the United States of America. 2005; 102: 10413-4.

6. Wang Y, Karlsson R, Lampa E, Zhang Q, Hedman ÅK, Almgren M, Almqvist C, McRae AF, Marioni RE, Ingelsson E. Epigenetic influences on aging: A longitudinal genome-wide methylation study in old Swedish twins. Epigenetics. 2018; 13:975-87.

7. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nature reviews genetics. 2018; 19: 129-47.

8. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013; 14: 3156.

9. Marioni R, Suderman MJ, Chen BH, Horvath S, Bandinelli S, Morris TJ, Becker S, Ferrucci L, Pedersen NL, Relton CL, Deary I, Hägg S. Tracking the epigenetic clock across the human life course. Journals of gerontology, series A. 2019: 74: 57-61.

10. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, Felton S, Matsuyama M, Lowe D, Kabacik S. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY). 2018; 10: 1758.

11. Grant CD, Jafari N, Hou L, Li Y, Stewart JD, Zhang G, Lamichhane A, Manson JE, Baccarelli AA, Whitsel EA. A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk. Geroscience. 2017; 39: 475-89.

12. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clinical epigenetics. 2016; 8: 64.

13. Lind L, Ingelsson E, Sundström J, Siegbahn A, Lampa E. Methylation-based estimated biological age and cardiovascular disease. European journal of clinical investigation. 2018; 48.

14. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR. DNA methylation age of blood predicts all-cause mortality in later life. Genome biology. 2015; 16: 25.

15. Fransquet PD, Wrigglesworth J, Woods RL, Ernst ME, Ryan J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. Clinical epigenetics. 2019; 11: 62.

16. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J-B, Gao Y. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell. 2013; 49: 359-67.

17. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018; 10: 573-91.

18. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019; 11: 303-27.

19. El Khoury LY, Gorrie-Stone T, Smart M, Hughes A, Bao Y, Andrayas A, Burrage J, Hannon E, Kumari M, Mill J. Systematic underestimation of the epigenetic clock and age acceleration in older subjects. Genome biology. 2019; 20: 283.

20. Tajuddin SM, Hernandez DG, Chen BH, Noren Hooten N, Mode NA, Nalls MA, Singleton AB, Ejiogu N, Chitrala KN, Zonderman AB, Evans MK. Novel age-associated DNA methylation changes and epigenetic age acceleration in middle-aged African Americans and whites. Clinical epigenetics. 2019; 11: 119.

21. Liu Z, Chen BH, Assimes TL, Ferrucci L, Horvath S, Levine ME. The role of epigenetic aging in education and racial/ethnic mortality disparities among older US Women. Psychoneuroendocrinology. 2019; 104: 18-24.

22. Levine ME. Assessment of epigenetic clocks as biomarkers of aging in basic and population research. The journals of gerontology: Series A. 2020; 75: 463-5.

23. Zhao W, Ammous F, Ratliff S, Liu J, Yu M, Mosley TH, Kardia SL, Smith JA. Education and lifestyle factors are associated with dna methylation clocks in older African Americans. International journal of environmental research and public health. 2019; 16: 3141.

24. Ryan J, Wrigglesworth J, Loong J, Fransquet PD, Woods RL. A systematic review and meta-analysis of environmental, lifestyle, and health factors associated with DNA methylation a The journals of gerontology: series A. 2019; 75: 481-94.

25. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, Colicino E, Dugué P-A, Artaud F, McKay GJ. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging (Albany NY). 2019; 11: 2045.

26. Quach A, Levine ME, Tanaka T, Lu AT, Chen BH, Ferrucci L, Ritz B, Bandinelli S, Neuhouser ML, Beasley JM, Snetselaar L, Wallace RB, Tsao PS, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. Aging (Albany NY). 2017; 9: 419-46.

27. Azab B, Camacho-Rivera M, Taioli E. Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects. PloS one. 2014; 9: e112361.

28. Forget P, Khalifa C, Defour J-P, Latinne D, Van Pel M-C, De Kock M. What is the normal value of the neutrophil-to-lymphocyte ratio? BMC research notes. 2017; 10: 12.

29. Kelsey KT, Wiencke JK. Immunomethylomics: A novel cancer risk prediction tool. Annals of the American Thoracic Society. 2018; 15: S76-80.

30. Teo K, Chow CK, Vaz M, Rangarajan S, Yusuf S. The prospective urban rural epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. American heart journal. 2009; 158: 1-7.

31. Cronjé HT, Elliott HR, Nienaber-Rousseau C, Pieters M. Replication and expansion of epigenome-wide association literature in a black South African population. Clinical epigenetics. 2020; 12: 6.

32. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018; 34: 3983-9.

33. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

34. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome biology. 2018; 19: 64.

35. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, Ramadori KE, Morin AM, Rider CF, Carlsten C, Quintana-Murci L, Horvath S, Kobor MS. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. Clinical epigenetics. 2018; 10: 123.

36. Fiorito G, Robinson O, Vineis P. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging (Albany NY). 2019; 11: 2045-70.

37. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

6.8 Supplementary material

Supplementary Table 1 Adjusted group means of aging-related phenotypes for current vs never smokers

Outcome	Model	Smoking status Group mean ± SE		
		Never (n = 56)	**Current** (n = 61)	**p**
GrimAge	1	62.2 ± 0.68	66.6 ± 0.65	2.0E–05
	2	62.0 ± 0.79	66.7 ± 0.76	3.5E–04
GrimAA	1	–2.27 ± 0.68	2.19 ± 0.65	2.0E–05
	2	–2.4 ± 0.79	2.31 ± 0.76	3.5E–04
Pack-years*	1	19.9 ± 1.6	35.1 ± 1.6	4.2E–09
	2	19.5 ± 1.9	35.5 ± 1.8	6.5E–07
Leptin*	1	6448 ± 541	8514 ± 520	1.1E–02
	2	6450 ± 634	8512 ± 604	4.5E–02
Plasminogen activator inhibitor 1*	1	16657 ± 503	18757 ± 483	5.5E–03
	2	16956 ± 586	18479 ± 559	1.1E–01
Adrenomedullin*	1	286 ± 2.69	291 ± 2.59	2.80E–01
	2	285 ± 3.15	292 ± 3.00	2.20E–01
Beta-2 microglobulin*#	1	1795 ± 26.1	1814 ± 25.0	6.20E–01
	2	1807 ± 30.4	1803 ± 29.0	9.40E–01
Cystatin C*#	1	651 ± 4.46	646 ± 4.29	4.50E–01
	2	648 ± 5.16	650 ± 4.92	7.50E–01
Growth differentiation factor 15*	1	1236 ± 41.7	1289 ± 40.1	3.90E–01
	2	1296 ± 47.44	1233 ± 45.2	4.10E–01
Tissue inhibitor metalloproteinase 1*	1	33548 ± 180	33482 ± 173	8.00E–01
	2	33442 ± 209	33580 ± 200	6.80E–01

Group N: Never = 56, Current = 61. AA: Age acceleration; *Methylation–derived surrogates as used for the GrimAge estimate. #Reported values are divided by 1000, therefore 286 reports >286000. Model 1: Outcome ~ smoking status + chronological age + BMI + education + WBC counts; Model 2: Outcome ~ smoking status + chronological age + BMI + education + WBC counts **+ alcohol consumption status**.

CHAPTER 7

DISCUSSION AND CONCLUSIONS

DISCUSSION AND CONCLUSIONS

7.1 Introduction

This thesis explored the relationship between DNA methylation (DNAm) and cardio-metabolic health in a group of black South African men from the North West province, South African arm of the Prospective Urban and Rural Epidemiology study (PURE-SA-NW). First, the urban-rural divide, as is experienced in developing countries such as South Africa, was evaluated as an epidemiological approach to investigate the role of DNAm in the association between urbanisation and non-communicable disease (NCD) risk, in the form of a review (Chapter 3). Empirically, methylation was investigated at an individual CpG level (Chapter 4) and as aggregate measures of systemic inflammation (methylation-derived cell count ratios, Chapter 5) and biological aging (DNAm clocks, Chapter 6). The data reported in this thesis contribute to the limited data available on continental African cohorts. Additionally, this thesis adds to the limited data available on the novel loci recently introduced in the Illumina® Infinium HumanMethylationEPIC bead chip (EPIC array). Multiple phenotypes related to cardio-metabolic health were considered, including chronological and biological age, alcohol consumption, smoking status, education, body composition, biochemical markers and markers of cardiovascular function (CVF). The aims of this thesis were to:

1. Evaluate the use of an urban-rural divide as an epidemiological approach when investigating the role of DNAm in the association between urbanisation-associated NCD risk;

2. Replicate findings from previously published NCD-related epigenome-wide association studies (EWASs) and contribute novel findings from the PURE-SA-NW cohort to the current literature;

3. Compare methylation-derived markers of cell-count and protein-based inflammatory markers in the PURE-SA-NW cohort in their associations with CVF markers and their literature-based portrayal of cardiovascular disease (CVD) risk; and

4. Characterise the behaviour of five DNAm clocks in the PURE-SA-NW cohort.

This final chapter summarises and integrates the main findings of this study to the current literature and concludes this thesis.

7.2 The role of urbanisation

Chapter 3 reports a proof of concept for future work in the PURE-SA-NW, international PURE or other similar cohorts. From the critical review of current literature, three investigative approaches proved to contribute to the understanding of amalgamated exposures and their contribution to disease risk. These were the migration, income-comparative and urban-rural study designs. It is, however, the integration of the knowledge gained from the different aspects of the perspective of each of these three models that will allow for a more holistic view of the different genetic and environmental origins of disease and the epigenetic mechanisms that bridge them. For the question of urban *vs* rural environment, lifestyle and health disparities, particularly in low- and middle-income countries, the urban-rural investigative approach is valuable. This approach provides the most controlled setting (least confounding) for investigating exposure-methylation-disease relationships. When, however, centring the question on genetic pre-disposition or the developmental origins of health and disease, the migration approach may be best suited. This approach allows for the investigation of migrants *vs* the population of origin, to disentangle the environmental contributors without genetic confounding. On the other hand, investigating migrants *vs* host populations allows for the investigation of the genetic origins of disease while reducing environmental confounding. Lastly, when the focus is shifted from individual populations or cohorts to a better global understanding of the extent and impact of urbanisation, this review showed that income-comparative models are particularly useful. Income-comparative models can provide valuable information on disease demographics stratified to stages of rural development. All three models are beneficial in their inclusion of typically under-represented cohorts (developing countries, migrant populations and low-income countries around the world). Apart from the benefit of understanding urbanisation and its health-related effects better, investigating understudied cohorts contributes to the lacking genetic/ethnic and environmental diversity of current epigenetic epidemiology literature.

7.3 Replication and expansion of EWAS literature

Chapter 4 contributes to the literature by providing access to novel EPIC array association data on commonly investigated NCD-related traits, in a non-European cohort. Furthermore, replication association statistics on loci represented on the HumanMethylation450K bead chip (450K array) that have already been published in other ethnic groups are provided. EWAS protocols were tailored to match that of the largest EWAS available in the EWAS catalog, per trait, to allow for comparative results, but also to allow for future replication of novel and/or population-specific associations. Consistent directionality was reported for most of the comparisons drawn between the PURE-SA-NW and reference cohort data, suggesting that many of the observed CpG

associations (with age, alcohol consumption, smoking, body mass index, waist circumference, C-reactive protein and blood lipids) are shared across different ethnic groups. More specifically, 86% of the findings overlapped between the PURE-SA-NW and reference cohorts, 13% did not overlap but were associated in the same direction and 1% showed clear differences in the direction and size of effect. This 1% consisted of 48 non-overlapping CpG-trait associations (44 unique CpGs). Thirty-six of these CpGs are associated with methylation quantitative trait loci (mQTLs) that have documented population differences in their minor allele frequencies (MAF) according to the 1000 Genomes project (Genomes Project Consortium, 2015). For example, two of the three CpGs identified to differ in their association with alcohol consumption between the PURE-SA-NW and reference cohorts (out of 361 tests), were linked to mQTLs reported to have substantial functional enrichment. Population differences in the MAF of these mQTLS, illustrated by the 1000 Genomes project, strengthened the evidence of population specificity in these differential associations (Bonder *et al.*, 2016; Genomes Project Consortium, 2015; Zhernakova *et al.*, 2016).

For the CpG associations that were directionally consistent but that did not overlap, wide confidence intervals of findings from the PURE-SA-NW data limited the ability to pursue any fine scale inference for possible explanations of the differences in effect sizes. Importantly, many of the confidence intervals of these findings spanned a zero effect. Consequently, although there is no evidence for differences in effects between the PURE-SA-NW and the reference cohorts, evidence of a specific directionality in the PURE-SA-NW findings is also not compelling and requires further exploration. Increased funding to generate methylation data for more PURE-SA-NW participants will allow for more robust evidence to support these findings. Apart from increasing DNAm data for the PURE-SA-NW cohort, generating data on similar comparisons of findings between ethnic groups from other large ethnically diverse cohorts will enable better global understanding of the generalisability of methylation associations of commonly investigated cardio-metabolic disease-related phenotypes. This will be particularly important as a surrogate data source for groups yet to generate DNAm data for their cohort of interest.

For this thesis, the successful replication of the numerous previously published EWASs strengthened confidence in the quality and statistical power of the methylation data generated for this PhD. It was therefore subsequently possible to report 19 novel genome-wide significant CpG associations with alcohol consumption, eight of which were 450K array probes present on the EPIC array and 11 were novel EPIC array probes. One genome-wide significant novel association with high-density lipoprotein was also reported (450K probe). Replication of particularly the CpG-alcohol associations will be beneficial for future epigenetic research. A methylation-based biomarker of alcohol consumption (Liu *et al.*, 2018) has proven successful in

identifying risky and heavy drinkers. Once replicated, the addition of these 11 EPIC probes to this biomarker could enhance its discriminatory potential. Alternatively, the addition of the eight 450K probes (which failed to reach significance in their associations with alcohol in other cohorts) to the biomarker could allow for its optimal use in the PURE-SA-NW and potentially other continental African cohorts that replicate these associations. Unfortunately, the limited sample size of the instigated PURE-SA-NW group did not allow for adequately powered stratification of the alcohol consumers to test the adapted alcohol consumption biomarker in this thesis. Generation of more DNAm data in the PURE-SA-NW cohort will, however, allow such an investigation in future research.

7.4 DNAm in the context of inflammation and cardiovascular risk

Chapter 5 reported the first use of the methylation-derived white blood cell (WBC) ratio estimates in an African cohort, and also the first investigation of methylation-derived WBC ratio estimates in the context of cardiovascular function/disease. The WBC ratios observed in the PURE-SA-NW cohort were comparable to other previously investigated ostensibly healthy ethnic groups. More favourable WBC ratios in black cohorts compared to white or Hispanic cohorts was also confirmed by the PURE-SA-NW data. Methylation-derived WBC ratios reflected a similar degree of CVD risk compared to protein-based inflammatory and CVD risk markers in the PURE-SA-NW group. The methylation-derived and protein-based inflammatory markers complemented one another in explaining variance in CVF, although when considered separately, stronger associations between CVF and the methylation-derived markers than between CVF and the protein-based markers were observed. Five CpGs, investigated as potential proxies for the methylation-derived neutrophil-to-lymphocyte ratio (mdNLR) proved not only to reflect the mdNLR in the PURE-SA-NW data, but also to complement the mdNLR by associating with CVF independently of the mdNLR.

The data generated for Chapter 5 will be useful for a range of researchers. Firstly, for epigenetic epidemiology groups, it provides the incentive to reconsider previously generated 450K or EPIC data, to investigate methylation-derived cell ratios as useful indicators of systemic inflammation. These groups can also retrieve data on the five reported myeloid CpGs, given the fact that such strong associations were observed in relation to CVF in the PURE-SA-NW cohort. Based on our confirmation of previous suggestions that the five myeloid CpGs surrogate for the mdNLR, groups without access to genome-wide data can consider sequencing only the five myeloid CpGs as proxies for the mdNLR, when investigating immunomodulation. This is, however, the first investigation of these CpGs as mdNLR proxies in a population-based, rather than cancer-based context, so replication in larger, more diverse population-based cohorts is warranted. Replication

of the findings presented in Chapter 5 in other ethnic groups in the context of cardio-metabolic disease will be particularly useful, given the limited methylation-based investigations of cell counts in low-grade inflammatory diseases.

The data presented in Chapter 5 is also of use to researchers of cardiovascular health/disease, in terms of the evidence that the characterisation of systemic inflammation should not be achieved by conventionally used inflammatory proteins only, but can be more comprehensive when including cell count distribution, and, if available, the five myeloid CpG sites. For researchers of immunology and inflammation, it is of interest that particularly C-reactive protein, frequently used as a general marker of inflammation either adjusted for or incorporated as an outcome, only associated with some, but not all of the CVF markers tested. C-reactive protein was also not strongly correlated to any of the cell count-related markers, suggesting that previous claims of it being a proxy for cell counts might not be true in cases of less overt inflammation.

7.5 DNAm in relation to aging

Chapter 6 reported the first investigation of DNAm clocks in a continental African cohort. To our knowledge, this was also the first comparison of the five predominantly used DNAm clocks in a single study population. Additionally, this was the first exploration of specific phenotypic age constituents and how these relate to the information provided by the aggregated methylation-based PhenoAge marker.

Generally, younger methylation-based biological ages were observed in the PURE-SA-NW cohort compared to their chronological age. These data therefore confirmed prior evidence of a tendency of the first-generation DNAm clocks to underestimate chronological age in older adults. When investigating the biophysiological decline portrayed by the next-generation clocks, similar observations of larger underestimations with age were observed, regardless of substantial increases in clock loci. The underestimation was, however, much more pronounced in the PhenoAge than in the GrimAge estimates. The data presented in Chapter 6 show that the PURE-SA-NW cohort, on average 63 years old (chronologically), is at the same average level of age-related mortality risk than a 47-year-old participant in the Levine *et al.* (2018) investigation (PhenoAge clock) while having, on average, the same estimated time-to-death than a 64-year-old in the Lu *et al.* (2018) investigation (GrimAge). Although the PURE-SA-NW cohort was apparently healthy, none of the other data presented in this thesis suggest that this group should have a particularly low mortality risk or decelerated biophysiological decline. These findings were analogous to those reported by the only other available study that compared the PhenoAge and GrimAge estimates of an African American cohort (Zhao *et al.*, 2019). The mean chronological

age of the cohort investigated by Zhao *et al.* was 57, while the PhenoAge was 44 (13 years younger) and the GrimAge 54 (three years younger). Although no further evaluation of these findings was reported in Zhao's investigation, for the PURE-SA-NW cohort, at least some of the variation between biological ages estimated by these clocks was attributed to the incorporation of smoking data by the GrimAge, but not by the PhenoAge algorithm. More than half of the PURE-SA-NW cohort were smokers and although previously reported to encapsulate the biophysiological effects of smoking, PhenoAge in the PURE-SA-NW data does not support this.

7.6 Limitations

All the manuscripts contained in this thesis discuss general limitations such as sample size, representation of only one sex and the inability to infer causality (lack of longitudinal and/or mortality data). One limitation of particular interest to the larger thesis is the lack of available genomic data for the PURE-SA-NW group. Genetic variation is associated with methylomic variance. For this reason, integrating genomic and DNAm data is always beneficial when attempting to account for ethnic confounding, or when attempting to infer causality (through methods such as Mendelian randomisation). The lack of genomic data available in the PURE-SA-NW group hindered our ability to investigate any hypotheses regarding ethnic-specific variance. Hypotheses generated regarding ethnic-specific variance relied on publicly available data from the 1000 Genomes project (Genomes Project Consortium, 2015). Limited genomic data on continental Africans are available in the 1000 Genomes project, which currently only represents individuals from one East African (Kenya) and three West African (Gambia, Nigeria and Sierra Leone) countries. In addition, the available data are hindered by the inability to extrapolate findings from one African sub-population to the next because of high levels of genetic diversity between African populations (Campbell & Tishkoff, 2008; Tishkoff *et al.*, 2009; Tishkoff & Williams, 2002). Although the diversity of the African genome is a well-known benefit to genomic research, the limitation related to increased inter-population diversity is the lack of generalisability of findings between groups. For the PURE-SA-NW cohort, specifically, individuals from a single, self-reported ethnic group named 'Batswana' South Africans were investigated. The lack of genetic data hinders the ability to explore the validity of self-reported classifications and also possible ancestral differences in this group. The impact of the findings reported in this thesis, although a worthy contribution to epigenetics research, can be expanded greatly and should be supplemented by numerous independent cohorts representing other ethno-linguistic groups in and around South Africa.

7.7 Future research

For the PURE-SA-NW cohort, this PhD represents only the start of the exploration of the epigenome's role in the relationship between the environment (specifically environmental exposures related to urbanisation) and NCDs. Many unexplored research questions remain in this cohort specifically, and also potentially in other PURE sub-cohorts internationally. The main advantage that should be leveraged in future PURE-SA-NW investigations is its longitudinal study design. The ten-year follow-up within the PURE-SA-NW cohort (and inclusion of mortality data for the foreseeable future) allows retrospective investigation of epigenetic and mortality risk. Ideally, the question pertaining to the role of DNAm in the relationship between urbanisation and cardio-metabolic disease/NCD-related mortality should be investigated longitudinally once there are sufficient mortality data to yield statistical power to do so. Until then, clinically relevant risk factors, other than those investigated in this thesis, such as metabolic syndrome, diabetes-related risk factors (insulin and glycated haemoglobin) and cardiac structure markers (intima-media thickness and plaque scores) can still be investigated. Genomic data will be an ideal addition to the current dataset, and could be valuable for cross-sectional Mendelian randomisation investigations, or in the retrospective setting where reverse causation is of less concern, for further investigation of ethnic-specific associations or confounding. One concern to keep in mind is that with the exception of the 2015 PURE-SA-NW samples, DNA-containing samples were not collected for the purpose of future epigenetic analyses. For example, for 2015, two main sources of DNA are available: whole blood collected in Tempus tubes and buffy coat preserved in RNAlater. For 2005, buffy coat and whole blood samples, each without preservative buffers, are available. Although DNA of sufficient quality is likely to be available from all of these samples, it should be noted that particular attention needs to be paid to the potential batch effects that result, not only from DNA collected 10 years earlier, but also from the influence of the contained buffer.

Furthermore, preliminary analyses conducted during the conceptualisation of this thesis revealed that while the rural and urban participants demonstrated clear differences in lifestyle and behavioural patterns in 2005, these became largely indistinguishable in 2015. Should a longitudinal research question concerning urbanisation be pursued, it will be critical to characterise the rural development comprehensively, because of the rapid changes occurring in these areas. Overall, the pilot data generated for this PhD resulted in many interesting findings that warrant larger scale validation. Generating methylation and genomic data for all the PURE-SA-NW participants (n = 2 010 at baseline) will allow maximal use of the phenotypic data generated over 10 years for further exploration of questions of causality not dealt with in this thesis.

7.8 Conclusion

This thesis contributes to the epigenetic epidemiology literature by providing a characterisation of the methylome of an ethnic population never investigated before. This is particularly relevant in Chapter 4, which evaluates the reproducibility of the current EWAS literature on frequently investigated cardio-metabolic traits and identifies where population differences occur. Data provided in Chapter 5 add to the current literature by providing normal ranges of the neutrophil-to-lymphocyte and lymphocyte-to-monocyte ratios of the PURE-SA-NW cohort, in response to previous acknowledgements of lack of ethnically-diverse data of these biomarkers (Azab *et al.*, 2014; Jhuang *et al.*, 2019; Lee *et al.*, 2018). The role of these ratios in the context of cardiovascular health and disease risk is also described for the first time. Lastly, Chapter 6 reports the first investigation of DNAmAge in an African population.

This thesis provides a starting point for more targeted intervention strategies in (South) Africa by investigating the ethnic-specific molecular aspects of disease through investigating associations of various methylomic traits with cardio-metabolic risk factors. The data used to inform current policies and interventions aimed at mitigating cardio-metabolic effects are skewed toward European data, where genetic and environmental variation are less pronounced (Petrovski & Goldstein, 2016). The lack of population-specific literature in combination with the known increase in cardio-metabolic disease prevalence, particularly in low- and middle-income countries such as South Africa, necessitates attempts to tailor strategies that promote health and combat disease progression better to the specific needs of these populations.

BIBLIOGRAPHY

Aapola, U., Shibuya, K., Scott, H.S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S. & Krohn, K. 2000. Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22. 3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics*, 65(3):293-298.

Adkins, R.M., Krushkal, J., Tylavsky, F.A. & Thomas, F. 2011. Racial differences in gene-specific DNA methylation levels are present at birth. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 91(8):728-736.

Agyemang, C., Beune, E., Meeks, K., Owusu-Dabo, E., Agyei-Baffour, P., Aikins, A., Dodoo, F., Smeeth, L., Addo, J., Mockenhaupt, F.P., Amoah, S.K., Schulze, M.B., Danquah, I., Spranger, J., Nicolaou, M., Klipstein-Grobusch, K., Burr, T., Henneman, P., Mannens, M.M., van Straalen, J.P., Bahendeka, S., Zwinderman, A.H., Kunst, A.E. & Stronks, K. 2014. Rationale and cross-sectional study design of the research on obesity and type 2 diabetes among African migrants: the RODAM study. *BMJ Open*, 4(3):e004877.

Akinyemiju, T., Do, A.N., Patki, A., Aslibekyan, S., Zhi, D., Hidalgo, B., Tiwari, H.K., Absher, D., Geng, X. & Arnett, D.K. 2018. Epigenome-wide association study of metabolic syndrome in African-American adults. *Clinical Epigenetics*, 10(1):49.

Alfano, R., Guida, F., Galobardes, B., Chadeau-Hyam, M., Delpierre, C., Ghantous, A., Henderson, J., Herceg, Z., Jain, P., Nawrot, T.S., Relton, C., Vineis, P., Castagné, R. & Plusquin, M. 2018. Socioeconomic position during pregnancy and DNA methylation signatures at three stages across early life: epigenome-wide association studies in the ALSPAC birth cohort. *International Journal of Epidemiology*, 48(1):30-44.

Ambatipudi, S., Langdon, R., Richmond, R.C., Suderman, M., Koestler, D.C., Kelsey, K.T., Kazmi, N., Penfold, C., Ho, K.M. & McArdle, W. 2018a. DNA methylation derived systemic inflammation indices are associated with head and neck cancer development and survival. *Oral Oncology*, 85:87-94.

Ambatipudi, S., Sharp, G.C., Clarke, S.L., Plant, D., Tobias, J.H., Evans, D.M., Barton, A. & Relton, C.L. 2018b. Assessing the Role of DNA Methylation-Derived Neutrophil-to-Lymphocyte Ratio in Rheumatoid Arthritis. *Journal of Immunology Research*, 2018:2624981.

Ambrosi, C., Manzo, M. & Baubec, T. 2017. Dynamics and context-dependent roles of DNA methylation. *Journal of Molecular Biology*, 429(10):1459-1475.

Angkananard, T., Anothaisintawee, T., Ingsathit, A., McEvoy, M., Silapat, K., Attia, J., Sritara, P. & Thakkinstian, A. 2019. Mediation Effect of Neutrophil Lymphocyte Ratio on Cardiometabolic Risk Factors and Cardiovascular Events. *Scientific Reports*, 9(1):2618.

Aslibekyan, S., Demerath, E.W., Mendelson, M., Zhi, D., Guan, W., Liang, L., Sha, J., Pankow, J.S., Liu, C., Irvin, M.R., Fornage, M., Hidalgo, B., Lin, L.A., Thibeault, K.S., Bressler, J., Tsai, M.Y., Grove, M.L., Hopkins, P.N., Boerwinkle, E., Borecki, I.B., Ordovas, J.M., Levy, D., Tiwari, H.K., Absher, D.M. & Arnett, D.K. 2015. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. *Obesity*, 23(7):1493-1501.

Azab, B., Camacho-Rivera, M. & Taioli, E. 2014. Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects. *PloS One*, 9(11):e112361.

Balta, S., Kurtoglu, E., Kucuk, U., Demirkol, S. & Ozturk, C. 2014. Neutrophil–lymphocyte ratio as an important assessment tool. *Expert Review of Cardiovascular Therapy*, 12(5):537-538.

Banerjee, S. 2015. Multimorbidity—older adults need health care that can count past one. *The Lancet*, 385(9968):587-589.

Bao, X., Borné, Y., Johnson, L., Muhammad, I.F., Persson, M., Niu, K. & Engström, G. 2018. Comparing the inflammatory profiles for incidence of diabetes mellitus and cardiovascular diseases: a prospective study exploring the 'common soil' hypothesis. *Cardiovascular Diabetology*, 17(1):87.

Barcelona, V., Huang, Y., Brown, K., Liu, J., Zhao, W., Yu, M., Kardia, S.L., Smith, J.A., Taylor, J.Y. & Sun, Y.V. 2019. Novel DNA methylation sites associated with cigarette smoking among African Americans. *Epigenetics*:14(4):383-391

Baubec, T., Ivánek, R., Lienert, F. & Schübeler, D. 2013. Methylation-dependent and-independent genomic targeting principles of the MBD protein family. *Cell*, 153(2):480-492.

Beard, C., Li, E. & Jaenisch, R. 1995. Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes & Development*, 9(19):2325-2334.

Bennett, J.E., Stevens, G.A., Mathers, C.D., Bonita, R., Rehm, J., Kruk, M.E., Riley, L.M., Dain, K., Kengne, A.P. & Chalkidou, K. 2018. NCD Countdown 2030: worldwide trends in non-

communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *The Lancet*, 392(10152):1072-1088.

Bird, A.P. & Wolffe, A.P. 1999. Methylation-induced repression—belts, braces, and chromatin. *Cell*, 99(5):451-454.

Bojesen, S.E., Timpson, N., Relton, C., Smith, G.D. & Nordestgaard, B.G. 2017. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*, 72(7):646-653.

Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. 2019. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *Epigenomics*, 11(13):1469-1486.

Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., Slieker, R.C., Jhamai, P.M., Verbiest, M., Suchiman, H.E.D., Verkerk, M., van der Breggen, R., van Rooij, J., Lakenberg, N., Arindrarto, W., Kielbasa, S.M., Jonkers, I., van 't Hof, P., Nooren, I., Beekman, M., Deelen, J., van Heemst, D., Zhernakova, A., Tigchelaar, E.F., Swertz, M.A., Hofman, A., Uitterlinden, A.G., Pool, R., van Dongen, J., Hottenga, J.J., Stehouwer, C.D.A., van der Kallen, C.J.H., Schalkwijk, C.G., van den Berg, L.H., van Zwet, E.W., Mei, H., Li, Y., Lemire, M., Hudson, T.J., the, B.C., Slagboom, P.E., Wijmenga, C., Veldink, J.H., van Greevenbroek, M.M.J., van Duijn, C.M., Boomsma, D.I., Isaacs, A., Jansen, R., van Meurs, J.B.J., t Hoen, P.A.C., Franke, L. & Heijmans, B.T. 2016. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, 49(1):131-138.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T. & Weber, M. 2010. Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genetics*, 42(12):1093.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S. & Jacobsen, S.E. 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, 317(5845):1760-1764.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B. & Bestor, T.H. 2001. Dnmt3L and the establishment of maternal genomic imprints. *Science*, 294(5551):2536-2539.

Brathwaite, R., Addo, J., Kunst, A.E., Agyemang, C., Owusu-Dabo, E., de-Graft Aikins, A., Beune, E., Meeks, K., Klipstein-Grobusch, K., Bahendeka, S., Mockenhaupt, F.P., Amoah, S.,

Galbete, C., Schulze, M.B., Danquah, I. & Smeeth, L. 2017. Smoking prevalence differs by location of residence among Ghanaians in Africa and Europe: The RODAM study. *PLoS One*, 12(5):e0177291.

Braun, K.V.E., Dhana, K., de Vries, P.S., Voortman, T., van Meurs, J.B.J., Uitterlinden, A.G., Hofman, A., Hu, F.B., Franco, O.H. & Dehghan, A. 2017. Epigenome-wide association study (EWAS) on lipids: The Rotterdam study. *Clinical Epigenetics*, 9:15.

Breiling, A. & Lyko, F. 2015. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & Chromatin*, 8(1):24.

Buck-Koehntop, B.A. & Defossez, P.-A. 2013. On how mammalian transcription factors recognize methylated DNA. *Epigenetics*, 8(2):131-137.

Campanero, M.R., Armstrong, M.I. & Flemington, E.K. 2000. CpG methylation as a mechanism for the regulation of E2F activity. *Proceedings of the National Academy of Sciences*, 97(12):6481-6486.

Campbell, M.C. & 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*, 9:403-433.

Cardenas, A., Lutz, S.M., Everson, T.M., Perron, P., Bouchard, L. & Hivert, M.-F. 2019. Mediation by placental DNA methylation of the association of prenatal maternal smoking and birth weight. *American Journal of Epidemiology*, 188(11):1878-1886.

Chambers, J.C., Loh, M., Lehne, B., Drong, A., Kriebel, J., Motta, V., Wahl, S., Elliott, H.R., Rota, F. & Scott, W.R. 2015. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *The Lancet Diabetes & Endocrinology*, 3(7):526-534.

Chen, B.H., Marioni, R.E., Colicino, E., Peters, M.J., Ward-Caviness, C.K., Tsai, P.C., Roetker, N.S., Just, A.C., Demerath, E.W., Guan, W., Bressler, J., Fornage, M., Studenski, S., Vandiver, A.R., Moore, A.Z., Tanaka, T., Kiel, D.P., Liang, L., Vokonas, P., Schwartz, J., Lunetta, K.L., Murabito, J.M., Bandinelli, S., Hernandez, D.G., Melzer, D., Nalls, M., Pilling, L.C., Price, T.R., Singleton, A.B., Gieger, C., Holle, R., Kretschmer, A., Kronenberg, F., Kunze, S., Linseisen, J., Meisinger, C., Rathmann, W., Waldenberger, M., Visscher, P.M., Shah, S., Wray, N.R., McRae, A.F., Franco, O.H., Hofman, A., Uitterlinden, A.G., Absher, D., Assimes, T., Levine, M.E., Lu, A.T., Tsao, P.S., Hou, L., Manson, J.E., Carty, C.L., LaCroix, A.Z., Reiner, A.P., Spector, T.D.,

Feinberg, A.P., Levy, D., Baccarelli, A., van Meurs, J., Bell, J.T., Peters, A., Deary, I.J., Pankow, J.S., Ferrucci, L. & Horvath, S. 2016. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*, 8(9):1844-1865.

Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y. & Li, E. 2007. Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. *Nature Genetics*, 39(3):391-396.

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z. & Li, E. 2003. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and Cellular Biology*, 23(16):5594-5605.

Chikowore, T., Conradie, K.R., Towers, G.W. & van Zyl, T. 2015. Common variants associated with type 2 diabetes in a black South African population of Setswana descent: African populations diverge. *Omics*, 19(10):617-626.

Chitrala, K.N., Hernandez, D.G., Nalls, M.A., Mode, N.A., Zonderman, A.B., Ezike, N. & Evans, M.K. 2019. Race-specific alterations in DNA methylation among middle-aged African Americans and Whites with Metabolic Syndrome. *Epigenetics,* DOI: 10.1080/15592294.2019.1695340

Christman, J.K. 2002. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35):5483-5495.

Clapier, C.R. & Cairns, B.R. 2009. The biology of chromatin remodeling complexes. *Annual Review of Biochemistry*, 78:273-304.

Corriere, T., Di Marca, S., Cataudella, E., Pulvirenti, A., Alaimo, S., Stancanelli, B. & Malatino, L. 2018. Neutrophil-to-Lymphocyte Ratio is a strong predictor of atherosclerotic carotid plaques in older adults. *Nutrition, Metabolism and Cardiovascular Diseases*, 28(1):23-27.

Cronjé, H.T., Elliott, H.R., Nienaber-Rousseau, C. & Pieters, M. 2020. Replication and expansion of epigenome-wide association literature in a black South African population. *Clinical Epigenetics*, 12(1):6.

Cronjé, H.T., Nienaber-Rousseau, C., Zandberg, L., Chikowore, T., de Lange, Z., van Zyl, T. & Pieters, M. 2017a. Candidate gene analysis of the fibrinogen phenotype reveals the importance of polygenic co-regulation. *Matrix Biology*, 60:16-26.

Cronjé, H.T., Nienaber-Rousseau, C., Zandberg, L., De Lange, Z., Green, F.R. & Pieters, M. 2017b. Fibrinogen and clot-related phenotypes determined by fibrinogen polymorphisms: Independent and IL-6-interactive associations. *PloS One*, 12(11):e0187712.

Davis Armstrong, N.M., Chen, W.-M., Brewer, M.S., Williams, S.R., Sale, M.M., Worrall, B.B. & Keene, K.L. 2018. Epigenome-wide analyses identify two novel associations with recurrent stroke in the vitamin intervention for stroke prevention clinical trial. *Frontiers in Genetics*, 9:358.

De Lange, Z., Pieters, M., Jerling, J.C., Kruger, A. & Rijken, D.C. 2012. Plasma clot lysis time and its association with cardiovascular risk factors in black Africans. *PLoS One*, 7(11):e48881.

Deaton, A.M. & Bird, A. 2011. CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010-1022.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C. & Fuks, F. 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6):771-784.

Department of Health (South Africa), Statistics South Africa, South African Medical Research Council, & International Coach Federation. 2019. *South Africa demographic and health survey 2016.* Pretoria, South Africa and Rockville, Maryland, USA.

Dimauro, I., Scalabrin, M., Fantini, C., Grazioli, E., Valls, M.R.B., Mercatelli, N., Parisi, A., Sabatini, S., Di Luigi, L. & Caporossi, D. 2016. Resistance training and redox homeostasis: Correlation with age-associated genomic changes. *Redox Biology*, 10:34-44.

Ding, R., Jin, Y., Liu, X., Ye, H., Zhu, Z., Zhang, Y., Wang, T. & Xu, Y. 2017. Dose-and time-effect responses of DNA methylation and histone H3K9 acetylation changes induced by traffic-related air pollution. *Scientific Reports*, 7:43737.

Dirks, R.A., Stunnenberg, H.G. & Marks, H. 2016. Genome-wide epigenomic profiling for biomarker discovery. *Clinical Epigenetics*, 8(1):122.

Dixon, M.A. & Chartier, K.G. 2016. Alcohol use patterns among urban and rural residents: demographic and social influences. *Alcohol Research: Current Reviews*, 38(1):69-77.

Dodge, J.E., Okano, M., Dick, F., Tsujimoto, N., Chen, T., Wang, S., Ueda, Y., Dyson, N. & Li, E. 2005. Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *Journal of Biological Chemistry*, 280(18):17986-17991.

Dugué, P.A., Bassett, J.K., Joo, J.E., Jung, C.H., Ming Wong, E., Moreno-Betancur, M., Schmidt, D., Makalic, E., Li, S. & Severi, G. 2018. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *International Journal of Cancer*, 142(8):1611-1619.

Egger, G., Liang, G., Aparicio, A. & Jones, P.A. 2004. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457-463.

El Khoury, L.Y., Gorrie-Stone, T., Smart, M., Hughes, A., Bao, Y., Andrayas, A., Burrage, J., Hannon, E., Kumari, M. & Mill, J. 2019. Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biology*, 20(1):283.

Elliott, H.R., Shihab, H.A., Lockett, G.A., Holloway, J.W., McRae, A.F., Smith, G.D., Ring, S.M., Gaunt, T.R. & Relton, C.L. 2017. The role of DNA methylation in Type 2 diabetes aetiology–using genotype as a causal anchor. *Diabetes*, 66(6):1713-1722

Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Smith, G.D., Hughes, A.D., Chaturvedi, N. & Relton, C.L. 2014. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clinical Epigenetics*, 6(1):4.

Fernandez-Jimenez, N., Allard, C., Bouchard, L., Perron, P., Bustamante, M., Bilbao, J.R. & Hivert, M.-F. 2019. Comparison of Illumina 450K and EPIC arrays in placental DNA methylation. *Epigenetics*, 14(12):1177-1182

Fiorito, G., Robinson, O. & Vineis, P. 2019. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. *Aging (Albany NY)*, 11(7):2045-2070.

Fong, Y.W., Cattoglio, C. & Tjian, R. 2013. The intertwined roles of transcription and repair proteins. *Molecular Cell*, 52(3):291-302.

Forouzanfar, M.H., Afshin, A., Alexander, L.T., Anderson, H.R., Bhutta, Z.A., Biryukov, S., Brauer, M., Burnett, R., Cercy, K. & Charlson, F.J. 2016. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, 388(10053):1659-1724.

Fortin, J-P., Triche, T. & Hansen, K. 2016. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array. *Bioinformatics*; 33(4):558-560.

Fransquet, P.D., Wrigglesworth, J., Woods, R.L., Ernst, M.E. & Ryan, J. 2019. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clinical Epigenetics*, 11(1):62.

Gakidou, E., Afshin, A., Abajobir, A.A., Abate, K.H., Abbafati, C., Abbas, K.M., Abd-Allah, F., Abdulle, A.M., Abera, S.F. & Aboyans, V. 2017. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1345-1422.

Galanter, J.M., Gignoux, C.R., Oh, S.S., Torgerson, D., Pino-Yanes, M., Thakur, N., Eng, C., Hu, D., Huntsman, S., Farber, H.J., Avila, P.C., Brigino-Buenaventura, E., LeNoir, M.A., Meade, K., Serebrisky, D., Rodriguez-Cintron, W., Kumar, R., Rodriguez-Santana, J.R., Seibold, M.A., Borrell, L.N., Burchard, E.G. & Zaitlen, N. 2017. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife*, 6:e20532.

Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W.L. & Ho, K. 2016. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, 17(1):61.

Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, 526(7571):68-74.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M. & Carell, T. 2010. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS One*, 5(12):e15367.

Goll, M.G. & Bestor, T.H. 2005. Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, 74:481-514.

Government Communications (South Africa). 2019. *South Africa Yearbook 2018/19*. https://www.gcis.gov.za/south-africa-yearbook-201819 Date of access: 15 Oct. 2019.

Grundberg, E., Meduri, E., Sandling, J.K., Hedman, Å.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J. & Sekowska, M. 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *American Journal of Human Genetics*, 93(5):876-890.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., Ritchie, G.R.S., Xue, Y., Asimit, J., Nsubuga, R.N., Young, E.H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A.P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S.A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E. & Sandhu, M.S. 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534):327-332.

Guthrie, G.J., Charles, K.A., Roxburgh, C.S., Horgan, P.G., McMillan, D.C. & Clarke, S.J. 2013. The systemic inflammation-based neutrophil–lymphocyte ratio: experience in patients with cancer. *Critical Reviews in Oncology/Hematology*, 88(1):218-230.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B. & Gao, Y. 2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359-367.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X. & Cheng, X. 2012. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Research*, 40(11):4841-4849.

Haybar, H., Pezeshki, S.M.S. & Saki, N. 2019. Evaluation of complete blood count parameters in cardiovascular diseases: An early indicator of prognosis? *Experimental and Molecular Pathology*, 110:104267.

Haycock, P.C., Burgess, S., Wade, K.H., Bowden, J., Relton, C.L. & Davey Smith, G. 2016. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. American Journal of Clinical Nutrition, 103(4):965-978.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z. & Li, L. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, 333(6047):1303-1307.

Hedman, A.K., Mendelson, M.M., Marioni, R.E., Gustafsson, S., Joehanes, R., Irvin, M.R., Zhi, D., Sandling, J.K., Yao, C., Liu, C., Liang, L., Huan, T., McRae, A.F., Demissie, S., Shah, S., Starr, J.M., Cupples, L.A., Deloukas, P., Spector, T.D., Sundstrom, J., Krauss, R.M., Arnett, D.K., Deary, I.J., Lind, L., Levy, D. & Ingelsson, E. 2017. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for

results from genome-wide association studies. *Circulation: Cardiovascular Genetics*, 10(1):e001487.

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E. & Lumey, L. 2008. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences*, 105(44):17046-17049.

Hermann, A., Goyal, R. & Jeltsch, A. 2004. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 279(46):48350-48359.

Holland, N. 2017. Future of environmental research in the age of epigenomics and exposomics. *Reviews on Environmental Health*, 32(1-2):45-54.

Horvath, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156.

Horvath, S. & Raj, K. 2018. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6):371-384.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schönfels, W., Ahrens, M., Heits, N., Bell, J.T., Tsai, P.-C. & Spector, T.D. 2014. Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences*, 111(43):15538-15543.

Horvath, S., Oshima, J., Martin, G.M., Lu, A.T., Quach, A., Cohen, H., Felton, S., Matsuyama, M., Lowe, D. & Kabacik, S. 2018. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany NY)*, 10(7):1758-1775.

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. & Kelsey, K.T. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86.

Houseman, E.A., Molitor, J. & Marsit, C.J. 2014. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431-1439.

Hunter, D.J., James, L., Hussey, B., Wadley, A.J., Lindley, M.R. & Mastana, S.S. 2019. Impact of aerobic exercise and fatty acid supplementation on global and gene-specific DNA methylation. *Epigenetics*, 14(3):294-309.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52-58.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M. & Webster, M. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178-186.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. & Zhang, Y. 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129-1133.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C. & Zhang, Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300-1303.

Iurlaro, M., von Meyenn, F. & Reik, W. 2017. DNA methylation homeostasis in human and mouse development. *Current Opinion in Genetics & Development*, 43:101-109.

Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J. & Ramsahoye, B. 2004. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Molecular and Cellular Biology*, 24(20):8862-8871.

Jacobs, A., Schutte, A.E., Ricci, C. & Pieters, M. 2019. Plasminogen activator inhibitor-1 activity and the 4G/5G polymorphism are prospectively associated with blood pressure and hypertension status. *Journal of hypertension*, 37(12):2361-2370.

Jaffe, A.E. & Irizarry, R.A. 2014. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31.

Jähner, D., Stuhlmann, H., Stewart, C.L., Harbers, K., Löhler, J., Simon, I. & Jaenisch, R. 1982. De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature*, 298(5875):623-628.

Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J. & Pinder, M. 2009. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics*, 41(6):657-665.

Jhuang, Y.-H., Kao, T.-W., Peng, T.-C., Chen, W.-L., Li, Y.-W., Chang, P.-K. & Wu, L.-W. 2019. Neutrophil to lymphocyte ratio as predictor for incident hypertension: a 9-year cohort study in Taiwan. *Hypertension Research*, 42(8):1209-1214.

Jhun, M.A., Smith, J.A., Ware, E.B., Kardia, S.L., Mosley Jr, T.H., Turner, S.T., Peyser, P.A. & Park, S.K. 2017. Modeling the causal role of DNA methylation in the association between cigarette smoking and inflammation in African Americans: A 2-Step epigenetic mendelian randomization study. *American Journal of Epidemiology*, 186(10):1149-1158.

Joehanes, R., Just, A., Marioni, R., Pilling, L., Reynolds, L., Mandaviya, P., Guan, W., Xu, T., Elks, C. & Aslibekyan, S. 2016. Epigenetic signatures of cigarette smoking. *Circulation: Cardiovascular Genetics*, 9(5):436-447.

Jones, P.A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484-492.

Jones, P.A. & Baylin, S.B. 2002. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6):415-428.

Jones, P.A. & Laird, P.W. 1999. Cancer-epigenetics comes of age. *Nature Genetics*, 21(2):163-167.

Jones, P.A. & Liang, G. 2012. The Human Epigenome. *(In* Michels, K.B., *ed.* Epigenetic Epidemiology. Dordrecht: Springer Netherlands. p. 5-20).

Jordahl, K.M., Phipps, A.I., Randolph, T.W., Tindle, H.A., Liu, S., Tinker, L.F., Kelsey, K.T., White, E. & Bhatti, P. 2019. Differential DNA methylation in blood as a mediator of the association between cigarette smoking and bladder cancer risk among postmenopausal women. *Epigenetics*, 14(11):1065-1073.

Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E., Markunas, C.A., Richmond, R.C. & Xu, C.J. 2016a. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *American Journal of Human Genetics*, 98(4):680-696.

Joubert, B.R., Herman, T., Felix, J.F., Bohlin, J., Ligthart, S., Beckett, E., Tiemeier, H., Van Meurs, J.B., Uitterlinden, A.G. & Hofman, A. 2016b. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nature Communications*, 7:10577.

Jylhävä, J., Pedersen, N.L. & Hägg, S. 2017. Biological age predictors. *EBioMedicine*, 21:29-36.

Kelsey, K.T. & Wiencke, J.K. 2018. Immunomethylomics: A novel cancer risk prediction tool. *Annals of the American Thoracic Society*, 15(Suppl 2):S76-S80.

Klesges, R.C., Debon, M. & Ray, J.W. 1995. Are self-reports of smoking rate biased? Evidence from the second national health and nutrition examination survey. *Journal of Clinical Epidemiology*, 48(10):1225-1233.

Koestler, D., Jones, M., Usset, J., Christensen, B., Butler, R., Kobor, M., Wiencke, J. & Kelsey, K. 2016. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics*, 17:120.

Koestler, D.C., Usset, J., Christensen, B.C., Marsit, C.J., Karagas, M.R., Kelsey, K.T. & Wiencke, J.K. 2017. DNA methylation-derived neutrophil-to-lymphocyte ratio: an epigenetic tool to explore cancer inflammation and outcomes. *Cancer Epidemiology and Prevention Biomarkers*, 26(3):328-338.

Krebs, A.R., Dessus-Babus, S., Burger, L. & Schübeler, D. 2014. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife*, 3: e04094.

Kriaucionis, S. & Heintz, N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929-930.

Kriebel, J., Herder, C., Rathmann, W., Wahl, S., Kunze, S., Molnos, S., Volkova, N., Schramm, K., Carstensen-Kirberg, M., Waldenberger, M., Gieger, C., Peters, A., Illig, T., Prokisch, H., Roden, M. & Grallert, H. 2016. Association between DNA Methylation in whole blood and measures of glucose metabolism: KORA F4 study. *PLoS One*, 11(3):e0152314.

Kurdyukov, S. & Bullock, M. 2016. DNA methylation analysis: Choosing the right method. *Biology (Basel)*, 5(1):3.

Ladd-Acosta, C. & Fallin, M.D. 2016. The role of epigenetics in genetic and environmental epidemiology. *Epigenomics*, 8(2):271-283.

Laubach, Z.M., Perng, W., Dolinoy, D.C., Faulk, C.D., Holekamp, K.E. & Getty, T. 2018. Epigenetics and the maintenance of developmental plasticity: extending the signalling theory framework. *Biological Reviews*, 93(3):1323-1338.

Lee, H.J., Hore, T.A. & Reik, W. 2014. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell*, 14(6):710-719.

Lee, J.S., Kim, N.Y., Na, S.H., Youn, Y.H. & Shin, C.S. 2018. Reference values of neutrophil-lymphocyte ratio, lymphocyte-monocyte ratio, platelet-lymphocyte ratio, and mean platelet volume in healthy adults in South Korea. *Medicine*, 97(26):e11138.

Lee, J.T. 2012. Epigenetic regulation by long noncoding RNAs. *Science*, 338(6113):1435-1439.

Levine, M.E., Hosgood, H.D., Chen, B., Absher, D., Assimes, T. & Horvath, S. 2015. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. Aging (Albany NY), 7(9):690-700.

Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D. & Li, Y. 2018. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4):573-591.

Li, E., Bestor, T.H. & Jaenisch, R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915-926.

Li, S., Wong, E.M., Bui, M., Nguyen, T.L., Joo, J.-H.E., Stone, J., Dite, G.S., Giles, G.G., Saffery, R. & Southey, M.C. 2018. Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study. *Clinical Epigenetics*, 10(1):18.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W. & Jones, P.A. 2002. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Molecular and Cellular Biology*, 22(2):480-491.

Ligthart, S., Marzi, C., Aslibekyan, S., Mendelson, M.M., Conneely, K.N., Tanaka, T., Colicino, E., Waite, L.L., Joehanes, R., Guan, W., Brody, J.A., Elks, C., Marioni, R., Jhun, M.A., Agha, G., Bressler, J., Ward-Caviness, C.K., Chen, B.H., Huan, T., Bakulski, K., Salfati, E.L., Fiorito, G., Wahl, S., Schramm, K., Sha, J., Hernandez, D.G., Just, A.C., Smith, J.A., Sotoodehnia, N., Pilling, L.C., Pankow, J.S., Tsao, P.S., Liu, C., Zhao, W., Guarrera, S., Michopoulos, V.J., Smith, A.K., Peters, M.J., Melzer, D., Vokonas, P., Fornage, M., Prokisch, H., Bis, J.C., Chu, A.Y., Herder, C., Grallert, H., Yao, C., Shah, S., McRae, A.F., Lin, H., Horvath, S., Fallin, D., Hofman, A., Wareham, N.J., Wiggins, K.L., Feinberg, A.P., Starr, J.M., Visscher, P.M., Murabito, J.M., Kardia, S.L., Absher, D.M., Binder, E.B., Singleton, A.B., Bandinelli, S., Peters, A., Waldenberger, M., Matullo, G., Schwartz, J.D., Demerath, E.W., Uitterlinden, A.G., van Meurs,

J.B., Franco, O.H., Chen, Y.I., Levy, D., Turner, S.T., Deary, I.J., Ressler, K.J., Dupuis, J., Ferrucci, L., Ong, K.K., Assimes, T.L., Boerwinkle, E., Koenig, W., Arnett, D.K., Baccarelli, A.A., Benjamin, E.J. & Dehghan, A. 2016. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biology*, 17(1):255.

Ligthart, S., Vaez, A., Võsa, U., Stathopoulou, M.G., de Vries, P.S., Prins, B.P., van der Most, P.J., Tanaka, T., Naderi, E., Rose, L.M., Schraut, K.E., Joshi, P., Campbell, H., Wilson, J., Marioni, R. & Deary, I. 2018. Genome-wide association analyses of >200,000 individuals identify 58 genetic loci for chronic inflammation and highlights pathways that link inflammation and complex disorders. *American Journal of Human Genetics*, 103(5):691-706.

Lind, L., Ingelsson, E., Sundström, J., Siegbahn, A. & Lampa, E. 2018. Methylation-based estimated biological age and cardiovascular disease. *European Journal of Clinical Investigation*, 48(2).

Lioznova, A.V., Khamis, A.M., Artemov, A.V., Besedina, E., Ramensky, V., Bajic, V.B., Kulakovskiy, I.V. & Medvedeva, Y.A. 2019. CpG traffic lights are markers of regulatory regions in human genome. *BMC genomics*, 20(1):102.

Liu, C., Marioni, R.E., Hedman, Å.K., Pfeiffer, L., Tsai, P.-C., Reynolds, L.M., Just, A.C., Duan, Q., Boer, C.G. & Tanaka, T. 2018. A DNA methylation biomarker of alcohol consumption. *Molecular Psychiatry*, 23(2):422-433.

Liu, J., Zhao, W., Ammous, F., Turner, S.T., Mosley, T.H., Zhou, X. & Smith, J.A. 2019a. Longitudinal analysis of epigenome-wide DNA methylation reveals novel smoking-related loci in African Americans. *Epigenetics*, 14(2):171-184.

Liu, Z., Chen, B.H., Assimes, T.L., Ferrucci, L., Horvath, S. & Levine, M.E. 2019b. The role of epigenetic aging in education and racial/ethnic mortality disparities among older US Women. *Psychoneuroendocrinology*, 104:18-24.

Lopez-Candales, A., Hernández Burgos, P.M., Hernandez-Suarez, D.F. & Harris, D. 2017. Linking chronic inflammation with cardiovascular disease: from normal aging to the metabolic syndrome. Journal of Nature and Science, 3(4):e341.

Lu, A.T., Quach, A., Wilson, J.G., Reiner, A.P., Aviv, A., Raj, K., Hou, L., Baccarelli, A.A., Li, Y. & Stewart, J.D. 2019. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)*, 11(2):303-327.

Mandy, M. & Nyirenda, M. 2018. Developmental origins of health and disease: the relevance to developing nations. *International Health*, 10(2):66-70.

Mansell, G., Gorrie-Stone, T.J., Bao, Y., Kumari, M., Schalkwyk, L.S., Mill, J. & Hannon, E. 2019. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics*, 20(1):366.

Marigorta, U.M., Rodríguez, J.A., Gibson, G. & Navarro, A. 2018. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in Genetics*, 34(7):504-517.

Marioni, R.E., Harris, S.E., Shah, S., McRae, A.F., von Zglinicki, T., Martin-Ruiz, C., Wray, N.R., Visscher, P.M. & Deary, I.J. 2016. The epigenetic clock and telomere length are independently associated with chronological age and mortality. *International Journal of Epidemiology*, 45(2):424-432.

Marioni, R.E., Shah, S., McRae, A.F., Chen, B.H., Colicino, E., Harris, S.E., Gibson, J., Henders, A.K., Redmond, P. & Cox, S.R. 2015. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, 16(1):25.

Martin, E.M. & Fry, R.C. 2018. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. *Annual Review of Public Health*, 39:309-333.

Martin, G.M. 2005. Epigenetic drift in aging identical twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10413-10414.

Mathers, C.D., Stevens, G.A., Boerma, T., White, R.A. & Tobias, M.I. 2015. Causes of international increases in older age life expectancy. *The Lancet*, 385(9967):540-548.

McEwen, L.M., Jones, M.J., Lin, D.T.S., Edgar, R.D., Husquin, L.T., MacIsaac, J.L., Ramadori, K.E., Morin, A.M., Rider, C.F., Carlsten, C., Quintana-Murci, L., Horvath, S. & Kobor, M.S. 2018. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clinical Epigenetics*, 10(1):123.

McRae, A.F., Powell, J.E., Henders, A.K., Bowdler, L., Hemani, G., Shah, S., Painter, J.N., Martin, N.G., Visscher, P.M. & Montgomery, G.W. 2014. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biology*, 15(5):R73.

Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L. & Bird, A.P. 1989. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell*, 58(3):499-507.

Meeks, K.A., Henneman, P., Venema, A., Addo, J., Bahendeka, S., Burr, T., Danquah, I., Galbete, C., Mannens, M.M. & Mockenhaupt, F.P. 2018. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. *International Journal of Epidemiology*, 48(1):58-70.

Meeks, K.A., Henneman, P., Venema, A., Burr, T., Galbete, C., Danquah, I., Schulze, M.B., Mockenhaupt, F.P., Owusu-Dabo, E. & Rotimi, C.N. 2017. An epigenome-wide association study in whole blood of measures of adiposity among Ghanaians: the RODAM study. *Clinical Epigenetics*, 9(1):103.

Mendelson, M.M., Marioni, R.E., Joehanes, R., Liu, C., Hedman, A.K., Aslibekyan, S., Demerath, E.W., Guan, W., Zhi, D., Yao, C., Huan, T., Willinger, C., Chen, B., Courchesne, P., Multhaup, M., Irvin, M.R., Cohain, A., Schadt, E.E., Grove, M.L., Bressler, J., North, K., Sundstrom, J., Gustafsson, S., Shah, S., McRae, A.F., Harris, S.E., Gibson, J., Redmond, P., Corley, J., Murphy, L., Starr, J.M., Kleinbrink, E., Lipovich, L., Visscher, P.M., Wray, N.R., Krauss, R.M., Fallin, D., Feinberg, A., Absher, D.M., Fornage, M., Pankow, J.S., Lind, L., Fox, C., Ingelsson, E., Arnett, D.K., Boerwinkle, E., Liang, L., Levy, D. & Deary, I.J. 2017. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: A Mendelian randomization approach. *PLoS Medicine*, 14(1):e1002215.

Mercer, T.R., Dinger, M.E. & Mattick, J.S. 2009. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155-159.

Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., Milosavljevic, A., Siegmund, K.D., Bock, C. & Irizarry, R.A. 2013. Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949-955.

Min, J.L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. 2018. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, 34(23):3983-3989.

Miranda, J.J., Barrientos-Gutiérrez, T., Corvalan, C., Hyder, A.A., Lazo-Porras, M., Oni, T. & Wells, J.C.K. 2019. Understanding the rise of cardiometabolic diseases in low- and middle-income countries. *Nature Medicine*, 25(11):1667-1679.

Moore, L.D., Le, T. & Fan, G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23-28.

Nakabayashi, K. 2017. Illumina HumanMethylation BeadChip for genome-wide DNA methylation profiling: Advantages and limitations. (*In* Handbook of nutrition, diet, and Epigenetics. Springer, Cham. p. 1-15).

Needham, B.L., Smith, J.A., Zhao, W., Wang, X., Mukherjee, B., Kardia, S.L., Shively, C.A., Seeman, T.E., Liu, Y. & Diez Roux, A.V. 2015. Life course socioeconomic status and DNA methylation in genes related to stress reactivity and inflammation: The multi-ethnic study of atherosclerosis. *Epigenetics*, 10(10):958-969.

Nelson, P.G., Promislow, D.E.L. & Masel, J. 2019. Biomarkers for aging identified in cross-sectional studies tend to be non-causative. *The Journals of Gerontology. Series A*. *Biological Sciences and Medical Sciences*, 75(3):466-472.

Nienaber-Rousseau, C., Ellis, S.M., Moss, S.J., Melse-Boonstra, A. & Towers, G.W. 2013. Gene-environment and gene-gene interactions of specific MTHFR, MTR and CBS gene variants in relation to homocysteine in black South Africans. *Gene*, 530(1):113-118.

Nienaber-Rousseau, C., Sotunde, O.F., Ukegbu, P.O., Myburgh, P.H., Wright, H.H., Havemann-Nel, L., Moss, S.J., Kruger, I.M. & Kruger, H.S. 2017. Socio-demographic and lifestyle factors predict 5-year changes in adiposity among a group of black South African adults. *International Journal of Environmental Research and Public Health*, 14(9):1089.

Nyirenda, M.J. & Byass, P. 2019. Pregnancy, programming, and predisposition. *The Lancet Global Health*, 7(4):e404-e405.

Okano, M., Bell, D.W., Haber, D.A. & Li, E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247-257.

Okano, M., Xie, S. & Li, E. 1998. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics*, 19(3):219-220.

Olden, K., Olden, H.A. & Lin, Y.S. 2015. The role of the epigenome in translating neighborhood disadvantage into health disparities. *Current Environmental Health Reports*, 2(2):163-170.

Olkhov-Mitsel, E. & Bapat, B. 2012. Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. *Cancer Medicine*, 1(2):237-260.

Ortiz, R., Joseph, J.J., Lee, R., Wand, G.S. & Golden, S.H. 2018. Type 2 diabetes and cardiometabolic risk may be associated with increase in DNA methylation of FKBP5. *Clinical Epigenetics*, 10(1):82.

Panning, B. & Jaenisch, R. 1996. DNA hypomethylation can activate Xist expression and silence X-linked genes. *Genes & Development*, 10(16):1991-2002.

Park, L. 2019. Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Scientific Reports*, 9(1):11380.

Park, S.L., Patel, Y.M., Loo, L.W., Mullen, D.J., Offringa, I.A., Maunakea, A., Stram, D.O., Siegmund, K., Murphy, S.E. & Tiirikainen, M. 2018. Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations. *Clinical Epigenetics*, 10(1):110.

Perna, L., Zhang, Y., Mons, U., Holleczek, B., Saum, K.U. & Brenner, H. 2016. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clinical Epigenetics*, 8(1):64.

Petrovski, S. & Goldstein, D.B. 2016. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biology*, 17(1):157.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C. & Clark, S.J. 2016. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208.

Pieters, M. & Vorster, H.H. 2008. Nutrition and hemostasis: a focus on urbanization in South Africa. *Molecular Nutrition and Food Research*, 52(1):164-172.

Popejoy, A.B. & Fullerton, S.M. 2016. Genomics is failing on diversity. *Nature*, 538(7624):161-164.

Popkin, B.M. 2015. Nutrition transition and the global diabetes epidemic. *Current Diabetes Reports*, 15(9):64.

Pranavchand, R. & Reddy, B. 2016. Genomics era and complex disorders: Implications of GWAS with special reference to coronary artery disease, type 2 diabetes mellitus, and cancers. *Journal of Postgraduate Medicine*, 62(3):188-198.

Prioreschi, A., Wrottesley, S.V., Cohen, E., Reddy, A., Said-Mohamed, R., Twine, R., Tollman, S.M., Kahn, K., Dunger, D.B. & Norris, S.A. 2017. Examining the relationships between body image, eating attitudes, BMI, and physical activity in rural and urban South African young adult females using structural equation modeling. *PloS One*, 12(11):e0187508.

Quach, A., Levine, M.E., Tanaka, T., Lu, A.T., Chen, B.H., Ferrucci, L., Ritz, B., Bandinelli, S., Neuhouser, M.L., Beasley, J.M., Snetselaar, L., Wallace, R.B., Tsao, P.S., Absher, D., Assimes, T.L., Stewart, J.D., Li, Y., Hou, L., Baccarelli, A.A., Whitsel, E.A. & Horvath, S. 2017. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)*, 9(2):419-446.

Ramsay, M. 2012. Africa: continent of genome contrasts with implications for biomedical research and health. *FEBS letters*, 586(18):2813-2819.

Ramsay, M., De Vries, J., Soodyall, H., Norris, S.A. & Sankoh, O. 2014. Ethical issues in genomic research on the African continent: experiences and challenges to ethics review committees. *Human Genomics*, 8(1):15.

Razin, A. & Cedar, H. 1991. DNA methylation and gene expression. *Microbiological Reviews*, 55(3):451-458.

Reddy, K.S. 2016. Global burden of disease study 2015 provides GPS for global health 2030. *The Lancet*, 388(10053):1448-1449.

Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D., Söderhäll, C., Scheynius, A. & Kere, J. 2012. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS One*, 7(7):e41361.

Relton, C.L. & Davey Smith, G. 2010. Epigenetic epidemiology of common complex disease. *PloS Medicine*, 7(10):e1000356.

Retshabile, G., Mlotshwa, B.C., Williams, L., Mwesigwa, S., Mboowa, G., Huang, Z., Rustagi, N., Swaminathan, S., Katagirya, E. & Kyobe, S. 2018. Whole-exome sequencing reveals

uncaptured variation and distinct ancestry in the southern African population of Botswana. *American Journal of Human Genetics*, 102(5):731-743.

Richard, M.A., Huan, T., Ligthart, S., Gondalia, R., Jhun, M.A., Brody, J.A., Irvin, M.R., Marioni, R., Shen, J., Tsai, P.-C., Montasser, M.E., Jia, Y., Syme, C., Salfati, E.L., Boerwinkle, E., Guan, W., Mosley, T.H., Bressler, J., Morrison, A.C., Liu, C., Mendelson, M.M., Uitterlinden, A.G., van Meurs, J.B., Franco, O.H., Zhang, G., Li, Y., Stewart, J.D., Bis, J.C., Psaty, B.M., Chen, Y.-D.I., Kardia, S.L.R., Zhao, W., Turner, S.T., Absher, D., Aslibekyan, S., Starr, J.M., McRae, A.F., Hou, L., Just, A.C., Schwartz, J.D., Vokonas, P.S., Menni, C., Spector, T.D., Shuldiner, A., Damcott, C.M., Rotter, J.I., Palmas, W., Liu, Y., Smith, J.A., Deary, I.J. & Consortium, B. 2017. DNA methylation analysis identifies loci for blood pressure regulation. *American Journal of Human Genetics*, 101(6):888-902.

Richardson, T.G., Zheng, J., Smith, G.D., Timpson, N.J., Gaunt, T.R., Relton, C.L. & Hemani, G. 2017. Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk. *American Journal of Human Genetics*, 101(4):590-602.

Richmond, R.C., Sharp, G.C., Ward, M.E., Fraser, A., Lyttleton, O., McArdle, W.L., Ring, S.M., Gaunt, T.R., Lawlor, D.A. & Smith, G.D. 2016. DNA methylation and body mass index: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes,* 65(5):1231-1244.

Rider, C.F. & Carlsten, C. 2019. Air pollution and DNA methylation: effects of exposure in humans. *Clinical Epigenetics*, 11(1):131.

Riggs, A.D. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9-25.

Roberts, M.E., Doogan, N.J., Kurti, A.N., Redner, R., Gaalema, D.E., Stanton, C.A., White, T.J. & Higgins, S.T. 2016. Rural tobacco use across the United States: how rural and urban areas differ, broken down by census regions and divisions. *Health & Place*, 39:153-159.

Rosen, A.D., Robertson, K.D., Hlady, R.A., Muench, C., Lee, J., Philibert, R., Horvath, S., Kaminsky, Z.A. & Lohoff, F.W. 2018. DNA methylation age is accelerated in alcohol dependence. *Translational Psychiatry*, 8(1):182.

Ryan, J., Wrigglesworth, J., Loong, J., Fransquet, P.D. & Woods, R.L. 2019. A systematic review and meta-analysis of environmental, lifestyle, and health factors associated with DNA

methylation age. *The Journals of Gerontology. Series A. Biological Sciences and Medical Sciences,* 75(3):481-494.

Sadakierska-Chudy, A., Kostrzewa, R.M. & Filip, M. 2015. A comprehensive view of the epigenetic landscape part I: DNA methylation, passive and active DNA demethylation pathways and histone variants. *Neurotoxicity Research*, 27(1):84-97.

Salas, L.A., Koestler, D.C., Butler, R.A., Hansen, H.M., Wiencke, J.K., Kelsey, K.T. & Christensen, B.C. 2018. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biology*, 19(1):64.

Schlebusch, C.M. & Jakobsson, M. 2018. Tales of human migration, admixture, and selection in Africa. *Annual Review of Genomics and Human Genetics*, 19:405-428.

Schübeler, D. 2015. Function and information content of DNA methylation. *Nature*, 517(7534):321-326.

Schulz, W., Steinhoff, C. & Florl, A. 2006. Methylation of endogenous human retroelements in health and disease. *Current Topics in Microbiology and Immunology*, 310:211-50.

Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F. & Qi, J. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature*, 463(7283):943-947.

Shah, S., Bonder, M.J., Marioni, R.E., Zhu, Z., McRae, A.F., Zhernakova, A., Harris, S.E., Liewald, D., Henders, A.K. & Mendelson, M.M. 2015. Improving phenotypic prediction by combining genetic and epigenetic associations. *American Journal of Human Genetics*, 97(1):75-85.

Shah, S., McRae, A.F., Marioni, R.E., Harris, S.E., Gibson, J., Henders, A.K., Redmond, P., Cox, S.R., Pattie, A. & Corley, J. 2014. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Research*, 24(11):1725-1733.

Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T. & Okamura, K. 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171):908-912.

Sharp, G.C. & Relton, C.L. 2017. Epigenetics and noncommunicable diseases. *Epigenomics*, 9(6):789-791.

Sharp, G.C., Arathimos, R., Reese, S.E., Page, C.M., Felix, J., Küpers, L.K., Rifas-Shiman, S.L., Liu, C., Cohorts for Heart and Aging Research in Genomic Epidemiology plus (CHARGE +) methylation alcohol working group., Burrows, K., Zhao, S., Magnus, M.C., Duijts, L., Corpeleijn, E., DeMeo, D.L., Litonjua, A., Baccarelli, A., Hivert, M.-F., Oken, E., Snieder, H., Jaddoe, V., Nystad, W., London, S.J., Relton, C.L. & Zuccolo, L. 2018. Maternal alcohol consumption and offspring DNA methylation: findings from six general population-based birth cohorts. *Epigenomics*, 10(1):27-42.

Siegfried, Z. & Simon, I. 2010. DNA methylation and gene expression. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3):362-371.

Singh, V., Sharma, P. & Capalash, N. 2013. DNA methyltransferase-1 inhibitors as epigenetic therapy for cancer. *Current Cancer Drug Targets*, 13(4):379-399.

Smith, J.A., Zhao, W., Wang, X., Ratliff, S.M., Mukherjee, B., Kardia, S.L., Liu, Y., Roux, A.V.D. & Needham, B.L. 2017. Neighborhood characteristics influence DNA methylation of genes involved in stress response and inflammation: the multi-ethnic study of atherosclerosis. *Epigenetics*, 12(8): 662–673.

Soriano-Tárraga, C., Giralt-Steinhauer, E., Mola-Caminal, M., Vivanco-Hidalgo, R.M., Ois, A., Rodríguez-Campello, A., Cuadrado-Godia, E., Sayols-Baixeras, S., Elosua, R. & Roquer, J. 2016. Ischemic stroke patients are biologically older than their chronological age. *Aging (Albany NY)*, 8(11):2655-2665.

Soriano-Tárraga, C., Mola-Caminal, M., Giralt-Steinhauer, E., Ois, A., Rodríguez-Campello, A., Cuadrado-Godia, E., Gómez-González, A., Vivanco-Hidalgo, R.M., Fernández-Cadenas, I. & Cullell, N. 2017. Biological age is better than chronological as predictor of 3-month outcome in ischemic stroke. *Neurology*, 89(8):830-836.

Stanton, W.R., McClelland, M., Elwood, C., Ferry, D. & Silva, P.A. 1996. Prevalence, reliability and bias of adolescents' reports of smoking and quitting. *Addiction*, 91(11):1705-1714.

Stats SA (Statistics South Africa). 2016. *Community survey 2016*. (Statistical release P0301). http://cs2016.statssa.gov.za/wp-content/uploads/2016/07/NT-30-06-2016-RELEASE-for-CS-2016-_Statistical-releas_1-July-2016.pdf Date of access: 19 Jan. 2019.

Stats SA (Statistics South Africa). 2018. *Mortality and causes of death in South Africa, 2016: Findings from death notification.* (Statistical release P0309.3). http://www.statssa.gov.za/?page_id=1854&PPN=P0309.3 Date of access: 19 Jan. 2019.

Suárez-Cuenca, J.A., Ruíz-Hernández, A.S., Mendoza-Castañeda, A.A., Domínguez-Pérez, G.A., Hernández-Patricio, A., Vera-Gómez, E., De la Peña-Sosa, G., Banderas-Lares, D.Z., Montoya-Ramírez, J. & Blas-Azotla, R. 2019. Neutrophil-to-lymphocyte ratio and its relation with pro-inflammatory mediators, visceral adiposity and carotid intima-media thickness in population with obesity. *European Journal of Clinical Investigation*, 49(5):e13085.

Sun, Z., Cunningham, J., Slager, S. & Kocher, J.P. 2015. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*, 7(5):813-828.

Suzuki, M.M. & Bird, A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465-476.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R. & Aravind, L. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929):930-935.

Tajuddin, S.M., Hernandez, D.G., Chen, B.H., Noren Hooten, N., Mode, N.A., Nalls, M.A., Singleton, A.B., Ejiogu, N., Chitrala, K.N., Zonderman, A.B. & Evans, M.K. 2019. Novel age-associated DNA methylation changes and epigenetic age acceleration in middle-aged African Americans and whites. *Clinical Epigenetics*, 11(1):119.

Takai, D. & Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, 99(6):3740-3745.

Teh, A.L., Pan, H., Chen, L., Ong, M.-L., Dogra, S., Wong, J., MacIsaac, J.L., Mah, S.M., McEwen, L.M. & Saw, S.-M. 2014. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Research*, 24(7):1064-1074.

Teo, K., Chow, C.K., Vaz, M., Rangarajan, S. & Yusuf, S. 2009. The prospective urban rural epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. *American Heart Journal*, 158(1):1-7.

Teo, Y., Small, K.S. & Kwiatkowski, D.P. 2010. Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics*, 11(2):149-160.

Tishkoff, S.A. & Williams, S.M. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics*, 3(8):611-621.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M. & Doumbo, O. 2009. The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035-1044.

Titus, A.J., Gallimore, R.M., Salas, L.A. & Christensen, B.C. 2017. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human Molecular Genetics*, 26(R2):R216-R224.

Tobi, E.W., Goeman, J.J., Monajemi, R., Gu, H., Putter, H., Zhang, Y., Slieker, R.C., Stok, A.P., Thijssen, P.E. & Müller, F. 2014. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature Communications*, 5:5592.

Tobi, E.W., Lumey, L., Talens, R.P., Kremer, D., Putter, H., Stein, A.D., Slagboom, P.E. & Heijmans, B.T. 2009. DNA methylation differences after exposure to prenatal famine are common and timing-and sex-specific. *Human Molecular Genetics*, 18(21):4046-4053.

Tobi, E.W., Slieker, R.C., Luijk, R., Dekkers, K.F., Stein, A.D., Xu, K.M., Slagboom, P.E., Van Zwet, E.W., Lumey, L. & Heijmans, B.T. 2018. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Science Advances*, 4(1):eaao4364.

Turkmen, K., Guney, I., Yerlikaya, F.H. & Tonbul, H.Z. 2012. The relationship between neutrophil-to-lymphocyte ratio and inflammation in end-stage renal disease patients. *Renal Failure*, 34(2):155-159.

Van Dongen, J., Nivard, M.G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C.V., Ehli, E.A., Davies, G.E., Van Iterson, M. & Breeze, C.E. 2016. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, 7:11115.

Venkatraghavan, L., Tan, T.P., Mehta, J., Arekapudi, A., Govindarajulu, A. & Siu, E. 2015. Neutrophil lymphocyte ratio as a predictor of systemic inflammation – A cross-sectional study in a pre-admission setting. *F1000Research*, 4:123.

Vorster, H. 2002. The emergence of cardiovascular disease during urbanisation of Africans. *Public Health Nutrition*, 5(1a):239-243.

Walsh, C.P., Chaillet, J.R. & Bestor, T.H. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, 20(2):116-117.

Wang, X., Fan, X., Ji, S., Ma, A. & Wang, T. 2018. Prognostic value of neutrophil to lymphocyte ratio in heart failure patients. *International journal of Clinical Chemistry*, 485:44-49.

Watt, F. & Molloy, P.L. 1988. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & Development*, 2(9):1136-1143.

Wentzel-Viljoen, E., Lee, S., Laubscher, R. & Vorster, H.H. 2018. Accelerated nutrition transition in the North West province of South Africa: results from the prospective urban and rural epidemiology (PURE-NWP-SA) cohort study, 2005 to 2010. *Public Health Nutrition*, 21(14):2630-2641.

West-Eberhard, M.J. 1989. Phenotypic plasticity and the origins of diversity. *Annual Review of Ecology and Systematics*, 20(1):249-278.

WHO (World Health Organization). 2017. *WHO report on the global tobacco epidemic, 2017: monitoring tobacco use and prevention policies.* https://www.who.int/tobacco/global_report/2017/en/ Date of access: 5 Dec. 2018.

WHO (World Health Organization). 2018a. *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases Date of access: 9 Aug. 2019.

WHO (World Health Organization). 2018b. *World health statistics 2018: monitoring health for the SDGs, sustainable development goals*. https://www.who.int/gho/publications/world_health_statistics/2018/en/ Date of access: 25 Feb. 2020.

WHO (World Health Organization). 2018c. *Global status report on alcohol and health 2018.* https://www.who.int/substance_abuse/publications/global_alcohol_report/en/ Date of access: 5 Dec. 2018.

WHO (World Health Organization). 2018d. *Prevalence of insufficient physical activity.* https://www.who.int/gho/ncd/risk_factors/physical_activity_text/en/ Date of access: 5 Dec. 2018.

Wiencke, J.K., Koestler, D.C., Salas, L.A., Wiemels, J.L., Roy, R.P., Hansen, H.M., Rice, T., McCoy, L.S., Bracci, P.M. & Molinaro, A.M. 2017. Immunomethylomic approach to explore the blood neutrophil lymphocyte ratio (NLR) in glioma survival. *Clinical Epigenetics*, 9(1):10.

Wolf, E.J., Maniates, H., Nugent, N., Maihofer, A.X., Armstrong, D., Ratanatharathorn, A., Ashley-Koch, A.E., Garrett, M., Kimbrel, N.A. & Lori, A. 2017. Traumatic stress and accelerated DNA methylation age: A meta-analysis. *Psychoneuroendocrinology*, 92:123-134.

World Bank. 2018. Urban population (% of total population) - South Africa. https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?locations=ZA Date of access: 14 Dec. 2018.

Wu, H. & Zhang, Y. 2014. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2):45-68.

Xiao, D., Dasgupta, C., Chen, M., Zhang, K., Buchholz, J., Xu, Z. & Zhang, L. 2013. Inhibition of DNA methylation reverses norepinephrine-induced cardiac hypertrophy in rats. *Cardiovascular Research*, 101(3):373-382.

Zannas, A., Carrillo-Roa, T., Iurato, S., Ressler, K., Nemeroff, C., Smith, A., Lange, J., Bradley, B., Heim, C. & Brückl, T. 2015. Lifetime stress accelerates epigenetic aging. *European Psychiatry*, 30:799.

Zemach, A., McDaniel, I.E., Silva, P. & Zilberman, D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916-919.

Zhang, Y., Florath, I., Saum, K.-U. & Brenner, H. 2016. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environmental Research*, 146:395-403.

Zhao, W., Ammous, F., Ratliff, S., Liu, J., Yu, M., Mosley, T.H., Kardia, S.L. & Smith, J.A. 2019. Education and lifestyle factors are associated with DNA methylation clocks in older African Americans. *International Journal of Environmental Research and Public Health*, 16(17):3141.

Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., Bonder, M.J., van Rooij, J., Verkerk, M., Jhamai, P.M., Moed, M., Kielbasa, S.M., Bot, J., Nooren, I., Pool, R., van Dongen, J., Hottenga, J.J., Stehouwer, C.D.A., van der Kallen, C.J.H., Schalkwijk, C.G., Zhernakova, A., Li, Y., Tigchelaar, E.F., de Klein, N., Beekman, M., Deelen, J., van Heemst, D., van den Berg, L.H., Hofman, A., Uitterlinden, A.G., van Greevenbroek, M.M.J., Veldink, J.H., Boomsma, D.I., van Duijn, C.M., Wijmenga, C., Slagboom, P.E., Swertz, M.A., Isaacs, A., van Meurs, J.B.J., Jansen, R., Heijmans, B.T., t Hoen, P.A.C. & Franke, L. 2016. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*, 49(1):139-145.

Zhou, W., Laird, P.W. & Shen, H. 2017. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Research*, 45(4):e22.

Zhu, X., Li, J., Deng, S., Yu, K., Liu, X., Deng, Q., Sun, H., Zhang, X., He, M., Guo, H., Chen, W., Yuan, J., Zhang, B., Kuang, D., He, X., Bai, Y., Han, X., Liu, B., Li, X., Yang, L., Jiang, H., Zhang, Y., Hu, J., Cheng, L., Luo, X., Mei, W., Zhou, Z., Sun, S., Zhang, L., Liu, C., Guo, Y., Zhang, Z., Hu, F.B., Liang, L. & Wu, T. 2016. Genome-wide analysis of DNA methylation and cigarette smoking in a Chinese population. *Environmental Health Perspectives*, 124(7):966-973.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A. & Bernstein, B.E. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477-481.

ANNEXURE A

Cronjé, H.T., Elliott, H.R., Nienaber-Rousseau, C. & Pieters, M. 2020. Replication and expansion of epigenome-wide association literature in a black South African population. *Clinical Epigenetics*, 12(1):6.

Cronjé et al. Clinical Epigenetics (2020) 12:6
https://doi.org/10.1186/s13148-019-0805-z

Clinical Epigenetics

RESEARCH Open Access

Replication and expansion of epigenome-wide association literature in a black South African population

H. Toinét Cronjé[1*], Hannah R. Elliott[2,3], Cornelie Nienaber-Rousseau[1] and Marlien Pieters[1]

Abstract

Background: DNA methylation is associated with non-communicable diseases (NCDs) and related traits. Methylation data on continental African ancestries are currently scarce, even though there are known genetic and epigenetic differences between ancestral groups and a high burden of NCDs in Africans. Furthermore, the degree to which current literature can be extrapolated to the understudied African populations, who have limited resources to conduct independent large-scale analysis, is not yet known. To this end, this study examines the reproducibility of previously published epigenome-wide association studies of DNA methylation conducted in different ethinicities, on factors related to NCDs, by replicating findings in 120 South African Batswana men aged 45 to 88 years. In addition, novel associations between methylation and NCD-related factors are investigated using the Illumina EPIC BeadChip.

Results: Up to 86% of previously identified epigenome-wide associations with NCD-related traits (alcohol consumption, smoking, body mass index, waist circumference, C-reactive protein, blood lipids and age) overlapped with those observed here and a further 13% were directionally consistent. Only 1% of the replicated associations presented with effects opposite to findings in other ancestral groups. The majority of these inconcistencies were associated with population-specific genomic variance. In addition, we identified eight new 450K array CpG associations not previously reported in other ancestries, and 11 novel EPIC CpG associations with alcohol consumption.

Conclusions: The successful replication of existing EWAS findings in this African population demonstrates that blood-based 450K EWAS findings from commonly investigated ancestries can largely be extrapolated to ethnicities for which epigenetic data are not yet available. Possible population-specific differences in 14% of the tested associations do, however, motivate the need to include a diversity of ethnic groups in future epigenetic research. The novel associations found with the enhanced coverage of the Illumina EPIC array support its usefulness to expand epigenetic literature.

Keywords: Ancestry, DNAm, EPIC, Epigenetic epidemiology, EWAS, Methylation, NCD, PURE

Background

The role of epigenetics in the aetiology of non-communicable diseases (NCDs) is of interest owing to its valuable addition to the limited variance of disease risk explained by genetics alone [1]. The modifiable nature of the epigenome also offers opportunities to predict, detect and prevent lifestyle-related diseases [2]. DNA methylation (DNAm) is the most intensively researched epigenetic modification, partly because of its ease of measurement from stored samples commonly collected in epidemiological studies. A number of robust associations between differentially methylated cytosine-guanine dinucleotides (CpGs) and NCD-related traits or exposures have been reported [3–7]. Epigenetic research has allowed for richer insight into the origin and progression of complex diseases, and is expected to continue doing so, thereby enhancing our ability to combat the continued rise in NCD prevalence [2, 8].

Despite its importance in the global context of NCDs, current epigenetic literature remains limited by the lack of ethnic diversity, with most investigating associations between DNAm and health outcomes/traits within European (EU) populations. Although several large-scale epigenome-wide association studies (EWASs) have used

* Correspondence: 23520825@nwu.ac.za
[1]Centre of Excellence for Nutrition at the North-West University Potchefstroom Campus, Potchefstroom 2520, South Africa
Full list of author information is available at the end of the article

data collected from African American (AA) individuals [4, 5, 9], information on continental African populations remain particularly limited. Sub-Saharan Africans are known to be genetically different from AA individuals, who typically stem from West African ancestors, with varying levels of admixture [10]. Because DNAm differences have been reported among ethnic groups [11–13], the degree to which current EWAS results can be extrapolated to other populations, including Sub-Saharan Africans, remains to be established. Understanding the degree of generalisability of EWAS results to different ethnicities informs one whether existing knowledge can be extrapolated to understudied ethnic groups or whether additional research is needed in these populations, where resources are often limited [14].

To this end, we replicated data extracted from the EWAS Catalogue (http://www.ewascatalog.org) on traits related to NCDs (alcohol consumption, smoking status, body mass index (BMI), waist circumference (WC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglycerides (TG), C-reactive protein (CRP) and age), in a subset of Batswana men from the North-West (NW) province of South Africa, who participated in the international Prospective Urban and Rural Epidemiology study (PURE-SA-NW). In doing so, we evaluated the reproducibility of previous EWAS findings in a Sub-Saharan African population that has never been investigated before. In addition, because the majority of EWASs to date have been conducted using the older Illumina 450 K BeadChip, our secondary aim was to report novel DNAm associations by using the new Illumina MethylationEPIC platform, to extend existing knowledge on methylation and traits related to NCDs in this population [15].

Results

For each trait, we report the degree of replication between EWAS findings in the PURE-SA-NW cohort and the reference studies identified, using the EWAS catalogue (complete test statistics in Additional file 1). In cases where the reference study included cohorts of different ancestries, the PURE-SA-NW cohort was compared to these ancestries separately. We first report the agreement between the effect sizes obtained in the PURE-SA-NW data and the reference studies for all the tested CpGs per trait, to evaluate the overall consensus between the studies (PURE-SA-NW vs. reference study). We then examine the similarity between studies at the individual CpG level by determining whether or not the individual PURE-SA-NW association's confidence intervals (CIs) overlap with those of the reference study. This allows us to identify systematic differences (e.g. attributable to exposure variation) between cohorts before investigating differences at an individual CpG level (e.g. attributable to site-specific genetic variation). To permit further investigation of individual CpG association differences, we inspect probes previously identified to measure methylation at polymorphic sites of which either the global minor allele frequency (MAF) is higher than 1% [16], or variation has been documented in Africans, specifically [17] (Additional file 1). Probes identified to hybridise to multiple genomic regions or to be cross-reactive are also noted [18]. Replication analyses are followed by a report of any methylation associations of newly investigated EPIC probes and novel 450K associations (of 450K probes present on the EPIC array used here), where applicable (Additional file 2). Table 1 provides the descriptive statistics for (i) the PURE-SA-NW cohort for traits used as covariates in the models, (ii) the trait of interest as reported by the EWAS catalogue reference study and (iii) the PURE-SA-NW cohort trait of interest reported in the same unit as in the specific reference study. For the different traits, the sample size here differs because we applied the specific inclusion criteria of the respective reference studies to our population to permit comparison (Additional file 1).

Comparatively, our study population had a more favourable body composition and blood lipid profile, but a much higher CRP concentration than those included in the reference studies [3, 9, 22]. The proportion of current smokers in our study population was twice as high as the reference cohort [4], and they consumed larger volumes of alcohol than the EU, but less than the AA reference cohorts [5]. The remaining traits were similar between our population and that of the reference studies.

Alcohol consumption

Ancestry-stratified (European American (EA) and AA) findings from the meta-analysis by Liu et al. [5] on the association of alcohol consumption (g/day) with differential methylation at individual CpGs were compared with those from the PURE-SA-NW (Fig. 1). In the study of Liu et al. [5], alcohol consumption was more strongly associated with DNAm in AA than EA individuals (regression slope = 3.2, $p = 8.6 \times 10^{-70}$). Effect sizes in the PURE-SA-NW cohort were larger than in either of the reference groups (regression slope = 0.12, $p = 3.2 \times 10^{-16}$ and 0.47, $p = 3.2 \times 10^{-17}$ for the AA and EA comparisons, respectively).

Individual association results showed stronger similarity between the PURE-SA-NW and the AA than with the EA findings. Overall, 361 CpGs were investigated (two unique AA, 131 unique EA and 228 associations reported for both ethnicities). Out of the 230 association tests to compare the AA reference cohort to the PURE-SA-NW data, 93% (213) of the regression CIs

Table 1 Descriptive characteristics of the study and reference cohorts

Trait	PURE-SA-NW	Reference study	Comparative PURE-SA-NW	Reference study citation
N	120	See Additional file 1		
Age (years)	64 [55–70]	62 [58 – 67]	64 [55–70]	[19]
BMI (kg/m^2)	22.5 ± 4.9	27.6 ± 4.4[a,e]	22.4 ± 5.0	[20]
		27.7 ± 4.5[b,e]		
WC (cm)	83.8 ± 12.8	101 ± 15.1[c,e]	83.6 ± 12.7	[9]
		97 ± 16[d,e]		[21]
Physical activity (index)	2.41 ± 0.94			
Smoking status [*N* (%)]				
Never smoker	56 (47)	6956 (74)[b,d,e]	56 (48)	[4]
Current smoker	61 (51)	2433 (26)[b,d,e]	61 (52)	
Ever smoker	64 (53)			
Alcohol use [*N* (%)]				
Never user	56 (47)			
Ever user	64 (53)			
Alcohol consumption (g/day)	16.7 ± 36.6	1.3 (0, 301)[c]	0 (0, 240)	[5]
		5.6 (0, 181)[d]		
CRP (mg/L)	9.7 ± 27.2	6.2 ± 8.8[c,e]	9.9 ± 27.5	[22]
		3.3 ± 5.6[d,e]		
TC (mg/dL)	171 ± 41.6	207 ± 37.1[d,e]	171 ± 41.6	[3]
LDL-C (mg/dL)	96.5 ± 35.7	125 ± 30.9[d,e]	96.5 ± 35.7	
HLD-C (mg/dL)	54.1 ± 22.7	57.0 ± 16.8[d]	54.1 ± 22.7	
TG (mg/dL)	48.5 ± 30.5	126 ± 69.0[d,e]	48.5 ± 30.5	
Education [*N* (%)]				
None	26 (22)			
1–7 years of schooling	66 (55)			
8–12 years of schooling	28 (23)			
Blood cell type proportions (%)				
B cells	0.04 ± 0.02			
CD4 T cells	0.11 ± 0.04			
CD8 T cells	0.11 ± 0.06			
Granulocytes	0.47 ± 0.11			
Monocytes	0.09 ± 0.02			
Natural killer cells	0.11 ± 0.03			

BMI body mass index, *CRP* C-reactive protein, *HDL-C* high-density lipoprotein cholesterol, *LDL-C* low-density lipoprotein cholesterol, *TC* total cholesterol, *TG* triglycerides, *WC* waist circumference. Values are presented as median [IQR], mean ± standard deviation, *N* (%) or median (minimum, maximum). Blood cell proportions were determined using methylation-based estimates [23]

[a]Indian Asian ancestry

[b]European American ancestry

[c]African American ancestry

[d]European ancestry

[e]Population means differ between the reference study and comparative PURE-SA-NW study population at $p < 0.05$ following Bonferroni adjustment

overlapped, compared to 80% (287) of the 359 comparisons between the EA and PURE-SA-NW. Where CIs did not overlap, directional consistency was nevertheless observed with the exception of the associations for cg15636519 (EA and AA comparisons), cg08471846 (EA comparison only) and cg21227253 (EA comparison only) with alcohol consumption (Additional file 1a). Data from the Biobank-based Integrative Omics Studies (BIOS) Consortium indicated that, apart from cg08471846, methylation quantitative trait loci (mQTLs) have been identified for each of these CpGs with absolute reported Z-scores ranging from 4.15 to 12.9 [24, 25]. Data from

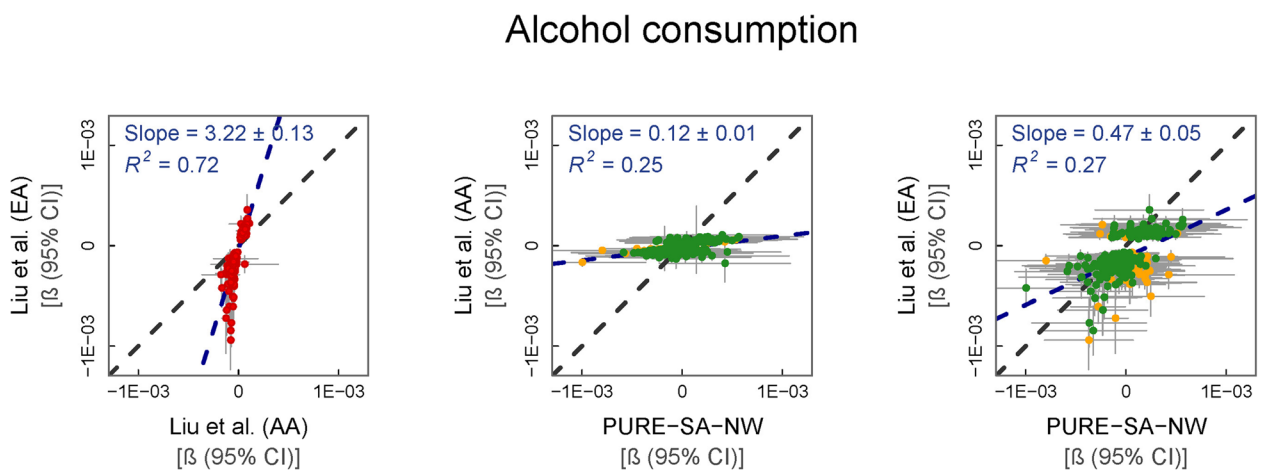


Fig. 1 % Methylation change per gram of alcohol intake. From left to right: (i) reference AA vs. EA data (247 CpGs), (ii) PURE-SA-NW vs. AA data (228 CpGs) and (iii) PURE-SA-NW vs. EA data (228 CpGs). Model used: methylation ~ alcohol consumption + age + BMI + cell counts + surrogate variables. Reference data: Liu et al. (2018). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line

the 1000 Genomes project support that the differences observed here could be partly influenced by ancestry-specific genetic variance; for example, the MAF of rs7153432 (*cis* mQTL for cg21227253) is 18% in Africans and 40% in Europeans [26].

The EWAS conducted on alcohol consumption in the PURE-SA-NW cohort resulted in 19 genome-wide significant findings ($p < 9.4 \times 10^{-8}$), 11 of which were newly investigated EPIC probes and eight were part of those previously investigated by 450K probes, that were present on the EPIC array, but failed to reach association thresholds in other cohorts (Additional file 2a). Table 2 provides the test statistics for these CpGs.

The proportion of methylation variance of these CpGs explained by including alcohol consumption in the model methylation ~ age + BMI + cell counts + smoking status, ranged from 10.3 to 43.8%. When alcohol consumption was used as the outcome variable, the addition of these 19 probes to the regression model, increased the percentage of alcohol consumption variance explained by 57% (adjusted $R^2 = 0.05$ before and 0.62 after including the CpGs, $p = 5.5 \times 10^{-26}$).

Smoking status

The association of smoking status with the DNAm of 3618 CpGs in the PURE-SA-NW cohort was compared to a multi-ethnic (EA and AA) EWAS conducted by Joehanes et al. [4]. 'Current' users in the PURE-SA-NW cohort included individuals regularly smoking any bought or self-made tobacco product (commercial cigarettes, bidis, pipes and cigars). Joehanes et al. [4], however, restricted the definition of 'current' smokers to those specifically reporting cigarette use. Regardless of the discrepancy in the product smoked, results from the respective EWASs were fairly similar. No ancestral comparisons were made by Joehanes et al. [4], who combined data from a number of different ethnic groups in a meta-analysis.

Effect sizes were generally larger in the PURE-SA-NW than in the reference data (regression slope = 0.34, $p = 1.7 \times 10^{-206}$). Of the 3618 CpGs tested for their independent association with smoking status, 3315 (92%) of the regression β 95% CIs overlapped and 269 were directionally consistent between cohorts (Fig. 2). Only 34 CpGs showed a difference in the direction of effect between the findings of Joehanes et al. [4] and the PURE-SA-NW cohort (Additional file 1b). Thirteen of these probes measure methylation at polymorphic sites, 20 have *cis*-mQTLs and five have *trans*-mQTLs, all of which with differing AA and EU ancestry MAFs, suggesting that genetic variation between cohorts could drive some of the dissimilarities observed [24–26]. No novel associations with smoking were identified and the only genome-wide significant CpG association was for a previously identified CpG (cg05575921) that was associated with a 17% ($p = 4.2 \times 10^{-10}$) reduction in DNAm in *current* smokers compared to participants who had *never* smoked (Additional file 2b).

Body mass index

We replicated findings from the largest EWAS on BMI conducted to date, that of Wahl et al. [20]. These authors investigated the relationship of methylation with BMI in individuals of Indian Asian (IA) and EU descent. Wahl et al. [20] observed larger effect sizes among the IA than the EU group (regression slope = 0.48, $p = 4.9 \times 10^{-72}$). PURE-SA-NW data reflected the IA better than

Table 2 EWAS CpG-alcohol consumption associations $p < 9.4 \times 10^{-8}$

ProbeID	Location	Gene	Region	β	SE	*p*	% Variance explained	X^2 *p* value
cg13153796[a]	14:101405628	*SNORD*113-6	TSS1500	− 6.78E−04	8.45E−05	2.2E−11	29.4 (38.4)	3.8E−12
cg00712390[a]	17:79373624	*BAHCC*1	1stExon	8.08E−04	1.14E−04	9.7E−10	37.5 (47.0)	5.6E−18
cg05706661	7:36134301	*LOC*101928618	TSS1500	− 1.05E−03	1.51E−04	2.0E−09	17.6 (57.1)	6.7E−12
cg24252287[a]	17:40250379			1.48E−04	2.21E−05	4.8E−09	36.5 (41.8)	7.7E−15
cg12177743[a]	11:113185079	*TTC*12	TSS200	1.59E−04	2.41E−05	7.5E−09	13.4 (23.4)	2.8E−05
cg19323439	17:9136232	*NTN*1	Body	5.06E−04	7.93E−05	1.9E−08	14.1 (59.0)	2.1E−08
cg19683675[a]	5:142077712	*FGF*1	TSS200	− 1.13E−03	1.78E−04	2.0E−08	35.1 (43.6)	2.7E−15
cg08333974	12:1956337	*CACNA2D*4	Body	− 1.24E−03	1.95E−04	2.2E−08	25.8 (38.3)	8.3E−12
cg12325997	15:59280148	*RNF*111	1stExon	9.84E−05	1.57E−05	3.2E−08	10.5 (58.4)	9.4E−08
cg19642811	13:95453039	*LOC*101927284	Body	− 6.02E−04	9.64E−05	3.4E−08	19.3 (37.3)	2.3E−08
cg06943216	8:102683096			− 1.33E−03	2.13E−04	3.5E−08	17.5 (33.9)	4.3E−08
cg26187237[a]	2:217498574	*IGFBP*2	1stExon	4.19E−04	6.72E−05	3.6E−08	15.5 (53.0)	2.2E−09
cg16358446[a]	1:1534984			8.10E−05	1.31E−05	4.4E−08	43.8 (52.4)	1.9E−21
cg08724692	6:133646558	*EYA*4	Body	− 6.26E−04	1.03E−04	6.4E−08	10.3 (43.4)	1.2E−06
cg08035774	9:136600662	*SARDH*	5′UTR	− 1.12E−03	1.85E−04	7.5E−08	23.6 (32.8)	6.3E−10
cg18780412[a]	3:179755086	*PEX5L*	TSS1500	6.36E−04	1.06E−04	8.6E−08	27.0 (33.5)	6.4E−11
cg15942324	1:38482118	*UTP*11*L*	Body	− 6.63E−04	1.10E−04	8.8E−08	23.3 (33.2)	3.4E−09
cg25278025	2:103378026	*TMEM*182	TSS1500	5.99E−04	9.98E−05	8.8E−08	15.4 (26.0)	5.0E−06
cg22572934[b]	5:173171061	*LINC*01484	Body	− 1.21E−03	2.02E−04	9.3E−08	13.3 (24.3)	5.5E−05

Model: methylation ~ alcohol consumption (g/d) + age + BMI + smoking + cell counts + surrogate variables
[a]450K probes
[b]Probe that should be interpreted with caution owing to the presence of genomic variance at probe measurement site [17]
The percentage variance explained reflects the added value of alcohol consumption to the variance in CpG methylation, reported as percentage explained by alcohol as an added exposure (percentage variance explained by the total model). X^2 *p* value = Chi-square *p* value when the regression models with and without alcohol consumption are compared

the EU data, but in both instances, PURE-SA-NW data showed larger effect sizes than either reference group (regression slope = 0.57, $p = 6.0 \times 10^{-7}$ and 0.37, $p = 1.8 \times 10^{-8}$ for IA and EU groups, respectively). However, when comparing the overlap between individual effect estimates, PURE-SA-NW mirrored findings from the EU group better. The 95% CIs of the 265 regression estimates between the cohorts overlapped 55% (147) and 77% (203) of the time when compared with IA data and EU data, respectively (Fig. 3). All regression CIs that did not overlap were directionally consistent between the PURE-SA-NW and reference cohorts. No genome-wide significant associations with BMI were identified (Additional file 2c).

Waist circumference

Eight previously reported associations of WC with DNAm in cohorts of AA and EA descent [9] were replicated in the PURE-SA-NW cohort (Fig. 4). The regression model used to quantify the relationship between WC and DNAm differed between the reference cohort subgroups. In addition to the covariates adjusted for in the EA regression model (age, smoking and white blood cell counts), the AA model also included alcohol consumption status, physical activity, education and household income. The use of the two different models was justified, as it resulted in highly comparable findings between the reference study's AA and EA groups ($r = 0.96$), with a slightly larger average effect size observed in the EA than in the AA data (regression slope = 0.56, $p = 0.0001$). Applying the fully adjusted (AA) model to the PURE-SA-NW data resulted in a 10.4% increase in average effect size compared to the model used for the EA group, justifying the use of the fully adjusted model in our cohort.

As for the previous traits, larger effect sizes were observed in the PURE-SA-NW than both the AA (regression slope = 0.38, $p = 0.03$) and EA (regression slope = 0.21, $p = 0.04$) cohorts with a closer resemblance to the AA than the EA data ($R^2 = 0.53$. vs. 0.47). When comparing individual effect estimates between the PURE-SA-NW and reference data, the 95% CIs overlapped in seven of the eight assessed associations in both groups (Additional file 1d). The non-overlapping associations were directionally consistent between studies, overall indicating strong comparability between WC's association with DNAm across the investigated ancestral groups. The single non-overlapping locus was the same in both ethnic groups compared. This

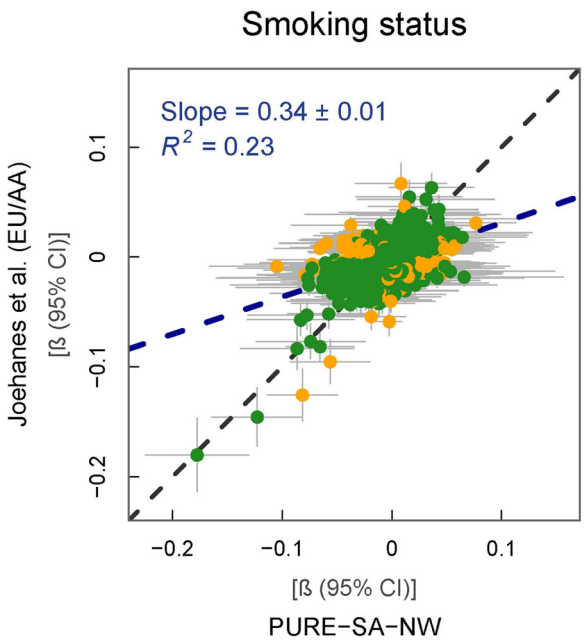


Fig. 2 % Methylation difference between current and never smokers in reference vs. PURE-SA-NW data. Model used for PURE-SA-NW EWAS: methylation ~ smoking + age + cell counts + surrogate variables. Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Black dashed line: line of equality. Blue dashed line: regression line

site, cg26403843, is associated with five *cis*-mQTLs and one *trans*-mQTL with absolute Z-scores ranging from 4.9 to 39.8. Population differences between the mQTL-associated SNPs were observed; rs6556405, for example, has a MAF of 26% in Europenas compared to a frequency of 66% in Africans [24–26].

Blood lipids

Findings from the largest TC, LDL-C, HDL-C and TG EWASs to date, reported by Hedman et al. [3], were compared to those of the PURE-SA-NW cohort. For each of the four lipids, larger effect sizes were observed in the PURE-SA-NW than in the EU reference cohort. The regression slopes when modelling the PURE-SA-NW effect sizes against those of the reference cohorts' were 0.12 (p = 0.18), 0.13 (p = 0.27), 0.19 ($p = 9.9 \times 10^{-06}$) and 0.30 (p = 0.01) for TC, LDL-C, HDL-C and TG, respectively (Fig. 5). Effect estimates and 95% CIs overlapped for 38/40 (95%) for TC, 18/21 (86%) for LDL-C, 96/102 (94%) for HDL-C and 15/16 (94%) for TG, of the associations tested (Additional file 1e). Ten of the 12 non-overlapping associations were directionally consistent, leaving only two associations divergent in the direction of effect: cg24939194-HDL-C and cg15878619-TC. Two mQTLs have been identified for cg24939194 (rs748097 and rs2969017), the strongest of which has a MAF of 6% in Africans and 37% in Europeans, indicating that genetic ancestry may be important for the association of cg24939194 with HDL-C [26].

Despite the consistency in the effect sizes between the PURE-SA-NW and the reference data, the large CIs observed in our data do not allow for further interpretation of these findings. There was one genome-wide significant

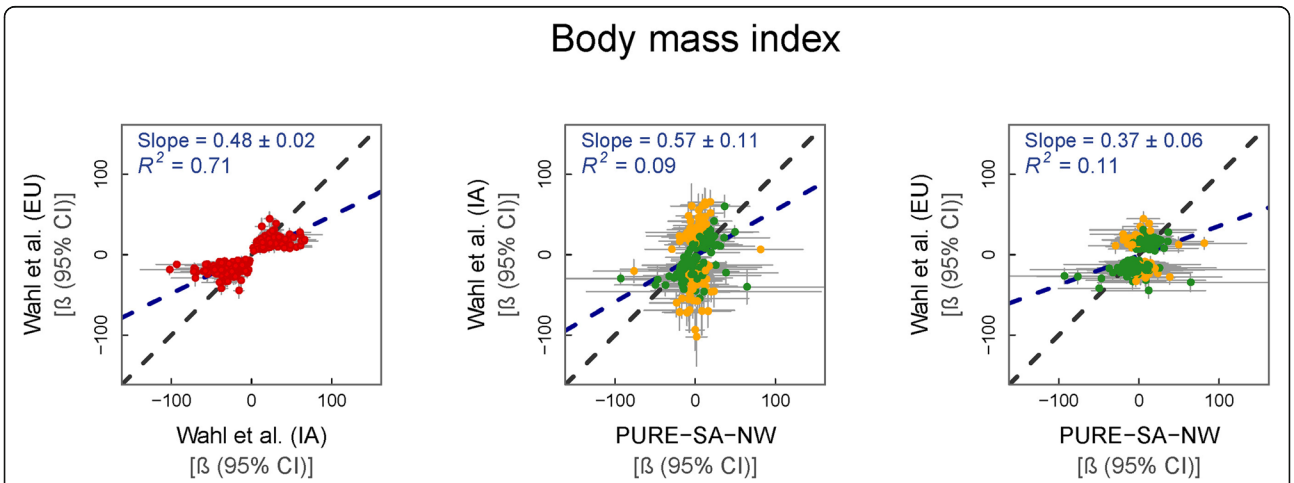


Fig. 3 Change in BMI (kg/m^2) per % methylation change. From left to right: (i) reference EU vs IA data, (ii) IA vs. PURE-SA-NW data and (iii) EU vs. PURE-SA-NW data. Model used for PURE-SA-NW EWAS: BMI ~ methylation + age + smoking status + alcohol consumption + physical activity + cell counts + surrogate variables. Reference data: Wahl et al. (2017). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line

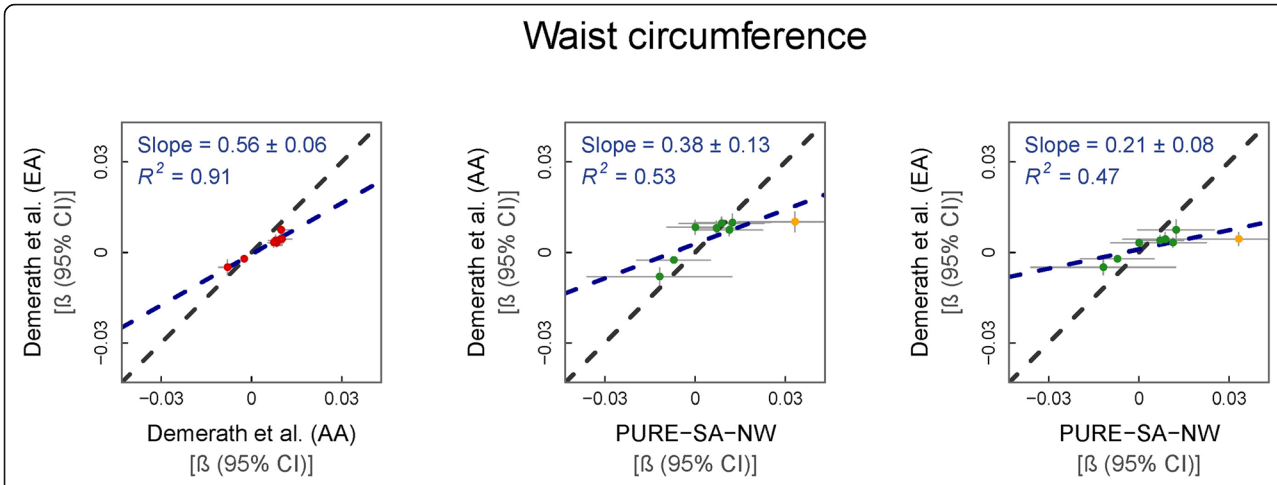


Fig. 4 % Methylation change per centimetre change in WC. From left to right: (i) reference EA vs. AA data, (ii) AA vs. PURE-SA-NW data and (iii) EA vs. PURE-SA-NW data. WC was normalised to have a mean of 0 and a standard deviation of 1. Model used for PURE-SA-NW EWAS: methylation ~ WC + age + alcohol consumption + smoking + physical activity + education + cell counts + surrogate variables. Reference data: Demerath et al. (2015). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line

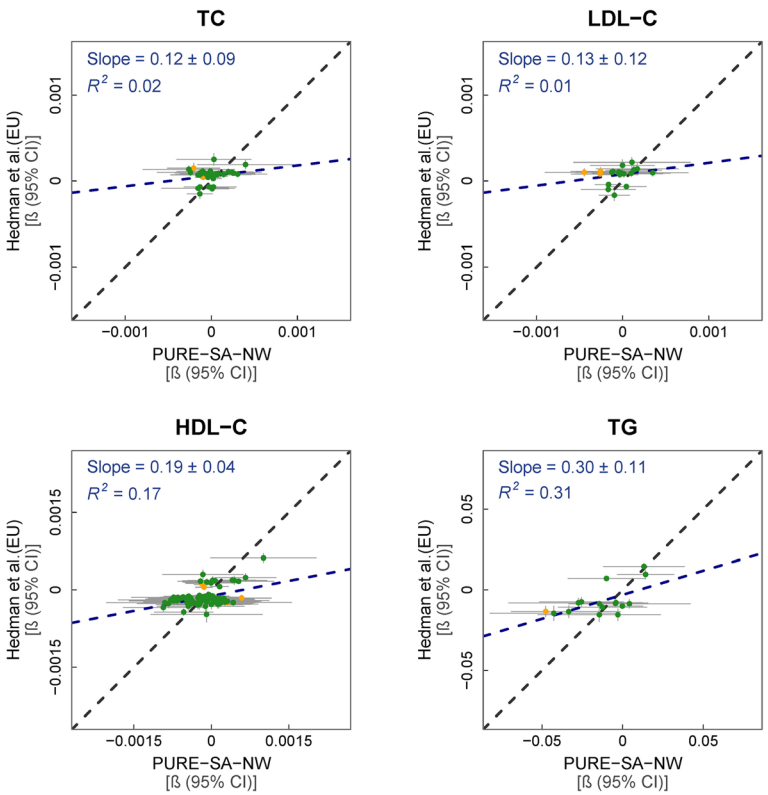


Fig. 5 % Methylation change per mg/dL change in lipid concentration in reference vs. PURE-SA-NW data. Models used: methylation ~ lipid (TC, LDL-C, HDL-C or TG) + age + cells + surrogate variables. Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line

lipid-DNAm association in our cohort (Additional file 2e). High-density lipoprotein cholesterol associated with cg23636606 at a regression β of $2.6 \times 10^{-04} \pm 4.4 \times 10^{-05}$ ($p = 4.8 \times 10^{-08}$).

CRP

Ancestry-stratified (AA and EU) data on the effect of CRP on the DNAm of 207 loci, by Ligthart et al. [22] were compared to PURE-SA-NW. The reference study reported highly comparable effect sizes between the AA and EU ancestral groups (regression slope = 0.82, $p = 1.25 \times 10^{-107}$), with slightly larger effects observed in the AA group. The comparison of the regression slope of effect sizes between the reference data and our own showed moderately larger effect sizes in the PURE-SA-NW findings than in the reference data, more so for the EU (regression slope = 0.25, $p = 2.5 \times 10^{-10}$) than the AA (regression slope = 0.22, $p = 1.3 \times 10^{-10}$) comparison (Additional file 1f). Confidence intervals of the individual effect estimates between the reference and PURE-SA-NW data overlapped for 192 out of the 207 tests (93%) in each ethnicity (Fig. 6). Twenty-two of the 30 non-overlapping associations were directionally consistent. Two CpGs had associations in opposing directions of effects compared to EU (cg01588592 and cg23740758) and three compared to the EU and AA (cg24174557, cg26846781, cg27184903) reference datasets. All the non-overlapping CpGs have *cis*-mQTLs with absolute reported Z-scores ranging from 4.06 to 22.95 [24, 25]. Data from the 1000 Genomes project support the notion that the differences observed here could be partly influenced by ancestry-specific genetic variance: for example, MAF of rs9791189 (*cis*-mQTL for cg23740758) is 12% in Africans and 23% in Europeans [26]. There were no genome-wide significant or novel CRP-DNAm associations in our cohort (Additional file 2f).

Age

Previous findings from EU-based research on the association of age with DNAm of 152 CpGs [19] were compared to those from the PURE-SA-NW cohort (Fig. 7). In contrast to all other traits, a much weaker association between age and DNAm was observed in our data than in the reference data (regression slope = 12.9, $p = 4.2 \times 10^{-31}$). Although the direction of effects was consistently similar between the two studies, none of the regression CIs overlapped when comparing the individual associations (Additional file 1g).

Formal data on disease diagnosis were not available for the PURE-SA-NW cohort and were, therefore, not included in regression models, as done by Florath et al. [19]. Furthermore, cell counts were not adjusted for in the reference study, but were included in our models, since cell counts are recognised confounders in our data. Sensitivity analyses were, however, performed by including data on chronic medication use (as a proxy for disease) as well as excluding cell count adjustments. These analyses did not result in any discernible differences in findings (inclusion of medication use: regression slope = 13.0, $p = 4.5 \times 10^{-30}$, exclusion of cells: regression slope = 12.4, $p = 1.7 \times 10^{-32}$). There were no genome-wide significant or novel age-DNAm associations in our cohort (Additional file 2g).

Discussion

Our primary analysis focussed on the replication of relevant EWAS literature in 120 Batswana men from the

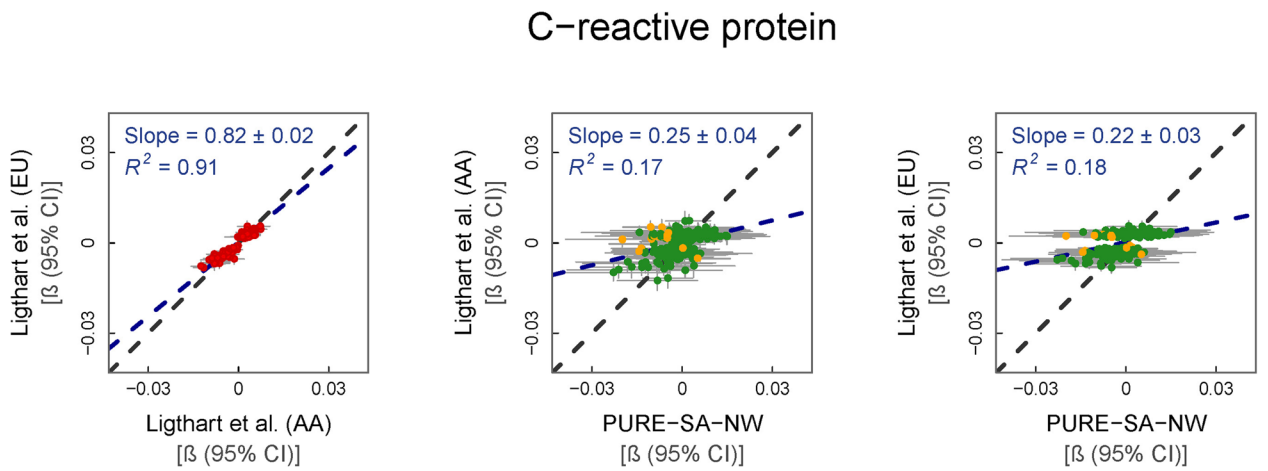


Fig. 6 Change in logarithmic CRP (mg/L) per % methylation change. From left to right: (i) reference EU vs. AA data, (ii) AA vs. PURE-SA-NW data and (iii) EU vs. PURE-SA-NW data. Model used for PURE-SA-NW EWAS: methylation ~ CRP + age + smoking + BMI + cells + surrogate variables. Reference data: Ligthart et al. (2016). Green data points represent CpGs where the 95% CIs for effect size estimates in each sample group overlap. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Red data points represent the comparison of effect sizes within the reference cohorts. Black dashed line: line of equality. Blue dashed line: regression line

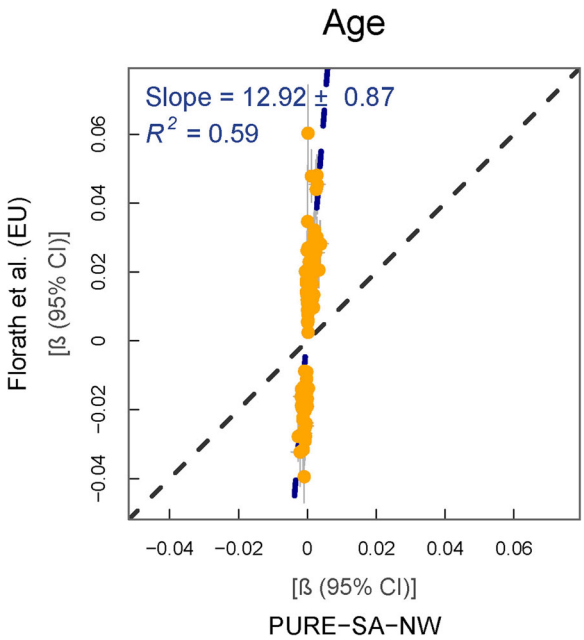


Fig. 7 % Methylation change per year of age in reference vs. PURE-SA-NW data. Model used for PURE-SA-NW EWAS: methylation ~ age + smoking + cell counts + surrogate variables. Yellow data points represent CpGs where the 95% CIs for effect size estimates in each sample group do not overlap. Black dashed line: line of equality. Blue dashed line: regression line

PURE-SA-NW cohort. Secondary analysis included the discovery of novel findings, either investigated for the first time on the EPIC array, or with the 450K probes incorporated in the EPIC array that had not previously been associated with these traits.

Overall, the 95% CI of effect estimates for 86% (4730 out of the 5498 CpG-trait association tests) of the PURE-SA-NW associations overlapped with previously reported findings, and a further 13% (720 out of the 5498 CpG-trait association tests) were directionally uniform. Generally, larger effect sizes were observed in the PURE-SA-NW data than those of the reference studies. Although the reason for differing effect sizes cannot be answered definitively, given the small sample size, the degree of association seems to be related to population-specific differences. Only ~ 1% of our findings (48 out of the 5498 CpG-trait association tests, including 44 unique CpGs) were directionally inconsistent with its compared association reported in the reference study. No data quality concerns were observed for any of these directionally contradicting findings. Of the 44 CpGs, 36 have mQTLs [24, 25] for which population differences in MAFs have been observed by the 1000 genomes project [26].

Overall, these results indicate general consistency in epigenome-wide associations among ethnicities, but ancestry may be important in up to 14% of the tested associations. This is supported by the fact that regardless of the similarity in traits measured among groups, the associations observed in PURE-SA-NW data consistently reflected those reported in AA better than in EU/EA cohorts and better in EU than IA in the case of methylation-BMI associations. Furthermore, eight novel associations between the methylation of 450K array probes, present on the EPIC platform, and alcohol consumption are reported in the Batswana South Africans that were not previously observed in populations of different ancestral origins. These population distinctions indicate the value of ethnic diversity in epigenetic research.

The only trait for which we were unable to replicate any associations was age. Apart from the reference study for age being the smallest of the reference studies included ($N = 498$), there were also clear differences in the pre-processing, data normalisation and EWAS approach followed between PURE-SA-NW and Florath et al. [19]. The reference cohort's analyses were restricted to a pre-selected set of 200 CpGs, the methylation levels of which were normalised using Box–Cox transformations. A mixed regression model with plate and BeadChip as random effects was used. For the PURE-SA-NW data, however, we employed a functional normalisation strategy on the raw methylation data of all the EPIC BeadChip probes, followed by linear regression where surrogate variables were adjusted for as fixed effects to control for possible unaccounted variance. Our findings remained directionally consistent with the reference study's, with the average difference in effect size amounting to 0.87% methylation change per year increase in age (calculated as the percentage difference between the average of the 152 tests' absolute regression βs of the PURE-SA-NW vs. Florath et al [19] results).

In terms of findings related to the EPIC array, 11 genome-wide significant alcohol associations are reported here. An additional eight genome-wide significant alcohol associations were observed for 450K probes present on the EPIC array. Alcohol consumption contributed to a large portion of the variance in the methylation of these probes, as well as, when reversed, the probes to the variance in alcohol consumption. Previous 450K CpG-alcohol associations have been used successfully to identify risky and heavy drinkers [5]. Our sample size did not allow stratification of alcohol intake, but we expect the addition of the alcohol-associated EPIC probes reported here to enhance the discriminatory potential of the current methylation-based biomarker of alcohol consumption [5]. The variance explained by these findings and their usefulness as potential biomarkers warrant replication in large and ethnically diverse cohorts. Larger sample sizes and ethnic diversity will also permit further exploration of the biological basis of these findings and their potential application in NCD-related epigenetic research.

The strengths of this study are the expansion of current data, both by using the EPIC array and investigating a novel study population, after first being able to observe similar findings to those from independent, highly powered, previously replicated literature. The overall consistency between effect sizes is reassuring, not only in terms of the comparability of the PURE-SA-NW data with previous findings, but also the consistency in the effect size and explained variability of novel associations compared to previous EWASs on similar traits [5, 9, 20, 21]. The efficacy of the enhanced coverage of the EPIC array, to uncover new associations with a range of traits, is shown in our study, even with our limited sample size. We motivate the use of this array in future large-scale analyses, as it is likely to add to the variance that can be explained using methylation markers and also to identify novel sites that may be important in prediction, risk stratification or understanding causal disease pathways.

In this study, however, the corresponding limitation to doubling the coverage of the 450K array was the relative loss of statistical power, given our sample size. The lack of power resulted in wide regression CIs for most association estimates that limited our capacity for the fine scale inference of findings. We were able to comment on general patterns and large differences, but we do not know whether more subtle differences between population subgroups exist. Furthermore, the unavailability of genomic data in our cohort and the absence of data on Southern African populations in the 1000 genomes' database restricted our ability to evaluate MAF differences between the reference and Batswana South African groups. We are, therefore, unable to quantify the relative contribution of genetic compared to environmental factors in the associations and association differences observed. The overall congruence in replication results between cohorts—even when large differences in phenotypes are demonstrated—does, however, suggest that these associations might be the result of genetic architecture rather than environmental differences, which we expect to affect the investigated traits as well.

The inclusion of only one sex also limits this study in that no assumptions can be made regarding the generalisability of these results to black South African women. However, because all the reference studies we replicated contained mixed-gender data, there are likely not major differences in these associations between the sexes.

Conclusions

This study reports that up to 86% of the previously reported epigenome-wide associations observed in other ethinicities are present in this black male South African population. While acknowledging the value of ethnic-specific genomic data, our results support the notion that current blood-based 450K EWAS findings can largely be extrapolated to under-represented ethnicities for whom epigenetic data are not yet available. However, the population-specific differences in up to 14% of the CpGs tested, together with the unique associations reported here, do motivate the inclusion of a diversity of ethnic groups in epigenetic association studies. Investigating multi-ethnic data in epigenome-wide studies should be considered the golden standard.

Methods

Study design

This study was performed on a sub-sample of individuals participating in the international PURE study [27]. The PURE study includes sub-cohorts across the world, including one comprising individuals residing in the NW province of South Africa. This sub-cohort represents a single, self-reported ethnicity, Batswana South Africans, who were born and still reside in the NW province of South Africa. Detailed descriptions of the international and PURE-SA-NW cohorts have been published previously [27, 28].

PURE-SA-NW data were collected in 2005, 2010 and 2015. A total of 126 participants were randomly selected for the current investigation, from a group of 990 individuals who took part in the 2015 PURE-SA-NW data collection. Eligibility depended on the following inclusion criteria: availability of bio-samples, testing negative for the human immunodeficiency virus at the time of data collection and male sex. These criteria were incorporated to eliminate confounding by sex and CD4 cell counts in a study with already limited power. The participants included in this study are referred to as the PURE-SA-NW cohort in this manuscript.

Data collection

Height and weight were quantified using a stadiometer and an electronic scale. BMI was calculated as weight per unit height squared (kg/m^2). WC was measured at the appropriate landmarks, by qualified anthropometrists using a steel tape.

An adapted physical activity index questionnaire was used to gather data to calculate a physical activity index [29]. Alcohol intake (g/day) was determined from a quantitative food frequency questionnaire adapted and validated for use in this population [30]. Participants reported the amount, frequency and any relevant description of the alcoholic drinks they had consumed in the preceding month. Data were processed to an amount in g/day, based on the South African food composition tables using FoodFinder3® software (available from http://foodfinder.mrc.ac.za). Smoking status (current, former or never) was self-reported, using a standardised questionnaire. When used as a covariate, smoking and drinking status were dichotomised into *never* and *ever* groups,

with former smokers/drinkers included in the *ever* category. When investigated as the EWAS exposure, smoking status and alcohol consumption were classified according to the classification used in the reference studies.

Fasting blood samples were collected and handled as described previously [31]. High-sensitivity CRP and fasting blood lipids (TC, LDL-C, HDL-C, TG) were quantified using the Cobas® Integra 400 (Roche® Clinical System, Roche Diagnostics, Indianapolis, IN, USA).

DNAm data generation and processing

Whole blood intended for the isolation of genomic DNA was collected in 9 mL Tempus tubes (Applied Biosystems™, Foster city, CA, USA) at the same time as blood used for the quantification of all other phenotypes. Tubes were vortexed for 10 s prior to storage in a − 20 °C freezer for up to 5 days, after which samples were transferred to cryostorage (− 80 °C) until analysis. DNA was isolated using QIAGEN Flexigene DNA extraction kits (QIAGEN® Valencia, CA, USA). The manufacturer's protocol was followed with minor modifications.

Upon extraction, the picoGreen® dsDNA quantitation assay (Invitrogen™, Carlsbad, CA, USA) was used to quantify DNA. Five hundred nanograms DNA from each participant was bisulphite-converted using the Zymo EZ DNAm™ kit (Zymo Research, Irvine, CA, USA), followed by genome-wide DNAm profiling on the Illumina Infinium MethylationEPIC BeadChip according to the manufacturer's protocol (Illumina®, San Diego, CA, USA).

Samples were randomised across slides to minimise the possibility of confounding by batch. Raw signal intensity data were processed from .idat files using functional normalisation as described by the R package *meffil* [32]. The quality threshold for samples and probes was set at 95%. All probes or samples with a detection p value > 0.01 for more than 5% of the evaluated measures were excluded. Six samples were removed on account of low quality: four samples because of a proportion of undetected probes above the quality control (QC) threshold and two with outlying control probes. Probes failing QC were removed prior to data normalisation ($N = 8343$). Eventually 857,516 probes and 120 individuals were included in subsequent data normalisation and analysis. Principal component analysis of the control probes identified 12 principal components to be included in the functional normalisation. In addition, *slide* was specified as a random effect to be included to address batch variance. Sample cell fractions (B cells, CD4 and CD8 T cells, neutrophils, monocytes and natural killer cells) were estimated using the IDOL optimised L-DMR library for whole blood samples [23].

Identification of reference data using the EWAS catalogue

Data we sought to replicate were extracted from the EWAS catalogue (http://www.ewascatalog.org, date of access: 27 April 2019). The EWAS catalogue indexes EWAS studies performed in a study sample of at least 100 individuals for whom at least 100,000 CpGs were available genome-wide. Only associations with $p < 1 \times 10^{-4}$ are included in the catalogue.

Data from the catalogue were pruned according to the following criteria: (i) the EWAS catalogue trait had to be available in the PURE-SA-NW study cohort in a comparable unit; (ii) methylation-trait associations had to be replicated (below a p value threshold of 1×10^{-4}) in at least one independent cohort, regardless of tissue, to reduce the possibility of including false positive findings from among the reference studies; (iii) the DNA had to have been extracted from a blood-based sample; (iv) DNAm had to be reported in Beta units; and (v) an effect estimate (β) and standard error had to be available for each association. Traits that fitted these criteria were age, alcohol consumption, smoking, BMI, WC, CRP, HDL-C, LDL-C, TG and TC. To simplify data analysis, we attempted to replicate results from the largest study indexed by the EWAS catalogue for each investigated trait only. The results reported in each replication subsection make reference to the particular study used for comparison, which would have been the largest EWAS included in the catalogue at the time of writing.

Statistical analysis

Statistical analysis was conducted using R 3.4.3 [33]. The normality of trait data was assessed using Shapiro-Wilks tests. Linear regression models were used to identify epigenome-wide associations using the *meffil* [32] and *ewaff* (https://github.com/perishky/ewaff) packages. DNAm was modelled as a β value between 0 and 1, representing the ratio of methylated to unmethylated probes. The relative contribution of exposures to the variance of outcome variables was determined using the *relaimpo* package's *lmg* metric from the *calc.relimp* function applied to linear models.

For the replication analysis, because of the small sample size of the PURE-SA-NW study population and, therefore, limited power, replication of previously published results focusses on the size and direction of effect sizes rather than comparison of p values. Associations were considered replicable when the 95% CI of the regression β of the reference and the PURE-SA-NW cohort overlapped. Most reference studies extracted from the EWAS catalogue adjust regression models for 'technical variation'. In PURE-SA-NW, surrogate variables were added to all models to reduce any unknown or unmeasured confounding [34]. The *sva* and *generate.confounders* functions within the *meffil* and *ewaff* packages estimated the surrogate variables that were included in each model based on the method described by Leek and Storey [34]. Annotation data were obtained from *meffil*.

For the investigation of novel findings, only associations with $p < 9.4 \times 10^{-8}$ were considered genome-wide significant [15]. Within our cohort, we estimated 80% power to detect a 5% difference in methylation at this threshold for 69% of the EPIC probes, assuming an alpha level of 0.05 and 530,639 independent tests [15]. Packages used in analyses, in addition to those already specified, include *BaseR*, *dplyr*, *FlowSorted.Blood.EPIC*, *ggplot*, *IlluminaHumanMethylationEPICanno.ilm10b2.hg19*, *minfi*, *readxl* and *xlsx*.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-019-0805-z.

Additional file 1. EWAS test statistics (PURE-SA-NW vs reference study) for: (a) alcohol consumption; (b) smoking status; (c) BMI; (d) WC; (e) lipids; (f) CRP and (g) age.

Additional file 2. Epigenome-wide associations with $p < 1 \times 10^{-4}$ in the PURE-SA-NW study for: (a) alcohol consumption; (b) smoking status; (c) BMI; (d) WC; (e) lipids; (f) CRP and (g) age.

Abbreviations

AA: African American; BIOS: Biobank-based Integrative Omics Studies; BMI: Body mass index; Chr: Chromosome; CpGs: Cytosine-phosphate-guanine sites; CRP: C-reactive protein; DNAm: DNA methylation; EA: European American; EU: European; EWAS: Epigenome-wide association study; HDL-C: High-density lipoprotein cholesterol; IA: Indian Asian; IQR: Inter-quartile range; LDL-C: Low-density lipoprotein cholesterol; MAF: Minor allele frequency; mQTLs: Methylation quantitative trait loci; NCD: Non-communicable disease; PURE-SA-NW: South Africa, North-West arm of the Prospective Urban and Rural Epidemiology study; QC: Quality control; TC: Total cholesterol; TG: Triglycerides; UTR: Untranslated region

Acknowledgements

The authors would like to thank all those that participated in the PURE-SA-NW study and those that made the PURE-SA-NW study possible, including the fieldworkers, researchers and staff of both the PURE-SA-NW (Africa Unit for Transdisciplinary Health Research (AUTHeR), Faculty of Health Sciences, NWU, Potchefstroom, South Africa) and PURE International (S Yusuf and the PURE project office staff at the Population Health Research Institute (PHRI), Hamilton Health Sciences and McMaster University, Ontario, Canada) teams. We thank the staff of the Bristol bioresource laboratories (Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK) who generated the DNAm data.

Authors' contributions

HTC isolated and curated the DNA, performed all statistical analysis and wrote the original draft. HRE oversaw the EPIC analysis, conceptualised the manuscript, supervised statistical analysis and critically reviewed and edited the manuscript. CN-R critically reviewed and edited the manuscript. MP is the principal investigator, acquired the funding for this project and critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

Financial support for the PURE-SA-NW study was provided by the North-West University, South African National Research Foundation (SANRF), Population Health Research Institute, South African Medical Research Council (SAMRC), the North West Province Health Department, and the South African Netherlands Partnerships in Development. Grants from the SANRF, Academy of Medical Sciences UK (Newton Fund Advanced Fellowship Grant) and the SAMRC supported the additional epigenetic work reported herein. HTC is supported by a PhD scholarship from the SANRF (SFH106264); HRE works in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/5). None of the funding bodies were involved in the design of the study, collection, analysis or interpretation of the data or the writing of this manuscript. Opinions expressed and conclusions arrived at are those of the authors and are not to be attributed to the funding sources.

Availability of data and materials

The data that support the findings of this study are available upon reasonable request and with the permission of the Health Research Ethics Committee of the North-West University and the principal investigator of the PURE-SA-NW study, Prof. I.M. Kruger (lanthe.kruger@nwu.ac.za) at the North-West University's Africa Unit for Transdisciplinary Health Research.

Ethics approval and consent to participate

Ethical approval for the 2015 data collection of the PURE-SA-NW study was granted by the Health Research Ethics Committee of the North-West University (NWU-00016-10-A1, NWU-00119-17-A1). All participants provided written informed consent, including consent for genetic/epigenetic analysis. All procedures described were performed in accordance with the revised version of the Helsinki Declaration of 1975 [35].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Centre of Excellence for Nutrition at the North-West University Potchefstroom Campus, Potchefstroom 2520, South Africa. [2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK. [3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, UK.

Received: 18 September 2019 Accepted: 30 December 2019
Published online: 07 January 2020

References

1. Rodger EJ, Chatterjee A. The epigenomic basis of common diseases. Clin Epigenetics. 2017;9:5.
2. Sharp GC, Relton CL. Epigenetics and noncommunicable diseases. Epigenomics. 2017;9:789–91.
3. Hedman AK, Mendelson MM, Marioni RE, Gustafsson S, Joehanes R, Irvin MR, Zhi D, Sandling JK, Yao C, Liu C, Liang L, Huan T, McRae AF, Demissie S, Shah S, Starr JM, Cupples LA, Deloukas P, Spector TD, Sundstrom J, Krauss RM, Arnett DK, Deary IJ, Lind L, Levy D, Ingelsson E. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. Circ Cardiovasc Genet. 2017;10:e001487.
4. Joehanes R, Just A, Marioni R, Pilling L, Reynolds L, Mandaviya P, Guan W, Xu T, Elks C, Aslibekyan S. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436–47.
5. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, Just AC, Duan Q, Boer CG, Tanaka T, Elks CE, Aslibekyan S, Brody JA, Kuhnel B, Herder C, Almli LM, Zhi D, Wang Y, Huan T, Yao C, Mendelson MM, Joehanes R, Liang L, Love SA, Guan W, Shah S, AF MR, Kretschmer A, Prokisch H, Strauch K, Peters A, Visscher PM, Wray NR, Guo X, Wiggins KL, Smith AK, Binder EB, Ressler KJ, Irvin MR, Absher DM, Hernandez D, Ferrucci L, Bandinelli S, Lohman K, Ding J, Trevisi L, Gustafsson S, Sandling JH, Stolk L, Uitterlinden AG, Yet I, Castillo-Fernandez JE, Spector TD, Schwartz JD, Vokonas P, Lind L, Li Y, Fornage M, Arnett DK, Wareham NJ, Sotoodehnia N, Ong KK, van Meurs JBJ, Conneely KN, Baccarelli AA, Deary IJ, Bell JT, North KE, Liu Y, Waldenberger M, London SJ, Ingelsson E, Levy D. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23:422–33.
6. Mendelson MM, Marioni RE, Joehanes R, Liu C, Hedman AK, Aslibekyan S, Demerath EW, Guan W, Zhi D, Yao C, Huan T, Willinger C, Chen B, Courchesne P, Multhaup M, Irvin MR, Cohain A, Schadt EE, Grove ML, Bressler J, North K, Sundstrom J, Gustafsson S, Shah S, AF MR, Harris SE, Gibson J, Redmond P, Corley J, Murphy L, Starr JM, Kleinbrink E, Lipovich L, Visscher PM, Wray NR, Krauss RM, Fallin D, Feinberg A, Absher DM, Fornage

M, Pankow JS, Lind L, Fox C, Ingelsson E, Arnett DK, Boerwinkle E, Liang L, Levy D, Deary IJ. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a mendelian randomization approach. PLoS Med. 2017;14:e1002215.
7. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, Volkova N, Schramm K, Carstensen-Kirberg M, Waldenberger M, Gieger C, Peters A, Illig T, Prokisch H, Roden M, Grallert H. Association between DNA methylation in whole blood and measures of glucose metabolism: Kora f4 study. PLoS One. 2016;11:e0152314.
8. World Health Organization. Noncommunicable diseases. 2018. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed 9 Aug 2019.
9. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, Hedman AK, Sandling JK, Li LA, Irvin MR, Zhi D, Deloukas P, Liang L, Liu C, Bressler J, Spector TD, North K, Li Y, Absher DM, Levy D, Arnett DK, Fornage M, Pankow JS, Boerwinkle E. Epigenome-wide association study (ewas) of bmi, bmi change and waist circumference in african american adults identifies multiple replicated loci. Hum Mol Genet. 2015;24:4464–79.
10. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O. The genetic structure and history of africans and african americans. Science. 2009;324:1035–44.
11. Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, Wahl S, Elliott HR, Rota F, Scott WR. Epigenome-wide association of DNA methylation markers in peripheral blood from indian asians and europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol. 2015;3:526–34.
12. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, Davey Smith G, Hughes AD, Chaturvedi N, Relton CL. Differences in smoking associated DNA methylation patterns in south asians and europeans. Clin Epigenetics. 2014;6:4.
13. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP. Accounting for population stratification in DNA methylation studies. Genet Epidemiol. 2014;38:231–41.
14. Teo Y, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in africa. Nat Rev Genet. 2010;11:149–60.
15. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, Hannon E. Guidance for DNA methylation studies: statistical insights from the illumina epic array. BMC Genomics. 2019;20:366.
16. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45:e22.
17. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationEPIC beadchip. Genomics Data. 2016;9:22–4.
18. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.
19. Florath I, Butterbach K, Muller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated cpg sites. Hum Mol Genet. 2014;23:1186–201.
20. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC, Ried JS, Zhang W, Yang Y. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6.
21. Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, Sha J, Pankow JS, Liu C, Irvin MR, Fornage M, Hidalgo B, Lin LA, Thibeault KS, Bressler J, Tsai MY, Grove ML, Hopkins PN, Boerwinkle E, Borecki IB, Ordovas JM, Levy D, Tiwari HK, Absher DM, Arnett DK. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity. 2015;23:1493–501.
22. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, Colicino E, Waite LL, Joehanes R, Guan W, Brody JA, Elks C, Marioni R, Jhun MA, Agha G, Bressler J, Ward-Caviness CK, Chen BH, Huan T, Bakulski K, Salfati EL, Fiorito G, Wahl S, Schramm K, Sha J, Hernandez DG, Just AC, Smith JA, Sotoodehnia N, Pilling LC, Pankow JS, Tsao PS, Liu C, Zhao W, Guarrera S, Michopoulos VJ, Smith AK, Peters MJ, Melzer D, Vokonas P, Fornage M, Prokisch H, Bis JC, Chu AY, Herder C, Grallert H, Yao C, Shah S, AF MR, Lin H, Horvath S, Fallin D, Hofman A, Wareham NJ, Wiggins KL, Feinberg AP, Starr JM, Visscher PM, Murabito JM, Kardia SL, Absher DM, Binder EB, Singleton AB, Bandinelli S, Peters A, Waldenberger M, Matullo G, Schwartz JD, Demerath EW, Uitterlinden AG, van Meurs JB, Franco OH, Chen YI, Levy D, Turner ST, Deary IJ, Ressler KJ, Dupuis J, Ferrucci L, Ong KK, Assimes TL, Boerwinkle E, Koenig W, Arnett DK, Baccarelli AA, Benjamin EJ, Dehghan A. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016;17:255.
23. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.
24. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, Slieker RC, Jhamai PM, Verbiest M, HED S, Verkerk M, van der Breggen R, van Rooij J, Lakenberg N, Arindrarto W, Kielbasa SM, Jonkers I, van 't Hof P, Nooren I, Beekman M, Deelen J, van Heemst D, Zhernakova A, Tigchelaar EF, Swertz MA, Hofman A, Uitterlinden AG, Pool R, van Dongen J, Hottenga JJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, van den Berg LH, van Zwet EW, Mei H, Li Y, Lemire M, Hudson TJ, the BC, Slagboom PE, Wijmenga C, Veldink JH, van Greevenbroek MMJ, van Duijn CM, Boomsma DI, Isaacs A, Jansen R, van Meurs JBJ, t Hoen PAC, Franke L, Heijmans BT. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics. 2016;49:131.
25. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van 't Hof P, Mei H, van Dijk F, Westra H-J, Bonder MJ, van Rooij J, Verkerk M, Jhamai PM, Moed M, Kielbasa SM, Bot J, Nooren I, Pool R, van Dongen J, Hottenga JJ, Stehouwer CDA, van der Kallen CJH, Schalkwijk CG, Zhernakova A, Li Y, Tigchelaar EF, de Klein N, Beekman M, Deelen J, van Heemst D, van den Berg LH, Hofman A, Uitterlinden AG, van Greevenbroek MMJ, Veldink JH, Boomsma DI, van Duijn CM, Wijmenga C, Slagboom PE, Swertz MA, Isaacs A, van Meurs JBJ, Jansen R, Heijmans BT, t Hoen PAC, Franke L. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics. 2016;49:139.
26. Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68.
27. Teo K, Chow CK, Vaz M, Rangarajan S, Yusuf S. The prospective urban rural epidemiology (pure) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries. Am. Heart J. 2009;158:1–7. e1
28. De Lange Z, Pieters M, Jerling JC, Kruger A, Rijken DC. Plasma clot lysis time and its association with cardiovascular risk factors in black africans. PLoS One. 2012;7:e48881.
29. Nienaber-Rousseau C, Sotunde OF, Ukegbu PO, Myburgh PH, Wright HH, Havemann-Nel L, Moss SJ, Kruger IM, Kruger HS. Socio-demographic and lifestyle factors predict 5-year changes in adiposity among a group of black south african adults. Int J Environ Res Public Health. 2017;14:1089.
30. MacIntyre UE, Venter CS, Vorster HH. A culture-sensitive quantitative food frequency questionnaire used in an african population: 1. Development and reproducibility. Public Health Nutr. 2001;4:53–62.
31. Pieters M, Kotze RC, Jerling JC, Kruger A, Ariens RA. Evidence that fibrinogen gamma' regulates plasma clot structure and lysis and relationship to cardiovascular risk factors in black africans. Blood. 2013;121:3254–60.
32. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;34:3983–9.
33. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017. URL https://www.R-project.org/
34. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:1724–35.
35. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310:2191–4.

Publisher's Note