Differential genome-wide DNA methylation in prostate tumours from South African men

by

Jenna Craddock

Submitted in partial fulfilment of the requirements for the degree
Master of Science in Human Genetics
In the Faculty of Health Sciences
University of Pretoria

Student number: 15011357

October 2021

Supervisor: Prof Vanessa M. Hayes v.hayes@garvan.org.au
Co-Supervisor: Prof M.S. Riana Bornman riana.bornman@up.ac.za

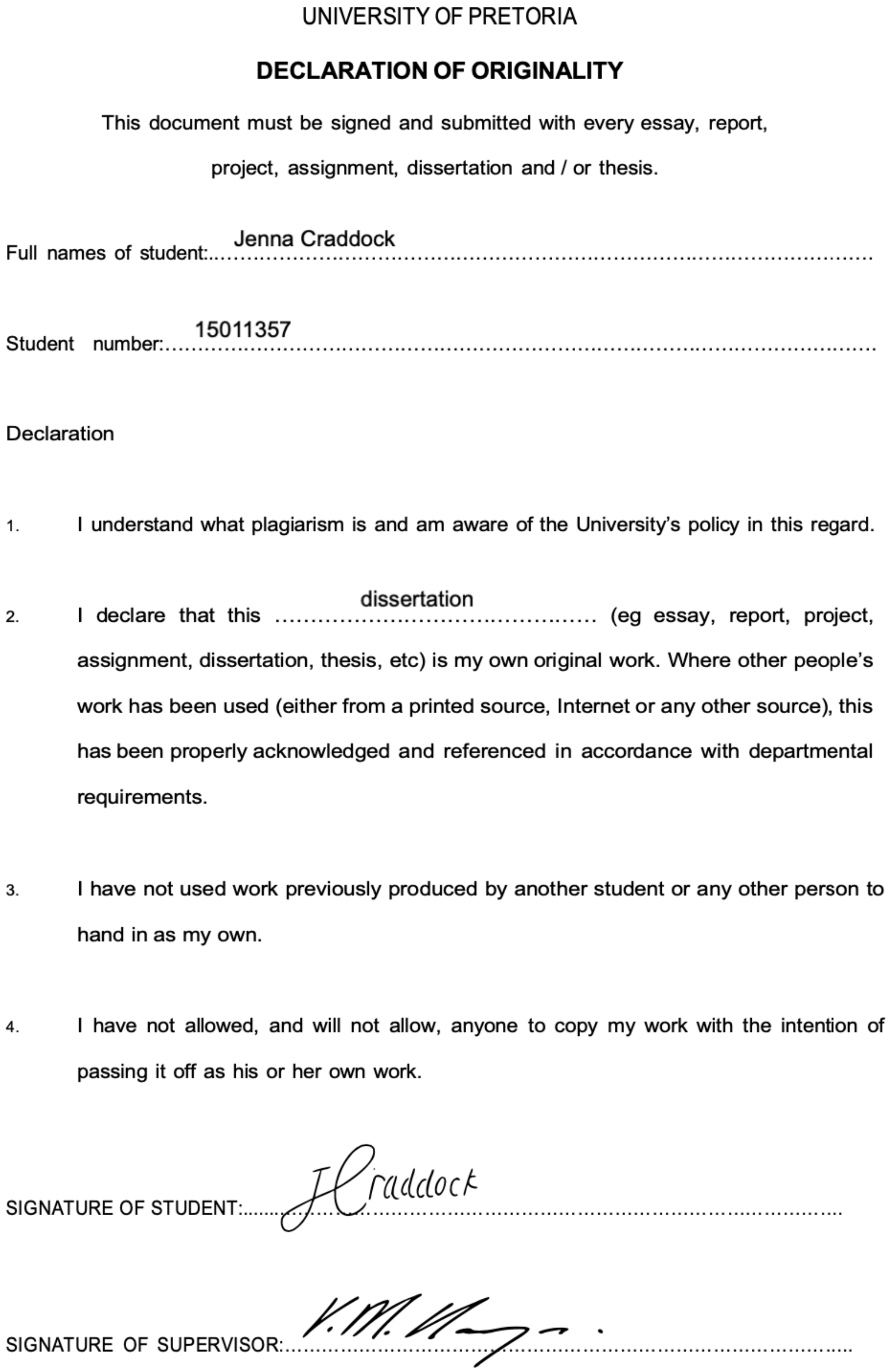
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Submission Declaration

I, **Jenna Craddock**, hereby declare that this dissertation, submitted in partial fulfilment of the requirements for the degree **Master of Science in Human Genetics**, in the Faculty of Health Sciences, at the University of Pretoria, is entirely my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Signed: J Craddock Date: 7 February 2022

Plagiarism Declaration



UNIVERSITY OF PRETORIA

DECLARATION OF ORIGINALITY

This document must be signed and submitted with every essay, report, project, assignment, dissertation and / or thesis.

Full names of student:........ Jenna Craddock ....

Student number:........ 15011357 ....

Declaration

1. I understand what plagiarism is and am aware of the University's policy in this regard.

2. I declare that this .... dissertation .... (eg essay, report, project, assignment, dissertation, thesis, etc) is my own original work. Where other people's work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with departmental requirements.

3. I have not used work previously produced by another student or any other person to hand in as my own.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

SIGNATURE OF STUDENT:........ J Craddock ....

SIGNATURE OF SUPERVISOR:........ V.M. ....

Acknowledgements

Throughout my MSc journey and the writing of this dissertation, I have received a great deal of support and assistance worth acknowledging.

I would first like to express my deep gratitude for my supervisors, Professor Vanessa Hayes and Professor Riana Bornman, for their unwavering support, guidance and countless hours eagerly invested in my project and academic career. Professor Hayes' research conceptualisations and insightful feedback continually pushed me to step out of my academic comfort zone, to broaden my knowledge and to sharpen my approach to scientific research. Additionally, I will forever be appreciative of the great personal support I received from Professor Bornman during challenging times. I have these two women to thank for every educational opportunity and enlightenment, every meaningful encouragement and every academic self-confidence boost that I experienced during my MSc pursuit. I admire them on both a personal and professional level.

I would like to thank the Human Comparative and Prostate Cancer Genomics Research team at the Garvan Institute of Medical Research, for openly welcoming me as their visiting student and from whom I have learnt so much during our interactions. A special thanks to Dr Weerachai Jaratlerdsiri and Jue Jiang for readily providing data and bioinformatic input.

A thank you to Dr Pavlo Lutsik, head of the Computational Cancer Epigenomics Group at Deutsches Krebsforschungszentrum, in Germany, for bioinformatic critical review. I would also like to thank Professor Clare Stirzaker, Dr Tim Peters, Dr Ruth Pidsley and Dr Elena Zotenko, of the Epigenetic Deregulation in Cancer Group at the Garvan Institute, for bioinformatic resources. A thanks to Professor Liza Bornman.

On a more personal note, I would like to extend a warm thanks to my family, whose unending support and encouragement motivates me to work hard and persevere, especially on days that prove challenging. I thank my parents for their wise advice and for always being there for me. I credit my mom for inspiring my curiosity, for having unfaltering faith in my capabilities and for always encouraging my ambitions and desired accomplishments. I would also like to extend a thanks to my siblings and friends, particularly Kirsten Channer, for being welcomed distractions allowing me to rest my mind outside of my research.

Finally, I would like to thank the University of Pretoria for the funding granted that made this academic pursuit a possibility as well as a reality for me. I am proud to be part of an institution that so determinedly cares for their students.

Executive Summary

Background: DNA methylation is an epigenetic mechanism known to aid the progression of cancer, including prostate cancer. It is part of a cluster of molecular processes that initiate tumorigenesis and drive its early evolution by altering other molecular processes. While studies have looked at DNA methylation in prostate cancer, most have been limited by targeted gene analysis, with further bias towards non-African cohorts. Considering the enhanced coverage of more recent genome-wide arrays, such as the Illumina Infinium HumanMethylationEPIC BeadChip, which measures DNA methylation over more than 850,000 CpG sites genome-wide, many studies that have employed a more global approach to DNA methylation analysis are further limited by frequently utilising lower-coverage arrays. Due to the bias against African cohorts, African-relevant bioinformatic tools for the processing of African DNA methylation data, particularly generated by the EPIC array, are scarce. As a result, the genomic mechanisms that underlie African prostate cancer as well as the contribution of DNA methylation alterations to African prostate cancer are poorly understood.

Results: Working with EPIC DNA methylation data, I present a novel established African-relevant genome-wide bioinformatic pipeline for the processing and normalisation of African tumour-derived genome-wide DNA methylation data. Pilot application of this pipeline on prostate tissue identified differentially methylated CpG dinucleotides that may contribute to aggressive prostate cancer in a small cohort of men of South African ancestry. Additionally, I identified top genes in South African prostate cancer that are significantly enriched for differentially methylated CpG sites. Finally, patient-matched genomic-epigenomic data integration revealed preliminary evidence for interplay between these two systems in African prostate cancer, although the identification of DNA methylation signatures would prove more insightful.

Conclusions: Ultimately, this work highlights the marginalization of Africans in scientific research. As a preliminary solution to this underrepresentation, this dissertation provides a novel toolset to appropriately handle African DNA methylation data with the ultimate goal of generating a deeper understanding of the genomic mechanisms harboured within African prostate cancer, a field with limited knowledge. Potential improvements to this tool, complications encountered when interpreting epigenome-wide results as well as the near future of cancer genomics is discussed.

Keywords: Prostate cancer, African, DNA methylation, epigenomics, EPIC, ethnic disparity

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
AGRF	Australian Genome Research Facility
AIC	Akaike information criterion
BH	Benjamini & Hochberg
BMIQ	Beta-mixture quantile
bp	Base pair
BPH	Benign prostatic hyperplasia
BPM	Bizagi process modeler
C	Cytosine
CGI	CpG island
ChAMP	Chip Analysis Methylation Pipeline for Illumina
chr	Chromosome
COSMIC	Catalogue Of Somatic Mutations In Cancer
CPU	Central processing unit
CSV	Comma-separated values
DBS	Doublet-base-substitution
dbSNP	The NCBI Short Genetic Variation database
DMP	Differentially methylated probe
DMR	Differentially methylated region
ENCODE	The Encyclopedia of DNA Elements
EPIC	Illumina Infinium HumanMethylationEPIC BeadChip
EWAS	Epigenome-wide association studies
ExonBnd	Exon boundary
FANTOM	Functional Annotation of the Mammalian Genome
FASTA	Fast-all
G	Guanine
GATK	Genome analysis toolkit

GCO	Global Cancer Observatory
GEO	Gene Expression Omnibus
GRCh37	Genome Reference Consortium Human genome build 37
GRCh38	Genome Reference Consortium Human genome build 38
GUI	Graphical user interface
HCC	Hepatocellular carcinoma
HCPCG	Human Comparative and Prostate Cancer Genomics
hg19	Genome Build 37
hg38	Genome Build 38
HPC	High-performance compute
HREC	Human Research Ethics Committee (University of Pretoria)
HRPCa	High-risk prostate cancer
ID	Insertion-and-deletion
IDAT	Intensity data
IGR	Intergenic region
IGSR	International Genome Sample Resource
Indels	Insertions and deletions
JDK™	Java™ Platform, Standard Edition Development Kit
kb	Thousand base pairs
KCCG	Kinghorn Centre for Clinical Genomics
M	Methylated probe signal intensity
MAF	Minor allele frequency
MANTIS	Microsatellite Analysis for Normal Tumor InStability
Mbp	Million base pair
MDS	Multidimensional scaling
meC	Methylated cytosine
MSI	Microsatellite instability
MSI-H	Microsatellite instability-high
MSS	Microsatellite stability

MTA	Material Transfer Agreement
NCBI	National Center for Biotechnology Information
NCI	National Compute Infrastructure
NHMRC	(Australian) National Health and Medical Research Council
NIH	National Institutes of Health
PBC	Peak-based correction
PCa	Prostate cancer
pd file	Phenotypes file
PGA	Percentage of genome alteration
PSA	Prostate-specific antigen
QC	Quality control
RefSeq	NCBI Reference Sequences database
SAPCS	Southern African Prostate Cancer Study
SBS	Single-base substitution
SCNA	Somatic copy number aberration
SD	Standard deviation
SIH	Sydney Informatics Hub
SNP	Single nucleotide polymorphism (present in > 1 % of a population)
SNV	Single nucleotide variant (present in < 1 % of a population)
SV	Structural variant
SVA	Surrogate variable analysis
SVD	Singular value decomposition
SVH	St. Vincent’s Hospital
SWAN	Subset-quantile within array normalisation
T	Thymine
TCGA	The Cancer Genome Atlas
TMB	Tumour mutational burden
TSS	Transcription start site

U	Unmethylated probe signal intensity
UCSC	The University of California, Santa Cruz
UTR	Untranslated region
VCF	Variant call format
WGBS	Whole-genome bisulfite sequencing
WGS	Whole-genome sequencing
27K array	Illumina Infinium HumanMethylation27 BeadChip
450K array	Illumina Infinium HumanMethylation450 BeadChip
850K array	Illumina Infinium HumanMethylationEPIC BeadChip

Chapter 1: General Introduction

Please note that throughout this dissertation, the mention of "South African" men makes specific reference to men of African descent.

1.1. Background and research problem

Continental Africans are significantly underrepresented in terms of genomic and epigenomic research despite knowledge of ethnic-related differences.[1] As a result, insight is limited regarding the factors that link numerous diseases to African ancestry. Of particular interest to my dissertation are the contributing factors that link prostate cancer to African ancestry. Prostate cancer is the most common urological cancer affecting aging men in South Africa.[2] South African men present with more aggressive disease and display higher incidence and mortality rates compared to their European counterparts.[3,4] In some cases, this presentation is true even in comparison to African Americans.[5] However, in light of the African marginalization just mentioned, it is unsurprising that the mechanisms that underlie African prostate tumorigenesis remain poorly understood. This poses a challenge for cancer pharmacologists and clinicians. To better understand the ethnic bias and aggressive nature of African prostate cancer, one cannot ignore the possible contribution of genomic factors related to African ancestry. In addition, it is crucial that such consideration not disregard the contribution of environmental factors which calls into question epigenomic mutational processes, given that the environment is capable of directly, epigenetically inferring disease susceptibility, including cancer.[6,7] Of course, epigenomic mutational processes may arise in response to intrinsic factors and are additionally known to influence genomic events, thereby alluding to a complex genomic-epigenomic interplay.[8,9]

DNA methylation has a well-established role in influencing cancer progression and this includes prostate cancer.[10] While studies have looked at DNA methylation in prostate cancer, most have been limited by targeted gene analysis, with further bias towards non-African cohorts. Presumably as a result of this bias, the availability of bioinformatic tools for DNA methylation data processing that accounts for South African cohorts at the necessary steps are scarce, if not absent. Ultimately, South African men are overlooked in terms of epigenetic prostate cancer research and appropriate African-relevant bioinformatic tools.[11] The subsequent limited knowledge on the genomics and epigenomics that underlie African prostate cancer limits disease screening, diagnostics and treatment. Considering (i) the lack of publicly-available African prostate cancer DNA methylation data (both targeted and genome-wide), (ii) the lack of published research on this topic, and (iii) prostate cancer incidence and mortality expected to rise over time, it is more apparent now than ever before that African-associated prostate cancer receive adequate scientific research attention.

1.2. Hypothesis, aim and specific objectives

I aimed to assess differential genome-wide DNA methylation for numerous variables, in the prostate tumours from South African men. My dissertation is built on the hypothesis that epigenetic alterations, such as DNA methylation, may (at least in part) explain the differences observed in prostate cancer pathogenesis between ethnicities. Additionally, I expect interplay between the prostate cancer genome and epigenome. To investigate this, I selected the Illumina Infinium HumanMethylationEPIC BeadChip (further detail presented in **Chapter 2**) as the DNA methylation data source used within the scope of this dissertation owing to its high genome-wide coverage.[12] To ensure suitable data processing, I chose ChAMP[13] (see **Chapter 2**) as the framework on which to develop a novel African-relevant bioinformatic workflow due to its comprehensive, user-friendly nature and its support of Illumina EPIC data. Performing analyses on prostate tissue from a cohort of South African men, my objectives were to establish a bioinformatic pipeline to interrogate African tumour-derived genome-wide DNA methylation (**Chapter 3**), to apply this novel pipeline to investigate DNA methylation in prostate tissue from men of African ancestry (**Chapter 4**) and finally, to integrate and analyse patient-matched prostate cancer genomic and epigenomic data (**Chapter 5**).

1.3. Rationale

The research presented in the following Chapters provides researchers with a novel bioinformatic African-relevant toolset to appropriately handle and analyse African biopsy-derived DNA methylation data. The applicability of this tool is not limited to prostate cancer; on the contrary, it is suitable for any African tissue-derived DNA methylation data. In the context of my dissertation, subsequent use of such a tool will allow for novel insights to be gained in the field of African prostate cancer genomics with the potential of answering clinically-relevant questions. Finally, pilot integration of matched genomic and epigenomic data provides preliminary evidence for the interaction of these two systems in African prostate cancer, which may fuel further investigation.

Although this dissertation is not presented in a publication format as I do not intend to submit the forthcoming research for publication, please note that **Chapters 3**, **4** & **5** below were written as though independent papers, each addressing a single research objective. As such, there is some repetition in the methods that were replicated from one Chapter to another.

1.4. References

1. Peedicayil J. Population pharmacoepigenomics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 511–7.
2. Alleyne G, Binagwaho A, Haines A, Jahan S, Nugent R, Rojhani A, et al. Embedding non-communicable diseases in the post-2015 development agenda. Lancet. 2013; 381(9866):566–74.
3. Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, et al. Whole-genome sequencing reveals elevated tumor mutational burden and initiating driver mutations in African men with treatment-naïve, high-risk prostate cancer. Cancer Res. 2018; 78(24):6736–46.
4. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71(3):209–49.
5. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014; 74(8):880–91.
6. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007; 8(4):253–62.
7. Arita A, Costa M. Environmental agents and epigenetics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 459–76.
8. Riley LB, Anderson DW. Cancer epigenetics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 521–34.
9. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009; 5(8):e1000602.
10. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet. 2020; 52(8):778–89.
11. Bishop OT, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, Ghouila A, et al. Bioinformatics education - perspectives and challenges out of Africa. Brief Bioinform. 2015; 16(2):355–64.
12. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016; 17(1):208.
13. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017; 33(24):3982–4.

Chapter 2: Literature Review

2.1. Overview of prostate cancer and epigenetics

Prostate cancer (PCa) is the leading cancer diagnosed in men in the developed world, and is the second most common cancer in men worldwide, following lung cancer.[1] However, it is true that an ethnic bias exists against the African population, whose ancestry is suggested to be a significant risk factor[2] contributing to their populations' elevated incidence and mortality rates where PCa is concerned and when compared to their counterparts of European or Asian descent.[1,3–5] In addition to this, PCa is often diagnosed at a younger age in African individuals, displays a higher mutational burden[6] and presents itself more aggressively (Gleason score ≥ 8), with the latter explaining the elevated mortality risk.[7] It is well-established that older age, family history and African ancestry are non-modifiable risk factors for PCa but lifestyle factors, such as diet, obesity and physical activity cannot alone account for ethnic differences in PCa risk[2], suggesting African ancestry to have an integrated genomic and epigenomic basis to explain the disparity.

Currently, a field of great interest in disease and genome evolution is epigenetics, which aims to shed light on gene-environment interactions. Because genetics cannot solely account for disease susceptibility and development, epigenetics has made clear that molecular factors and processes exist around DNA, which are able to regulate genome activity independent of DNA sequence and are mitotically stable.[8] These molecular factors include DNA methylation, histone modifications, non-coding RNAs, chromatin structure and RNA methylation[9], and have well-documented roles in various diseases, including cancer, when epigenetic regulation is disrupted.[10] When the epigenetic marks these molecular factors leave are aberrant and heritable (i.e. in the germline), they are called epimutations.[11] More specifically, epimutations refer to altered epigenetic marks at specific DNA sites that result in response to an environmental factor[12], such as toxicants[13], nutrition and stress.[12] These epimutations are capable of altering genome activity, including gene expression[9,12], thereby having the potential to introduce disease susceptibility. Conversely, should developing somatic tissue be directly exposed to an environmental toxicant, the somatic genome, subsequent cell signalling and resultant phenotype will be altered in that individual[14], also introducing disease susceptibility. Epimutations and epigenetic alterations differ from genetic variations in that genetic variations are permanent alterations in the DNA sequence[15], whereas epigenetic modifications are reversible.[16,17] While epigenetic changes are able to activate, silence or influence gene expression, genetic variations are able to do this as well as change protein structure and function; either alteration has the ability to introduce vulnerability to disease.

DNA methylation is a common epigenetic signalling tool that cells use to influence gene expression. Typically, DNA methylation is most commonly associated with downregulation of gene expression. As a normal and vital component of cell processing, DNA methylation is involved in embryonic development, genomic imprinting, X-chromosome inactivation and chromosome stability preservation.[18] However, as previously mentioned, epigenetic regulation may go awry, resulting in aberrant methylation which is characteristic of numerous diseases. Commonly, cytosine residues that lie within a CpG site are subject to methylation, resulting in 5-methylcytosine (5-mC).[19] However, although CpG dinucleotides remain the primary site for DNA methylation in mammals, it is also true that cytosines may be methylated in other contexts, such as within CH or CHG sites (where H may be A/T/C)[20,21], referred to as non-CpG methylation. Methylated cytosine residues are subject to spontaneous deamination to a thymine residue, resulting in a mC > T point mutation. This displays a direct link between an epigenetic modification and a more permanent sequence change. For this reason, methylated cytosine residues and by extension, CpG sites, are considered mutational hotspots in germline and somatic cells.[19,22,23] The high mutability of CpG dinucleotides not only drives its own genetic variation, but studies have shown that methylation plays a role in increasing the mutability of neighbouring nucleotides.[24] This is demonstrated by findings of methylated CpG sites having ~1.5 times more SNPs (single nucleotide polymorphisms) around them (±10 bp) compared to unmethylated CpG's.[25] Additionally, a recent study showed C > T variations at CpG sites and T > C variations to be common in the germline of individuals, while also noting that these very same mutational signatures are known to generate somatic variations.[22] This led authors to suggest that these mutational signatures operating in the germline underlie those in somatic cells. Ultimately, there appears to be a link between DNA methylation and SNP prevalence, at least under normal conditions; and therefore, it may be reasonable to suggest that in the presence of aberrant methylation, one may expect an altered or even higher incidence of variants under disease conditions. Additionally, these alterations underlie a known link between germline and somatic cells, suggesting DNA methylation at CpG sites play a significant role in both inherited and acquired disease susceptibility (in response to an environmental factor) as well as in genome evolution.

A common and well-known feature of human cancer, and specifically PCa, is the epigenetic silencing of cancer-associated genes.[26] Typically, these genes undergo hypermethylation of CpG islands (CGIs) in their promoter regions, resulting in a partial or complete block of gene expression.[27] This is another mechanism for achieving gene silencing besides gene mutation or deletion and occasionally hypomethylation renders the same effect.[28] Additionally, even non-CpG methylation (i.e. CH or CHG sites) has been found in various tumorigenic contexts, including PCa[29,30], although it is unclear whether non-CpG methylation contributes to or is a consequence of cancer. For CGI regions, the CpG island refers to regions of the genome, usually

300-3,000 bp in length, that contain a large number of CpG dinucleotide repeats[31], providing numerous opportunities for aberrant methylation since these are the sites typically subject to methylation. The CGI shores lie 2,000 bp upstream and downstream from the nearest boundary of the CGIs (**Fig. 2-1**). Beyond the shores, a further 2,000 bp upstream and downstream from the nearest shore boundary, lie the CGI shelves. Regions beyond the shelves are referred to as the open sea, or inter-CGI.

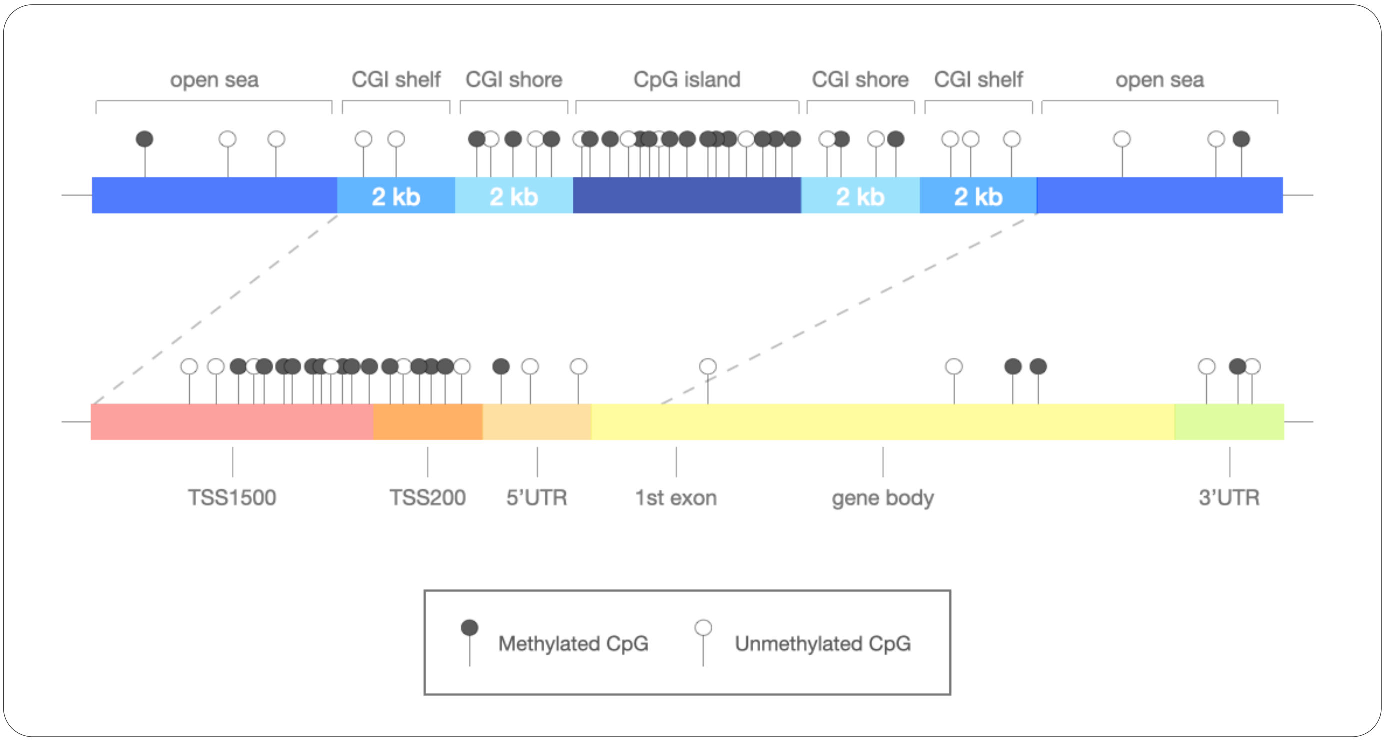


Fig. 2-1 Illustration of a CpG island (CGI) with surrounding CGI shores, CGI shelves and open sea regions. Where CGI regions typically map to gene regions is also depicted. CpG density and CpG methylation is most abundant within CGIs. Increasing distance from the border of CGIs is associated with decreasing CpG density and CpG methylation within CGI shores and CGI shelves.

Adapted from "DNA methylation/hydroxymethylation in melanoma," by Fu et al., 2017, *Oncotarget*, *8(44)*, p. 78164. Copyright 2017 by The Authors.
CGI: CpG island | TSS: transcription start site | UTR: untranslated region

Methylation changes occur predominantly early in cancer development and are also believed to occur in non-malignant cells contiguous with cancerous tissue, leading to a field effect (aka field cancerization).[26,33] Another early event in PCa is the *TMPRSS2:ERG* gene fusion, in which an individual's fusion status (i.e. positive/negative) has been shown to be associated with changes in DNA methylation.[34–36] Intriguingly, *TMPRSS2:ERG* gene fusions have been reported to be less common in prostate tumours derived from men of African ancestry from South Africa (13 %)[37], compared with men of European (49 %) or Asian (27 %) ancestry, and half that observed for African American men (25 %).[38] Additionally, a large body of evidence has shown that a number of cancer-associated genes are not only hypermethylated in PCa, but are methylated to an even higher degree in African American PCa compared to patient-matched tumours from men of European ancestry.[39–41] This suggests that these epigenetic alterations may (at least partly) explain the differences observed in PCa pathogenesis between ethnicities.

It is clear that PCa, as with any cancer, is both a genetic and an epigenetic disease. However, the mere activation of an oncogene and accompanying inactivation of a tumour suppressor gene, whether by genetic or epigenetic mechanisms or both, does not account for the full spectrum of alterations responsible for carcinogenesis. In reality, cancer is the result of a complex network of dysregulated (epi)genomic interactions. Thus, to begin to truly understand the depth underlying aggressive African-associated PCa, a deeper insight into African-relevant epigenomic variation is necessary, a matter of limited understanding, in particular at the genome-wide level.

One such method, whole-genome bisulfite sequencing (WGBS), whereby input DNA is treated with sodium bisulfite and sequenced, allows for high-resolution, genome-wide measurement of DNA methylation (see **Section 2.2.** for more detail). However, although comprehensive, WGBS is plagued by several limitations. Currently, the NIH Roadmap Epigenomics Project recommends the use of two replicates with a combined total coverage of 30x. This requires approximately 800 million aligned, high quality reads for human samples, rendering WGBS cost prohibitive, particularly for large scale studies.[42] However, even at this recommended 30x coverage, up to 50 % of informative methylation data may be lost[43] and as such, the use of WGBS for high-resolution feature analysis (e.g. differentially methylated probes) is limited. Although a number of data recovery methods are available, advertising up to 12 % data rescue[43], it remains that multiple replicates are necessary for accurate feature identification, reiterating WGBS's unsuitability for large scale studies. In solution to these limitations, a number of array-based platforms for genome-wide DNA methylation analysis are available, with popular technologies offered by Illumina.

2.2. Genome-wide methylation array technology

The Illumina Infinium HumanMethylation BeadChip arrays are based on a popular genome-wide CpG methylation profiling technology that is commonly used in large-scale population-based methylation studies.[44] These studies may be based on thousands of human individuals owing to the arrays' comprehensive coverage and high throughput. The earliest of this technology is the Illumina Infinium HumanMethylation27 BeadChip (27K array), which measures the methylation status of 27,578 CpG sites across the human genome at single nucleotide resolution.[45] Following the 27K array was the release of the Illumina Infinium HumanMethylation450 BeadChip (450K array), which offers an even higher genome-wide coverage, assessing methylation levels at 485,577 individual CpG sites.[46,47] These CpG sites span all chromosomes; cover CGIs, shores and shelves; as well as genomic regions including transcription start sites (TSSs), gene bodies, first exons and 3'/5' untranslated regions (UTRs) of 99 % RefSeq (NCBI Reference Sequences database) genes. However, considering that the human genome harbours approximately 28 million individual CpG sites[48], the 27K and 450K arrays can hardly be considered truly genome-wide CpG methylation profiling technologies due to their low coverage. The 450K array only accounts for about 1.7 % of all CpGs in the human genome. To address this limitation, the more recent Illumina Infinium HumanMethylationEPIC BeadChip (EPIC/850K array) was introduced, which boasts a much wider genome-wide coverage of 863,904 CpG sites and 2,932 CNG sites on important regulatory regions.[49] These regions include FANTOM5 (Functional Annotation of the Mammalian Genome) enhancers, ENCODE (The Encyclopedia of DNA Elements) open chromatin and enhancers, DNase hypersensitivity sites and miRNA promoter regions that previously were not captured by the 450K array. In fact, the EPIC array covers more than 90 % of the CpG sites covered by the 450K array.[49] The final probes contained in the EPIC array includes 59 probes targeting SNP sites to allow for sample matching and 636 probes for sample-dependent and sample-independent quality control, totalling to 866,836 EPIC probes.

The Illumina Infinium arrays are based on bisulfite conversion, whereby unmethylated cytosine bases are converted to uracil bases (read as thymine bases after PCR) and methylated cytosine bases remain unconverted. As such, WGBS enables identification of methylated cytosine bases at single base-pair resolution. As per Illumina's protocol, the bisulfite converted DNA is subject to whole genome amplification, enzymatic end-point fragmentation, precipitation and resuspension before hybridizing to the array.[47] The methylation level at each CpG on the array is then measured using one of two probe types, namely Infinium type I and Infinium type II probes. Each of the two probe types have different designs with different hybridization chemistries[49] and as a result, they display different beta-value distributions (beta-values discussed in more detail below). The purpose of having two types of probes is to ensure the full spectrum of DNA methylation is captured; the two probe types also offer complementary strengths.[50]

On the EPIC array, Infinium type I probes measure methylation at approximately 16 % of the CpGs and Infinium type II probes cover the remaining 84 % of the CpGs.[44] Each probe is designed to hybridize a 50 bp DNA sequence, downstream of the target CpG site. Infinium type I probes use two probes (beads) per CpG site, one corresponding to the methylated allele and the other to the unmethylated allele (**Fig. 2-2a**).[44,49] The methylated probe sequence is designed to match the bisulfite-converted DNA sequence of the methylated locus; the methylated probe has a G (guanine) at its 3' end which will bind to a C (cytosine) at

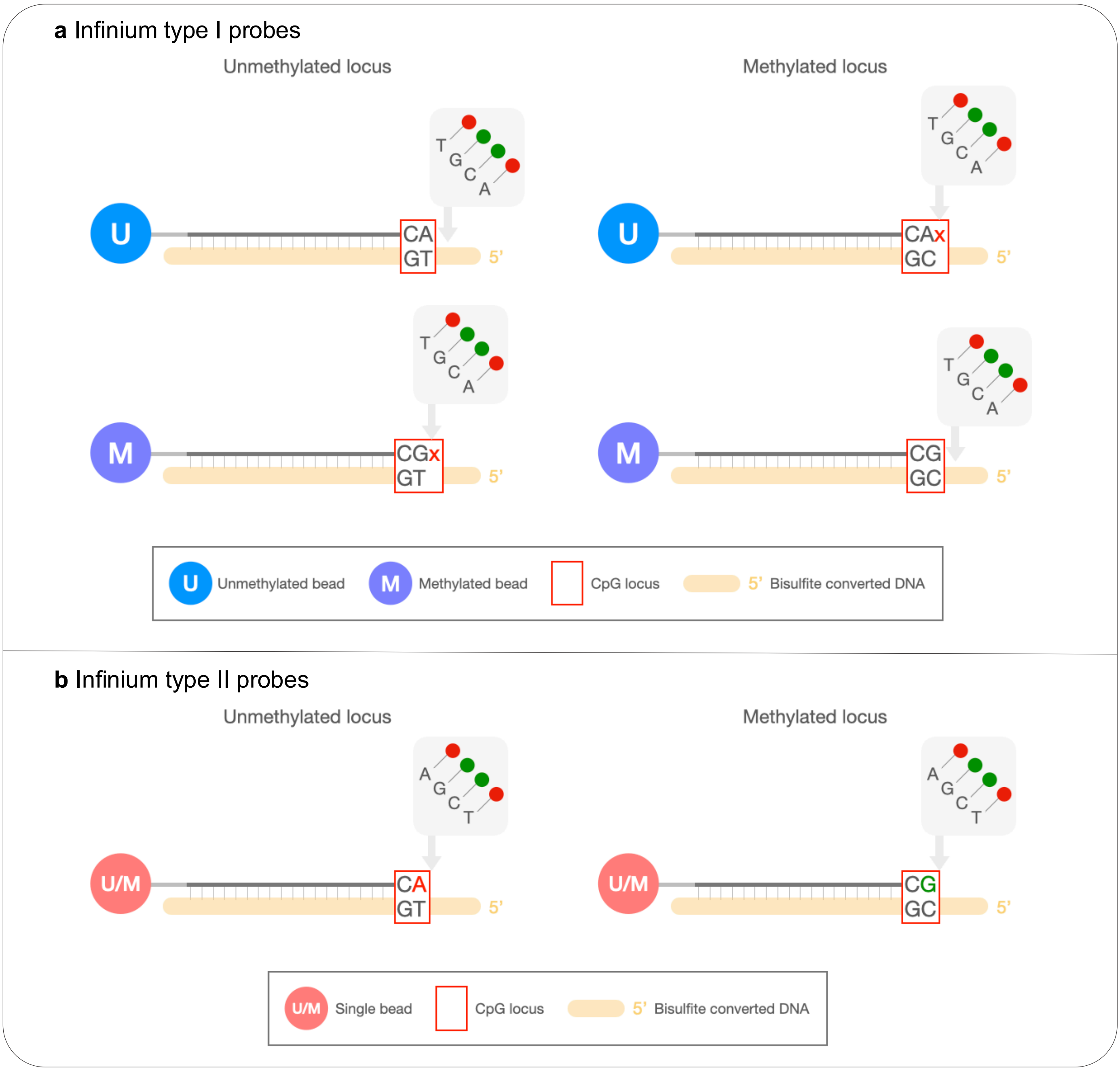


Fig. 2-2 Illumina Infinium methylation probe design. **a** The Infinium type I assay uses two bead types per CpG locus, one for the unmethylated (U) state and one for the methylated (M) state. **b** The Infinium type II assay uses a single bead type per CpG locus, which detects both unmethylated and methylated states (U/M). The state of methylation is determined after hybridization, at the single base extension step.

Adapted from technical note, "Illumina Methylation BeadChips Achieve Breadth of Coverage Using 2 Infinium® Chemistries," by Illumina, Inc., 2015, retrieved November 20, 2020, from https://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf. Copyright 2015 by Illumina, Inc.

a methylated locus. On the other hand, the unmethylated probe sequence is designed to match the bisulfite-converted DNA sequence of the unmethylated locus; the unmethylated probe has an A (adenine) at its 3' end which will bind to a T (thymine) at an unmethylated locus. Other CpG sites bound by the 50 bp probe are assumed to share the methylation status of the target CpG. Once the probe has bound to a bisulfite-converted DNA fragment, incorporation of a single labelled nucleotide is enabled at the probe's 3' end. This labelled nucleotide matches the nucleotide immediately upstream of the target CpG site and this single base extension event allows the signal detection of a methylated or unmethylated site. Should a methylated probe hybridize an unmethylated locus (or vice versa), mismatch at the 3' end of the probe would occur, thereby inhibiting single base extension.

Infinium type II probes use a single probe (bead) per CpG site, and use different dye colours (red/green) to differentiate methylated alleles from unmethylated alleles (**Fig. 2-2b**).[44,49] In this case, the probe sequence is designed to match the bisulfite-converted DNA sequence of both the methylated locus and the unmethylated locus. To achieve this, the cytosine of the target CpG site is made to act as the single base extension site. Therefore, in Infinium type II probes, cytosines of all other CpG sites within the probe sequence are replaced with degenerate R bases. As such, these probes may hybridize to both T (representing unmethylated and converted cytosine) and C (representing methylated and protected cytosine) bases. Probe hybridization to a bisulfite-converted DNA fragment enables single base extension and incorporation of a single labelled nucleotide. Should a green-labelled G be incorporated (opposite a methylated and protected C), signal detection is on the green (methylated) channel. Conversely, should a red-labelled A be incorporated (opposite an unmethylated and converted C i.e. T), signal detection is on the red (unmethylated) channel. Because Infinium type II probes make use of a single bead type, it increases the capacity for the number of CpG sites that can be queried and so, they are applied whenever possible.[50]

2.3. Normalising genome-wide methylation data

As previously mentioned, the use of two different probes types results in different beta-value distributions. This technical variability needs to be corrected for by normalising the data.[44] Normalisation is used to reduce the variability that exists between Infinium type I and Infinium type II probe designs; in other words, it makes the beta distributions of the two probe types comparable thereby preventing a decrease in data quality. Oftentimes, this is essential for downstream data analysis. For example, region-based analyses assume that probes within a shared region are comparable, which is only true if probe bias has been corrected for.[51] Alternatively, when performing clustering, variability that exists between the two probe types may drive the clustering rather than variability contributed by a factor of interest. Typically, the Infinium type II probes are normalised to the Infinium type I probes, owing to the fact that type II probes are considered to be less reproducible and less sensitive than type I probes.[52] Therefore, by extension, normalisation of Infinium type II probes to Infinium type I probes improves reproducibility.

There are a number of normalisation methods that may be applied to Illumina methylation array data, namely, but not limited to, the beta-mixture quantile (BMIQ)[53], subset-quantile within-array normalisation (SWAN)[51], peak-based correction (PBC)[52] and FunctionalNormalization[54] methods. PBC was the first correction method proposed for adjusting probe type bias and performs correction by rescaling the methylation values of Infinium type II probes to the same bimodal distribution of that for Infinium type I probes.[52] However, this method is sensitive to the shape of beta-value density curves, making it less robust when the methylation density distribution does not exhibit well-defined peaks.[53] To address this limitation, the more recent SWAN and BMIQ normalisation methods were proposed.

For subset-quantile within-array normalisation, a subset of biologically similar probes (based on similarities in CpG content) are used to define an average quantile distribution which is then used to normalise Infinium type I and type II probes together.[51] Similarly, the BMIQ method makes use of quantiles to normalise the Infinium type II probe values into a distribution that is comparable to the Infinium type I probes by fitting a beta-mixture model.[53] However, the BMIQ method differs from the SWAN method in that it does not depend on biological characteristics in order to normalise the data, making it the favourable choice for correcting probe type bias.[55,56]

2.4. Quantifying methylation

Quantifying methylation of a particular CpG site involves calculating the beta-value, which is the raw methylation level at each CpG site. The beta-value is the ratio of the methylated probe intensity and the overall intensity[44], and is defined as:

$$\beta = \frac{M}{(M + U + 100)} \tag{1}$$

where M is the intensity measured by the methylated probe and U is the intensity measured by the unmethylated probe. The alpha value of 100 stabilizes the beta-values when the intensities are low. The beta-value approximately represents the percentage of cells for which that particular CpG is methylated and it falls on a spectrum between 0 and 1 (or 0 and 100 %). Under ideal conditions, a value of zero would indicate that all copies of the CpG site in the sample were completely unmethylated, and a value of one would indicate that all copies of the CpG site in the sample were methylated. Of course, such extremes rarely occur, in which case a beta-value ≤ 0.2 may be defined as hypomethylation, a beta-value ≥ 0.8 may be defined as hypermethylation and beta-values intervening ($0.2 < \beta < 0.8$), particularly those ~0.5, represent sites that are partially methylated.[57]

An alternative method to using beta-values is the M-value method for quantifying DNA methylation. Although beta-values are more widely used and is the recommended method by Illumina[58], M-values offer a number of benefits for the differential analysis of methylation levels. Firstly, the M-value method displays approximate homoscedasticity for highly methylated and unmethylated CpG sites.[57] This is in contrast to the beta-value method which displays quite the opposite; heteroscedasticity violates a number of assumptions for various statistical tests, rendering beta-values inappropriate for several statistical analyses e.g. violation of the Gaussian distribution assumption for *t*-tests.[57] It is also true that M-values perform better than beta-values in terms of detection rate and true positive rate for both highly methylated and unmethylated CpG sites.[57] Overall, beta-values are generally preferred when modelling underlying biological effects because these values have a direct biological interpretation i.e. the beta-value approximately represents the percentage of cells for which that particular CpG is methylated, as mentioned above. This is not true for M-values although M-values are more statistically valid in differential and other statistical analyses owing to their approximate homoscedasticity. Ultimately, the two statistics each have their own benefits and limitations and are interconvertible[57] thus the more appropriate one may be chosen where applicable. However, the debate on whether or not to transform beta-values is on-going[44] and it has been shown that whether the data has been transformed or not, does not seriously affect analysis results.[59]

2.5. Bioinformatic tools for processing and normalising genome-wide methylation data

There are a number of bioinformatic tools available for the processing and normalisation of Illumina Infinium DNA methylation array data. **Table 2-1** provides a very brief overview of some of the current packages and pipelines currently available although many more exist. From the table, it is evident that these tools are somewhat new and that their use is limited to a very particular investigative field. The majority of these tools are compatible with Unix/Linux, Mac OS and Windows systems and many are conveniently implemented within the R statistical environment, oftentimes available through Bioconductor. While some tools provide isolated functions (e.g. *minfi*[60] may be used for data preprocessing and differentially methylated region (DMR) analysis), other tools offer full workflows for data preprocessing, differentially methylated probe (DMP) and DMR analysis, data visualisation throughout the workflow and gene ontology and pathway analysis. Such an extensive workflow is offered by the Bioconductor ChAMP (Chip Analysis Methylation Pipeline for Illumina) pipeline.[61] An advantage of ChAMP is the tool's integration of a number of existing analysis methods (such as *minfi*) to make up a comprehensive workflow, and whose outputs may be saved and incorporated with other pipelines. However, a downfall of many of these tools is their complex usage which poses a challenge to researchers without proficient programming skills. An additional limitation of these tools is their oversight of African-relevance. The latter is particularly evident when addressing polymorphisms at methylation probe sites (discussed below). Bioinformatic tools like ChAMP do allow for users to specify the population with which they're working, subsequently accounting for population polymorphism differences. However, the list of populations from which to choose are limited to western and eastern Africans with no reference available for southern Africans. As such, there is a pressing need for the development of a southern African-relevant DNA methylation data processing and analysis tool or for the tailoring of an existing workflow to render it southern African-relevant. Establishing such a workflow is critical for the analysis of DNA methylation data derived from southern African cancer cohorts.

Evidently, numerous tools exist for the processing and normalisation of Illumina DNA methylation data, each possessing their own benefits and limitations. However, there is no standardized approach, particularly when it pertains to African-relevant studies. With the wide use of DNA methylation data for the exploration of associations between DNA methylation and complex diseases, there is an urgency for more efficient and population-appropriate tools for processing Illumina DNA methylation array data.

2.6. Cause for methylation error: polymorphisms

It is broadly accepted that the presence of variants affects the performance of the Illumina Infinium arrays and that they influence results, so should be considered during filtering.[44,62–65] If a SNP is present at or near the target CpG site, the methylation value may actually capture the profile of the SNP rather than that of the CpG methylation; and SNPs that lie closer to the target CpG site are more influential.[64] Typically, SNPs at methylation probe sites display a characteristic "methylation" pattern of three discreet levels (modes) of methylation that correspond to underlying SNP genotype frequencies rather than actual methylation.[64,65] For example, full methylation would correspond to a methylated CC genotype, partial methylation would correspond to a methylated CT genotype and no methylation would correspond to a TT genotype; the beta-values of these genotypes would fall into three separate levels when plotted on a continuous scale of 0-1 (**Fig. 2-3**). This tri-modal distribution of beta-values differs from that of a polymorphism-free site, which would display beta-values within a narrow range or across a continuum, lacking those distinct tiers. Other SNP "methylation" patterns that have been observed are bi-modal beta-value distributions and cloud-like beta-value distributions[64], the latter of which shows no clear correlation between beta-values and SNP genotypes. Although it is true that SNPs located in the interrogated CpG site (both in the first and second position of the target CpG) have a stronger potential to influence DNA methylation quantification at those particular sites compared to SNPs that lie within the body of the probe, it must be noted that these probe body SNPs are able to affect the stability of a probe's hybridization and extension efficiency.[64] In addition to this, while Illumina cautions against retaining SNPs within 10 bp of the interrogated CpG site, it has also been shown that the effect of a SNP within a methylation probe site is present and evident throughout the entire length of the 50 bp probe.[66] Therefore, it is essential to consider the impact of SNPs on beta-value quantification and interpretation whether they lie within the target CpG, the single base extension site (for Infinium type I probes) or within the body of the probe.

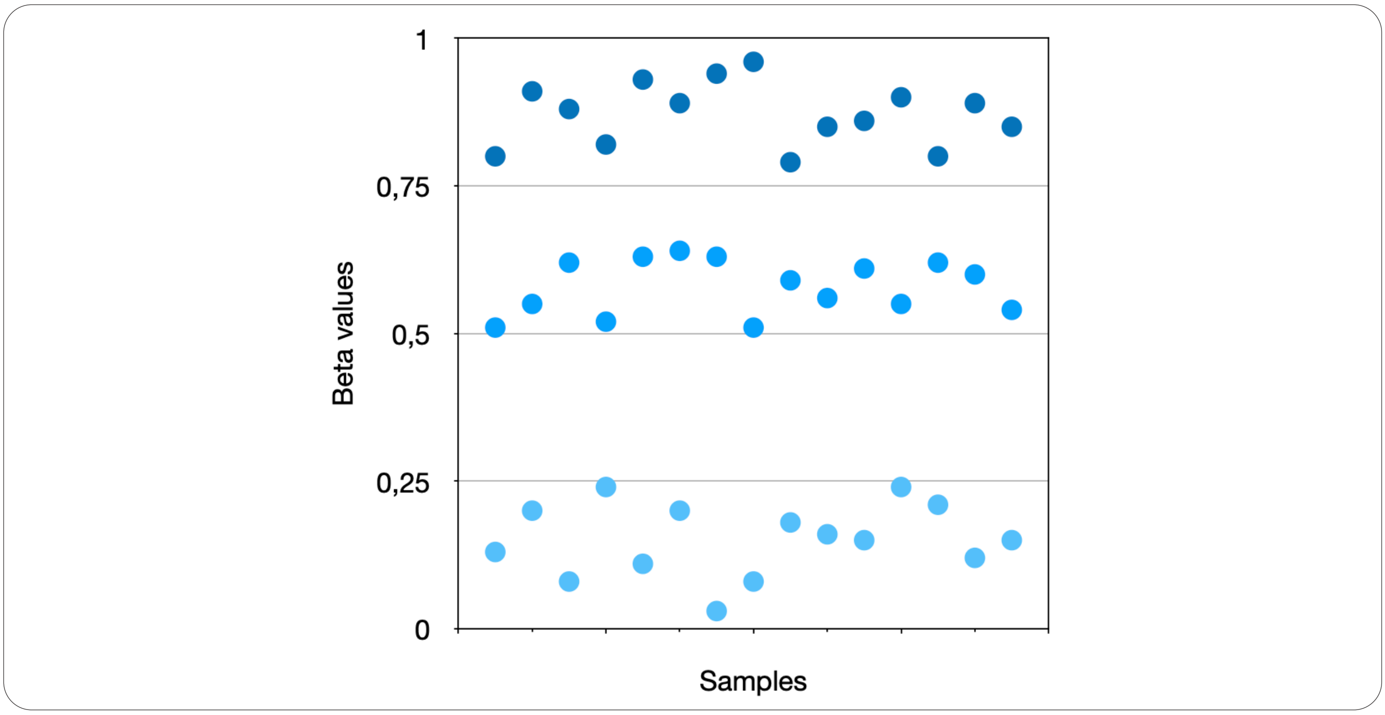


Fig. 2-3 Simulated DNA methylation beta-values at a single CpG site containing a SNP, plotted across 15 simulated samples. The presence of a SNP clearly distributes the methylation data into three discreet levels which correspond with the underlying SNP genotypes.

Adapted from "MethylToSNP: identifying SNPs in Illumina DNA methylation array data," by LaBarre et al., 2019, *Epigenetics & Chromatin*, *12(1)*, p. 5. Copyright 2019 by The Authors.

When performing SNP-affected probe filtering, it is recommended that the SNP reference be similar in ethnicity and population genetic structure to the study population.[44] Generalized references include the 1000 Genomes Project data[67] and dbSNP[68]; lists of recommended probes for filtering have also been annotated for the Illumina EPIC array.[62,69] However, while these references account for numerous populations, the fact remains that they are largely European-relevant and therefore, unsuitable for use in an

Table 2-1 A number of bioinformatic tools available for Illumina BeadChip data processing and analysis.

Tool	*Description*	*Analysis platform*	*Supported Illumina microarray data*	*Limitations*	*Year published*	*Number of Citations**	*Ref*
Illumina GenomeStudio Methylation Module	Illumina software that supports the analysis of Infinium methylation array data.	Illumina GenomeStudio software	27K, 450K, EPIC	During data loading, IDAT files are converted to plain-text files, causing large data loss.		1993	
minfi	A Bioconductor package that provides tools for analysing and visualizing Illumina methylation array data.	R	450K, EPIC (with some modifications)	Can only control for cell type heterogeneity on whole blood data.	2014	2059	[60]
				Cannot handle large datasets on a personal computer.			
RnBeads	A comprehensive package that can analyse single-CpG resolution DNA methylation data.	R	27K, 450K, EPIC	Does not offer pathway analysis.	2014	503	[73]
ChAMP	A pipeline that integrates existing algorithms and functions from numerous sources for microarray data processing and analysis.	R	450K, EPIC	Can only control for cell type heterogeneity on whole blood data.	2017	193	[61]
Ewastools	A pipeline for DNA methylation analysis using the Infinium HumanMethylation BeadChip.	Galaxy	27K, 450K, EPIC	SNP-affected probe filtering conducted with minfi package is European-specific.	2020	0	[71]
MADA	Methylation array data analysis. A web-based service for analysing DNA methylation array data.	MADA web server	450K, EPIC	Public web services have limited operations on throughput.	2020	1	[72]

*According to Google Scholar, August 2021.

African cohort due to potential excessive, unnecessary data loss. This is particularly true considering that a number of polymorphisms are population-specific.[70] Additionally, this approach to SNP-affected probe filtering is not standardized, resulting in published lists that agree on 89 % of problematic probes but differ on another 11 % (considering lists published by Zhou et al. 2017[62] & McCartney et al. 2016[69]). Conversely, by combining DNA methylation data with genomic data, it would be possible to remove sequence variants that coincide with methylated positions. However, oftentimes genomic data is not available to accompany epigenetic studies due to extensively high costs, in which case it can be difficult to distinguish true methylation patterns at a CpG site from the presence of a SNP at that very site. The importance of this lies in accurately distinguishing epigenetic effects that are independent of genetic effects; this is particularly important in distinct populations.

To address this issue, a fairly recent R Bioconductor package, MethylToSNP[63], was introduced by LaBarre et al. (2019) which detects polymorphisms at methylation probe sites in Illumina 450K and EPIC array data, as well as in Illumina 27K array data. MethylToSNP is based on a method called "gap hunting", in which methylation data is parsed to flag locations with characteristic clustered distributions of data points that suggest there to be potential problems in the underlying data. In doing so, MethylToSNP detects methylation data generated specifically at C or T SNP positions in the Illumina Infinium methylation arrays. One alternative approach is to remove all probe locations that are known to harbour human genetic variants (e.g. annotation using dbSNP), in which case polymorphisms that may not even be present in the sequences of the studied individuals are flagged for overlapping probe removal, resulting in unnecessary exclusion of a large amount of methylation data, some of which may have been informative to one's research question. In contrast, MethylToSNP only considers SNPs specific to the individuals in the study thereby preventing extensive data loss. Additional benefits include the potential identification of novel variants in underrepresented populations and the annotation of identified variants in functional regions. LaBarre et al. (2019)[63] further highlight the need for SNP-affected probe filtering, in that considering the principle behind bisulfite conversion (i.e. an unmethylated C is converted to a T), common C > T polymorphisms may be misinterpreted as differential DNA methylation between individuals. For this reason, the majority of sites that the MethylToSNP algorithm detects are (methylated) meC > T SNPs, whereby a reliability threshold ≥ 0.5 is indicative of a true SNP.[63] However, this bias towards meC > T SNPs likely creates an overall underestimation of SNP-affected probes within a dataset because other types of SNPs are overlooked. Indeed, LaBarre et al. (2019)[63] confirm that if a C is always unmethylated, this SNP's location will go unpredicted, with the reason being that it will never appear as differential methylation. Additionally, MethylToSNP is not able to identify SNP-associated patterns other than the three-tier pattern[64], further illustrating the underestimation in SNP-affected probes that is suggested by this tool. A final limitation is the recommendation of a minimum of 50 samples to accurately rely on results generated by MethylToSNP. Due to the high costs for generating Illumina Infinium DNA methylation array data as well as limited sample numbers, especially when considering understudied populations, this sample recommendation may render this tool inappropriate for smaller epigenetic studies. While MethylToSNP is a promising tool for SNP-affected probe identification, improvements are necessary going forward.

Ultimately, based on the discussion above, it is clear that a standardized approach to SNP-affected probe filtering is needed.

2.7. Cause for methylation error: cell-type heterogeneity

A common issue that arises in DNA methylation studies is the confounding introduced by cell-type heterogeneity.[44,74,75] It is true that all cells contain the same genetic code[76]; however, methylation plays a substantial role in cell differentiation[77], in that methylation patterns determine cell-type specific functions. Ideally, DNA methylation analysis would be conducted exclusively on the tissue of interest. In reality, most, if not all, patient samples (e.g. tumour, whole blood, saliva etc.) contain a mixture of different cell types, in different proportions.[44,75] As a result, differential methylation may be driven by underlying changes in cell type composition. Should cell-separated data be available for samples, this confounding may be easily corrected for. Unfortunately, researchers often don't have this information at their disposal, requiring the use of analytical tools that can account for this potential confounding in the absence of cell-separated data. Two popular approaches for such correction includes using a reference-based method[74] or alternatively, a reference-free method such as surrogate variable analysis.[78] By correcting for cell type composition in tumour samples, for example, one can be sure that methylation signals from any non-cancerous cells present would be accounted for, thus ensuring differential methylation analysed in said tumour samples is related to disease phenotype (of course assuming other necessary confounders have been controlled for as well). A commonly used term to reflect this cell-type heterogeneity is tumour purity, which is the proportion of cancer cells in the tumour tissue.

The reference-based method[74] for controlling cell-type heterogeneity involves using an appropriate reference dataset containing methylation measurements of already-separated cell types in order to directly estimate the cell type composition and proportions within one's own samples. This approach is based on the principle that different cell types cluster according to similar or shared methylation patterns.[75] The estimated cell-type proportions can then be included as covariates for further analysis. This method is recommended for when a complete set of the required cell-separated methylation profiles are available.[75] Often, it is applied on methylation data derived from whole blood owing to the availability of an appropriate reference sample; this is offered within the ChAMP pipeline (RefbaseEWAS).[61] The *minfi* package (**Table 2-1**) also offers a cell type heterogeneity correction method, with available references of whole blood, cord blood and the frontal cortex.[60] However, a notable drawback of this approach is the limited availability of appropriate reference samples[44,75] which, in part, may be due to difficulty in extracting certain cell types e.g. syncytiotrophoblast cells in placenta.[79] A commonly used dataset provided by Reinius et al. (2012)[80] contains the methylation profiles of cell-sorted blood samples from adult Swedish men, although the appropriateness of this dataset has been questioned.[44] Due to the difficulty in identifying appropriate references, reference-free approaches have been introduced.

One example of a reference-free approach is termed surrogate variable analysis (SVA).[78] Although SVA was not initially intended for use on DNA methylation data (it was originally developed for gene expression data), it has since become a popular method, being well-suited for controlling cell-type heterogeneity.[75] In fact, in a comprehensive study that evaluated some of the more popular methods for correcting cell-type heterogeneity, SVA was recommended by authors for adjustment owing to its adequate performance under multiple simulations and reasonable computation time.[75] This recommendation has been reiterated in more recent publications.[81,82] Briefly, SVA constructs unmodeled confounders (i.e. surrogate variables i.e. cell-type proportions) directly from high-dimensional data, such as DNA methylation data, which can then be used as covariates in subsequent analyses to adjust for unknown or unmodeled sources of noise.[78]

Ultimately, these methods construct covariates for adjustment, after which one may infer phenotype-associated changes that are not driven by changes in cell-type composition. However, a number of concerns have been raised as to the necessity of these adjustments.[44] For example, should cell type sit along the causal pathway, then adjusting for cell type may adjust out the signal of interest due to collinearity. Additionally, it has been suggested that adjustment by methods like SVA may introduce confounding in a dataset that did not require any adjustment to begin with.[44] It has even been shown that certain methods (not discussed here) are guilty of overcorrecting for cell-type composition.[75] Overall, it is recommended that analyses be conducted both with adjustment and without adjustment.[44] In conclusion, there may not be any one adjustment method whose performance is uniformly the best, this includes an adjustment-free approach, as all available methods have both benefits and limitations.[75]

2.8. References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer.

2015; 136:E359–86.

2. Park SY, Haiman CA, Cheng I, Park SL, Wilkens LR, Kolonel LN, et al. Racial/ethnic differences in lifestyle-related factors and prostate cancer risk: The multiethnic cohort study. Cancer Causes Control. 2015; 26(10):1507–15.

3. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014; 64(1):9–29.

4. Abouassaly R, Thompson I, Platz E, Klein E. Epidemiology, etiology, and prevention of prostate cancer. In: Wein AJ, editor. Campbell-Walsh urology. 10th ed. Philadelphia: Elsevier Inc.; 2012. p. 2704–25.

5. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014; 74(8):880–91.

6. Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, et al. Whole-genome sequencing reveals elevated tumor mutational burden and initiating driver mutations in African men with treatment-naïve, high-risk prostate cancer. Cancer Res. 2018; 78(24):6736–46.

7. Powell IJ, Bock CH, Ruterbusch JJ, Sakr W. Evidence supports a faster growth rate and/or earlier transformation to clinically significant prostate cancer in black than in white American men, and influences racial progression and mortality disparity. JURO. 2010; 183:1792–7.

8. Skinner MK. Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. Epigenetics. 2011; 6(7):838–42.

9. Nilsson EE, Sadler-Riggleman I, Skinner MK. Environmentally induced epigenetic transgenerational inheritance of disease. Environ Epigenetics. 2018; 4(2):1–13.

10. Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011.

11. Oey H, Whitelaw E. On the meaning of the word "epimutation." Trends Genet. 2014; 30(12):519–20.

12. Skinner MK. Endocrine disruptor induction of epigenetic transgenerational inheritance of disease. Mol Cell Endocrinol. 2014; 398(0):4–12.

13. Arita A, Costa M. Environmental agents and epigenetics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 459–76.

14. Skinner MK. Environmental epigenetics and a unified theory of the molecular aspects of evolution: A neo-Lamarckian concept that facilitates neo-Darwinian evolution. Genome Biol Evol. 2015; 7(5):1296–302.

15. U.S. National Library of Medicine [Internet]. What is a gene mutation and how do mutations occur? National Institutes of Health; [cited 2020 Jun 12]. Available from: https://ghr.nlm.nih.gov/primer/mutationsanddisorders/genemutation

16. Martin DIK, Ward R, Suter CM. Germline epimutation: A basis for epigenetic disease in humans. Ann N Y Acad Sci. 2005; 1054:68–77.

17. Venza M, Visalli M, Beninati C, Catalano T, Biondo C, Teti D, et al. Involvement of epimutations in meningioma. Brain Tumor Pathol. 2015; 32(3):163–8.

18. Cheng X, Hashimoto H, Horton JR, Zhang X. Mechanisms of DNA methylation, methyl-CpG recognition and demethylation in mammals. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 9–24.

19. Shen JC, Rideout WM, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 1994; 22(6):972–6.

20. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462(7271):315–22.

21. Patil V, Ward RL, Hesson LB. The evidence for functional non-CpG methylation in mammalian cells. Epigenetics. 2014; 9(6):823–8.

22. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation Europe PMC funders group. Nat Genet. 2016; 48(2):126–33.

23. Walser JC, Furano AV. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res. 2010; 20(7):875–82.

24. Kusmartsev V, Drozdz M, Schuster-Böckler B, Warnecke T. Cytosine methylation affects the mutability of neighboring nucleotides in germline and soma. Genetics. 2020; 214:809–23.

25. Qu W, Hashimoto SI, Shimada A, Nakatani Y, Ichikawa K, Saito TL, et al. Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns. Genome Res. 2012; 22(8):1419–25.

26. Narayan VM, Konety BR, Warlick C. Novel biomarkers for prostate cancer: An evidence-based review for use in clinical practice. Int J Urol. 2017; 24(5):352–60.

27. Esteller M. CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. Oncogene. 2002; 21(35):5427–40.

28. Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. 2012; 22(2):246–58.

29. Kinoshita H, Shi Y, Sandefur C, Meisner LF, Chang C, Choon A, et al. Methylation of the androgen receptor minimal promoter silences transcription in human prostate cancer. Cancer Res. 2000; 60:3623–30.

30. Truong M, Yang B, Wagner J, Desotelle J, Jarrard DF. Analysis of promoter non-CG methylation

in prostate cancer. Epigenomics. 2013; 5(1):65–71.

31. Janitz K, Janitz M. Assessing epigenetic information. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 173–81.

32. Fu S, Wu H, Zhang H, Lian CG, Lu Q. DNA methylation/hydroxymethylation in melanoma. Oncotarget. 2017; 8(44):78163–73.

33. Stewart GD, Van Neste L, Delvenne P, Delrée P, Delga A, McNeill SA, et al. Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: Results of the MATLOC study. J Urol. 2013; 189:1110–6.

34. Geybels MS, Alumkal JJ, Luedeke M, Rinckleb A, Zhao S, Shui IM, et al. Epigenomic profiling of prostate cancer identifies differentially methylated genes in TMPRSS2:ERG fusion-positive versus fusion-negative tumors. Clin Epigenetics. 2015; 7(1):128.

35. Börno ST, Fischer A, Kerick M, Fälth M, Laible M, Brase JC, et al. Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. Cancer Discov. 2012; 2(11):1025–35.

36. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. 2011; 21(7):1028–41.

37. Blackburn J, Vecchiarelli S, Heyer EE, Patrick SM, Lyons RJ, Jaratlerdsiri W, et al. TMPRSS2-ERG fusions linked to prostate cancer racial health disparities: A focus on Africa. Prostate. 2019; 79(10):1191–6.

38. Zhou CK, Young D, Yeboah ED, Coburn SB, Tettey Y, Biritwum RB, et al. TMPRSS2:ERG gene fusions in prostate cancer of West African men and a meta-analysis of racial differences. Am J Epidemiol. 2017; 186(12):1352–61.

39. Woodson K, Hayes R, Wideroff L, Villaruz L, Tangrea J. Hypermethylation of GSTP1, CD44, and E-cadherin genes in prostate cancer among US blacks and whites. Prostate. 2003; 55(3):199–205.

40. Kwabi-Addo B, Wang S, Chung W, Jelinek J, Patierno SR, Wang B-D, et al. Identification of differentially methylated genes in normal prostate tissues from African American and Caucasian men. Clin Cancer Res. 2010; 16(14):3539–47.

41. Devaney J, Wang S, Furbert-Harris P, Apprey V, Ittmann M, Wang BD, et al. Genome-wide differentially methylated genes in prostate cancer tissues from African-American and Caucasian men. Epigenetics. 2015; 10(4):319–28.

42. Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nat Methods. 2015; 12(3):230–2.

43. Libertini E, Heath SC, Hamoudi RA, Gut M, Ziller MJ, Herrero J, et al. Saturation analysis for

whole-genome bisulfite sequencing data. Nat Biotechnol. 2016; 34(7):691–3.

44. Wu MC, Kuan PF. A guide to Illumina BeadChip data analysis. In: Tost J, editor. DNA methylation protocols. Methods in molecular biology. New York: Humana Press; 2018. p. 303–30.

45. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium assay. Epigenomics. 2009; 1(1):177–200.

46. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011; 98(4):288–95.

47. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011; 6(6):692–702.

48. Youk J, An Y, Park S, Lee JK, Ju YS. The genome-wide landscape of C:G > T:A polymorphism at the CpG contexts in the human population. BMC Genomics. 2020; 21(1):1–11.

49. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016; 17(1):208.

50. Illumina [Internet]. Illumina methylation BeadChips achieve breadth of coverage using 2 Infinium® chemistries. Illumina; [cited 2020 Nov 20]. Available from: https://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf

51. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 2012; 13(6):R44.

52. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium methylation 450K technology. Epigenomics. 2011; 3(6):771–84.

53. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013; 29(2):189–96.

54. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014; 15(11):503.

55. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013; 8(3):333–46.

56. Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 2018; 19(S5):115.

57. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010; 11(1):587.

58. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. 2006; 16(3):383–93.

59. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12(1):R10.

60. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–9.

61. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017; 33(24):3982–4.

62. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017; 45(4):1–12.

63. LaBarre BA, Goncearenco A, Petrykowska HM, Jaratlerdsiri W, Bornman MSR, Hayes VM, et al. MethylToSNP: Identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin. 2019; 12(1):79.

64. Daca-Roszak P, Pfeifer A, Żebracka-Gala J, Rusinek D, Szybińska A, Jarząb B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip array: Implications for comparative population studies. BMC Genomics. 2015; 16(1):1003.

65. Naeem H, Wong NC, Chatterton Z, Hong MKH, Pedersen JS, Corcoran NM, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics. 2014; 15(1):51.

66. Zhi D, Aslibekyan S, Irvin MR, Claas SA, Borecki IB, Ordovas JM, et al. SNPs located at CpG sites modulate genome-epigenome interaction. Epigenetics. 2013; 8(8):802–6.

67. Siva N. 1000 genomes project. Nat Biotechnol. 2008; 26(3):256–7.

68. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11.

69. 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.

70. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamieldien J, et al. Population-specific common SNPs reflect demographic histories and highlight regions of genomic

plasticity with functional relevance. BMC Genomics. 2014; 15(1):437.

71. Murat K, Grüning B, Poterlowicz PW, Westgate G, Tobin DJ, Poterlowicz K. Ewastools: Infinium human methylation BeadChip pipeline for population epigenetics integrated into Galaxy. Gigascience. 2020; 9(5).

72. Hu X, Tang L, Wang L, Wu F-X, Li M. MADA: A web service for analysing DNA methylation array data. BMC Bioinformatics. 2020; 21(S6):403.

73. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014; 11(11):1138–40.

74. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13(1):86.

75. Mcgregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016; 17(84):1–17.

76. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002; 16:6–21.

77. Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of cellular differentiation. Cell Cycle. 2010; 9(19):3880–3.

78. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3(9):e161.

79. Kaspi T, Nebel L. Isolation of syncytiotrophoblast from human term placentas. Obstet Gynecol. 1974; 43(4):549–57.

80. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. PLoS One. 2012; 7(7):e41361.

81. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, et al. Correcting for cell-type heterogeneity in epigenome-wide association studies: Revisiting previous analyses. Nat Methods. 2017; 14(3):216–7.

82. Kaushal A, Zhang H, Karmaus WJJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. BMC Bioinformatics. 2017; 18(1):216.

Chapter 3: Establishing a bioinformatic pipeline to interrogate genome-wide DNA methylation in African-derived tumour tissue

As discussed in **Chapters 1** and **2**, there is a pressing need for the development of an African-relevant genome-wide DNA methylation bioinformatic workflow that is applicable to the South African population. To the best of my knowledge, no such workflow has been established. As such, the overall aim of **Chapter 3** was to establish a novel bioinformatic workflow for the processing and normalisation of South African DNA methylation data. Additionally, this workflow should allow genome-wide DNA methylation to be interrogated in African-derived tumour tissue.

In this Chapter, I would like to acknowledge the patients who consented to donating their tissue for the purposes of this research, as well as the many clinicians and staff associated with the Southern African Prostate Cancer Study. I would also like to thank Dr Pavlo Lutsik, head of the Computational Cancer Epigenomics Group at Deutsches Krebsforschungszentrum, in Germany, for offering me the opportunity to present my proposed workflow to him for critical review. I would like to extend an acknowledgement to the Human Comparative and Prostate Cancer Genomics Research team at the Garvan Institute of Medical Research, in Australia, for data availability. Finally, I would like to thank Prof Clare Stirzaker and select members of her Epigenetic Deregulation in Cancer Group at the Garvan Institute, namely Dr Tim Peters, Dr Ruth Pidsley and Dr Elena Zotenko, for providing me with bioinformatic resources and support.

Abstract

The emergence of the Illumina Infinium BeadChips has provided researchers with comprehensive, user-friendly platforms to interrogate genome-wide DNA methylation in human samples. The most recent of these technologies, the Illumina Infinium HumanMethylationEPIC BeadChip, measures methylation at over 850,000 CpG sites throughout the human genome. However, assay design and bioinformatic tools for the suitable processing, normalisation and analysis of Illumina array-generated DNA methylation data, is biased towards non-African cohorts. This creates a challenge for researchers working on African-derived data because population-specific genomic diversity affects probe hybridization, methylation quantification and subsequent data filtering. Although these challenges exist for European-derived data too, it is to a lesser extent, as such platforms and bioinformatic tools are largely designed to account for the genomic diversity that exists for these more highly-represented populations. Consequently, as far as I am aware, there are no available bioinformatic tools that consider South African-relevance at critical points in the workflow. Such a tool is necessary to ensure minimal appropriate African-specific data is lost and minimal confounding African-specific data is retained. Here, I present a novel established pipeline that allows researchers to appropriately process and analyse tumour-derived southern African DNA methylation data while accounting for confounding African-relevant polymorphisms and unnecessarily eliminating African-relevant data.

3.1. Introduction

In recent years, epigenome-wide association studies (EWAS) have gained popularity for allowing researchers to investigate variation in the epigenome, with a particular focus on DNA methylation. Similar to genetic epidemiology, epigenetic epidemiology is concerned with understanding the molecular basis for disease risk. Array-based approaches are a cost-effective means to assess the DNA methylation status across tumour genomes, especially for large studies aimed at identifying biomarkers of cancer progression. As early as 2006, bead arrays had been developed for high-throughput DNA methylation profiling, although only for 1,536 CpG sites.[1] The earliest effort to conduct EWAS can be credited to the Illumina Infinium HumanMethylation27 BeadChip (27K array), released in 2009, which measures the methylation status of 27,578 CpG sites across the human genome at single nucleotide resolution.[2] However, considering the human genome contains over 28 million CpG sites, the 27K array falls terribly short from being considered "genome-wide".

More recently and aiming to address the low coverage of the 27K array, the Illumina Infinium HumanMethylation450 BeadChip (450K array) was developed and made available in 2011. This array offers an even higher genome-wide coverage, assessing methylation levels at 485,577 individual CpG sites[3,4], although this accounts for less than 2 % of the CpG sites in the genome. Regardless, one can reveal with a simple literature search that the 450K array is still the most widely-used platform for studies reporting EWAS. Finally, the most recent of these technologies and only in its fifth year of use is the Illumina Infinium HumanMethylationEPIC BeadChip (EPIC/850K array). Introduced in 2016 and covered in detail in **Chapter 2**, the EPIC array interrogates 863,904 CpG sites and 2,932 CNG sites on important regulatory regions.[5,6] As such, the genome-wide EPIC array was selected as the platform of choice for my thesis.

A notable and broadly accepted limitation of array-based methylation screening is the impact of genomic variation on probe hybridization (discussed in **Chapter 2**).[7–11] Given that many SNPs are population-specific[12], these arrays are limited by selecting content according to databases that are heavily reflective of Europeans, thereby biasing assay design towards European populations. Additionally, probes annotated for SNPs by Illumina reference Genome Build 37, which notoriously underrepresents populations whose genetic makeup is not commonly shared in European and North American nations. Of particular interest to this work is the underrepresentation of (southern) African populations, which display vast within and between genetic diversity.[13] Due to this high diversity and the lack of African inclusion in genomic data, echoed by Cronjé et al. (2020)[14], one cannot be confident in the broad African-relevance of the Illumina arrays and one may even be less confident in their applicability to genetically diverse subpopulations within Africa. As a result, one might expect African variants, not accounted for in assay design, to greatly affect probe hybridization. This could potentially cause a loss of informative sites in African EWAS or could affect the accuracy of their DNA methylation quantification. Even within European populations, studies have reported erroneous calling as a result of genomic variation impacting probe hybridization.[15]

Although the above-mentioned assays cannot simply be tailored for the purposes of this study, it is clear that polymorphisms at methylation probe sites need to be addressed when processing DNA methylation data to reduce the risk of false discoveries, and this can be tackled for the scope of this research. A number of bioinformatic tools are currently available for Illumina DNA methylation data processing and analysis (see **Chapter 2**), a number of which provide functions for filtering SNP-affected probes. Though the obstacle of European bias persists. The identification of SNP-affected probes is typically conducted using generalized references, including the 1000 Genomes Project data[16] and dbSNP.[17] Additionally, lists of recommended probes for filtering have been annotated for the Illumina EPIC array.[7,18] However, while these references account for numerous populations, the fact remains that they are largely European-relevant and therefore, unsuitable for use in an African cohort due to potential excessive, unnecessary data loss.

Presently to my knowledge, no such tool accounts for confounding southern African-relevant polymorphisms and for unnecessarily eliminating African-relevant data, presumably due to the lack of African inclusion in (epi)genomic data and the impact of European-biased array design.

Based on the above discussion, the aim of this study is to establish an African-relevant genome-wide bioinformatic pipeline for the processing and normalisation of African DNA methylation data. In addition, this novel pipeline should allow genome-wide DNA methylation to be interrogated in prostate tissue from men of African ancestry as well as in tumour tissue from other cancer types. This novel toolset will be a significant and unique contribution to the field of epigenetic epidemiology.

3.2. Materials & Methods

3.2.1. Resource & ethics

Data was made available for eight South African men who consented upon enrolment in the Southern African Prostate Cancer Study (SAPCS)[19] (further outlined in **Chapter 4**). Patients were of African ethnicity, confirmed using ancestry markers, and self-identified as such. A total of eight patients were recruited at diagnosis and clinicopathologically confirmed as either presenting with high-risk prostate cancer (HRPCa, 7 patients), defined by a Gleason score of ≥ 8, or with benign prostatic hyperplasia (BPH, 1 patient). The previous SAPCS as well as the current study outlined here was reviewed and approved by the University of Pretoria's Human Research Ethics Committee (HREC #43/2010 and #37/2021, respectively).

3.2.2. DNA methylation data generation and quality control

Raw DNA methylation data was generated at the Australian Genome Research Facility (AGRF, Melbourne, Australia) and subsequently provided by the Human Comparative and Prostate Cancer Genomics (HCPCG) Research team at the Garvan Institute of Medical Research, for the eight above-mentioned South African patients. DNA methylation was quantified using the Illumina Infinium HumanMethylationEPIC BeadChip (hereafter referred to as the EPIC (micro)array) following the Illumina Infinium HD Methylation Assay (Illumina, CA, USA). The EPIC array quantifies DNA methylation at around 860,000 individual CpG sites and just under 3,000 CNG sites on important regulatory regions.[5]

The data produced and subsequently provided by the AGRF included raw Illumina intensity data (IDAT) files, the Illumina manifest file (v1.0 B5, BPM format), a sample sheet (CSV format) and a genotyping service report from the research facility. For each sample, a “red” and a “green” IDAT file is supplied, representing the intensities of the methylated and unmethylated probes. These IDAT files contain the actual DNA methylation measurements for each probe, represented by a beta-value and corresponding detection *p*-value for each probe. The *p*-value is a confidence measure for the reported beta-value. The Illumina manifest file contains a description of the probes including probe IDs (cg-00000000), chromosome, location, relation to epigenetically relevant features, gene membership, nearby SNPs etc. and it references Genome Build 37 (hg19). A CSV format of the manifest can be downloaded from Illumina’s website. The sample sheet stores phenotypic data associated with the EPIC BeadChip including sample information and metadata associated with a given experiment. Finally, the genotyping service report provided by AGRF contains a project description, details on project data and a quality control report for the EPIC BeadChip. All samples were within the Illumina expectations of ≥ 96 % of CpG sites having been detected ($p < 0.01$).

3.2.3. Germline variant data

Variant called germline data (VCFv4.2 format) for the eight patients was made available by the HCPCG Research team. The VCF files reference Genome Build 38 (hg38) and chromosome notation is of the UCSC style (e.g. chr1) versus the NCBI/Ensembl style (e.g. 1). The tools used for germline variant extraction includes Java[20] (JDK™, v.1.8.0_111), GATK[21] (v.4.1.4.1), HTSlib[22] (v.1.10.2) and VCFtools[23] (v.0.1.14). Data provided is currently unpublished and funded by the Australian National Health and Medical Research Council (NHMRC).

R[24] ≥ v.4.0.2 and RStudio[25] ≥ v.1.3 were used in this study.

3.3. Results

As discussed above, relevant bioinformatic tools for the suitable processing of African data are scarce. As such, an African-relevant genome-wide bioinformatic pipeline for DNA methylation data processing and analysis had to be established.

3.3.1. Selecting the bioinformatic backbone

After a thorough review of the literature (see **Chapter 2** for a brief overview), the Chip Analysis Methylation Pipeline for Illumina[26] (**Fig. 3-1**), or ChAMP, was chosen for modification (see **Fig. 3-2** for the novel workflow). Selection criteria included: (i) it supports the processing and analysis of EPIC microarray data; (ii) it is a comprehensive and fairly complete analysis pipeline that supports a number of existing DNA methylation microarray data analysis packages; (iii) it allows that each function may be run individually and resultant datasets thereof saved individually, to optionally integrate data with other pipelines and finally, (iv) it is user-friendly. ChAMP is an R package available from Bioconductor.

3.3.2. Preparing an African-relevant EPIC array methylation dataset

3.3.2.1. Extracting the data with champ.load()

Beginning with the raw DNA methylation data in the form of IDAT files, the first step in the pipeline is data extraction. Data is loaded into R using the champ.load() function, which imports the beta-value for each probe and the corresponding detection *p*-value. The sample sheet provided with the DNA methylation data is also imported, hereafter referred to as the phenotypes (pd) file. The champ.load() function performs some preliminary probe filtering upon data loading, essentially combining the champ.import() and champ.filter() functions available in the original ChAMP pipeline. I found this first step in the pipeline is already where one needs to consider and implement African-relevance. This may be achieved with careful consideration of function parameters.

3.3.2.2. Considering and modifying the champ.load() function parameters

The champ.load() function parameters are numerous (**Table 3-1**). Standard filtering involves excluding probes that fail in individual samples; typically, this denotes probes with detection *p*-values greater than 0.05[9], although a detection *p*-value threshold of 0.01 was also employed for comparison. I found probe rejection between the *p*-value thresholds of 0.05 and 0.01 to be highly comparable, which motivated the use of the recommended threshold of 0.05[9] to proceed. Probes that fail the detection *p*-value threshold in more than 20 % of the samples or had a bead count < 3 in at least 5 % of the samples should also be removed.[9] Probes with a low bead count are not very informative. When considering the African data, while a number of parameters were set to default, others needed modification, each discussed further.

"population" and "filterSNPs" parameters. Given that many SNPs are population-specific[12], specifying the appropriate population is critical to achieve the most accurate SNP filtering. Therefore, to optimize accurate filtering, it is essential to consider one's cohort to ensure minimal appropriate data is lost and minimal confounding data is retained. The populations available from the International Genome Sample Resource[27] (IGSR) may be classified as a "Super Population" (e.g. African, European) or a more specific population (e.g. Yoruba, Nigerian; Finnish). However, at the time of this study, the specific African populations to choose from did not include southern Africans, only western and eastern Africans. Considering the vast genetic diversity that exists between different African populations[13], to specify a "nonsouthern" African population at this step would not be suitable. To overcome this limitation, the "population" parameter was set to "NULL" and the "filterSNPs" parameter to "FALSE".

Table 3-1 champ.load() function arguments, indicating both default and modified parameters.

Argument	*Default Setting*	*Chosen Setting*	*Description*
directory	getwd()	getwd()	Location of the IDAT files.
method	ChAMP	minfi	Specifies the method with which to load the data.
methValue	B	B	Specifies whether M or beta values are loaded.
autoimpute	TRUE	TRUE	Any missing data present after filtering will be imputed using the k-nearest neighbour algorithm.
filterDetP	TRUE	TRUE	Probes above the detection p-value threshold will be filtered out.
ProbeCutoff	0	0	The missing data ratio threshold for probes.
SampleCutoff	0.1	0.2	The failed p-value or missing data ratio threshold for samples.
detPcut	0.01	0.05	The detection p-value threshold.
filterBeads	TRUE	TRUE	Probes with a beadcount less than 3 will be removed depending on the beadCutoff value.
beadCutoff	0.05	0.05	The fraction of samples that must have a beadcount less than 3 before the probe is removed.
filterNoCG	TRUE	FALSE	Non-CG probes are removed.
filterSNPs	TRUE	FALSE	SNP-affected probes according to Nordlund et al. (2013) are removed.
population	TRUE	NULL	Specifies the population with which user is working.
filterMultiHit	TRUE	FALSE	Cross-reactive probes according to Nordlund et al. (2013) are removed.
filterXY	TRUE	FALSE	Probes from X and Y chromosomes are removed.
force	FALSE	FALSE	Allows user to select common probes from different batches.
arraytype	450K	EPIC	Specifies whether the microarray type is 450K or EPIC.

Descriptions adapted from *Bioconductor*, ChAMP reference manual by Tian et al., 2020, https://bioconductor.org/packages/release/bioc/html/ChAMP.html. Copyright 2020 by The Authors.

To further address an oversight at this step, when the default "population" parameter is used, SNP-affected probes will be filtered out according to either Nordlund et al. (2013)[28] or Zhou et al. (2016)[7], references which only provide general probe filtering recommendation lists. Although Zhou et al. (2016)'s list considers population differences, it remains non-specific to the southern African cohort in this study. Wu & Kuan (2018)[9] recommend that when filtering SNP-affected probes, the reference should be similar in ethnicity and population genetic structure to the study population. Using a generalised list to filter SNP-affected probes calls into question how many relevant probes would be removed as well as how many confounding probes would be retained. For this concern, the "filterSNPs" parameter was set to "FALSE" and population-specific SNP-affected probe filtering was carried out downstream (discussed below).

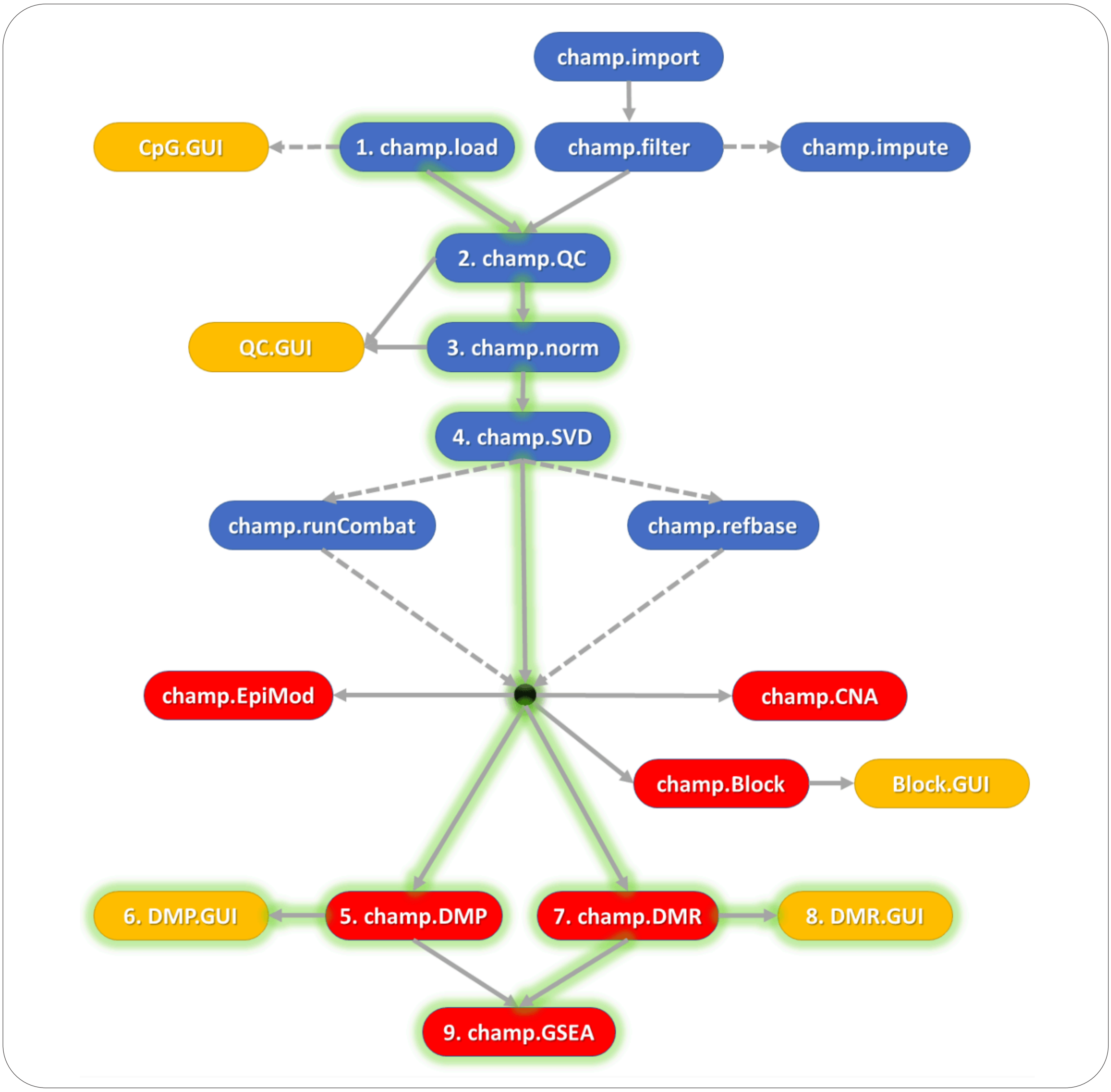


Fig. 3-1 The original, unmodified and complete Chip Analysis Methylation Pipeline for Illumina (ChAMP) workflow. The green highlight represents a standard analysis workflow that is most likely to be used for various datasets. Blue blocks represent functions for methylation data preparation. Yellow blocks represent functions for graphical user interface i.e. for dataset and analysis result visualisation. Red blocks represent functions for generating analysis results. Solid grey arrows denote the workflow of the pipeline while dashed grey arrows are optional functions. The black circle symbolises a fully prepared methylation dataset.

Reprinted from *Bioconductor*, by Tian et al., 2017, https://bioconductor.org/packages/release/bioc/vignettes/ChAMP/inst/doc/ChAMP.htht#section-citing-champ. Copyright 2017 by The Authors.

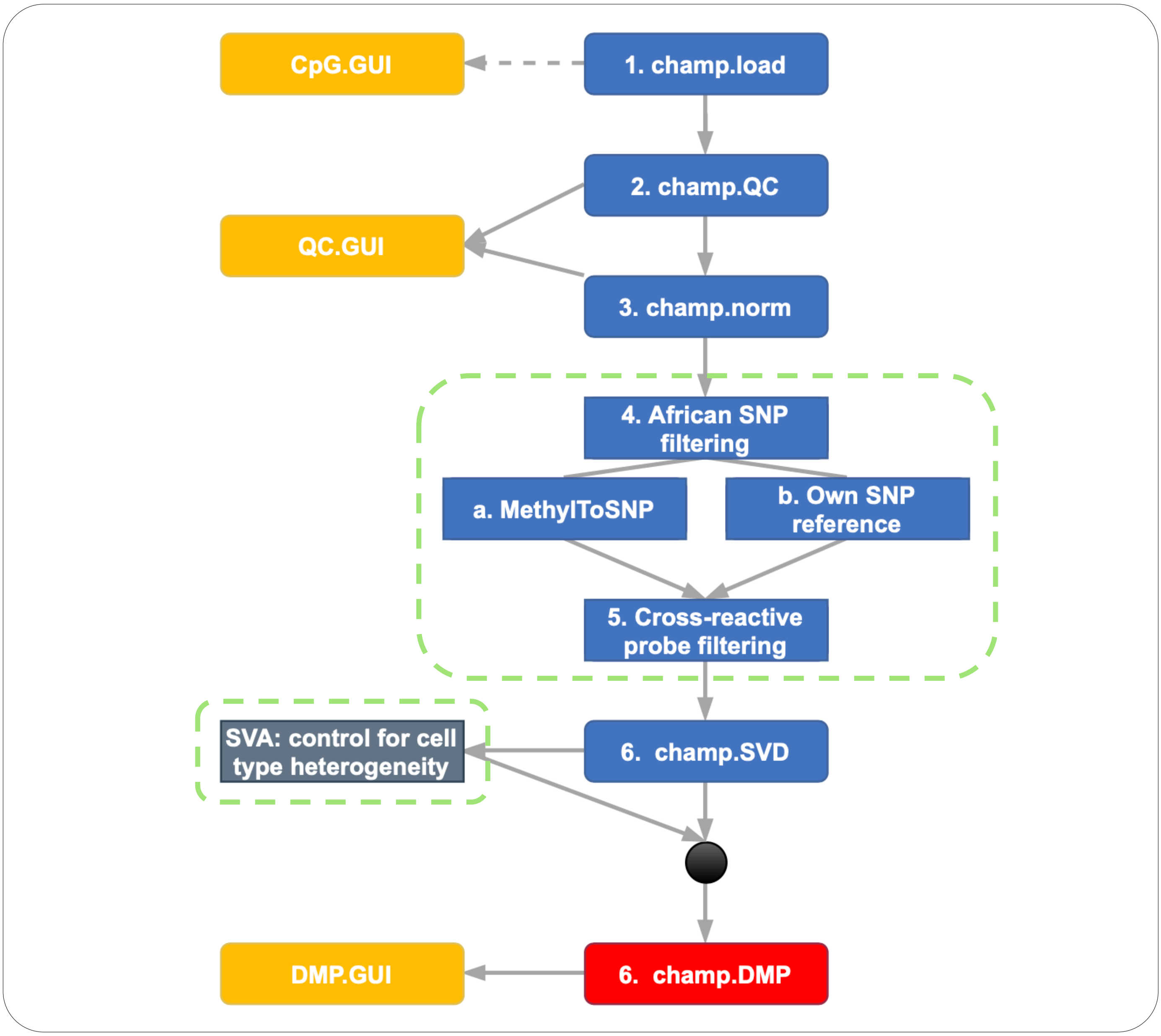


Fig. 3-2 The novel African-relevant methylation data processing, normalization and basic analysis pipeline developed in this study. Blue blocks represent functions for methylation data preparation. Yellow blocks represent functions for graphical user interface i.e. for dataset and analysis result visualisation. Red blocks represent functions for generating analysis results. Grey blocks represent functions that may improve this novel pipeline but were not successfully run. Rounded-edge blocks represent functions that were selected from ChAMP for integration with this pipeline. Sharp-edge blocks circled by dashed green lines represent novel steps relevant for southern African data that are unique to this study. Solid grey arrows denote the workflow of the pipeline while dashed grey arrows are optional functions. The black circle symbolises a fully prepared methylation dataset, signifying the end-point of this developed workflow.

DMP: differentially methylated probes | GUI: graphical user interface | QC: quality control | SNP: single nucleotide polymorphism | SVA: surrogate variable analysis | SVD: singular value decomposition

Adapted from *Bioconductor*, by Tian et al., 2017, https://bioconductor.org/packages/release/bioc/vignettes/ChAMP/inst/doc/ChAMP.htht#section-citing-champ. Copyright 2017 by The Authors.

"method" parameter. For data loading, I chose the classic "*minfi*" method over the "ChAMP" method for downstream purposes. Essentially, the two methods return the same data objects, only differing in that the *minfi*[29] Bioconductor package loads data to additionally produce "mset" and "rgSet" data objects. The "mset", or MethylSet, object contains the methylated and unmethylated probe signals and the "rgSet", or RedGreenChannelSet, object contains red and green channel intensities as well as phenotype and manifest data. These objects store the very same data as the object loaded by the "ChAMP" method; all that differs is the format of such objects. Particular functions and analyses require this methylation data to be stored in "mset" and "rgSet" object format.

"filterMultiHit" parameter. The "filterMultiHit" parameter removes cross-reactive probes i.e. probes that align to multiple genomic locations. However, this filtering is performed according to the Nordlund et al. (2013)[28] multi-hit probe list, which was determined based on the 450K array. Although the EPIC array covers more than 90 % of the CpG sites covered by the 450K array[9], a more recent multi-hit probe list is available, one that is applicable for the EPIC array.[5] For this reason, "filterMultiHit" was set to "FALSE".

"filterXY" parameter. The "filterXY" parameter removes probes located on the X and Y chromosomes. This removal should be performed when analysing both male and female samples; this is to prevent sex from being the largest source of variation in a methylation dataset.[30] However, since the samples in this dataset were all derived from male donors, sex chromosome probes need not be removed. Thus, "filterXY" was set to "FALSE".

"methValue" parameter. For the "methValue" parameter, one may consider the use of either beta or M-values (see **Chapter 2** for further details). However, I made use of beta values (rather than M-values) in this study, as recommended by Illumina, as well as for their direct biological interpretation and evidence supporting the negligible effect of beta-value transformation on analysis results.[31]

Finally, after loading the data coupled with some initial filtering as discussed, a suitable dataset was available for further processing in R.

3.3.2.3. Visualising probe distribution

The CpG.GUI function (see **Fig. 3-2**) is an optional step in the pipeline that allows a user to visualise the distribution of probes in a dataset. This CpG distribution may be analysed in the context of chromosomes, CGI regions, gene regions and Infinium probe types. Annotations for CpG islands includes CGIs, CGI shores (<2 kb upstream and downstream of CGIs), CGI shelves (2-4 kb upstream and downstream of CGIs) and open sea (non-CGI-related sites). Gene region annotations includes TSS1500 (200-1500 bp upstream of the transcription start site, TSS), TSS200 (up to 200 bp upstream of the TSS), 5'UTR (5' untranslated region), 1st exon, Body (gene body), ExonBnd (exon boundaries), 3'UTR and IGR (intergenic regions). One may return to this function at any point in the pipeline to check probe distribution e.g. before and after normalisation. Probe distribution was examined after data loading and initial filtering (**Fig. 3-3a**).

3.3.2.4. Performing quality control & normalising the data to correct for probe bias

Quality control is an important step during data processing as it allows a user to check whether or not their dataset is suitable for downstream analysis. One should perform a quality control check both before and after normalising the data; this provides a visual confirmation for satisfactory data normalisation. Normalisation of the data itself is an essential step. Typically, the goal of normalisation is to remove any technical and systematic variability from the data to ensure measurements are comparable across samples.[9] However, within the context of Illumina methylation analysis, it is true that these procedures have an emphasis on within-sample normalization. Because the Illumina EPIC array uses two different probe types, i.e. Infinium type I & II probes, each of which have different designs with different hybridization chemistries, the two different probe types display different beta-value distributions. This is a form of technical variability. Normalisation is a means of correcting for this difference, to reduce variability or to make the distributions of these two probe types comparable. Oftentimes, downstream analyses assume that data has been normalized. Quality control can be visualised with the champ.QC() and QC.GUI() functions; normalisation is performed with the champ.norm() function (see **Fig. 3-2**).

Normalisation methods offered by the ChAMP pipeline include the beta-mixture quantile (BMIQ)[32], subset-quantile within array normalisation (SWAN)[33], peak-based correction (PBC)[34] and FunctionalNormalization[35] methods (see **Chapter 2** for more detail on these methods). For this particular pipeline, I chose to apply the BMIQ normalisation method to the data because it was previously suggested to be the optimal normalisation method for the reduction of technical variability in comparison to SWAN.[36] Consequently, the established workflow includes a quality control check both before and after BMIQ normalisation (**Fig. 3-4**) (see **Fig. S1** for individual sample density plots).

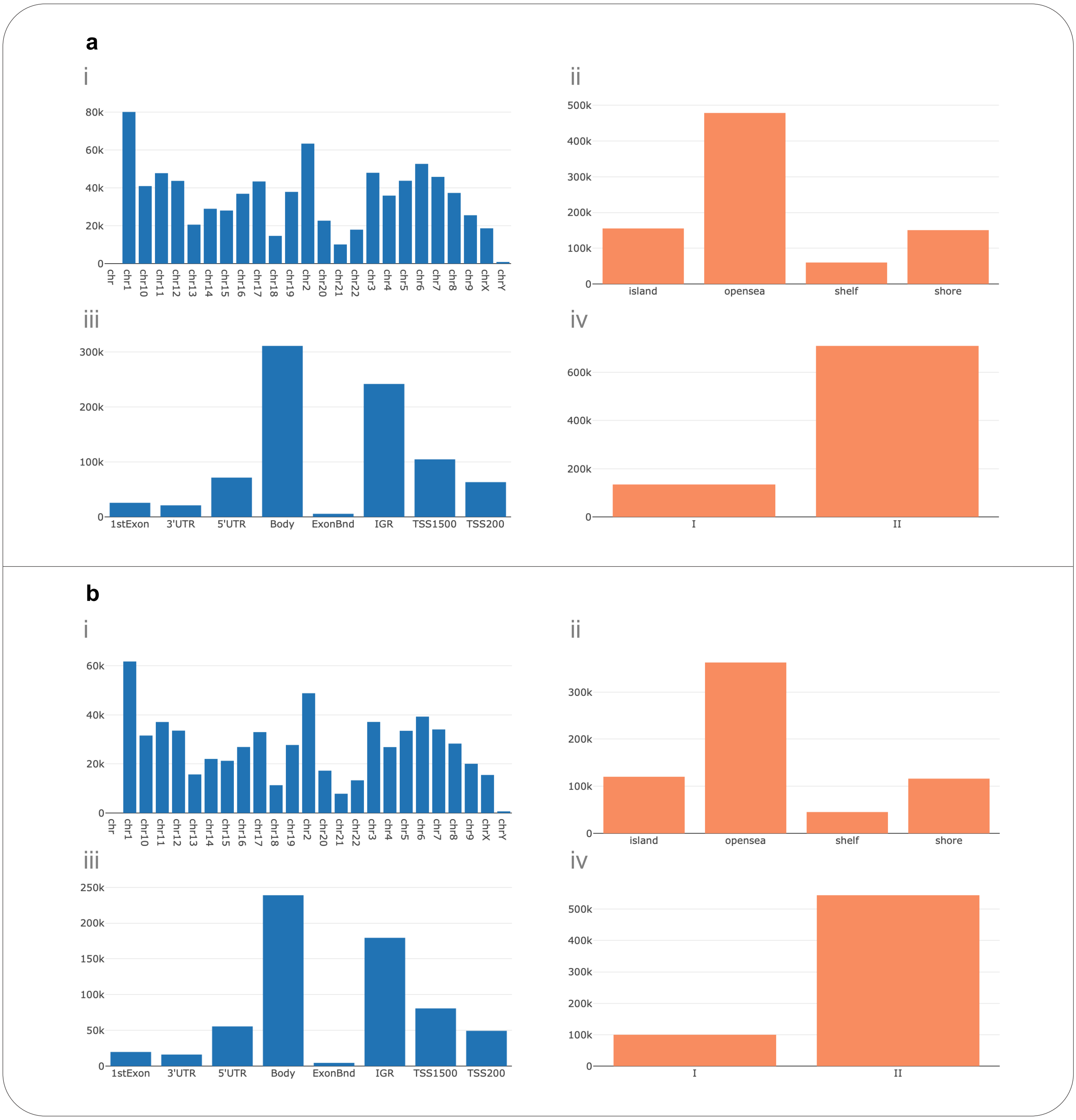


Fig. 3-3 Probe (CpG) distribution in the context of (i) chromosomes, (ii) CpG island features, (iii) gene features and (iv) Infinium probe types. **a** Probe distribution after data loading and initial filtering. **b** Probe distribution after complete filtering and normalisation i.e. distribution of the fully prepared methylation dataset in this study. The y-axis indicates number of probes.

chr: chromosome | ExonBnd: exon boundaries | IGR: intergenic region | TSS: transcription start site | UTR: untranslated region | I: Infinium type I probes | II: Infinium type II probes

3.3.2.5. Novel African-relevant filtering of SNP-affected probes

As previously discussed, many bioinformatic tools do not offer appropriate population-relevant filtering steps for addressing polymorphisms at DNA methylation probe sites, and this is particularly true when considering southern African populations. To overcome this limitation, I discovered two such filtering approaches which I then tested for their efficacy on African-relevant data. The approaches outlined below are unique to the standard ChAMP workflow and as far as I understand, are novel to any existing Illumina DNA methylation data processing pipelines, highlighting the necessity for the integration of such an African-relevant filtering tool.

3.3.2.5.1. MethylToSNP

My initial approach to filter SNP-affected probes from the dataset was performed using the R Bioconductor package, MethylToSNP.[8] MethylToSNP operates by searching for tri-modal beta-value distribution patterns that are commonly characterised by underlying polymorphisms. Detecting this characteristic data pattern allows the identification (or inference) of SNPs (mostly meC > T polymorphisms) from the methylation data itself. MethylToSNP was performed with gap sum ratio 0.5, gap ratio 0.75 and without outlier removal to detect polymorphisms in this southern African cohort's methylation data. These parameters are a replication of those used by LaBarre et al. (2019)[8] on southern African data. This tool requires an "mset" data object as input hence use of the *minfi* method for data loading (see **Section 3.3.2.2.**). A reliability score ≥ 0.5 is considered a high-confidence call.[8] SNP-affected probes called using MethylToSNP were annotated using dbSNP[17] release 147. At the time of this study, dbSNP 147 was the latest dbSNP release available for SNP annotation provided by the Bioconductor package *IlluminaHumanMethylationEPICanno.ilm10b4.hg19*.[37] This tool annotates SNPs from various releases of dbSNP as represented on the UCSC Common SNP table. For limitations discussed previously (see **Chapter 2**), including a recommended minimum number of 50 samples and underlying SNP underestimation called by this tool, I chose not to use the list of probes generated by MethylToSNP (**Table S1**) for filtering. Instead, I proceeded with the second approach in which DNA methylation data was combined with patient-matched germline variant data to ensure only African-relevant probes were filtered out.

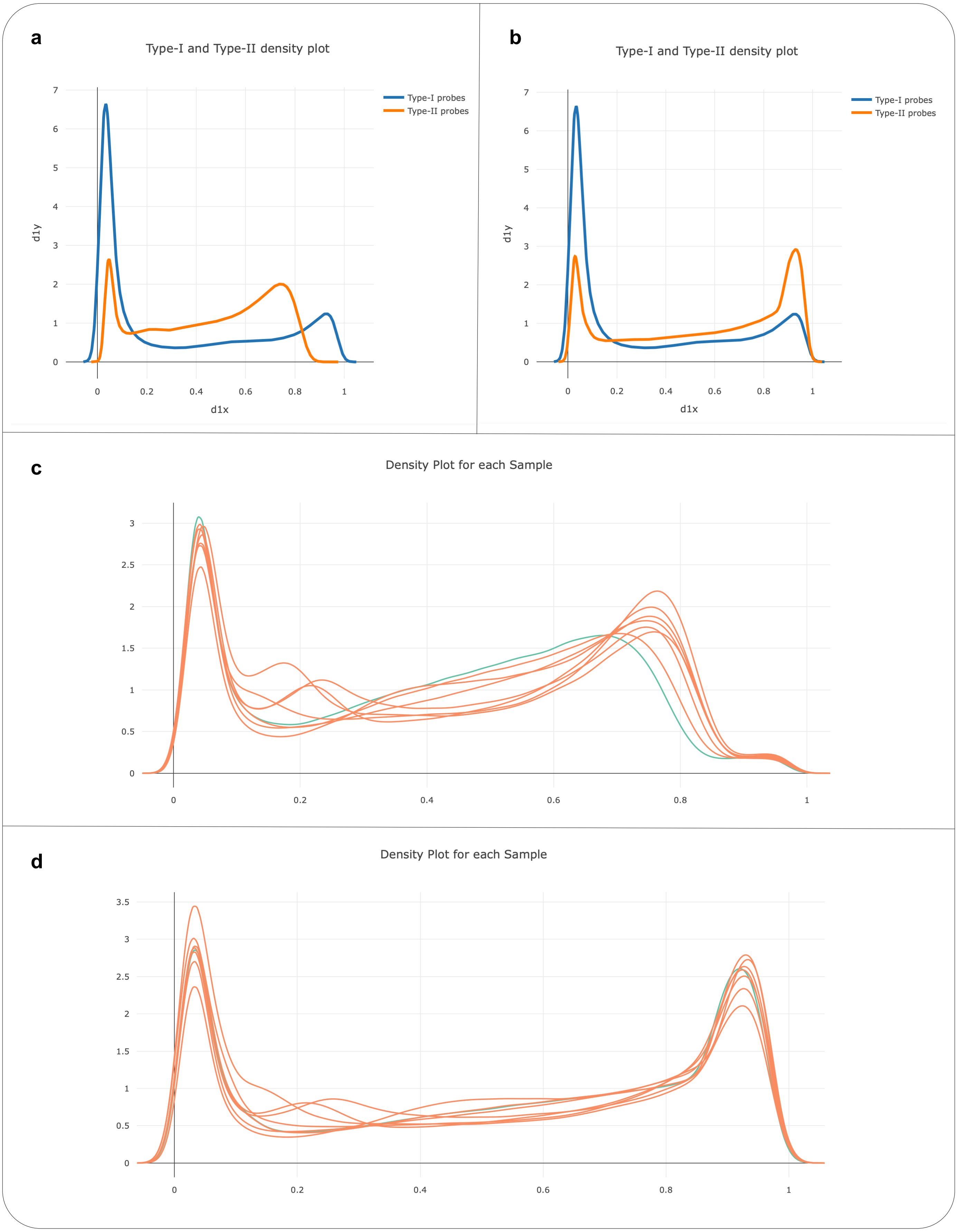


Fig. 3-4 (See legend on next page.)

(See figure on previous page.)

Fig. 3-4 Density plots representing probe beta-value distributions. **a** Beta distribution for Infinium type I and type II probes in the raw dataset. **b** Beta distribution for Infinium type I and type II probes in the normalised dataset. **c** Sample beta distribution in the raw dataset. **d** Sample beta distribution in the normalised dataset. In c & d, orange curves represent African prostate cancer patients; the green curve represents the single African benign prostatic hyperplasia individual.

3.3.2.5.2. African patient-matched germline variant data method

For the second approach, I assessed the use of patient-matched germline variant data in conjunction with the EPIC DNA methylation data to identify polymorphisms at methylation probe sites. The advantage of this method is ensuring only African-relevant and patient cohort-relevant SNP-affected probes would be identified for filtering. The method developed consists of two parts; first, the preparation of the African germline VCF files and second, parsing these germline VCF files for the actual calling of African SNP-affected probes (**Fig. 3-5**). The objective, to parse a single reference VCF file to extract all SNP and indel variants overlapping EPIC probes and as such, ensuring a streamlined process. Probe coordinates are contained in the Illumina EPIC manifest file and in the context of these probe coordinates, variants were examined according to three categories: (1) variants overlapping target CpG sites; (2) variants overlapping single base extension (SBE) sites for Infinium type I probes; and (3) variants overlapping the rest of the probe body, 48 bp for Infinium type I probes and 49 bp for Infinium type II probes (see **Chapter 2** for a description of these sites). The method of extracting probes overlapping genetic variants has been described previously.[5]

VCF file preparation. Part one was completed on the command line (macOS Terminal v.2.11) and began with the raw patient germline VCF files. Using VCFtools[23] (v.0.1.16), VCF files were filtered to only contain "PASS" variants i.e. variants that passed all necessary filters; this filters out redundant information. Next, I used BCFtools[38] (v.1.12) to combine all eight African germline VCF files without any duplications (bcftools merge), thereby creating a single reference VCF file. As previously mentioned, these African VCF files reference the GRCh38 (hg38) genome assembly. However, the Illumina manifest file references the GRCh37 (hg19) genome assembly. Therefore, when one parses the VCF file to extract SNPs overlapping probe coordinates contained within the manifest file, both the VCF file and the manifest file must reference the same genome assembly i.e. hg19. To achieve this, the single reference VCF file needs to be "lifted over" from the hg38 to the hg19 genome assembly. I discovered a number of tools available for this. In essence, these lift over tools adjust the coordinates of variants in a VCF file to match a new reference.

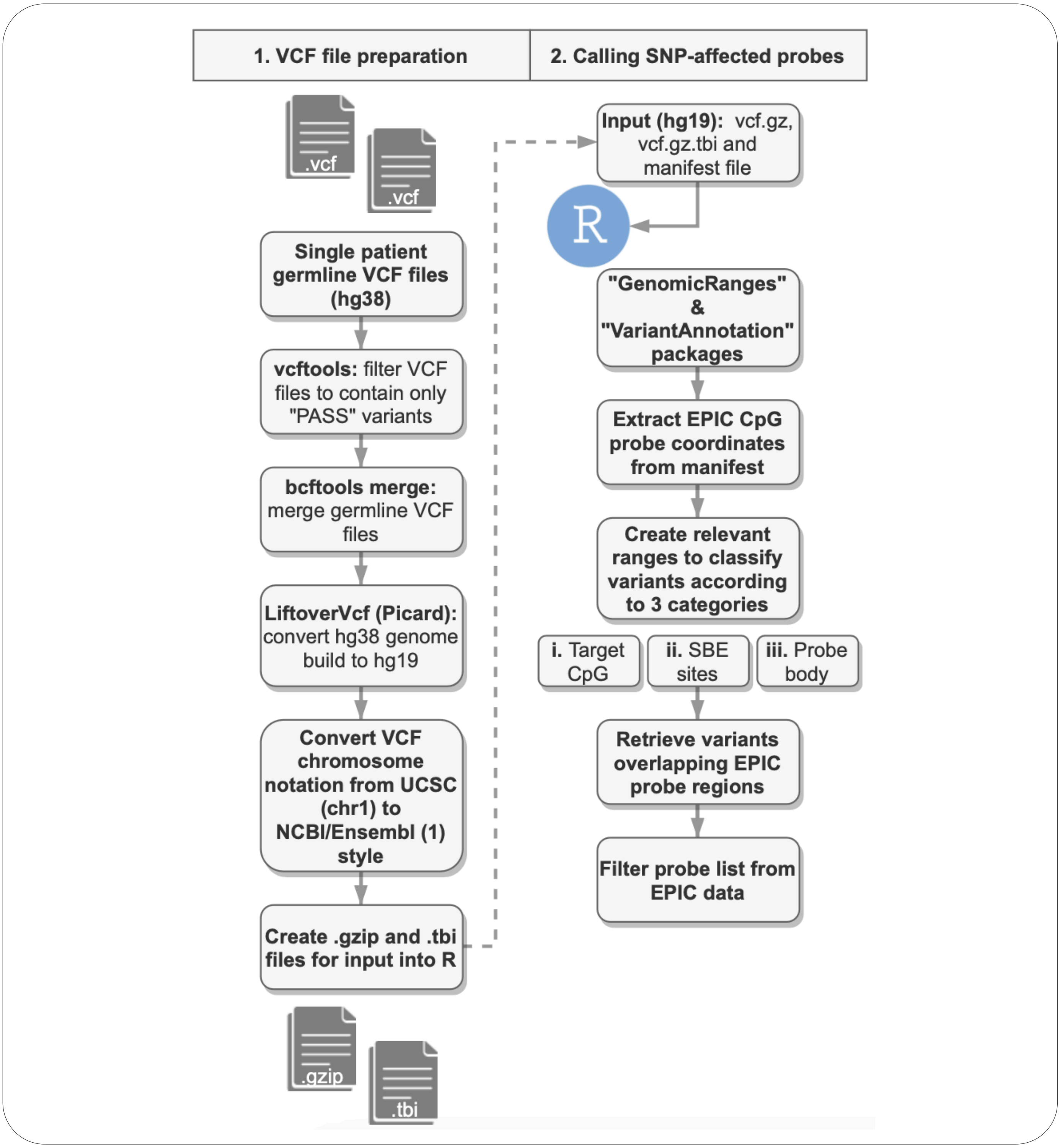

1. VCF file preparation
2. Calling SNP-affected probes
.vcf
.vcf
Single patient germline VCF files (hg38)
vcftools: filter VCF files to contain only "PASS" variants
bcftools merge: merge germline VCF files
LiftoverVcf (Picard): convert hg38 genome build to hg19
Convert VCF chromosome notation from UCSC (chr1) to NCBI/Ensembl (1) style
Create .gzip and .tbi files for input into R
.gzip
.tbi
Input (hg19): vcf.gz, vcf.gz.tbi and manifest file
R
"GenomicRanges" & "VariantAnnotation" packages
Extract EPIC CpG probe coordinates from manifest
Create relevant ranges to classify variants according to 3 categories
i. Target CpG
ii. SBE sites
iii. Probe body
Retrieve variants overlapping EPIC probe regions
Filter probe list from EPIC data

Fig. 3-5 Developed workflow for the preparation (1) and parsing (2) of African germline variant files to extract all African SNP and indel variants overlapping Illumina Infinium HumanMethylationEPIC probes.

BCF: binary variant call format | NCBI: National Center for Biotechnology Information | SBE: single base extension | SNP: single nucleotide polymorphism | UCSC: University of California Santa Cruz | VCF: variant call format

I chose to perform a liftover of the VCF file using Picard (v.2.25.6) LiftoverVcf with the UCSC chain file (hg38ToHg19.over.chain available from UCSC's GoldenPath).[39] Liftover also requires the reference genome sequence file for the target build, hg19 (obtained via UCSC's GoldenPath, hg19.fa) as well as an accompanying FASTA sequence dictionary file and FASTA index file (created from the FASTA file using GATK CreateSequenceDictionary and SAMtools[38] faidx, respectively). This procedure and particularly Picard LiftoverVCF requires GATK[21] (v.4.2.0.0) and Java[20] (JDK™ v.16.0.1). Finally, to prevent a Java "OutOfMemoryError", I found it was necessary to specify a higher memory allocation (Xmx8G) than the default allocation. Only SNPs that were successfully lifted over to the same chromosome were retained. Rejected variants were stored in a separate VCF file. This method rendered a single African germline VCF file that references the hg19 genome assembly and is properly headered, sorted and indexed.

An alternative tool, CrossMap[40], was initially implemented at this step for liftover, requiring Python[41] (v.3.9). Essentially, the purpose of this tool is the very same as that of Picard LiftoverVcf and both tools require a number of the same files as input e.g. the UCSC chain file and the target build reference genome sequence file. However, the advantage of using Picard LiftoverVcf rather than CrossMap is that it produces a properly headered, sorted and indexed VCF file. Conversely, I found that CrossMap produces an output file that is not sorted which then creates a problem when trying to create a tabix index for this VCF file, which is necessary for part two of this workflow.

Once liftover had been completed, an additional feature of the VCF file had to be addressed. To reiterate, when parsing the VCF file to extract SNPs overlapping probe coordinates contained within the manifest file, the VCF file and the manifest file must be in the same format where applicable. In this case, the feature being referred to is chromosome notation, which may be of the UCSC style (e.g. chr1) or the NCBI/Ensembl style (e.g. 1). I discovered that the Illumina EPIC manifest file details chromosomes using the NCBI/Ensembl style whereas the VCF file uses the UCSC style. Therefore, the VCF file needed to be edited to remove the term "chr" prior to the chromosome number, thereby creating a VCF file with an NCBI/Ensembl chromosome notation that matches that of the manifest file. This was simply achieved using the "awk" command to substitute "" (i.e. no text) in place of "chr"; "awk" is a command-line text manipulation tool. Finally, this single African germline VCF file (hg19, NCBI/Ensembl chromosome notation) then needed to be compressed by bgzip (creating a vcf.gz file) and indexed by Tabix[42] (creating a vcf.gz.tbi file). These files were then suitable for input into R. This concludes part 1 of the established workflow.

Calling SNP-affected probes. Part 2 is completed in the R environment, using the *GenomicRanges*[43] and *VariantAnnotation*[44] Bioconductor packages. Input includes the bgzip and tabix files as well as the Illumina EPIC manifest file (CSV format). I adapted a method presented in the Pidsley et al. (2016)[5] paper ("Identification of probes overlapping genetic variants") to achieve this part of the analysis. Authors from this paper shared the relevant R script for me to perform the method on my own data. Essentially, the single reference African germline VCF file was parsed to extract all African SNP and indel variants that overlap with EPIC probe coordinates, according to the manifest file. Relevant ranges were created to further classify variants according to the three probe-region categories discussed above and the results were filtered to only include genetic variants with a maximum minor allele frequency (MAF) > 0.05. The adapted R script required the removal of some features: (i) references to variant type as this was not a feature of the African germline VCF file; (ii) references to other populations (e.g. European, Asian); and (iii) missing data values (i.e. those variants that did not meet the MAF threshold). This produced a list of EPIC probes that contain overlapping African variants (**Tables S2**, **S3** and **S4**), which I then filtered from the main EPIC dataset to reduce the risk of false discoveries. This new filtered dataset object was extracted by embedding a filter within the unfiltered object; this way, I could ensure that both the unfiltered and filtered objects could be reused. Ultimately, the established workflow presented here (**Fig. 3-5**), which exists within the larger developed pipeline (**Fig. 3-2**), is a novel and standalone approach to identifying EPIC probes affected by southern African polymorphisms.

Finally, commonality was assessed between the SNP-affected probe lists generated by the two approaches outlined above (presented in **Chapter 4**).

3.3.2.6. Novel filtering of cross-reactive probes

As previously mentioned, the standard function offered by ChAMP for filtering cross-reactive probes is performed according to a 450K array-derived list. To ensure filtering was performed exclusively relevant to the EPIC array, I chose to perform more appropriate cross-reactive probe filtering. This filtering step is novel to the standard ChAMP workflow. A list of cross-reactive probes (n = 43,254) has been annotated for the Illumina EPIC array and is provided in the supplementary data from Pidsley et al. (2016).[5] In order to further reduce the risk of false discoveries, this list was chosen to be filtered from the main EPIC dataset in the same manner as just mentioned for the SNP-affected probe list. Consequently, this is the cross-reactive probe filtering method that was integrated into the African-relevant developed pipeline (**Fig. 3-2**).

3.3.2.7. Correcting for batch effects and biological variables

In any dataset, it is common to investigate possible sources of variation, especially those that may largely and significantly confound results. It is true that methylation is sensitive to a wide range of factors including technical issues such as batch effects (e.g. running samples at different times or using different batches of reagents) as well as certain biological variables[9] such as age or tumour purity. For this reason, batch effects should be avoided and biological sources of variation should be corrected for, where applicable.

Correction functions. In order to identify significant components of variation within one's dataset, singular value decomposition (SVD) may be implemented. The method of SVD applied within the ChAMP package is that by Teschendorff et al. (2009).[45] The champ.SVD() function is able to identify components of variation that correlate with both technical and biological factors of interest within a dataset, provided (epidemiological) data has been collected for biological factors of interest. Covariates should contain at least two values to be tested. Numeric covariates are calculated using linear regression whilst factor and character covariates are calculated using a Kruskal-Wallis test. Batch effects may be corrected for by applying the ComBat[46] function, which only corrects for technical variation, is embedded within the original ChAMP pipeline (champ.runCombat()) and is implemented within the sva package[47] in R. ComBat is a method of further normalisation. It is essential to understand one's own data intimately, especially in terms of confounders, which is why it is critical to identify possible sources of variation. For this reason, an SVD analysis was integrated as part of the novel African-relevant pipeline (**Fig. 3-2**).

Finally, following the SVD analysis, the pipeline reaches the position represented by the black circle in **Figure 3-2**, which symbolises a fully prepared southern African DNA methylation dataset. This marks the point between data processing and data analysis, rendering this suitably processed and normalised southern African dataset ready for analysis.

3.4. Discussion

Here I introduce a novel developed bioinformatic workflow for the processing and normalisation of southern African DNA methylation data. This workflow may be applied to interrogate genome-wide DNA methylation in prostate tissue from men of African ancestry (presented in **Chapter 4**) as well as in tumour tissue from other cancer types. The African-relevant pipeline thoroughly accounts for population-specific genomic diversity that may affect appropriate data filtering. Two distinct approaches were tested for their efficacy in identifying African polymorphisms overlapping EPIC probes, for which I identified MethylToSNP as the lesser of the two approaches for a number of reasons discussed elsewhere (see **Section**

3.3.2.5.1., **Chapter 2** and **Chapter 4**). The SNP-affected probe filtering offered by this novel pipeline is both African-relevant and cohort-specific. Additionally, the developed workflow incorporates EPIC array-relevant cross-reactive probe filtering, which is not offered by the standard ChAMP pipeline. In short, the novel pipeline allows researchers to appropriately process and analyse southern African DNA methylation data while accounting for (i) confounding African-relevant polymorphisms and (ii) unnecessarily eliminating African-relevant data by filtering according to an inappropriate reference population.

The development of a bioinformatic tool such as the one designed in this Chapter, addresses an aspect that is overlooked within epigenetic epidemiology, i.e. the inclusion of African populations. The problem being, African-relevant tools are scarce, if not entirely absent. However, I acknowledge that several limitations exist for this novel pipeline. The small African sample size upon which I developed this pipeline may require modification for processing larger cohorts. I refer specifically to Java memory allocation as well as CPU core allocation. Additionally, the *minfi* package has been described as ideal for analysing small datasets.[48] Future work could assess the suitability of applying *minfi* to a larger dataset. A further notable limitation refers to the SNP-affected probe filtering. While I identified the patient-matched germline variant data method as preferred for such filtering, this approach would require that researchers be in possession of patient-matched genomic data, often absent in studies due to high costs.[14] Finally, although included in the novel pipeline (**Fig. 3-2**), I did not successfully implement the adjustment for potentially confounding biological variables (see **Chapter 4** for more detail), calling for further future studies.

3.5. References

1. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. 2006; 16(3):383–93.
2. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium ® assay. Epigenomics. 2009; 1(1):177–200.
3. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011; 98(4):288–95.
4. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011; 6(6):692–702.
5. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016; 17(1):208.

6. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016; 8(3):389–99.

7. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017; 45(4):1–12.

8. LaBarre BA, Goncearenco A, Petrykowska HM, Jaratlerdsiri W, Bornman MSR, Hayes VM, et al. MethylToSNP: Identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin. 2019; 12(1):79.

9. Wu MC, Kuan PF. A guide to Illumina BeadChip data analysis. In: Tost J, editor. DNA methylation protocols. Methods in molecular biology. New York: Humana Press; 2018. p. 303–30.

10. Daca-Roszak P, Pfeifer A, Żebracka-Gala J, Rusinek D, Szybińska A, Jarząb B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip array: Implications for comparative population studies. BMC Genomics. 2015; 16(1):1003.

11. Naeem H, Wong NC, Chatterton Z, Hong MKH, Pedersen JS, Corcoran NM, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics. 2014; 15(1):51.

12. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamieldien J, et al. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC Genomics. 2014; 15(1):437.

13. Hayes VM, Bornman MSR. Prostate cancer in southern Africa: Does Africa hold untapped potential to add value to the current understanding of a common disease? J Glob Oncol. 2018; (4):1–7.

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

15. Price EM, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013; 6(1):4.

16. Siva N. 1000 genomes project. Nat Biotechnol. 2008; 26(3):256–7.

17. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11.

18. 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.

19. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014; 74(8):880–91.

20. Arnold K, Gosling J, Holmes D. The Java programming language. Addison Wesley Professional;

2005.

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303.

22. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021; 10(2).

23. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8.

24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Available from: http://www.r-project.org/.

25. RStudio Team. RStudio: Integrated development environment for R. RStudio, Inc., Boston, MA. 2020. Available from: http://www.rstudio.com/.

26. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017; 33(24):3982–4.

27. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The international genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 genomes project data. Nucleic Acids Res. 2017; 45(D1):D854–9.

28. Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol. 2013; 14(9):r105.

29. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–9.

30. Maksimovic J, Phipson B, Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data. F1000Research. 2017; 5(1281):1–53.

31. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12(1):R10.

32. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013; 29(2):189–96.

33. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 2012; 13(6):R44.

34. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium

methylation 450K technology. Epigenomics. 2011; 3(6):771–84.

35. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014; 15(11):503.

36. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013; 8(3):333–46.

37. Kasper DH. IlluminaHumanMethylationEPICanno.ilm10b4.hg19: Annotation for Illumina's EPIC methylation arrays. 2017. Available from: https://bitbucket.com/kasperdanielhansen/Illumina_EPIC

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9.

39. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002; 12(6):996–1006.

40. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: A versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014; 30(7):1006–7.

41. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.

42. Li H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011; 27(5):718–9.

43. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9(8):e1003118.

44. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014; 30(14):2076–8.

45. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS One. 2009; 4(12):e8274.

46. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8(1):118–27.

47. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28(6):882–3.

48. Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 2018; 19(S5):115.

Chapter 4: Application of a novel African-relevant genome-wide bioinformatic pipeline to investigate DNA methylation in prostate tissue from men of African ancestry: a pilot study

The underrepresentation of African men in prostate cancer research and the lack of subsequent data and accompanying bioinformatic African-relevant tools has been discussed throughout this dissertation. Although a pilot study, this Chapter aimed to remedy the insufficiency of African-associated prostate cancer epigenomic research and knowledge. By applying the novel African-relevant pipeline established in **Chapter 3**, **Chapter 4** presents an evaluation of genome-wide DNA methylation in prostate tissue from men of African ancestry.

In this Chapter, I would like to acknowledge the patients who consented to donating their tissue for the purposes of this research, as well as the many clinicians and staff associated with the Southern African Prostate Cancer Study. I would also like to thank the Human Comparative and Prostate Cancer Genomics Research team at the Garvan Institute of Medical Research, in Australia, for data availability.

Abstract

Prostate cancer is the second most common cancer in men worldwide and the most prevalent urological cancer affecting South African men. Known risk factors for prostate cancer include increasing age, a family history of prostate cancer and African ancestry. However, there is a severe underrepresentation of Africans in prostate cancer research and as such, contributing factors that link prostate cancer to African ancestry remains elusive. Neither genetics nor environment can solely enlighten on such uncertainty but their interaction may prove revealing. However, research and knowledge on the African prostate cancer epigenome is scarce. DNA methylation has a well-established role in prostate cancer pathogenesis and has been shown to differ amongst prostate tumours derived from different ethnicities. Thus, it may be reasonable to suggest that differential DNA methylation, at least in part, underlies the African ancestral contribution to prostate cancer. Using the Illumina Infinium HumanMethylationEPIC BeadChip, I profiled genome-wide DNA methylation from prostate tissue derived from eight South African men. While appreciating the small cohort, I applied the novel pipeline established in **Chapter 3**, as a result of which, I identified differentially methylated CpG sites that potentially contribute to aggressive prostate cancer in this South African cohort. The novelty of this research is evident in its African-relevance and genome-wide approach.

4.1. Introduction

Globally, prostate cancer (PCa) is the second most common cancer in men, following lung cancer.[1] According to the Global Cancer Observatory (GCO) 2020 estimates[2], 1,414,259 new cases of PCa were reported worldwide, with this figure estimated to increase by 37 % by 2040 (2,235,568). Additionally, the GCO reports southern Africa to display one of the highest mortality rates for PCa in the world, along with a number of other African regions. By 2040, the African PCa mortality rate is expected to rise by 53 %. These estimates illustrate native Africans to be a high-risk population regarding PCa incidence and mortality and this is also true when compared with other ethnicities. With PCa incidence and mortality expected to rise over time and African ancestry being a known PCa risk factor[3], one would expect an abundance of available African-relevant PCa research and accompanying data. However, this is not the case. In fact, African men are notoriously underrepresented in PCa research. Therefore, there is an urgent need to analyse the African PCa (epi)genome to gain insight into the pathogenic nature of African PCa, a field with limited understanding.

DNA methylation is an epigenetic mechanism known to aid the progression of cancer, including PCa.[4,5] DNA methylation is part of a cluster of molecular processes that initiate tumorigenesis and drive its early evolution by altering other molecular processes.[6] While studies have looked at DNA methylation in PCa, most have been limited by targeted gene analysis, with further bias towards non-African cohorts, as just mentioned. Considering the enhanced coverage of more recent genome-wide arrays, such as the Illumina Infinium HumanMethylationEPIC BeadChip (EPIC array), which measures DNA methylation over more than 850K CpG sites genome-wide[7] (see **Chapter 2** for more detail), many studies that have employed a more global approach to DNA methylation analysis are further limited by frequently utilising lower-coverage arrays (e.g. Illumina Infinium HumanMethylation450 and Illumina Infinium HumanMethylation27 BeadChips, see **Chapter 2**). Nevertheless, the 450K array is still the most widely-used platform to investigate and report on epigenome-wide studies. For instance, The Cancer Genome Atlas' (TCGA) repository contains 498 prostate case files generated using the 450K array but zero such files generated using the EPIC array. Further highlighting the scarcity of this data, publicly-available EPIC DNA methylation array data from NCBI's GEO is limited to only 7 studies (as of August 2021), of which none are African-relevant. Due to this bias against African cohorts, African-relevant bioinformatic tools for the processing of African DNA methylation data are limited (addressed in **Chapter 3**).

Overall, it is evident that Africa as a continent is overlooked in PCa (epi)genomic research, thereby necessitating expanded knowledge on this topic. Such research is crucial for ultimately improving clinical approaches to African PCa disease screening, diagnostics and treatment. Therefore, this research aims to apply the novel African-relevant pipeline established in **Chapter 3** on a pilot study of prostate tissue-derived genome-wide DNA methylation data from eight African-ancestral patients from South Africa. Application of this novel pipeline will identify differentially methylated CpG sites that contribute to aggressive prostate cancer in this southern African cohort. While appreciating the small study size, certainly no definitive conclusions can be drawn from the data but given the argument presented above, the novelty of this study cannot be overlooked. The undertaking of a global approach rather than a targeted one, as well as an African focus rather than a European one, has the potential to provide an in-depth understanding of southern African PCa, and at the very least, will act as a foundation on which more sizable studies can be built.

4.2. Materials & Methods

4.2.1. Resource & ethics

Data was made available for eight South African men who consented upon enrolment in the Southern African Prostate Cancer Study (SAPCS). Initiated in 2008, the SAPCS is a unique study that provides an epidemiological, genetic and prostate tissue resource to ultimately define the contributing factors that link

PCa to African ancestry.[8] The previous SAPCS as well as the current study outlined here was reviewed and approved by the University of Pretoria's Human Research Ethics Committee (HREC #43/2010 and #37/2021, respectively). A total of eight patients were recruited at diagnosis and clinicopathologically confirmed as either presenting with high-risk prostate cancer (HRPCa, 7 patients), defined by a Gleason score of ≥ 8, or with benign prostatic hyperplasia (BPH, 1 patient). Prostate tissue was taken at biopsy and all patients self-identified as being of African ethnicity. African ethnicity was further confirmed using ancestry markers. The age distribution of the patients ranged from 54-99 (**Table 4-1**). Tissue-blood pairs were snap frozen and shipped to the Garvan Institute of Medical Research (Sydney, Australia) in accordance with institutional Material Transfer Agreements (MTA) with the University of Pretoria. DNA was extracted from blood and tissue using the commercially available Qiagen DNeasy blood and tissue kit protocol (Qiagen, Maryland, USA). Genomic screening and analysis were performed in accordance with approval granted by St. Vincent's Hospital HREC (SVH/15/227) and governance review authorisation granted for human research at the Garvan Institute of Medical Research (GHRP1522).

4.2.2. Germline & somatic data

DNA extracted from tumour-blood pairs underwent whole genome sequencing (60x, 30x coverage) on the Illumina NovaSeq platform at the Garvan Institute's Kinghorn Centre for Clinical Genomics (KCCG) and was analysed by the Human Comparative and Prostate Cancer Genomics (HCPCG) Research team using in-house pipelines and high-performance compute (HPC) infrastructure provided by the University of Sydney Informatics Hub (SIH) and the National Compute Infrastructure (NCI) in Canberra. Data provided is currently unpublished and funded by the Australian National Health and Medical Research Council (NHMRC).

For each tumour sample, high-confidence somatic variants (single nucleotide variants and indels) were called against patient-matched blood samples using GATK's Mutect2[9] (v.2.2). Calls were additionally filtered to label false positives with a list of failed filters and true positives with "PASS". Variant called somatic data (VCFv4.2 format) for the eight patients was made available and I further filtered the VCF files using VCFtools[10] (v.0.1.16) to only contain true positive variants. Additionally, the HCPCG Research team used Mutect2 to extract C:G > T:A somatic variants that lie within a CpG context, providing flanking bases (3 bp) and referencing hg38. I then filtered and counted these CpG C > T variants (**Table 4-1**) within the R statistical environment using the *dplyr* package.[11] In addition, variant called germline data (VCFv4.2 format) for the eight patients was made available. The VCF files reference Genome Build 38 (hg38) and chromosome notation is of the UCSC style (e.g. chr1) versus the NCBI/Ensembl style (e.g. 1). The tools used for germline variant extraction includes Java[12] (JDK™, v.1.8.0_111), GATK[13] (v.4.1.4.1), HTSlib[14] (v.1.10.2) and VCFtools[10] (v.0.1.14).

In addition to raw variant data, summary data was provided for the eight patients, as summarised in **Table 4-1**, including (besides age and pathology) genomic-derived features such as: tumour purity, tumour mutational burden (TMB), percentage of genome alteration (PGA), structural variant (SV) calls and microsatellite (in)stability (MSI/MSS) status. TMB refers to small somatic mutations; it is defined by the total number of small somatic variants, divided by genome size 3,088 Mbp, as previously described by Jaratlerdsiri et al. (2018).[15] Small variants include single nucleotide variants (SNVs) and indels < 50bp; SVs refer to alterations (gain or loss events) > 50bp. PGA refers to the sum of the number of base pairs altered by SVs for each patient, divided by genome size 3,088 Mbp.[15] Tumour purity was estimated for each patient based on WGS data (a combination of somatic SNV and SCNA data), using Sequenza software and the THetA2 program.[16,17] A five-tooled MetaSV analysis was used to detect high-confidence somatic SVs, as previously described.[18] Finally, MANTIS v1.0.5 was used to detect MSI versus MSS calls.[19]

Table 4-1 Clinicopathological, sequencing and somatic variation data for African patients.

		UP2037	UP2048	UP2039	UP2099	UP2113	UP2116	UP2119	UP2133
Clinical presentation at diagnosis									
	Age (years)	65	59	71	76	88	99	54	58
	Pathology	BPH	HRPCa	HRPCa	HRPCa	HRPCa	HRPCa	HRPCa	HRPCa
	Gleason score	N/A	9	8	8	8	10	8	9
Somatic variants (total number)									
	TMB per Mbp	0.04	0.05	2.56	2.62	59.61	2.10	0.79	4.13
	SVs	6	3	5	134	59	61	56	492
	C > T variants	13	15	1055	1006	33507	921	443	1758
	CpG C > T variants	5	8	354	330	15960	361	235	680
Somatic alterations									
	PGA	0.004	0.005	0.168	0.233	0.013	0.164	0.009	0.252
	MSI	No	No	No	No	Yes (MSI-H)	No	No	No
WGS statistics									
	Tumour purity	0.37	0.44	0.64	0.79	0.38	0.43	0.45	0.40

BPH: benign prostatic hyperplasia | HRPCa: high-risk prostate cancer | MSI-H: microsatellite instability-high | PGA: percentage of genome alteration | SV: structural variant | TMB: tumour mutational burden | UP0000: African patient identifier | WGS: whole-genome sequencing

4.2.3. DNA methylation data & processing

Raw DNA methylation data was generated from tissue DNA for the eight African patients at the Australian Genome Research Facility (AGRF, Melbourne, Australia) and subsequently provided by the HCPCG Research team at the Garvan Institute. DNA methylation was quantified using the Illumina Infinium HumanMethylationEPIC BeadChip (hereafter referred to as the EPIC (micro)array) following the Illumina Infinium HD Methylation Assay (Illumina, CA, USA). The data provided by the AGRF included raw Illumina intensity data (IDAT) files, the Illumina manifest file (v1.0 B5, BPM format), a sample sheet (CSV format) and a genotyping service report from the research facility.

I processed and analysed the African EPIC data using the novel African-relevant bioinformatic pipeline developed in **Chapter 3** (**Fig. 3-2**). Briefly, raw data was loaded into R using the *minfi* method[20] accompanied by a number of function parameter modifications (see **Section 3.3.2.2.** for more detail). During initial filtering, some probe exclusion criteria was specified: (i) probes with detection *p*-values greater than 0.05; (ii) probes that failed the detection *p*-value threshold in more than 20 % of samples; and (iii) probes with a bead count < 3 in at least 5 % of samples. Sex chromosomes were retained and no missing values were present in the data thus no imputation for missing values was performed. Quality control tests confirmed all 8 samples suitable for inclusion. Normalisation of data was performed using the selected beta-mixture quantile (BMIQ)[21] method and beta-values were chosen for their direct biological interpretation. I identified SNP-affected probes using both the MethylToSNP and patient-matched germline variant approaches (see **Section 3.3.2.5.**), tested the performance of both approaches and assessed the commonality between the probe lists generated by each of the methods. For the latter approach, I filtered results to only include genetic variants with a maximum minor allele frequency (MAF) > 0.05. Consequently, I filtered the probe list generated by the patient-matched germline variant method. As per the novel African-relevant pipeline, cross-reactive probes were filtered according to Pidsley et al. (2016).[7] Finally, a singular value decomposition (SVD) analysis was chosen for application to identify significant components of technical and/or biological variation within the African dataset.

4.2.4. Identifying differentially methylated probes

The analyses detailed below were carried out by myself and using the hg19 genome assembly. Initial visualisation of the processed EPIC dataset was generated using the QC.GUI() function integrated in the African-relevant pipeline (**Fig. 3-2**) to produce multidimensional scaling (MDS) plots. Differential methylation analysis was then performed between African HRPCa patients versus the single BPH patient, MSI-H versus MSS tumours as well as among a range of ages, tumour purity predictions, TMB measurements, PGA measurements, SV counts, Gleason scores, C > T mutation counts and CpG C > T mutation counts (as presented in **Table 4-1**). I used the champ.DMP() function to identify significantly differential methylated probes (DMPs) for a particular variable of interest. It was necessary to recode categorical variables to numeric variables (i.e. HRPCa = 1 and BPH = 0) and I then used DMP.GUI() to visualise the results. Linear regression was conducted on each CpG site within the African dataset to identify covariate-related CpG sites and for visualisation, the function grouped numeric variables into intervals. It was decided that DMPs be selected based on a BH-adjusted threshold of $p < 0.05$. The Benjamini-Hochberg *p*-value adjustment controls the false discovery rate.[22] DMPs were categorized as displaying hypermethylation (beta ≥ 0.8), partial methylation (beta ~ 0.5) or hypomethylation (beta ≤ 0.2), as per recommendations from Du et al. (2010).[23] For each covariate, I chose the top three genes (most abundant for significant CpGs) for closer DNA methylation pattern analysis.

DMPs were annotated according to CpG island (CGI) and gene regions. Annotations for CpG islands includes CGIs, CGI shores (<2 kb upstream and downstream of CGIs), CGI shelves (2-4 kb upstream and downstream of CGIs) and open sea (non-CGI-related sites), (see **Fig. 2-1**). Gene region annotations includes TSS1500 (200-1500 bp upstream of the transcription start site, TSS), TSS200 (up to 200 bp upstream of the TSS), 5'UTR (5' untranslated region), 1st exon, Body (gene body), ExonBnd (exon boundaries), 3'UTR and IGR (intergenic regions).

All data processing and analyses were performed with R[24] ≥ v.4.0.2 and RStudio[25] ≥ v.1.3 statistical software.

4.3. Results

4.3.1. Processing the African dataset using a novel African-relevant bioinformatic pipeline

4.3.1.1. Assessing the extracted dataset and normalisation

Prior to any initial filtering, the EPIC microarray consists of 867,531 probes for genome-wide DNA methylation quantification. After data loading and initial filtering using the champ.load() function (see **Fig. 3-2**), 13,683 probes were filtered out, rendering an extracted dataset made up of 853,848 probes. Probes that were filtered out did not meet the selected criteria specified by the developed pipeline (see **Table 3-1**) including: (i) probes that failed in individual samples (detection *p*-value > 0.05, n = 2,024 probes); (ii) probes that failed the detection *p*-value threshold in greater than 20 % of the samples (n = 1,293 probes); and (iii) probes that had a bead count < 3 in at least 5 % of the samples (n = 10,366 probes). A detection *p*-

value threshold of 0.01 removed 2,650 probes. Since this number of rejected probes was so similar to the number rejected by the detection *p*-value threshold of 0.05, I proceeded with the recommended threshold of 0.05.[26] Of course, it is also possible that entire samples fail due to inadequate input DNA concentrations or other processing issues[26]; however, all 8 samples in this analysis were retained. Additionally, there were no missing values in this data matrix, thus imputation was not necessary.

Of the 853,848 probes initially loaded into R, most lie within the open sea when considering CGI annotations (**Fig. 3-3a.ii**) and within gene bodies when considering gene region annotations (**Fig. 3-3a.iii**). Evidently, probe distribution is largely similar before (**Fig. 3-3a**) and after (**Fig. 3-3b**) complete filtering and normalisation. One must inspect these distributions to ensure evenly distributed probe filtering was performed i.e. filtering was not biased towards any particular region of the genome. Of course, the difference between probe distributions before and after complete filtering and normalisation can be observed in probe quantities. Additionally, I found that BMIQ normalisation of the data across the eight samples reduced the variability seen in the beta-value distributions of Infinium type I and Infinium type II probes (**Fig. 3-4b**), and although normalisation of Illumina methylation data is usually focused on within-sample correction, it is evident that the BMIQ normalisation method selected and employed here reduced variability between samples too (**Fig. 3-4d**).

4.3.1.2. Identifying and filtering SNP-affected probes: MethylToSNP versus the novel patient-matched germline variant data method

MethylToSNP. Application of the tool MethylToSNP for SNP-affected probe filtering identified 14,474 potential SNPs (**Table S1**) underlying methylation probes in the southern African cohort, 3,572 of which MethylToSNP labelled as high-confidence predictions (reliability score ≥ 0.5, see **Table 4-2**). Notably, I found the median reliability score of the 14,474 SNPs to be 0, suggesting that most of these sites are not viable meC > T SNP candidates. Of the high-confidence predictions, 2,963 SNPs were identified as being potentially novel in their absence of dbSNP[27] release 147.

The patient-matched germline variant data method. Using the developed patient-matched germline variant data workflow (see **Fig. 3-5**), I identified a substantially larger number of 179,918 EPIC probes as having SNP or indel variants overlapping them (maximum MAF > 0.05) (**Tables S2**, **S3** and **S4**). More specifically, the number of SNPs affecting EPIC probes were further classified according to identification at a target CpG site (n = 30,483 SNPs) (**Table S2**), at an SBE site (for Infinium type I probes, n = 918 SNPs) (**Table S3**) and within the probe body (n = 174,316 SNPs) (**Table S4**). In considering the total number of SNPs identified, I found that a number of EPIC probes overlapped more than one polymorphism (n = 25,799 probes). To note, as part of the patient-matched germline variant data SNP-affected probe filtering method, liftover of the single reference germline VCF file from hg38 to hg19 using Picard LiftoverVcf (see **Section 3.3.2.5.2.**) resulted in 91,994 variants being rejected. This amounts to a minimal 0.66 % loss of SNPs and indels that could not be lifted over to the target genome assembly. This is expected considering sequence incompatibilities between the source (hg38) and target (hg19) reference genomes.

Table 4-2 MethylToSNP predictions in the southern African cohort.

Description	*Probes tested (#)*	*SNP predictions (#)*	*Overlap with dbSNP 147 (#)*	*Potential novel SNPs (#)*	*Median reliability score of predicted SNPs*
All probes sites	844,706	14,474	2,555	11,894	0.00
MethylToSNP identified probes with reliability scores ≥ 0.5	14,474	3,572	609	2,963	0.56

Adapted from "MethylToSNP: identifying SNPs in Illumina DNA methylation array data," by LaBarre et al., 2019, *Epigenetics & Chromatin*, *12*(1), p. 7. Copyright 2019 by The Authors.

4.3.1.2.1. Assessing commonality between the two SNP-affected probe filtering approaches

I investigated the degree of overlap between the high-confidence SNP-affected probe lists generated by the two above methods and found a total of 1,089 shared probes. Considering African patient-matched germline variant data was used in one approach, one would expect all MethylToSNP-identified probes to be common in the African reference probe list assuming accuracy of the MethylToSNP tool. However, it appears as though while MethylToSNP identified a number of true SNPs, another 2,483 SNPs weren't accounted for in the African germline variant data. As previously mentioned, MethylToSNP filtering was not selected for application to the African dataset; limitations are discussed in **Chapter 2**. Instead, the African reference approach was selected to proceed in order to ensure only cohort-specific, African-relevant probes were filtered out. Of course, a number of probes present on this filtering list may have already been removed from the dataset during data extraction.

4.3.1.3. Filtering SNP-affected probes and cross-reactive probes

Once I had performed SNP-affected probe filtering using the African reference approach, as well as cross-reactive probe filtering, a total of 653,337 probes remained for data analysis. From beginning (raw data) to end, this amounts to a loss of approximately 25 % of the total probes.

4.3.1.4. Identifying components of variation

An SVD analysis was run as per the developed workflow (**Fig. 3-2**) to identify the source and nature of both technical and biological variation within the African DNA methylation dataset (**Fig. 4-1**). It was necessary to add epidemiological data to phenotypic data in the pd file in order to include biological variables in the analysis. The analysis revealed two principal components to explain the majority of the variance (> 98 %) observed in the dataset. In **Figure 4-1a**, pale pink blocks ($p < 0.05$) indicate a significant correlation between the deconvoluted components and select covariates. To elaborate, I found that principal component 1 significantly correlated with PGA ($p = 0.0352$) and explained ~77 % of the variation in the dataset whilst principle component 2 significantly correlated with tumour purity ($p = 0.0327$) and was able to explain ~23 % of the variation (**Fig. 4-1b**). Although no batch effects were identified as part of the SVD analysis, this finding highlights substantial contributions from confounding biological factors that ideally should be adjusted for. The absence of batch effects in this dataset was expected. Because this African cohort is comprised of only eight individuals, samples were all run on the same array plate at the same time, minimising potential batch effects and negating the need for ComBat implementation.

Although the standard ChAMP package offers a function to correct for cell-type heterogeneity, it may only be implemented on samples derived from whole blood and as a result, is not applicable for tumour-derived data. However, beyond the scope of the ChAMP pipeline, I suggest that the sva package may be applied to model and correct for biological sources of variation such as tumour purity (see **Chapter 2** for further discussion). Regrettably, I was not able to successfully perform an adjustment for the confounding biological variables due to several reasons including project time constraints (see Discussion).

4.3.2. Pilot interrogation of DNA methylation in prostate tissue from eight southern African men

Once data processing and normalisation had been successfully completed with the novel African-relevant pipeline, analysis of this dataset identified significant DMPs for a number of covariates within the small African cohort (**Table 4-3**). Of particular interest were DMPs I identified between the single BPH individual and the HRPCa patients. In this case, only 4 DMPs were identified, with the BPH individual (UP2037) displaying overall partial methylation at these 4 CpG sites (β = ~0.4) and the HRPCa patients displaying hypermethylation at 2 CpG sites and hypomethylation at the other 2 CpG sites (**Fig. 4-2**). Probe cg04295700 overlaps with the TSS1500 of the *CMSS1* gene whereas the remining 3 probes fall within IGRs in the open sea. Although only 4 probes (i.e. CpG sites) were identified here, it is still evident that a difference in DNA methylation exists between HRPCa and BPH (albeit small). It is worth mentioning that I expected a greater number of DMPs.

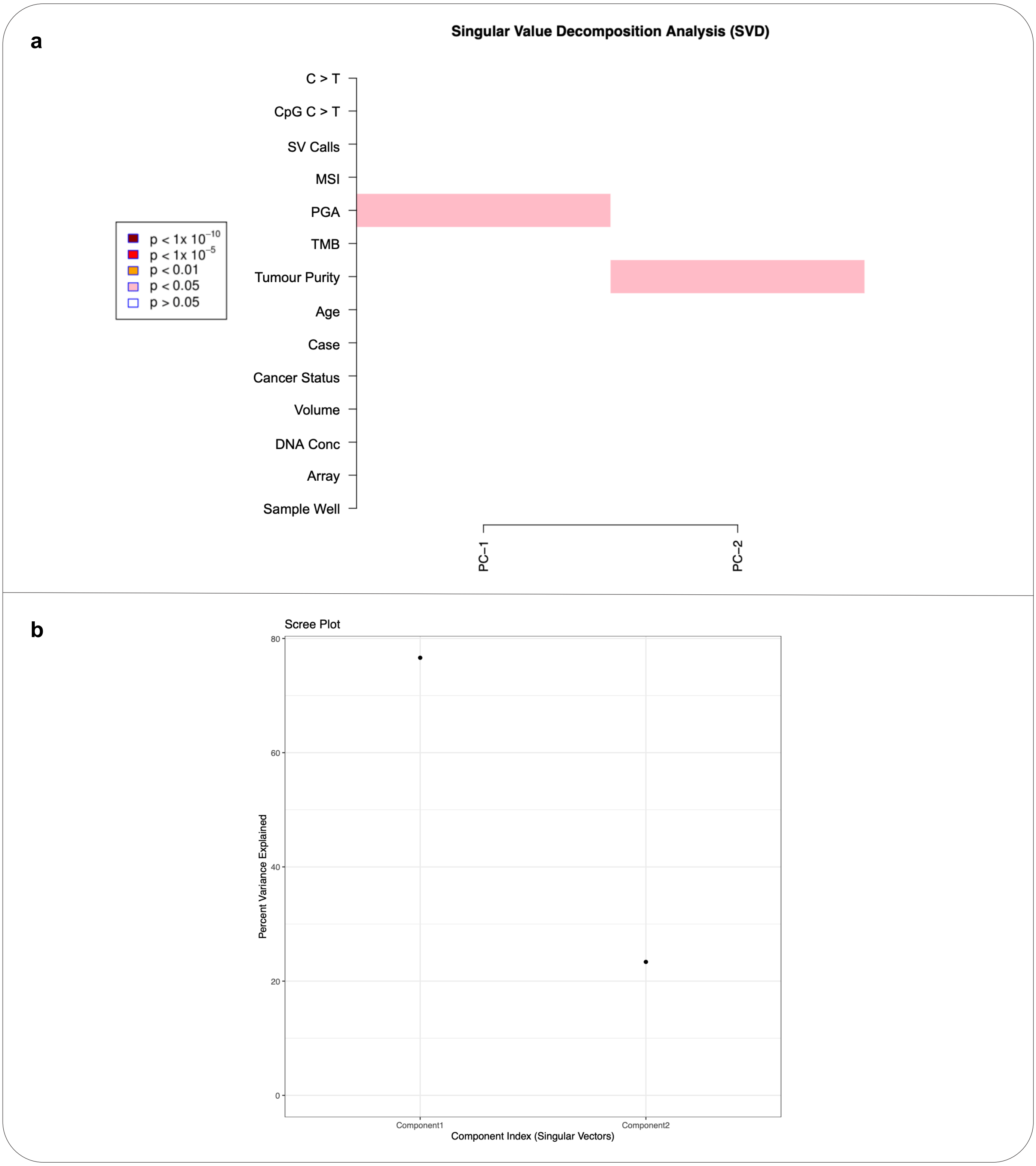


Fig. 4-1 Singular value decomposition analysis identifying the **a** source of and **b** percentage contributed by significant components of variation observed in the African DNA methylation dataset. Two principal components were identified. A *p*-value < 0.05 was regarded as statistically significant.

Cancer status: Gleason score | Case: PCa or BPH | DNA conc: DNA concentration | MSI: microsatellite (in)stability | PC: principal component | PGA: percentage of genome alteration | SV: structural variant | TMB: tumour mutational burden | Volume: DNA volume

An MDS plot was selected to initially visualise the prepared African dataset (**Fig. 4-3**). The MDS plot is a visual representation of the level of similarity of individual cases of a dataset. In terms of methylation patterns, while we would expect the BPH individual to map distinctly from the HRPCa patients, **Figure 4-3** shows that the BPH individual maps with two of the HRPCa patients, suggesting the three individuals share similar methylation patterns. This finding is interesting considering DNA methylation profiles have been shown previously to accurately distinguish between PCa and BPH tissue samples.[28,29] The DMP results related to tumour purity could shed light on this finding.

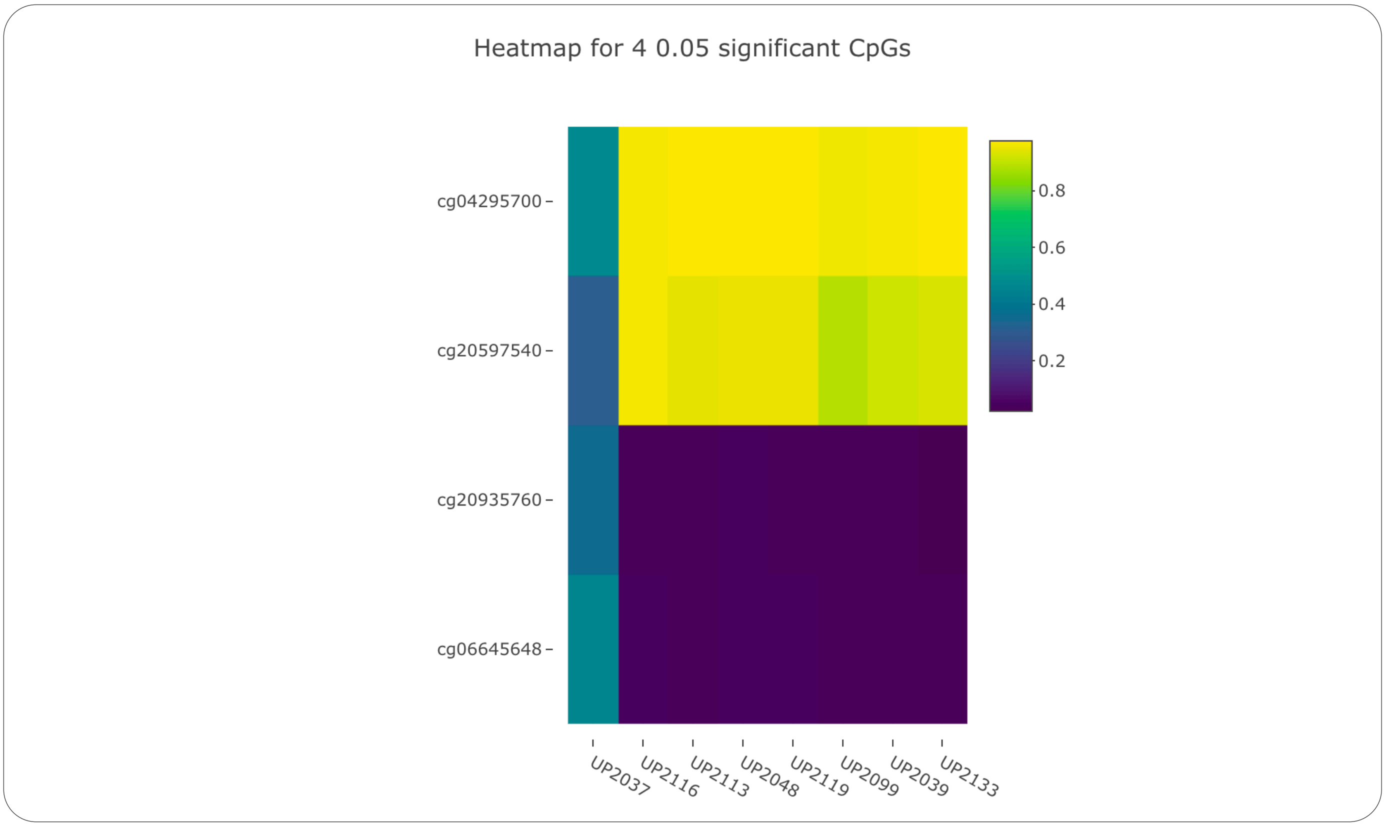


Fig. 4-2 Heatmap displaying significant differentially methylated probes (cg00000000) between a single African BPH individual (UP2037) and seven African HRPCa patients ($p < 0.05$).

I identified a large number of tumour purity-related CpG sites within this African cohort (**Table 4-3**). The DMPs identified relating to this covariate were the most abundant of all covariates tested, highlighting the confounding nature of this variable. Once again, the BPH individual was not distinct from the HRPCa patients; rather, I found that the BPH sample grouped with the HRPCa samples in the lower tumour purity interval, all of which displayed similar methylation patterns for the identified CpG sites (**Fig. S2a**). The fact that the BPH sample has a tumour purity estimation and that I found it to

Table 4-3 Differentially methylated probes identified by the novel African-relevant pipeline for a number of covariates in African prostate cancer.

Covariate	*Number of significant CpGs identified ($p < 0.05$)*	*Overall methylation pattern for significant CpGs*	*Gene feature & CGI region for the highest proportion of CpGs*	*Top genes enriched for significant CpGs*
HRPCa versus BPH	4	BPH: partial-to-low methylation. HRPCa: one probe hypermethylated, two probes hypomethylated.	IGR & TSS1500. Open sea.	*CMSS1*
Age	None	-	-	-
Tumour Purity	14704	Lower TP group (.37-.45): hyper-to-partial methylation & few hypomethylated probes. Higher TP group (.64-.79): partial-to-hypomethylation.	Gene body, IGR & TSS1500. Open sea & shore.	*ADARB2, NTM, BCL11A*
TMB	4078	Outlier TMB (UP2113, 59.61): partial methylation. Non-outlier TMBs: three quarters hypermethylated, one quarter hypomethylated.	Body, IGR & TSS200. Open sea & island.	*TCF4, MECOM, XYLT1*
PGA	996	Lower PGA group (.004-.013): distinct regions of hypo- & hypermethylation. Higher PGA group (.233-.252): distinct regions of hypo- & hypermethylation, opposite to that of lower PGA group. Intermediate group (.164-.168): partial methylation.	Gene body & IGR. Open sea.	*DSCAML1, CAMTA1, GLT1D1*
SV Calls	4810	High SV call individual (UP2133, 492): partial methylation. Other individuals: two thirds hypermethylated, one third hypomethylated.	Body, IGR, TSS200. Open sea & island.	*GABBR1, RASA3, ACACB*
Gleason Score	None	-	-	-

Table 4-3 (continued) Differentially methylated probes identified by the novel African-relevant pipeline for a number of covariates in African prostate cancer.

MSI-H versus MSS	4128	MSI-H tumour (UP2113): partial-to-hypomethylation. MSS tumours: three quarters hypermethylated, one quarter hypomethylated.	Body, IGR & TSS200. Open sea & island.	*TCF4, MECOM, XYLT1*
CpG C > T Count	4112	Outlier (UP2113, 15,960): partial methylation. Non-outliers: three quarters hypermethylated, one quarter hypomethylated.	Body, IGR & TSS200. Open sea & island.	*TCF4, MECOM, XYLT1*
C > T Count	4115	Outlier (UP2113, 33,507): partial methylation. Non-outliers: three quarters hypermethylated, one quarter hypomethylated.	Body, IGR & TSS200. Open sea & island.	*TCF4, MECOM, XYLT1*

BPH: benign prostatic hyperplasia | CGI: CpG island | HRPCa: high-risk prostate cancer | IGR: intergenic region | indel: insertions and deletions | MSI-H: microsatellite instability-high | MSS: microsatellite stability | PGA: percentage of genome alteration | SV: structural variant | TMB: tumour mutational burden | TP: tumour purity | TSS: transcription start site consistently map with HRPCa samples (for all covariates tested, **Fig. S2**), strongly suggests that this histopathologically normal sample harbours underlying aberrant DNA methylation that may be predictive of PCa.

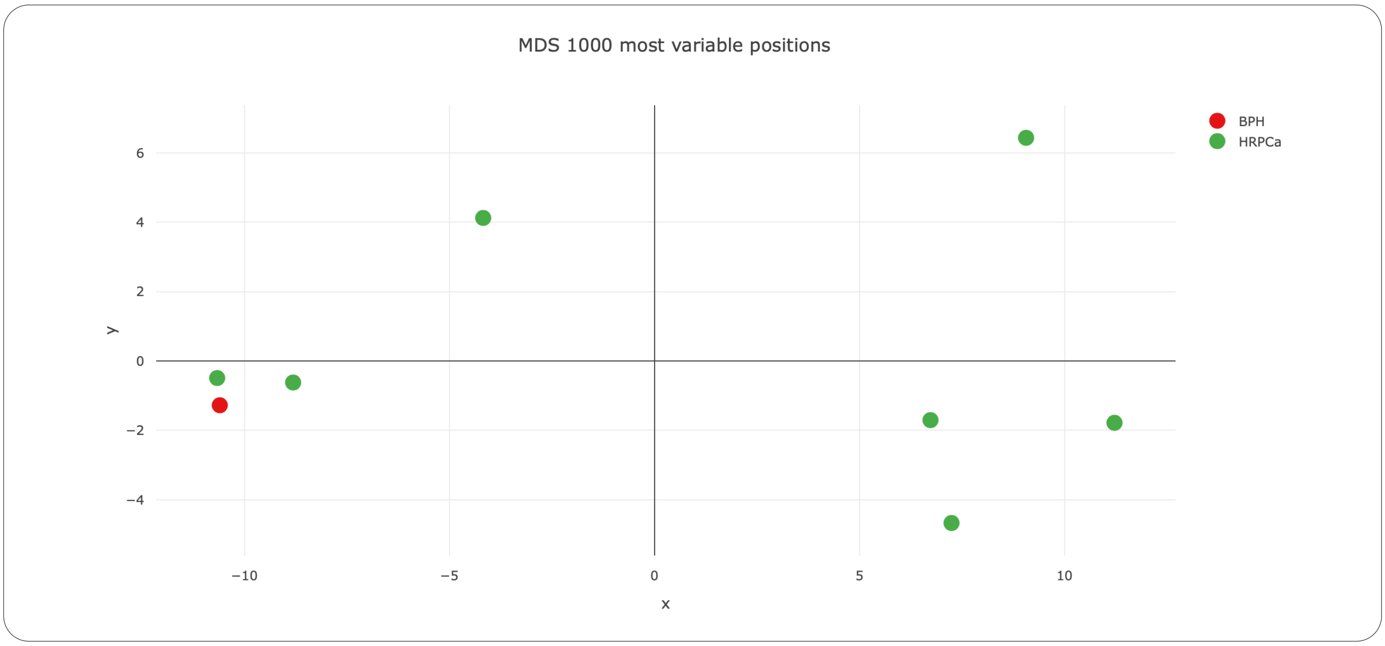


Fig. 4-3 Multidimensional scaling plot of the 1,000 most variably methylated CpG sites between a single African BPH individual and African HRPCa patients.

BPH: benign prostatic hyperplasia | HRPCa: high-risk prostate cancer

While no significant DMPs were identified for the covariates age and Gleason score, it is evident that TMB, MSI-H versus MSS, the CpG C > T variant count and the C > T variant count are all highly correlated, as expected. This can be seen in the number of DMPs identified and the top genes enriched for significant CpG sites (**Table 4-3**). Additionally, a number of DMPs were identified for the covariates SV calls and PGA.

4.4. Discussion

In this pilot study, I successfully applied the novel African-relevant genome-wide bioinformatic pipeline, established in **Chapter 3**, to tissue-derived DNA methylation data generated from eight South African men. The pipeline suitably processed African data and generated a fully prepared methylation dataset for analysis. I assessed the performance of two SNP-affected probe filtering approaches, which consequently revealed the patient-matched germline variant data method to produce a more comprehensive probe list for filtering. The workflow was able to identify differentially methylated CpG sites contributing to aggressive prostate cancer in South African men, although no real conclusions can be drawn from findings presented here due to the small study size. However, this tool is a first-of-its-kind and provides a foundation upon which larger studies may expand in future.

In terms of SNP-affected probe filtering and considering some methods remove as much as 60 % of array probes, an advantage MethylToSNP claims to offer is the prevention of excessive unnecessary probe removal. Comparing the two approaches outlined in this study, namely MethylToSNP and the patient-matched germline variant data method, the former method identified a substantially lower number of African-specific SNP-affected probes (3,572 vs. 179,918, respectively). This may be explained by the principle underlying the algorithm of MethylToSNP, in which there is a bias towards identifying methylated C > T polymorphisms in methylation data.[30] As such, a number of other variant types, particularly C > T polymorphisms where the C is never methylated, are likely to be overlooked or completely missed. Additionally, SNP-associated patterns other than the three-tier pattern (discussed in **Chapter 2**), are not currently supported by MethylToSNP.[30] Overall, I believe MethylToSNP offers an underestimation of the true number of Illumina array probes that overlap polymorphisms. In highlighting this, it is my opinion that MethylToSNP is not a sufficient tool if used exclusively. Because of its underestimation in identified SNP-affected probes within a dataset, should this tool be used for DNA methylation data filtering, it should be used in conjunction with other, more thorough SNP-affected probe identification methods. Conversely, using the patient-matched germline variant data method developed in **Chapter 3**, a comprehensive list of African-relevant, cohort-specific variants were identified as overlapping with EPIC probes. I found the number of SNP-affected probes identified in this African cohort to be slightly higher than but still consistent with alternative methods using general references such as dbSNP[30] (~100-144K) or 1000 Genomes[7] (~110K) as a variant reference. Because polymorphic sites are known to influence DNA methylation quantification, it is essential that one performs comprehensive filtering of SNP-affected probes.

To further comment on results generated by MethylToSNP, I identified a total of 2,963 SNPs as potentially novel in their absence of dbSNP[27] release 147. As new information is obtained by dbSNP, vast amounts of new variants are incorporated, released and validated in an updated "build".[27] The dbSNP build 147 was made available in 2016 and more recently, the dbSNP build 153 was made available in 2019. Therefore, should a more updated dbSNP annotation reference for EPIC array data become available, the total number of novel SNPs identified by MethylToSNP would likely decrease.

The SVD analysis revealed PGA and tumour purity to be significant confounders within the African DNA methylation dataset. Interestingly, genomic alterations have been suggested to play a role in mediating changes in the PCa epigenome, specifically with regard to DNA methylation.[31] However, if this is true in the context of PCa development and progression, to adjust for this may actually remove African PCa-relevant methylation signals that are of interest to this study. On the other hand, it is not surprising that I found tumour purity to be a significant confounding biological factor within this DNA methylation dataset. After all, cell-type heterogeneity is an issue that confounds all DNA methylation studies.[26,32,33] It is unfortunate though that as part of this study, I was unable to implement an adjustment for these confounders using the sva package in R, that I later discovered. In part, this was due to project time constraints i.e. I would have needed a few more weeks to successfully run the adjustment on the African methylation dataset. However, I also believe the filtering performed throughout this pipeline to have been rather thorough, so much so that I would fear any further corrections may result in a loss of significant probes when analysing DMPs. Due to the small cohort size, further adjustments may be too harsh for this dataset. As previously discussed, corrections of this nature may actually adjust out the signal one is searching for[26] i.e. should significant components of variation be highly correlated with the phenotype of interest, one may choose to ignore said variation.

An objective for this research was to assess whether the novel African-relevant pipeline would be able to identify differentially methylated CpG sites in the aggressive PCa African cohort. My analyses revealed a number of significant DMPs for several covariates within the African cohort (**Table 4-3**), providing very preliminary evidence for the role of DNA methylation in aggressive African PCa. Again, I must note that these findings should be interpreted with great caution due to the low power of this study. Comparison of DNA methylation at CpG sites between the BPH individual and HRPCa patients identified only 4 significant DMPs. In addition, I found that the BPH individual consistently displayed DNA methylation patterns similar to that of the HRPCa patients for all covariates (barring BPH versus HRPCa). DNA methylation changes are said to occur early in cancer development, are present in non-malignant cells contiguous with cancerous tissue, leading to a field effect[5,34] and can even distinguish BPH from PCa samples.[28] In light of this, this lack of distinction between the two groups seen here is interesting given that BPH is not considered to be a precursor of PCa nor does it increase your risk of developing prostate cancer.[35,36] However, it is the role of the histopathologist to classify samples as either BPH or cancerous, thereafter allocating malignant samples with a Gleason score, although this is subjectively done. Therefore, I believe it is possible that this particular BPH sample, although not displaying any abnormal cell histomorphology, is in fact characterised by underlying aberrant methylation that may be predictive of PCa development. This very occurrence of epigenomic alterations in benign tissue being able to act as a marker for PCa prediction has been reported previously.[37]

My analysis of DMPs related to tumour purity revealed an overwhelming number of significant CpG sites that are associated with this covariate. The number of DMPs identified was roughly three times as many DMPs that were identified for the next most abundant covariate (i.e. SV Calls). Ideally, one would expect few-to-no CpG sites associating with tumour purity and this is especially true for BPH, in which case we'd expect an absence of cancerous cells. However, this observation highlights the immense influence this confounding variable wields on this African dataset. It is worth noting that tumour purity estimates for the 8 African individuals were highly variable (0.37-0.79, *Mdn* = 0.44), and it is true that DNA methylation studies typically include samples with high tumour purities, such as those greater than 95 %[33] (as measured by a pathologist), in order to limit potential confounding. However, it should be noted that tumour purities measured by pathologists are usually higher than sample-matched estimates derived from WGS data. Interestingly, the BPH individual was assigned a tumour purity estimation (0.37), which was predicted based on WGS data. This alone indicates underlying somatic and epigenetic alterations in this sample that precedes any visual cell abnormalities. One may question the age of this particular BPH individual and whether or not the tumour purity estimation could be explained by age-accumulated mutations. However, the BPH individual was the fourth-youngest patient in the cohort at just 65 years old, making it unlikely that his tumour purity estimation, which is similar to that of HRPCa patients, is due to age-accumulated alterations alone. To further support the malignant nature of this BPH-classified sample is the MDS plot of the 1,000 most variably methylated CpG sites between the BPH individual and HRPCa patients (**Fig. 4-3**), in which it is evident that the BPH individual maps closely with two HRPCa patients, rather than mapping distinctly from them.

In a recent study by Parry et al. (2019)[29], DNA methylation analysis was conducted on adjacent benign and prostate tumour cores for six patients. In terms of DNA methylation, they found clear distinctions between benign and tumour cores with the exception of a single core in two separate cases. They observed these two tumour cores appearing more similar to benign cores in heatmaps similar to that shown in **Figure 4-2** and in an MDS analysis similar to that shown in **Figure 4-3**, said tumour cores mapped closely to benign cores. Owing to the fact that in each case, these cores were sampled from the same patient, authors suggest this finding to be explained by a cancer-proximity field effect or field cancerization i.e. aberrant methylation present in the contiguous benign cores. However, apart from these two exceptions and as shown in a previous study[28], DNA methylation was able to distinguish between benign and tumour cores[29], highlighting the existence of distinct methylation profiles between these two tissue states. Although a field effect cannot be cited in this current study to explain the observation of the African BPH individual, the findings by Parry et al. (2019)[29] further supports the notion that the African BPH individual included in this current study does in fact contain underlying epigenomic changes with a likeness to that of HRPCa. Should a true BPH individual have been included in this African study (i.e. one with a tumour purity close or equal to zero), it is likely that I would have observed a higher number of significant DMPs between BPH and HRPCa. I believe the finding in **Figure 4-2** to be limited by the particular BPH individual chosen for inclusion in this study; in my opinion, this BPH individual is not enough of a contrast to the HRPCa patients in terms of DNA methylation so I was not able to identify any more significantly associated CpGs in this case. Furthermore, to sufficiently identify aggressive PCa-associated CpG sites, I would suggest a comparison of true controls (i.e. non-cancerous, non-BPH samples) with HRPCa samples. Of course, this may be difficult to achieve considering healthy men are not likely to provide a biopsy sample should it not be directly necessary for their own healthcare.

Evidence presented in **Table 4-3** suggests high correlation between the covariates TMB, MSI-H versus MSS, CpG C > T count and C > T count. All covariates share similar numbers of DMPs identified as well as the top genes enriched for significant CpG sites. The correlation between TMB and variant counts is expected considering TMB is defined as the total number of small somatic variants, divided by genome size 3,088 Mbp.[15] Moreover, a correlation between TMB and MSI-H may also be expected since MSI refers to cells that have a high number of mutations within microsatellites, caused by the loss of DNA mismatch repair activity.[38] Deficient DNA mismatch repair in PCa can result from mutational inactivation or epigenetic silencing of any genes within the mismatch repair pathway. Interestingly, this finding for MSI provides a direct link between DNA methylation in African HRPCa and deficient DNA mismatch repair activity. Although only a single African HRPCa individual was of MSI-H status, MSI is used as a biomarker indicative of deficient DNA mismatch repair.[38] However, it has been suggested that low tumour purity (< 70 %) could confound the identification of MSI status in gastric and colon cancer.[39] The single MSI-H individual displayed a low tumour purity (0.38). Thus, these findings are likely insignificant.

The DMP analysis further revealed SVs and PGA to have a number of significant CpG sites associated with these covariates. As mentioned previously, genomic alterations have been suggested to play a role in mediating changes in the epigenome in PCa and this is particularly true when considering DNA methylation.[31] Dhingra et al. (2017)[31] propose that SVs dysregulate transcription factor hubs within the prostate regulatory network. Subsequent crosstalk between the dysregulated transcription factor hub expression can lead to DNA methylation changes, thereby promoting global expression changes in the network. One model to explain this proposes that upregulated transcription factor expression leads to increased recruitment of histone H3 lysine 4 (H3K4) methyltransferase, which will protect bound regions from methylation.[40] Conversely, upregulated transcription factor expression may be associated with increased recruitment of DNA methyltransferases, which promotes methylation at bound regions.[41] Either way, SVs may be the initiators that disrupt the expression of transcription factor hubs and DNA methylation changes promote gene expression changes in the regulatory network.[31] Overall, these changes support prostate tumorigenesis. Ultimately, the true carcinogenic effect of SVs in PCa is reflected not by the number of SV events that take place in a genome, but rather the cumulative effect of the number of base pairs affected by such events, as defined by the PGA.

A large number of the top genes enriched for significant CpGs that were associated with the respective covariates, as outlined in **Table 4-3**, have not been mentioned in relation to PCa in existing literature (as of July 2021) and if mentioned, I have found that it is not typically in the context of aberrant DNA methylation; furthermore, no mentions were made in regard to African (American) PCa. *CMSS1* was a top gene identified in the BPH versus HRPCa DMP analysis; under normal conditions, *CMSS1* (CMS1 ribosomal small subunit homolog) has been shown to interact with *MDM2*[42], a negative regulator of the tumour suppressor p53, thereby potentially promoting cell proliferation if dysregulated in cancer. However, no mention is made of the role of DNA methylation. It appears as though the closest connection is offered by two identified genes; *BCL11A*, whose paralog, *BCL11B*, has been shown to be aberrantly methylated in treatment-naïve, (mostly) high-risk PCa[43], and *RASA3*, in which case other *RAS*-family genes have similarly been shown to be aberrantly methylated.[43] When aberrant DNA methylation is noted, it is often in relation to hepatocellular carcinoma (HCC, e.g. *ADARB2*, *RASA3*).[44,45] It is worth noting that several of the top genes that are usually mentioned when aberrant DNA methylation in PCa is discussed (such as *GSTP1*, *APC* and *RARβ*) were not identified in this African analysis; however, these genes are usually identified when comparing PCa versus normal samples or even African American verses Caucasian PCa.[34,46–48] In general, methylation patterns observed along the length of the top genes in **Table 4-3** included declines in DNA methylation over CGIs that frequently overlapped with functional gene regions e.g. TSS and UTRs, indicative of an open chromatin conformation for gene expression. Genes that displayed this pattern were *MECOM*, *GABBR1* and *ACACB*, suggesting potential upregulated gene expression. Notably, overexpression of the oncoprotein *MECOM* is well-documented in acute myeloid leukemia.[49] Conversely, there was also evidence for higher levels of methylation over certain CGIs, most often occurring within gene bodies. The genes *DSCAML1* and *RASA3* displayed this methylation pattern. While gene promoter CGI hypermethylation has been well-documented in PCa[50], gene body CGI hypermethylation has been shown to be associated with gene overexpression in HCC.[51] In fact, hypermethylation of gene body CGIs is considered to be predictive of elevated oncogene levels in HCC. Either way, aberrant methylation is capable of reprogramming gene expression, thereby potentially promoting tumorigenesis. Further analyses should be conducted to determine whether aberrant methylation of the particular genes mentioned here correlates with altered gene expression. From what has been presented above and what one may reveal in a simple literature search, it is overwhelmingly clear that there is a lack of published literature on the PCa epigenome with the depth achieved by Illumina's EPIC array and frankly, an absence of published literature on the African PCa epigenome, making this study unquestionably novel. Next to no information is available on African PCa epigenomics, highlighting the profound contribution of this work, despite the small sample size, and the potential future research that may be built thereon.

There are a number of limitations to this current study that I recognize and are necessary to point out. Firstly, the small African cohort upon which this study was based is not truly representative of the whole southern African PCa population and as such, caution must be taken when interpreting results based on the eight individuals' DNA methylation patterns. Furthermore, this study lacked suitable controls i.e. one of these eight individuals' samples was histopathologically classified as BPH, possibly erroneously so in terms of underlying (epi)genomic alterations indicative of HRPCa, which of course cannot be seen with the naked eye. As a result, an objective of this study has gone partly unanswered. That is to say I was only able to identify 4 DMPs associated with African BPH versus African HRPCa, a value I presume would be much higher should a true BPH individual have been included for analysis. In future, I recommend that inclusion criteria for BPH samples should refer to a tumour purity cut-off. It remains largely unanswered just how distinct African BPH is from African HRPCa in terms of DNA methylation. A further limitation is having only one BPH individual to draw conclusions on as well as to make a comparison on. Additionally, the HRPCa African individuals displayed a range of tumour purities, with some as low as 38 %. In future, only high tumour purity samples should be chosen for analysis to minimise potential confounding from cell type compositions. Even so, I believe these confounders should ideally be adjusted for which I did not implement on this African dataset, unfortunately. However, I stand by my stating any further corrections on this small African cohort would have been too harsh on the data resulting in a loss of significant DMPs during data analysis. Overall, there is scope for improvement but the novelty of this current study remains.

4.5. References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015; 136:E359–86.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71(3):209–49.
3. Park SY, Haiman CA, Cheng I, Park SL, Wilkens LR, Kolonel LN, et al. Racial/ethnic differences

in lifestyle-related factors and prostate cancer risk: The multiethnic cohort study. Cancer Causes Control. 2015; 26(10):1507–15.

4. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet. 2020; 52(8):778–89.

5. Narayan VM, Konety BR, Warlick C. Novel biomarkers for prostate cancer: An evidence-based review for use in clinical practice. Int J Urol. 2017; 24(5):352–60.

6. Kanwal R, Gupta K, Gupta S. Cancer epigenetics: An introduction. In: Verma M, editor. Cancer epigenetics. Methods in molecular biology (methods and protocols). Vol 1238. New York: Humana Press; 2015. p. 3–21.

7. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016; 17(1):208.

8. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014; 74(8):880–91.

9. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and indels with Mutect2. bioRxiv. 2019; 861054.

10. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8.

11. Wickham H, François R, Henry L, Müller K. dplyr: A grammar of data manipulation. R package version 1.0.0. 2020. Available from: https://cran.r-project.org/package=dplyr.

12. Arnold K, Gosling J, Holmes D. The Java programming language. Addison Wesley Professional; 2005.

13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303.

14. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021; 10(2).

15. Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, et al. Whole-genome sequencing reveals elevated tumor mutational burden and initiating driver mutations in African men with treatment-naïve, high-risk prostate cancer. Cancer Res. 2018; 78(24):6736–46.

16. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015; 26(1):64–70.

17. Oesper L, Mahmoody A, Raphael BJ. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013; 14(7):R80.

18. Jaratlerdsiri W, Chan EKF, Petersen DC, Yang C, Croucher PI, Bornman MSR, et al. Next generation mapping reveals novel large genomic rearrangements in prostate cancer. Oncotarget. 2017; 8(14):23588–602.

19. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017; 8(5):7452–63.

20. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–9.

21. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013; 29(2):189–96.

22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. JR Statist Soc. 1995; 57:289–300.

23. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010; 11(1):587.

24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Available from: http://www.r-project.org/.

25. RStudio Team. RStudio: Integrated development environment for R. RStudio, Inc., Boston, MA. 2020. Available from: http://www.rstudio.com/.

26. Wu MC, Kuan PF. A guide to Illumina BeadChip data analysis. In: Tost J, editor. DNA methylation protocols. Methods in molecular biology. New York: Humana Press; 2018. p. 303–30.

27. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11.

28. Geybels MS, Zhao S, Wong CJ, Bibikova M, Klotzle B, Wu M, et al. Epigenomic profiling of DNA methylation in paired prostate cancer versus adjacent benign tissue. Prostate. 2015; 75(16):1941–50.

29. Parry MA, Srivastava S, Ali A, Cannistraci A, Antonello J, Barros-Silva JD, et al. Genomic evaluation of multiparametric magnetic resonance imaging-visible and -nonvisible lesions in clinically localised prostate cancer. Eur Urol Oncol. 2019; 2(1):1–11.

30. LaBarre BA, Goncearenco A, Petrykowska HM, Jaratlerdsiri W, Bornman MSR, Hayes VM, et al. MethylToSNP: Identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin. 2019; 12(1):79.

31. Dhingra P, Martinez-Fundichely A, Berger A, Huang FW, Forbes AN, Liu EM, et al. Identification

of novel prostate cancer drivers using RegNetDriver: A framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. Genome Biol. 2017; 18(1):141.

32. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13(1):86.

33. Mcgregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016; 17(84):1–17.

34. Stewart GD, Van Neste L, Delvenne P, Delrée P, Delga A, Alan McNeill S, et al. Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: Results of the MATLOC study. J Urol. 2013; 189:1110–6.

35. Malins DC, Polissar NL, Gunselman SJ. Models of DNA structure achieve almost perfect discrimination between normal prostate, benign prostatic hyperplasia (BPH), and adenocarcinoma and have a high potential for predicting BPH and prostate cancer. Proc Natl Acad Sci USA. 1997; 94(1):259–64.

36. National Cancer Institute [Internet]. Understanding prostate changes: A health guide for men. National Cancer Institute at the National Institutes of Health; [cited 2021 Sep 2]. Available from: https://www.cancer.gov/types/prostate/understanding-prostate-changes

37. Trujillo KA, Jones AC, Griffith JK, Bisoffi M. Markers of field cancerization: Proposed clinical applications in prostate biopsies. Prostate Cancer. 2012; 2012:1–12.

38. Hempelmann JA, Lockwood CM, Konnick EQ, Schweizer MT, Antonarakis ES, Lotan TL, et al. Microsatellite instability in prostate cancer by PCR or next-generation sequencing. J Immunother Cancer. 2018; 6(29):1–7.

39. Cheng J, He J, Wang S, Zhao Z, Yan H, Guan Q, et al. Biased influences of low tumor purity on mutation detection in cancer. Front Mol Biosci. 2020; 7:343.

40. Blattler A, Farnham PJ. Cross-talk between site-specific transcription factors and DNA methylation states. J Biol Chem. 2013; 288(48):34287–94.

41. Di Croce L, Raker V, Corsaro M, Fazi F, Fanelli M, Faretta M, et al. Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. Science. 2002; 295(5557):1079–82.

42. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex network: A systematic exploration of the human interactome. Cell. 2015; 162(2):425–40.

43. Mahapatra S, Klee EW, Young CYF, Sun Z, Jimenez RE, Klee GG, et al. Global methylation profiling for risk prediction of prostate cancer. Clin Cancer Res. 2012; 18(10):2882–95.

44. Gentilini D, Scala S, Gaudenzi G, Garagnani P, Capri M, Cescon M, et al. Epigenome-wide association study in hepatocellular carcinoma: Identification of stochastic epigenetic mutations through an innovative statistical approach. Oncotarget. 2017; 8(26):41890–902.

45. Lin H, Fan X, He L, Zhou D. Methylation patterns of RASA3 associated with clinicopathological factors in hepatocellular carcinoma. J Cancer. 2018; 9(12):2116–22.

46. Kwabi-Addo B, Wang S, Chung W, Jelinek J, Patierno SR, Wang BD, et al. Identification of differentially methylated genes in normal prostate tissues from African American and Caucasian men. Clin Cancer Res. 2010; 16(14):3539–47.

47. Jerónimo C, Henrique R, Hoque MO, Mambo E, Ribeiro FR, Varzim G, et al. A quantitative promoter methylation profile of prostate cancer. Clin Cancer Res. 2004; 10:8472–8.

48. Woodson K, Hanson J, Tangrea J. A survey of gene-specific methylation in human prostate cancer among black and white men. Cancer Lett. 2004; (205):181–8.

49. Ogawa S, Mitani K, Kurokawa M, Matsuo Y, Minowada J, Inazawa J, et al. Abnormal expression of Evi-1 gene in human leukemias. Hum Cell. 1996; 9(4):323–32.

50. Bastian PJ, Palapattu GS, Lin X, Yegnasubramanian S, Mangold LA, Trock B, et al. Preoperative serum DNA GSTP1 CpG island hypermethylation and the risk of early prostate-specific antigen recurrence following radical prostatectomy. Clin Cancer Res. 2005; 11(11):4037–43.

51. Arechederra M, Daian F, Yim A, Bazai SK, Richelme S, Dono R, et al. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. Nat Commun. 2018; 9(1):3164.

Chapter 5: Pilot interrogation of the association between somatic mutational signatures and genome-wide DNA methylation in prostate tissue from men of African ancestry

As discussed in **Chapter 2**, a complex interaction exists between genomic and epigenomic processes in cancer. A dysregulating factor for the epigenome may result in indirect dysregulation for the genome, and vice-versa. Findings presented in **Chapter 4** provide evidence for the role of aberrant DNA methylation in African prostate cancer, although the driving factors for such observations are unknown. In this Chapter, I present a pilot investigation of the association between mutational signatures and global DNA methylation to speculate on the contribution of intrinsic and extrinsic factors to signatures observed in prostate tissue from men of African ancestry.

In this Chapter, I would like to acknowledge the patients who consented to donating their tissue for the purposes of this research, as well as the many clinicians and staff associated with the Southern African Prostate Cancer Study. I would also like to extend a special thanks to Dr Weerachai Jaratlerdsiri, of the Human Comparative and Prostate Cancer Genomics Research team at the Garvan Institute of Medical Research, for generating the mutational signature data utilized in this Chapter.

Abstract

An emerging field in cancer genomics is the identification of mutational signatures that provide novel insights into individual cancer aetiology. The power harnessed by these mutational signatures lies in their ability to reveal both endogenous and exogenous factors that contribute to cancer development. While a number of mutational signature classes have been identified, namely single-base-substitution, doublet-base-substitution, clustered-base-substitution, small insertion-and-deletion, and genome rearrangement signatures, DNA methylation signatures are absent from these catalogues. Given the genomic-epigenomic interaction that exists not only for normal cellular processing, but also tumorigenesis, as well as the fact that epigenetic mechanisms offer environmental agents a direct link to mediate their carcinogenic properties on the human genome, resulting in altered DNA methylation, the addition and analysis of DNA methylation signatures will prove invaluable for providing further insight to the endogenous and exogenous contributors to cancer. The identification of DNA methylation signatures is beyond the scope of the work presented in this Chapter. However, herein I present a novel pilot investigation of the association between single-base-substitution signatures and genome-wide DNA methylation to provide evidentiary support for the interaction that exists between these two processes. Additionally, I present a brief discussion on the value of DNA methylation signatures for the future of cancer genomics.

5.1. Introduction

The somatic mutations in a cancer genome are the result of multiple mutational processes.[1] Each of these processes gives rise to a characteristic pattern of mutations, termed a mutational signature, and are caused by the activity of endogenous and/or exogenous mutational processes.[2] Some of these processes have been active throughout one's lifetime whilst others have been sporadically triggered by exogenous factors such as lifestyle choices.[1] Recently, mathematical methods have been developed to decipher these distinct mutational signatures from large sets of somatic mutations.[3,4] Such models extract these signatures by identifying the minimal set of mutational signatures that is able to best explain the proportion of each mutation type found in each cancer sample. Thereafter, the model estimates the contribution of each mutational signature to each sample. The value of identifying mutational signatures rests in their ability to provide new insights into the causes of individual cancers and are able to propose endogenous and exogenous factors that have influenced oncogenesis. By characterising the mutational processes that contribute to cancer, mutational signatures provide researchers with a tool to better understand cancer aetiology.

Mutational signatures are classed by type, namely (i) single-base-substitution (SBS), (ii) doublet-base-substitution (DBS), (iii) clustered-base-substitution, (iv) small insertion-and-deletion (ID), and genome rearrangement signatures. SBSs are examined using 96 mutation types i.e. for each of the six types of somatic substitutions (C>A, C>G, C>T, T>A, T>G and T>C), the base immediately 5' before the somatic mutation and the base immediately 3' after the somatic mutation is included, resulting in 16 mutation types for each somatic substitution.[4] SBS signatures are the most prevalent of the mutational signatures, with former studies having identified more than 30, as previously reported in COSMIC v.2. To date, this has been expanded to 49 identified SBS signatures.[4] While many of these signatures are of known cause, there are numerous signatures of unknown aetiology. Some signatures are common, while others are rare and some represent normal biology, while others are more sinister, reflective of carcinogenic exposure or tumorigenic processes.[4] For instance, SBS1 and SBS5 are universally-present signatures whereas SBS3, SBS6, SBS8 and SBS21 represent rarer mutational processes. Mutational signatures SBS1, SBS5 and SBS40 have previously been shown to associate with age, with SBS1 arising in response to 5-methylcytosine deamination, a normal cellular event, and the mechanism by which SBS5 and SBS40 arises being unknown.[4] Notably, SBS5 and SBS40 are flat signatures that share a high degree of similarity, sparking debate as to whether or not they represent the same signature. A number of signatures, such as SBS6 and SBS21, have been attributed to defective DNA maintenance processes but for a number of signatures, such as SBS8, mere speculation exists to explain the signature's origins (DNA damage to guanine in response to an unknown, possibly external, DNA-damaging agent).

Similar to somatic mutations, cancer-associated DNA methylation alterations may also be due to endogenous and/or exogenous mutational processes.[5,6] Epigenetic dysregulation may arise in response to methylation alterations driven by factors such as age, innate susceptibility, the tumour microenvironment, toxicants, nutrition and stress.[7–9] While epigenetic changes are reversible, if this dysregulation is not corrected for, they may accumulate over time. Therefore, the epigenome holds clues about one's life stage and previous exposures, given the ability of the epigenome to be replicated during somatic cell mitosis.[10] DNA methylation is well-established as a key regulator of gene expression, thus alterations to this epigenetic mechanism can result in the activation of oncogenes and/or silencing of tumour suppressor genes, when considering cancer. Although once considered independent mechanisms contributing to cancer progression, it is now well-understood that a complex interaction exists between genomic and epigenomic mechanisms to aid tumorigenesis (see **Chapter 2** for further detail).[8] Ultimately, accumulated genomic and epigenomic aberrations result in the evolution of a malignant cell.

The genomic-epigenomic interplay has been demonstrated for prostate cancer (PCa); however, there is currently no known modifiable risk factor for this disease.[11] Although great insights have been gained through deciphering the mutational signatures described above, there is much to be discovered by identifying and incorporating DNA methylation signatures with these signature catalogues given the genomic-epigenomic interaction that exists for oncogenesis. Perhaps the identification of a modifiable risk factor for PCa lies in the proposed aetiology of an as-yet unidentified DNA methylation signature. Such an accomplishment is beyond the scope of this current work and would require the development of bioinformatic tools and novel machine learning methods. However, to the best of my knowledge, researchers have yet to correlate matched mutational signature and DNA methylation data to confirm the role of DNA methylation in signature aetiology. Using African prostate tissue-derived data, the genomic-epigenomic interaction will be explored in this pilot analysis. Furthermore, in light of DNA methylation profiles established for available signatures, I will address the applicability of previously-proposed aetiologies for SBS mutational signatures present in this African cohort.

5.2. Materials & Methods

5.2.1. Resource & ethics

Data was made available for eight South African men who consented upon enrolment in the Southern African Prostate Cancer Study (SAPCS).[12] Patients were of African ethnicity, confirmed using ancestry markers, and self-identified as such. The previous SAPCS as well as the current study outlined here was reviewed and approved by the University of Pretoria's Human Research Ethics Committee (HREC #43/2010 and #37/2021, respectively). The age distribution of the patients ranged from 54-99 and further summary data provided for the eight patients can be viewed in **Table 4-1**. Such data was provided by the Human Comparative and Prostate Cancer Genomics (HCPCG) Research team at the Garvan Institute of Medical Research, located in Sydney, Australia. Data provided is currently unpublished and funded by the Australian National Health and Medical Research Council (NHMRC).

5.2.2. Single-base-substitution signature data

Single-base-substitution (SBS) signature data was made available for the eight African samples by the HCPCG Research team. Somatic mutational signatures were identified using SigProfiler[4,13], whereby the number of somatic mutations associated with each mutational signature was estimated for each sample. *De novo* extraction of SBS mutational signatures and existing global COSMIC (Catalogue of Somatic

Mutations in Cancer v3.2) signatures were used for analysis. Single-base-substitution signatures identified in this cohort includes SBS1, SBS3, SBS5, SBS6, SBS8, SBS21 and SBS40 (see **Fig. 5-1**).

5.2.3. DNA methylation data & processing

Raw DNA methylation data was generated for the eight African patients at the Australian Genome Research Facility (AGRF, Melbourne, Australia) and subsequently provided by the HCPCG Research team at the Garvan Institute. DNA methylation was quantified using the Illumina Infinium HumanMethylationEPIC BeadChip following the Illumina Infinium HD Methylation Assay (Illumina, CA, USA). The data provided by the AGRF included raw Illumina intensity data (IDAT) files, the Illumina manifest file (v1.0 B5, BPM format), a sample sheet (CSV format) and a genotyping service report from the research facility. DNA methylation data was processed and analysed by myself using the novel African-relevant pipeline established in **Chapter 3** and applied as previously described in **Chapter 4**.

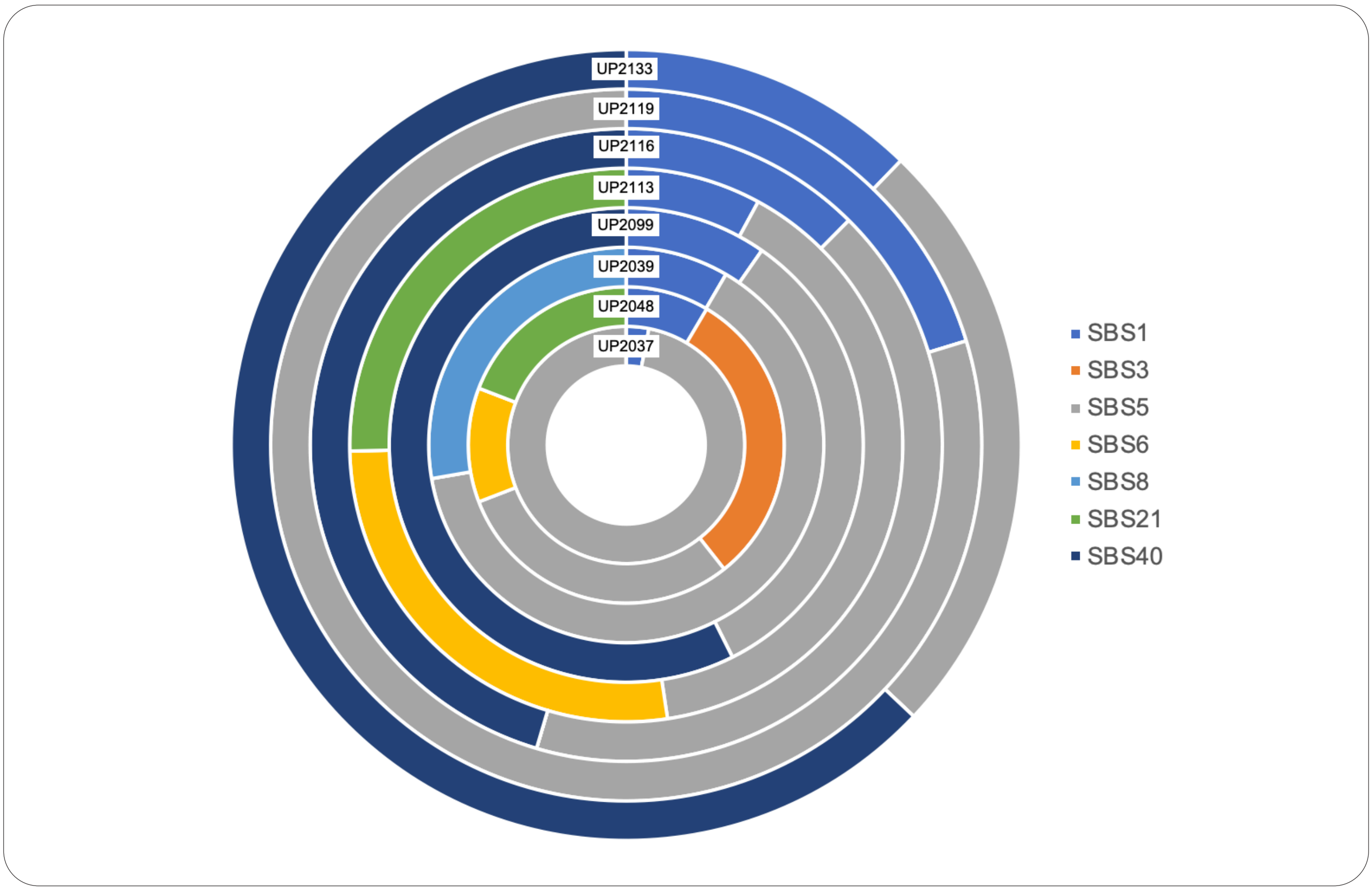


Fig. 5-1 Single-base-substitution (SBS) signature contributions present in each African sample (UP0000).

5.2.4. *Identifying differentially methylated probes*

I performed differential methylation analysis among the African samples to identify significant differential methylated probes (DMPs) that correlate with each of the above-mentioned mutational signatures. I used the champ.DMP() function to identify significant DMPs and the DMP.GUI() function to visualise the results. Linear regression was conducted on each CpG site within the African dataset to identify signature-related CpG sites. It was decided that DMPs be selected based on a BH-adjusted threshold of $p < 0.05$. The Benjamini-Hochberg *p*-value adjustment controls the false discovery rate.[14] DMPs were categorized as displaying hypermethylation (beta ≥ 0.8), partial methylation (beta ~ 0.5) or hypomethylation (beta ≤ 0.2), as per recommendations from Du et al. (2010).[15] For each signature, I chose the top three (where applicable) genes (most abundant for significant CpGs) for closer DNA methylation pattern analysis. Analyses were conducted using the hg19 genome assembly.

5.2.5. *Statistical analysis*

Spearman's rank correlation was computed to assess the relationship between mutational signatures and clinical/genomic variables (as summarised in **Table 4-1**). Spearman's correlation was selected for being robust to non-normally distributed data and for not assuming a distribution of the data. A *p*-value < 0.05 was considered to be statistically significant.

All data processing and analyses were performed with R[16] ≥ v.4.0.2 and RStudio[17] ≥ v.1.3 statistical software.

5.3. Results

5.3.1. *Investigating the association between single-base-substitution signatures and genome-wide DNA methylation*

I found a number of significant DMPs to be associated with each of the SBS signatures identified within the small African cohort (**Table 5-1**). Additionally, the DNA methylation profiles associated with each SBS signature (**Fig. 5-2**) was analysed, each discussed further, below. However, the pilot nature of this analysis must be emphasised, in that no definitive conclusions can be drawn from results presented below.

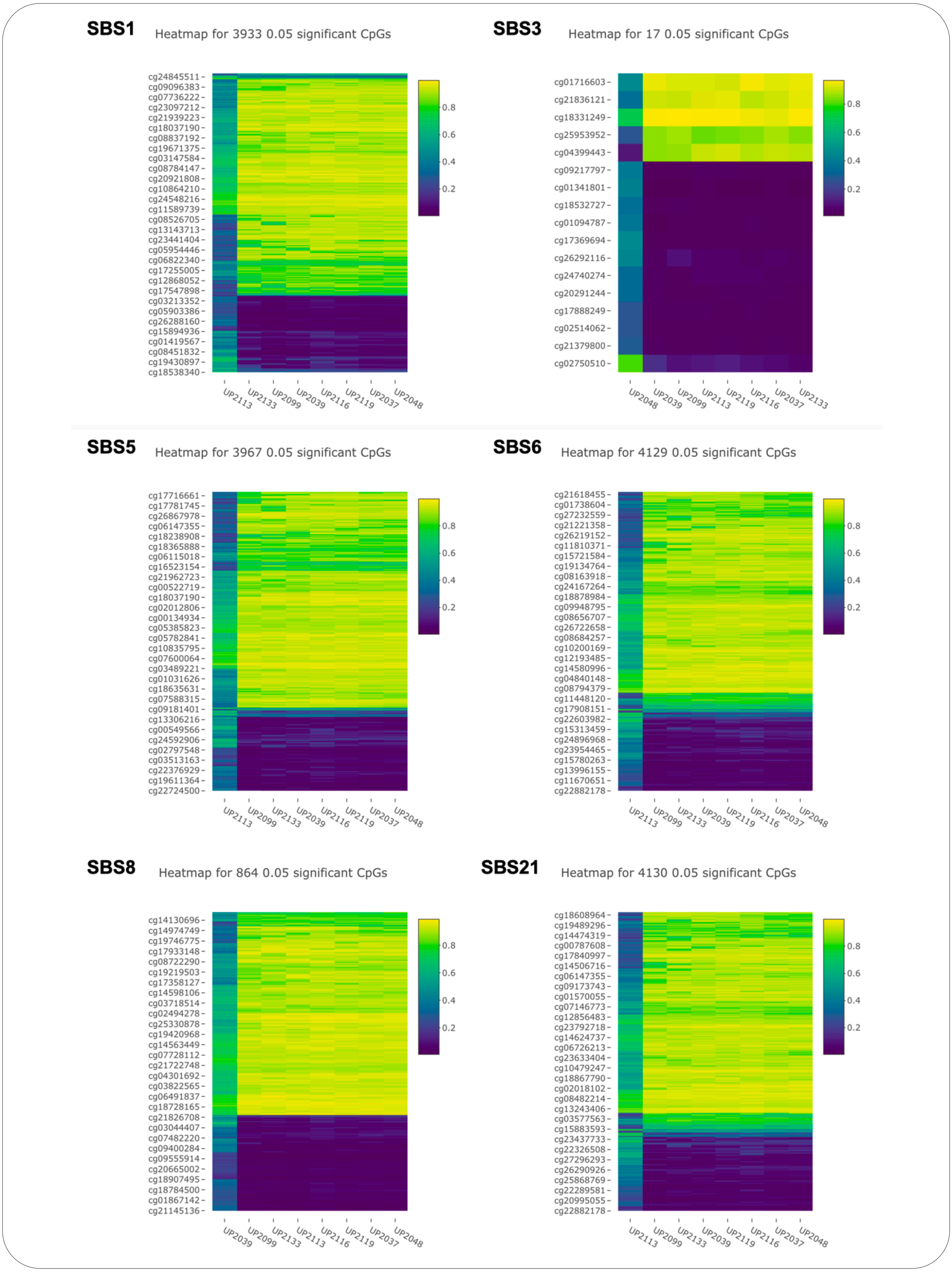
SBS1 and SBS5. All African patients in this cohort displayed mutational signatures SBS1 and SBS5, which are traditionally age-related.[4] **Figure 5-2** shows that for SBS1 and SBS5, patient UP2113 exhibited a distinct methylation profile compared to the other African patients. If these signatures were related to age, one would expect all African individuals to display similar DNA methylation profiles that were reflective of age-accumulated DNA methylation changes. However, the vastly distinct DNA methylation profile displayed by UP2113 cannot be attributed to age but perhaps could be explained by an environmental exposure. Interestingly, I noted that this individual is an outlier for TMB (see **Table 4-1**).

Table 5-1 Differentially methylated probes identified by the novel African-relevant pipeline to be associated with single-base-substitution signatures in the African cohort.

Mutational signature	*Number of significant CpGs identified ($p < 0.05$)*	*Top genes enriched for significant CpGs*
SBS1	3933	*TCF4, MECOM, XYLT1*
SBS3	17	*HLA-DRB5, NEK11*
SBS5	3967	*TCF4, MECOM, NFIX*
SBS6	4129	*TCF4, MECOM, XYLT1*
SBS8	864	*CELSR3, PCTP*
SBS21	4130	*TCF4, MECOM, XYLT1*
SBS40	414	*NFYC, ECHDC1, LOC100133991, TIAM1*

SBS3. A single African patient, UP2048, displayed mutational signature SBS3. The proposed aetiology for SBS3 is homologous recombination deficiency (HRD) due to germline and/or somatic mutations, frequently in *BRCA1* and *BRCA2*.[4] As evidenced by a single patient displaying SBS3, HRD is not a common occurrence in PCa (more so in ovarian and breast cancer) and I only identified 17 DMPs to be associated with this mutational signature (**Table 5-1**). Aberrant *BRCA1* promoter methylation has been suggested as a possible mechanism for HRD in ovarian cancer.[18] Although *BRCA1* (or *BRCA2*) was not enriched for any significant DMPs in this African prostate tissue-derived cohort, UP2048 did display a distinct methylation profile in comparison to the individuals that did not display SBS3 (**Fig. 5-2**).

SBS6 and SBS21. Mutational signatures SBS6 and SBS21 are proposed to be the result of defective DNA mismatch repair.[4] African patients UP2113 and UP2048 both displayed SBS6 and SBS21. In **Figure 5-2**, patient UP2113 clearly displays a distinct DNA methylation profile compared to the other patients, for both SBS6 and SBS21. Considering the defective DNA mismatch repair nature of these SBS signatures, I noted that UP2113's tumour sample displayed MSI-H, which itself is characterised by a DNA mismatch repair deficiency. However, UP2048 showed MSS and a methylation profile for SBS6 and SBS21 that was


SBS1
Heatmap for 3933 0.05 significant CpGs
cg24845511
cg09096383
cg07736222
cg23097212
cg21939223
cg18037190
cg08837192
cg19671375
cg03147584
cg08784147
cg20921808
cg10864210
cg24548216
cg11589739
cg08526705
cg13143713
cg23441404
cg05954446
cg06822340
cg17255005
cg12868052
cg17547898
cg03213352
cg05903386
cg26288160
cg15894936
cg01419567
cg08451832
cg19430897
cg18538340
UP2113
UP2133
UP2099
UP2039
UP2116
UP2119
UP2037
UP2048
0.8
0.6
0.4
0.2
SBS3
Heatmap for 17 0.05 significant CpGs
cg01716603
cg21836121
cg18331249
cg25953952
cg04399443
cg09217797
cg01341801
cg18532727
cg01094787
cg17369694
cg26292116
cg24740274
cg20291244
cg17888249
cg02514062
cg21379800
cg02750510
UP2048
UP2039
UP2099
UP2113
UP2119
UP2116
UP2037
UP2133
0.8
0.6
0.4
0.2
SBS5
Heatmap for 3967 0.05 significant CpGs
cg17716661
cg17781745
cg26867978
cg06147355
cg18238908
cg18365888
cg06115018
cg16523154
cg21962723
cg00522719
cg18037190
cg02012806
cg00134934
cg05385823
cg05782841
cg10835795
cg07600064
cg03489221
cg01031626
cg18635631
cg07588315
cg09181401
cg13306216
cg00549566
cg24592906
cg02797548
cg03513163
cg22376929
cg19611364
cg22724500
UP2113
UP2099
UP2133
UP2039
UP2116
UP2119
UP2037
UP2048
0.8
0.6
0.4
0.2
SBS6
Heatmap for 4129 0.05 significant CpGs
cg21618455
cg01738604
cg27232559
cg21221358
cg26219152
cg11810371
cg15721584
cg19134764
cg08163918
cg24167264
cg18878984
cg09948795
cg08656707
cg26722658
cg08684257
cg10200169
cg12193485
cg14580996
cg04840148
cg08794379
cg11448120
cg17908151
cg22603982
cg15313459
cg24896968
cg23954465
cg15780263
cg13996155
cg11670651
cg22882178
UP2113
UP2099
UP2133
UP2039
UP2119
UP2116
UP2037
UP2048
0.8
0.6
0.4
0.2
SBS8
Heatmap for 864 0.05 significant CpGs
cg14130696
cg14974749
cg19746775
cg17933148
cg08722290
cg19219503
cg17358127
cg14598106
cg03718514
cg02494278
cg25330878
cg19420968
cg14563449
cg07728112
cg21722748
cg04301692
cg03822565
cg06491837
cg18728165
cg21826708
cg03044407
cg07482220
cg09400284
cg09555914
cg20665002
cg18907495
cg18784500
cg01867142
cg21145136
UP2039
UP2099
UP2133
UP2113
UP2116
UP2119
UP2037
UP2048
0.8
0.6
0.4
0.2
SBS21
Heatmap for 4130 0.05 significant CpGs
cg18608964
cg19489296
cg14474319
cg00787608
cg17840997
cg14506716
cg06147355
cg09173743
cg01570055
cg07146773
cg12856483
cg23792718
cg14624737
cg06726213
cg23633404
cg10479247
cg18867790
cg02018102
cg08482214
cg13243406
cg03577563
cg15883593
cg23437733
cg22326508
cg27296293
cg26290926
cg25868769
cg22289581
cg20995055
cg22882178
UP2113
UP2099
UP2133
UP2039
UP2119
UP2116
UP2037
UP2048
0.8
0.6
0.4
0.2

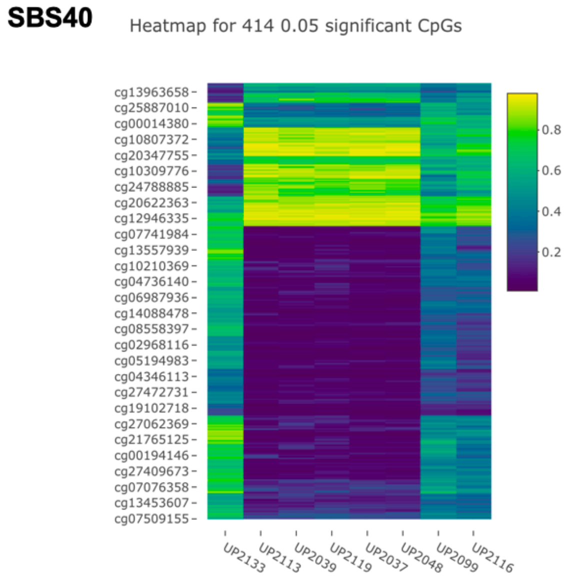


Fig. 5-2 Heatmaps displaying significant differentially methylated probes (cg00000000) between African individuals for a number of single-base substitution (SBS) signatures ($p < 0.05$).

comparable to the other MSS samples (**Fig. 5-2**). This may be explained by the contribution of somatic mutations that UP2048 displayed for SBS6 and SBS21 compared to contributions displayed by UP2113 (11 and 18 versus 39,133 and 36,686, respectively). It appears as though a higher mutational contribution is indicative of a more distinct methylation profile.

SBS8. A single African patient, UP2039, displayed mutational signature SBS8, for which the aetiology is currently unknown.[4] **Figure 5-2** shows UP2039 to display a distinct DNA methylation profile for SBS8 compared to the other African patients, although which factor(s) may be driving this differential methylation associated with SBS8 is unclear.

SBS40. African patients UP2133, UP2099 and UP2116 displayed mutational signature SBS40. Currently the aetiology of SBS40 is unknown.[4] As evidenced in **Figure 5-2**, patients UP2133, UP2099 and UP2116 all display DNA methylation profiles that are unique in contrast to patients that did not display SBS40. Notably, UP2133 exhibits its own methylation profile that differs somewhat from UP2099 and UP2166. This may be explained by UP2133's higher mutational contribution to SBS40 compared to UP2099's and UP2116's (6,903 versus 3,764 and 2,483, respectively). This suggests that different SBS mutational contributions associate with different DNA methylation profiles. Which factors potentially account for different SBS40 mutational contributions is unknown, as previously mentioned. Furthermore, although SBS40 often appears similar to SBS5 in multiple cancer types, I found the two mutational signatures to be rather different from one another in this small African cohort for a number of reasons: (i) SBS5 has substantially more associated DMPs than SBS40 (although this may be reflective of more patients displaying SBS5), (ii) SBS5 and SBS40 show different top genes enriched for significant CpGs (**Table 5-1**), (iii) SBS5 and SBS40 display different DNA methylation profiles (**Fig. 5-2**), and (iv) SBS5 and SBS40 correlate with different genomic variables (see below, **Section 5.3.2.**).

5.3.2. Assessing the correlations of mutational signatures

From the DMP analysis presented above, I aimed to investigate whether or not SBS signatures within this African cohort associated with the clinical and genomic variables they had traditionally been shown to in previous studies.

5.3.2.1. Correlations with age

Mutational signatures SBS1 and SBS5 traditionally correlate with age.[4] However, I did not find a statistically significant correlation between age and SBS1 (r(6) = 0.45, *p* = 0.268) nor between age and SBS5 (r(6) = 0.45, *p* = 0.268) despite a range of ages amongst the Africans at diagnosis (54-99, *Mdn* = 68). Due to the small cohort size, one cannot conclude that no correlation exists between these variables but rather that in this study, I do not have sufficient evidence to suggest that there is a correlation between age and SBS1/SBS5. However, as noted above, patient UP2113, whose sample displayed a distinct DNA methylation profile for both SBS1 and SBS5 (**Fig. 5-2**), was an outlier for TMB.

5.3.2.2. Correlations with tumour mutational burden

Interestingly, I found that both mutational signatures SBS1 and SBS5 showed a significantly strong positive correlation with TMB (r(6) = 0.93, *p* = 0.002 and r(6) = 0.86, *p* = 0.011, respectively) in the African cohort.

5.3.2.3. Correlations with high microsatellite instability

The mutational signatures SBS6 and SBS21 are known to associate with deficiency in DNA mismatch repair.[4] A single African sample (UP2113) in this cohort displayed MSI-H cancer cells (see **Table 4-1**), characterized by impaired DNA mismatch repair. Consistent with previous findings, I found a significant positive correlation between MSI-H status (i.e. deficient DNA mismatch repair) and SBS6 (r(6) = 0.76, *p* = 0.030) as well as between MSI-H status (i.e. deficient DNA mismatch repair) and SBS21 (r(6) = 0.76, *p* = 0.030) in the African cohort.

5.3.2.4. Correlations with percentage of genome alteration

Mutational signature SBS40, for which the aetiology is currently unknown, showed a significant positive correlation with PGA ($r(6) = 0.79$, $p = 0.019$) in the African cohort.

5.4. Discussion

Here I present a novel investigation, albeit pilot, of the association between genomic mutational signatures and genome-wide differential DNA methylation in prostate tissue from South African men. Additionally, I assessed correlations between these mutational signatures and known clinical and genomic variables to ascertain whether proposed aetiologies for single-base-substitution signatures in previous studies were explicable for findings generated within this small African cohort. However, it must be noted that no significant conclusions can be drawn from findings presented here owing to the limited cohort size.

Investigating the association between SBS signatures and genome-wide DNA methylation, I identified a wealth of significant CpG sites to be associated with each of the SBS signatures of relevance to this study (**Table 5-1**). This link between SBS signatures and DNA methylation profiles is perhaps reflective of the genomic-epigenomic interplay known to underlie not only normal cell processes, but also oncogenesis.[19,20] In other words, DNA methylation at significant CpG sites identified for an SBS signature may interact with the genomic and/or environmental processes that ultimately give rise to that signature.

Mutational signatures SBS1 and SBS5 have previously been suggested to be reflective of age-accumulated alterations.[4] However, in this study, the distinct methylation profile demonstrated by UP2113 for both these signatures (**Fig. 5-2**) was likely reflective of an extreme TMB. This observation was supported by a significantly strong positive correlation between SBS1 and TMB as well as between SBS5 and TMB. The proposed aetiology for SBS1 is 5-methylcytosine deamination whereas the mechanism by which SBS5 arises is unknown. Interestingly, a high TMB is suggestive of an environmental/carcinogenic exposure[21,22], such as UV radiation in melanoma and tobacco smoke in lung cancer.[23,24] Given that African prostate tumours display significantly higher TMBs than European prostate tumours[25], it may be that an environmental exposure underlies this observation and by extension, SBS1 and SBS5 mutational signatures in African prostate tumours. Indeed, it has been suggested that SBS5 may be the result of DNA damage to adenine in response to an unknown (external) DNA-damaging agent.[4] Furthermore, it may be that the degree of exposure is associated with the extent to which DNA is aberrantly methylated. Further investigation on a much larger cohort accompanied by patient exposure data is needed to confirm this.

Some contention exists as to whether or not mutational signatures SBS5 and SBS40 represent distinct signature profiles.[4] The uncertainty is attributed to the great similarity that exists for these two signatures. However, the aetiology for both is currently unknown. A number of existing identified signatures are split into several constituent signatures, reflecting several distinct mutational processes. Such processes may be initiated by the same exposure that have closely, but not perfectly, correlated activities (e.g. SBS7a, SBS7b, SBS7c and SBS7d all due to UV light exposure). In agreement with findings for SBS5, I found no correlation between age and SBS40 (not shown), despite previous studies showing otherwise.[4] However, I did find SBS40 showed a significant positive correlation with PGA. Intriguingly, a high PGA may be indicative of an external exposure[26,27], which makes one question whether the same could be said for SBS40. As discussed just above, SBS5 too may possibly arise in response to an external exposure. Should it be that the similarity observed between SBS5 and SBS40 could be credited to the same exposure initiating processes that have closely correlated activities, it would support the classification of these signatures as constituent signatures (e.g. SBS5a and SBS5b) rather than distinct signature profiles. However, this is pure speculation and a much larger, comprehensive investigation would be required to explore this further.

MSI is a hypermutable phenotype characterised by the loss of DNA mismatch repair activity and can arise through CpG island hypermethylation.[28] Signatures SBS6 and SBS21 are proposed to result in response to defective DNA mismatch repair.[4] A single African patient (UP2113) displayed MSI-H and showed a unique DNA methylation profile for both SBS6 and SBS21 (**Fig. 5-2**). Additionally, I found a significant positive correlation between MSI-H and SBS6 as well as between MSI-H and SBS21. Consistent with the previously proposed aetiology for SBS6 and SBS21, it appears as though these signatures are associated with defective DNA mismatch repair in this African cohort.

Finally, neither mutational signature SBS3 nor SBS8 correlated with any of the variables measured in this study. SBS3 is characterised by HRD due to *BRCA1* and/or *BRCA2* mutations whereas for SBS8, the aetiology is currently unknown. However, it has been suggested that SBS8 is associated with DNA damage to guanine in response to an unknown (external) DNA-damaging agent.[4] Because so many signatures are still of unknown cause and assuming intrinsic associations have been investigated for said signatures (with insignificant findings), it may be reasonable to assume that environmental (as yet, unmeasured) factors are at play.

Although I identified SBS signature-associated DNA methylation profiles, the question remains as to which of those significant CpG sites are representative of a genome-wide DNA methylation *signature* i.e. a generic pattern of aberrant DNA methylation that arises during tumorigenesis. A greater spectrum and understanding of the mutational processes that contribute to cancer could be attained from identifying such DNA methylation signatures. To achieve this, large sample numbers and tools to extract DNA methylation signatures would be required. Furthermore, and beyond the scope of this present study, the true insight to be gained from genomic-epigenomic assimilation would be to integrate genomic signatures with epigenomic signatures. Such integration would require the development of novel machine learning methods.

Genomic signatures provide novel insights into the causes of individual cancers and reveal intrinsic and extrinsic factors (where known) that have contributed to cancer development.[4] A number of mutational signatures, SBS signatures aside, have been identified in human cancer, namely doublet-base-substitution (DBS), clustered-base-substitution and small insertion-and-deletion (ID) signatures.[4] However, many of these signatures are still of unknown cause. An individual cancer may be characterised by a number of these signatures and the addition of epigenomic signatures would undoubtedly provide further insight into cancer aetiologies. Of particular interest for PCa, for which there is no known modifiable risk factor[11], the identification of possible external driving factors would be an invaluable discovery, allowing clinicians to advise at-risk individuals against such exposures. The key to such a discovery may very well lie in the identification of DNA methylation signatures further associated with patient external exposure data. Indeed, it is well-established that environmental exposures are capable of influencing epigenomic changes.[6] To revisit the integration of genomic and epigenomic signatures, should an extrinsic contributing factor be identified for a DNA methylation signature and should said DNA methylation signature correlate with a known genomic signature, a potential cascade of oncogenic processes may be identified. I propose a mechanism in which an environmental exposure drives aberrant DNA methylation, which in turn causes genetic alterations i.e. a carcinogen directly associates with the epigenome and indirectly associates with the genome. Such investigations may propose aetiologies for genomic signatures currently of unknown cause.

From the discussion above, it is evident that there is much to be gained from investigating cancer with a genome-wide versus targeted approach. Should the environment be credited for a role in tumorigenesis, it seems only logical to assume such an influence would be genome-wide. Given that cancer arises in response to complex cooperation between a multitude of events, a greater potential for uncovering cancer aetiology lies in genomic, versus genetic, investigations. This underscores the very power harnessed by mutational signatures. Although currently an emerging field in cancer genomics, advances in mutational signature knowledge promise to uncover new insights into the causes of individual cancers and the addition of DNA methylation signatures will surely aid this endeavour.

To my knowledge, the integration of DNA methylation data with existing genomic signatures is a novel concept that has not yet been investigated, and is challenged by the rarity of researchers possessing matched genomic and epigenomic data due to high costs. This highlights the unique nature of the research and discussions presented in this Chapter. However, I acknowledge that the limited cohort size is a considerable limitation that exists for this present study and as such, interpretations of findings presented here are mere speculation. Evidently, there is a need for large volumes of such matched data to identify global DNA methylation signatures and to ultimately answer clinically-relevant questions as a direct achievement of genomic-epigenomic integration. No doubt future research will aim to address this.

5.5. References

1. Alexandrov LB, Stratton MR. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014; 24:52–60.
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239):719–24.
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013; 3(1):246–59.
4. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020; 578(7793):94–101.
5. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet. 2020; 52(8):778–89.
6. Arita A, Costa M. Environmental agents and epigenetics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 459–76.
7. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009; 5(8):e1000602.
8. Riley LB, Anderson DW. Cancer epigenetics. In: Tollefsbol T, editor. Handbook of epigenetics: The new molecular and medical genetics. London: Elsevier Inc.; 2011. p. 521–34.
9. Skinner MK. Endocrine disruptor induction of epigenetic transgenerational inheritance of disease. Mol Cell Endocrinol. 2014; 398(0):4–12.
10. Skinner MK. Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. Epigenetics. 2011; 6(7):838–42.
11. Tindall EA, Bornman MR, Van Zyl S, Segone AM, Monare LR, Venter PA, et al. Addressing the contribution of previously described genetic and epidemiological risk factors associated with

increased prostate cancer risk and aggressive disease within men from South Africa. BMC Urol. 2013; 13:74.

12. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014; 74(8):880–91.
13. Islam SMA, Wu Y, Díaz-Gay M, Bergstrom EN, He Y, Barnes M, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. bioRxiv. 2021; 2020.12.13.422570.
14. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. JR Statist Soc. 1995; 57:289–300.
15. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010; 11(1):587.
16. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Available from: http://www.r-project.org/.
17. RStudio Team. RStudio: Integrated development environment for R. RStudio, Inc., Boston, MA. 2020. Available from: http://www.rstudio.com/.
18. da Cunha Colombo Bonadio RR, Fogace RN, Miranda VC, Diz MDPE. Homologous recombination deficiency in ovarian cancer: A review of its epidemiology and management. Clinics. 2018; 73(suppl 1):e450s.
19. Brena RM, Costello JF. Genome–epigenome interactions in cancer. Hum Mol Genet. 2007; 16(R1):R96–105.
20. Achinger-Kawecka J, Taberlay P. Alterations in three-dimensional organization of the cancer genome and epigenome. Cold Spring Harb Symp Quant Biol. 2017; 81:41–51.
21. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–8.
22. Halbert B, Einstein DJ. Hot or not: Tumor mutational burden (TMB) as a biomarker of immunotherapy response in genitourinary cancers. Urology. 2021; 147:119–26.
23. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485(7399):502–6.
24. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010; 463(7278):184–90.
25. Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, et al. Whole-genome sequencing reveals elevated tumor mutational burden and initiating driver mutations in African men

with treatment-naïve, high-risk prostate cancer. Cancer Res. 2018; 78(24):6736–46.

26. Pös O, Radvanszky J, Buglyó G, Pös Z, Rusnakova D, Nagy B, et al. Copy number variation: Characteristics, evolutionary and pathological aspects. Biomed J. 2021.

27. Hovhannisyan G, Harutyunyan T, Aroutiounian R, Liehr T. DNA copy number variations as markers of mutagenic impact. Int J Mol Sci. 2019; 20(19):4723.

28. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010; 138(6):2073–87.e3.

Chapter 6: Conclusions & Future Directions

As part of this research, I successfully established a novel African-relevant genome-wide bioinformatic pipeline for the processing and normalisation of African DNA methylation data (**Chapter 3**). I then applied this pipeline, enabling the identification of differentially methylated CpG sites (DMPs) that potentially contribute to aggressive prostate cancer (PCa) in a small cohort of South African men (**Chapter 4**). Finally, I assessed the association between mutational genomic signatures and DNA methylation to confirm whether or not there is evidence that epigenomic alterations interact with genomic processes that ultimately give rise to such signatures (**Chapter 5**). In **Chapters 4** and **5**, although a number of significant DMPs and genes enriched therewith were identified, caution should be taken in interpreting these results due to the small African cohort analysed. Furthermore, the novel pipeline established herein is not only relevant within the context of PCa, but other cancer types too.

Considering that African-relevant tools are extremely scarce, the development of a bioinformatic tool such as this was necessary. Such a scarcity may be explained by the vast lack of published African PCa epigenomic literature, suggesting African men to be underrepresented in this field. Hence, should there be no African DNA methylation (EPIC) data to process and analyse, it follows that there would be no appropriate tools to suitably do so. It is even true that European EPIC PCa data is scarce, with only 7 studies providing such publicly-accessible data via NCBI's Gene Expression Omnibus (GEO, as of August 2021; no African data is available), making this African study a first of its kind. Therefore, despite the small African cohort upon which DNA methylation analysis was conducted, the novelty of this current study cannot be overlooked. Even more importantly, the findings of this research provide a glimpse of the epigenomics that underlies African PCa, which no doubt holds vital insights to expand our understanding of African PCa and which ultimately, will hopefully motivate more comprehensive, sizable work in future.

For future research, naturally a much larger African cohort would be ideal for data processing and analysis. This would allow for more sound interpretation of DNA methylation's influence on PCa for the larger southern African population. If possible, only samples with a high tumour purity (e.g. greater than 90 %, as estimated by a histopathologist) should be chosen for analysis and suitable controls should be included; that is, either non-BPH, non-malignant samples or true BPH samples with very low tumour purity estimates. Additionally, a greater number of controls should be included. Analysis on a larger cohort may even motivate the use of M-values rather than beta-values, where appropriate, for statistical validity.

As discussed in **Chapter 4**, confounding variables should ideally be adjusted for within the data. In this study, PGA (percent genome alteration) and tumour purity were identified as such. Necessity for this adjustment is particularly true with regards to tumour purity because confounding introduced by cell-type heterogeneity is a common issue that arises in DNA methylation studies[1–3] (see **Chapter 2** for a discussion on this topic), and as seen in this current study, it had a significant effect on the African dataset (presented in **Chapter 4**). Under ideal circumstances, one should be sure that DMPs inferred are not driven by underlying changes introduced by confounders. On that note, an improvement I could suggest for this study would be to utilize the 18 red and green internal control probes included within Illumina's EPIC array. In doing so, one could assess bisulfite conversion efficiency as well as identify any possible variability in the data that was introduced by the control probes. An SVD (singular value decomposition) analysis using the champ.SVD() function could achieve this. Should significant components of variation correlate with the control probes, such variability could be adjusted for in a similar manner to that of biological confounders. A method introduced in **Chapter 2**, namely surrogate variable analysis (SVA), appears to be an appropriate manner for performing such corrections, having been recommended numerous times.[3–5] This may be implemented within the sva package in R.[6]

In terms of SNP-affect probes, for future, it may even be sufficient to flag and filter SNPs that lie in the first and second CpG positions for these being the SNPs that exert the strongest effect on influencing methylation value callouts.[7] In contrast to this study, SNPs that lie along the body of the probe may just be flagged for the purpose of being cautious when interpreting results from these probes. In light of these suggestions, the full SNP-affected probe filtering used here may have been too harsh. Furthermore, a limitation I touched on in **Chapter 3** is that the established novel African-relevant pipeline requires that researchers be in possession of patient-matched genomic data to filter SNP-affected probes, which I acknowledge is often absent in studies due to high costs. Future studies may consider developing a consensus panel of African SNP-affected probes recommended for filtering, that is designed according to population genetics rather than individual patients. Development of such a panel would require large volumes of African germline variant data but would universalize SNP-affected probe filtering for researchers working on any African-related DNA methylation dataset.

For data analysis, should time have allowed for it, an intriguing comparison would be between African versus European HRPCa for the identification of significant ethnicity-associated DMPs. As mentioned above, NCBI's GEO contained a limited number of publicly-accessible European EPIC PCa data and of those studies, only one or two were suitable for comparison with the African cohort in terms of matching for treatment-naïve and appropriate Gleason score patients. However, an issue that exists with the publicly-available data is the absence of a sample sheet (see **Section 3.2.2.**). This file, which stores phenotypic data associated with the EPIC BeadChip, is needed by methods like ChAMP and *minfi* for the data extraction step in the bioinformatic pipeline. Should this file have been available, the accompanying European EPIC PCa data could be processed and analysed within the ChAMP pipeline utilizing many of the default parameters that assume European ethnicity of a cohort (e.g. SNP filtering according to dbSNP). Upon recent investigation into the sample sheet issue, I came across a Python-based package, *methylprep*[8], which is available for Illumina methylation array processing of public datasets from NCBI's GEO, and includes a function for creating a sample sheet from a public dataset. Therefore, this tool may offer a promising solution.

Further analyses that may be performed on this dataset or on a larger African cohort would be to confirm the effects of aberrant DNA methylation on gene expression for a number of identified top candidate genes. In findings presented here (**Chapter 4**), potential candidate genes may be *MECOM*, *GABBR1*, *ACACB*, *DSCAML1* and *RASA3* due to suspected gene expression changes in response to aberrant DNA methylation in African HRPCa (high-risk prostate cancer) and for their respective roles, although not all epigenomic, in numerous cancer types. Targeted bisulfite sequencing could then be performed to confirm DNA methylation patterns at single-base resolution along the length of these genes. Following this, the DNA methylation data of said genes could be correlated with expression data to assess potential functional impacts.

As discussed in **Chapter 5**, mutational signatures are an emerging field in cancer genomics that are able to shed light on the mutational processes that contribute to cancer development. There is great potential to expand on knowledge within this field by identifying and integrating DNA methylation signatures with existing genomic signatures. Large patient numbers and appropriate bioinformatic tools would be required to decipher DNA methylation as a global signature. I believe the addition of such signatures would prove invaluable to better understanding cancer aetiology, including PCa, and particularly where external exposures are concerned, given that epigenetics provides a molecular mechanism for the environment to directly infer disease susceptibility in an individual.[9]

Overall, the research presented throughout this dissertation is novel in its African-relevance and contributes to the African genomic knowledge economy. Although based on a small cohort, I believe this work provides a secure foundation on which the improvements and possible advances, discussed above, may be built. Additionally, future prospects addressed in this Chapter encompass budding fields within cancer genomics that promise to deliver on new and exciting discoveries. It is imperative that researchers grasp the immense value to be gained should future research efforts be focussed within the context of the African continent.

6.1. References

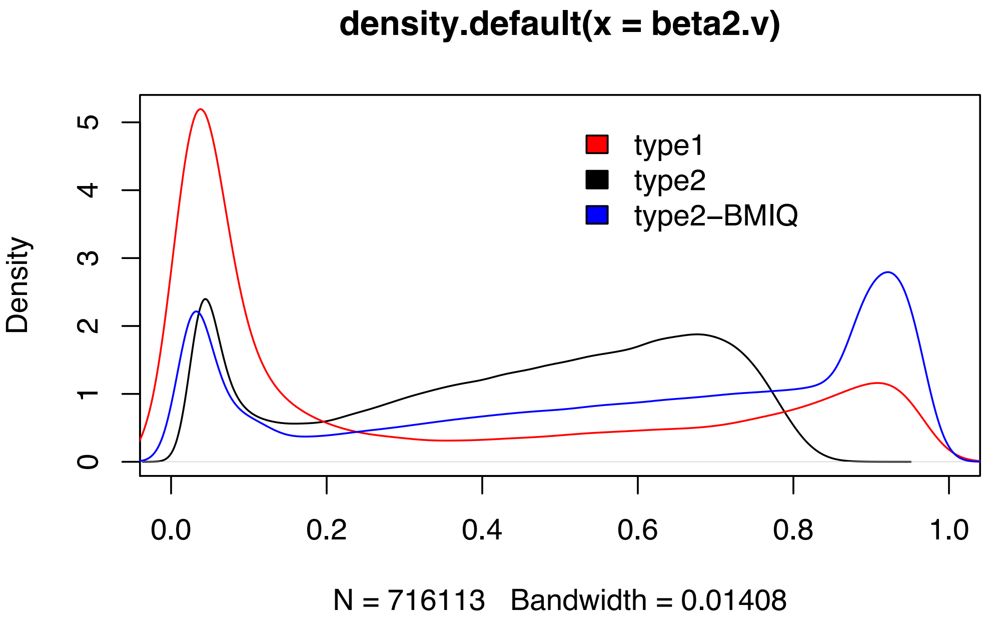
1. Wu MC, Kuan PF. A guide to Illumina BeadChip data analysis. In: Tost J, editor. DNA methylation protocols. Methods in molecular biology. New York: Humana Press; 2018. p. 303–30.
2. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13(1):86.
3. Mcgregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016; 17(84):1–17.
4. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, et al. Correcting for cell-type heterogeneity in epigenome-wide association studies: Revisiting previous analyses. Nat Methods. 2017; 14(3):216–7.
5. Kaushal A, Zhang H, Karmaus WJJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. BMC Bioinformatics. 2017; 18(1):216.
6. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28(6):882–3.
7. Daca-Roszak P, Pfeifer A, Żebracka-Gala J, Rusinek D, Szybińska A, Jarząb B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip array: Implications for comparative population studies. BMC Genomics. 2015; 16(1):1003.
8. Wang W, Auer P, Spellman SR, Carlson KSB, Nazha A, Maiers M, et al. Epigenomic signatures in myelodysplastic syndrome patients as predictors of donor compatibility and transplant outcome. Blood. 2019; 134:4557.
9. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007; 8(4):253–62.

Appendices

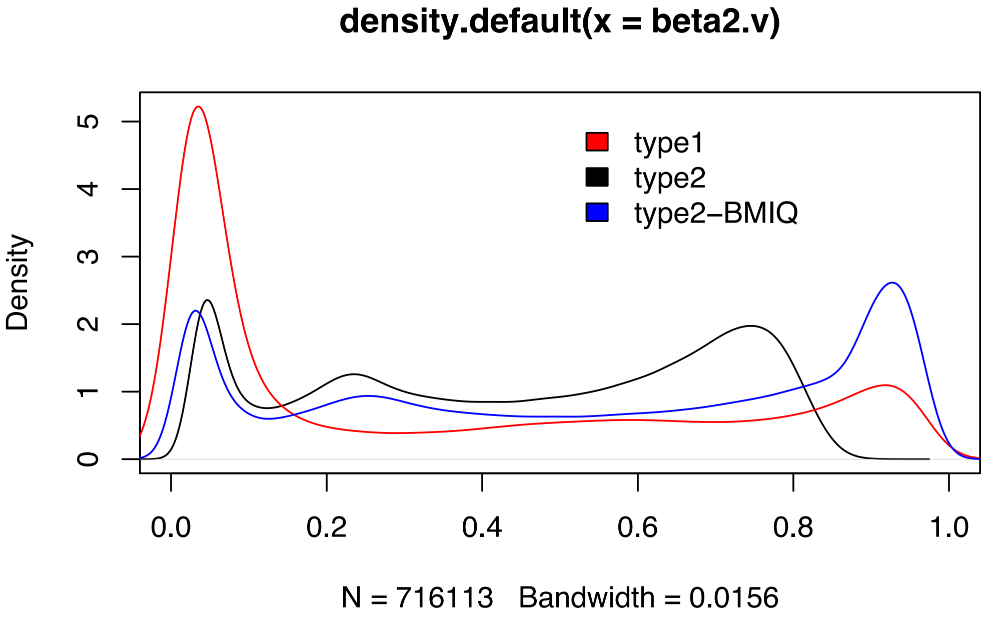
Appendix 1: Figure S1

Individual sample density plots displaying Infinium type I and Infinium type II probe beta-value distributions before and after BMIQ normalisation. Patients **a** UP2037, **b** UP2039, **c** UP2048, **d** UP2099, **e** UP2113, **f** UP2116, **g** UP2119 and **h** UP2133 are displayed.

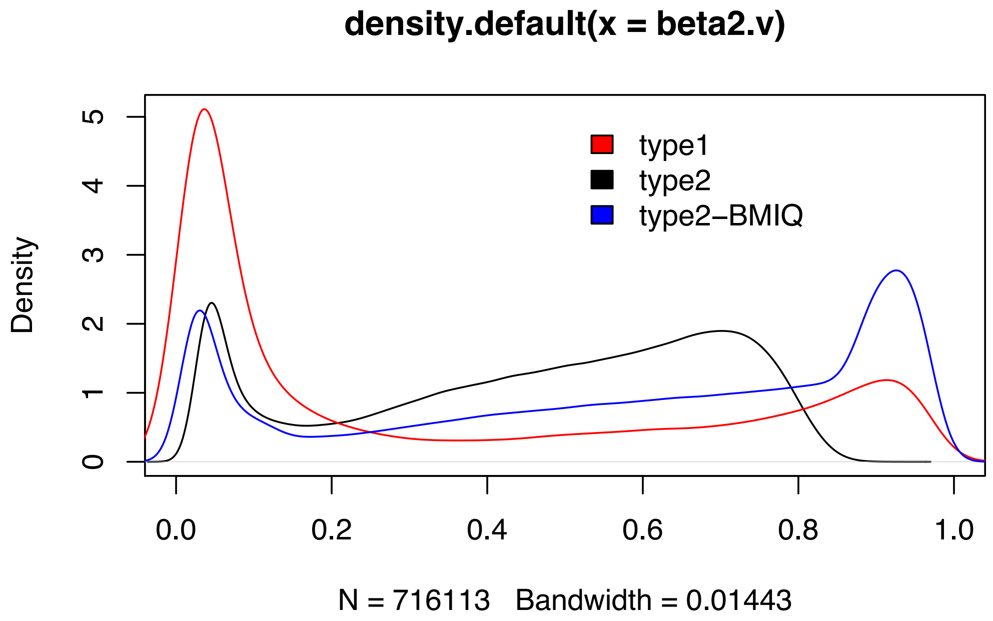
a



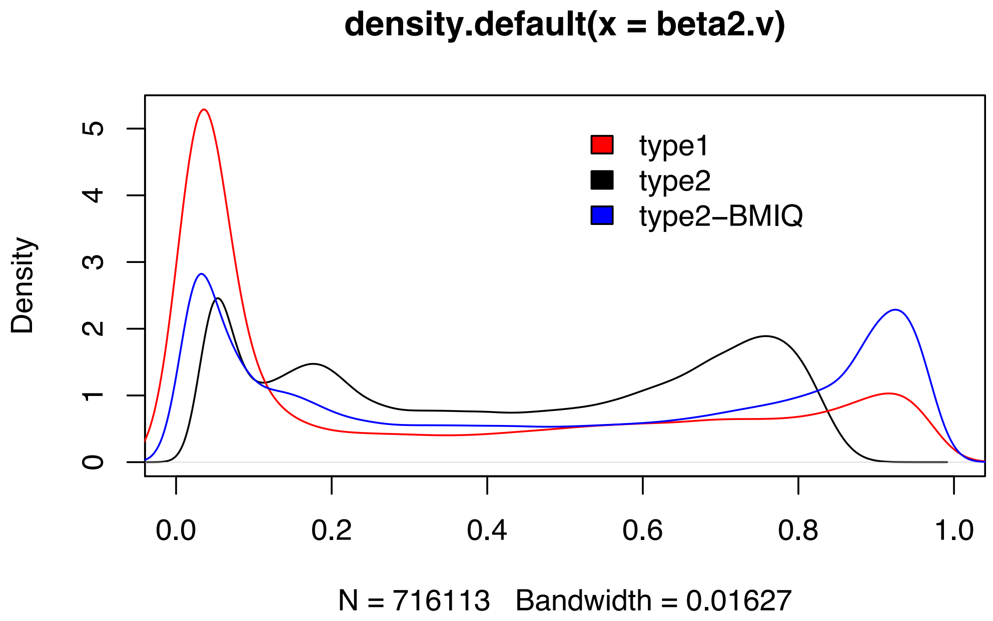
b



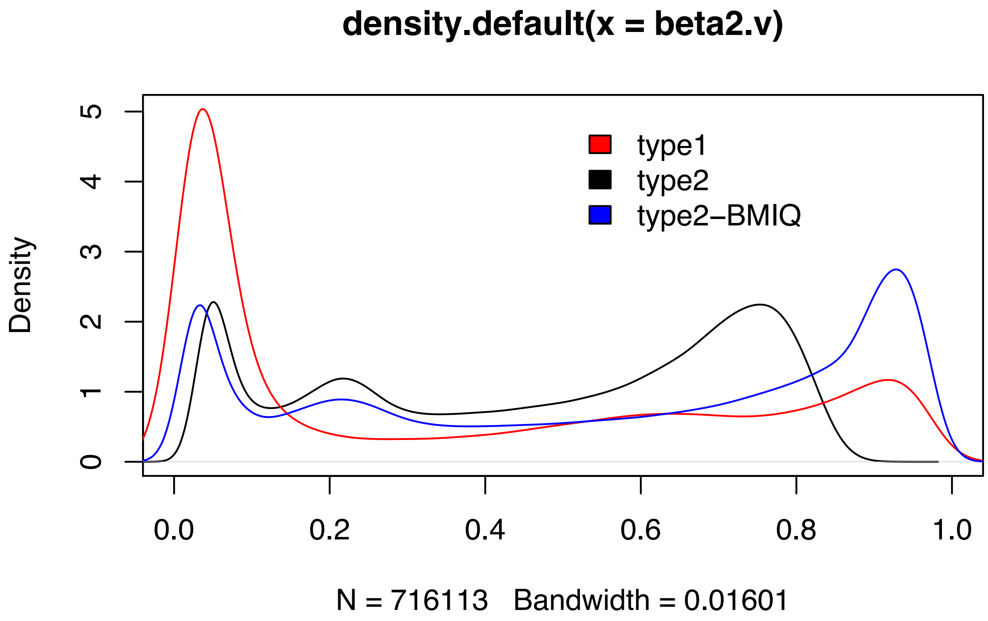
c


density.default(x = beta2.v)
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
🟥 type1
■ type2
🟦 type2–BMIQ
N = 716113   Bandwidth = 0.01443

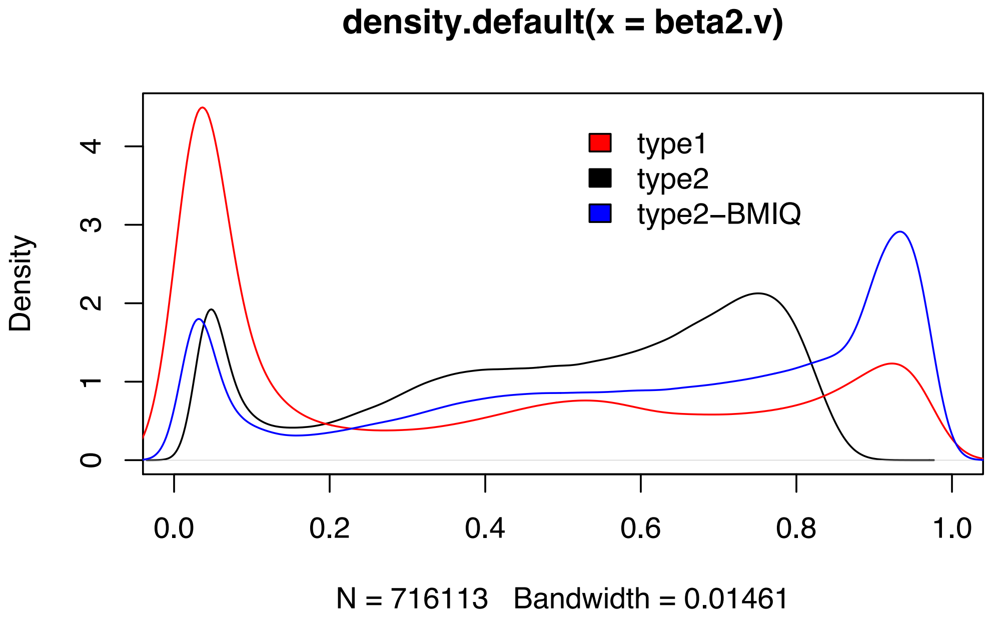
d


density.default(x = beta2.v)
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
🟥 type1
■ type2
🟦 type2–BMIQ
N = 716113   Bandwidth = 0.01627

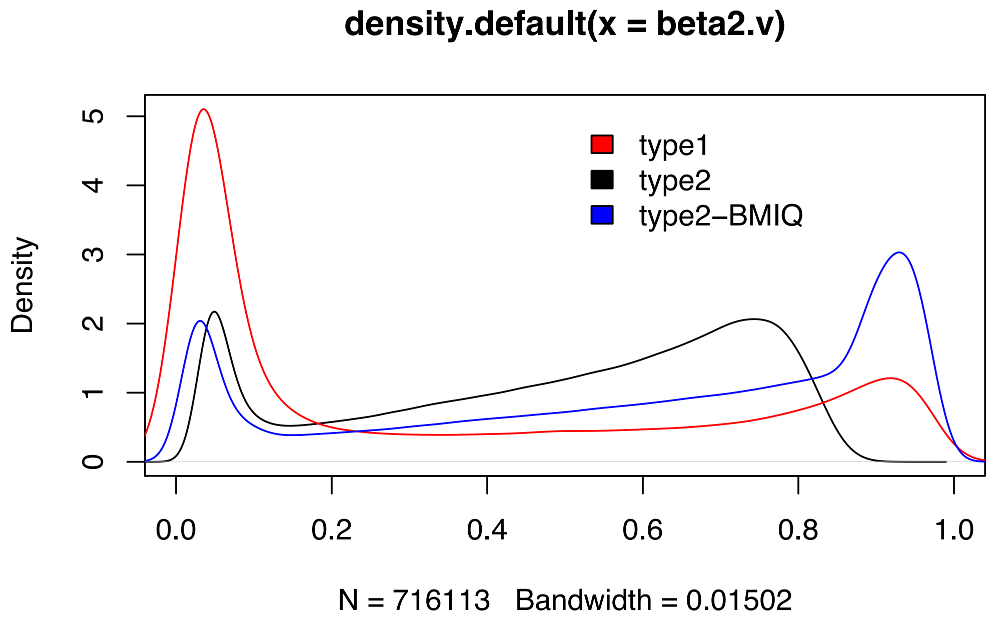
e


density.default(x = beta2.v)
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
🟥 type1
⬛ type2
🟦 type2–BMIQ
N = 716113 Bandwidth = 0.01601

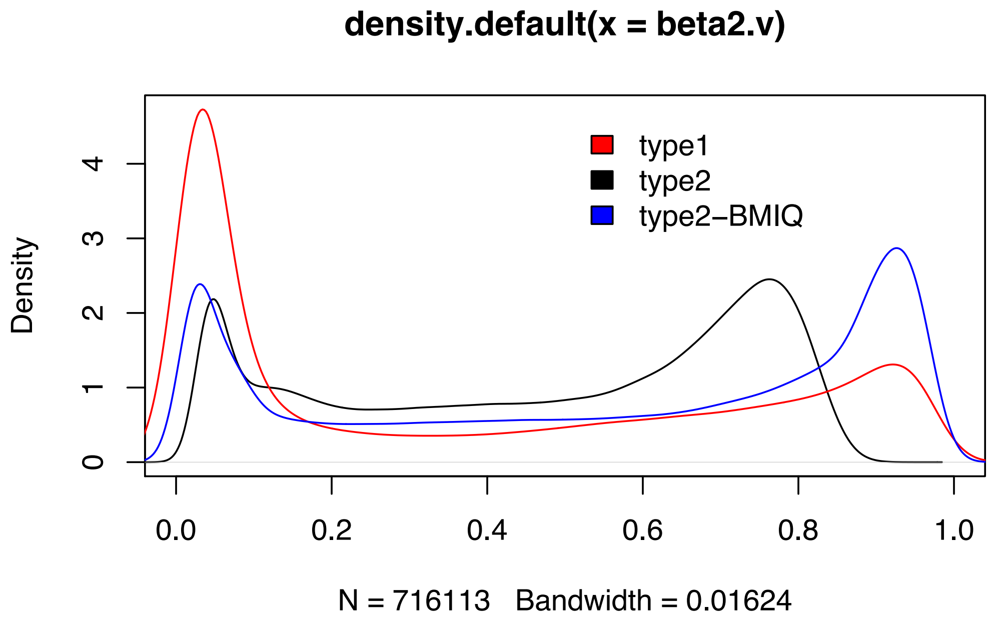
f


density.default(x = beta2.v)
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
🟥 type1
⬛ type2
🟦 type2–BMIQ
N = 716113 Bandwidth = 0.01461

g


density.default(x = beta2.v)
Density
0
1
2
3
4
5
🟥 type1
⬛ type2
🟦 type2–BMIQ
0.0
0.2
0.4
0.6
0.8
1.0
N = 716113   Bandwidth = 0.01502

h


density.default(x = beta2.v)
Density
0
1
2
3
4
🟥 type1
⬛ type2
🟦 type2–BMIQ
0.0
0.2
0.4
0.6
0.8
1.0
N = 716113   Bandwidth = 0.01624

Appendix 2: Table S1 (Additional File)

Probes overlapping African genetic variants, as identified by MethylToSNP, at targeted CpG sites, at single base extension sites (for Infinium type I probes) and within the body of the probe (48 base pairs for Infinium type I probes and 49 base pairs for Infinium type II probes). Annotation according to dbSNP release 147.

Attached and accessible via https://drive.google.com/drive/folders/1EGYbH3z5XXeBrpRpCnY7P73f7Aql84Mp?usp=sharing.

Filename: A2_MethylToSNP_variants.csv

Appendix 3: Table S2 (Additional File)

Probes overlapping genetic variants at targeted CpG sites, as identified by the established African patient-matched germline variant data method.

Attached and accessible via https://drive.google.com/drive/folders/1EGYbH3z5XXeBrpRpCnY7P73f7Aql84Mp?usp=sharing.

Filename: A3_EPIC_variants_CpG.csv

Appendix 4: Table S3 (Additional File)

Probes overlapping genetic variants at single base extension sites (for Infinium type I probes), as identified by the established African patient-matched germline variant data method.

Attached and accessible via https://drive.google.com/drive/folders/1EGYbH3z5XXeBrpRpCnY7P73f7Aql84Mp?usp=sharing.

Filename: A4_EPIC_variants_SBE.csv

Appendix 5: Table S4 (Additional File)

Probes overlapping genetic variants within the body of the probe (48 base pairs for Infinium type I probes and 49 base pairs for Infinium type II probes), as identified by the established African patient-matched germline variant data method.

Attached and accessible via https://drive.google.com/drive/folders/1EGYbH3z5XXeBrpRpCnY7P73f7Aql84Mp?usp=sharing.

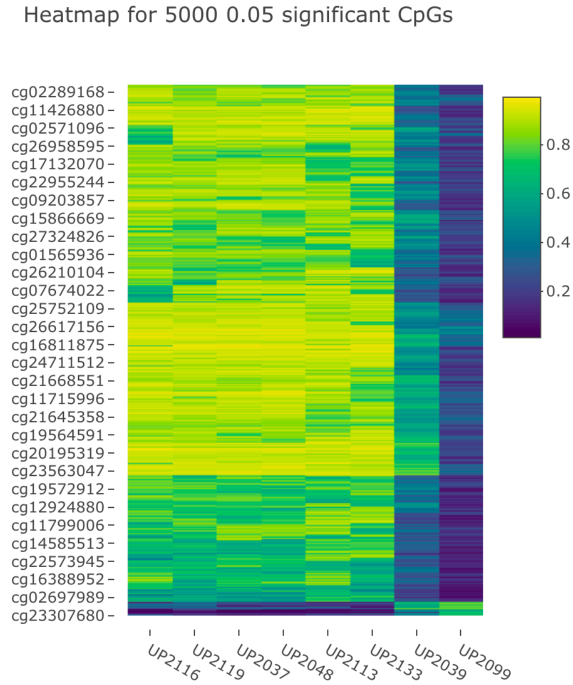
Filename: A5_EPIC_variants_Body.csv

Appendix 6: Figure S2

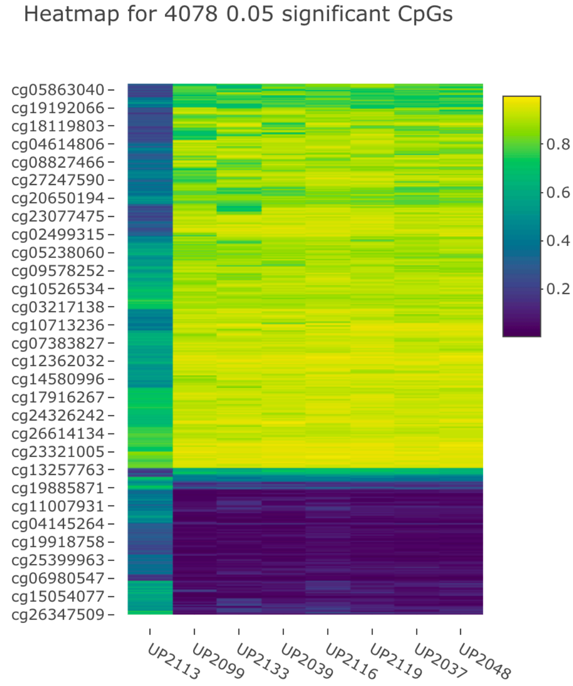
Heatmaps displaying significant differentially methylated probes (cg00000000) between African individuals for a number of clinical and genomic variables ($p < 0.05$). Displayed are DMPs associated with

a tumour purity, **b** TMB, **c** PGA, **d** SV calls, **e** MSI-H versus MSS, **f** CpG C > T variant count and **g** C > T variant count. Heatmaps show a maximum of 5,000 significant DMPs.

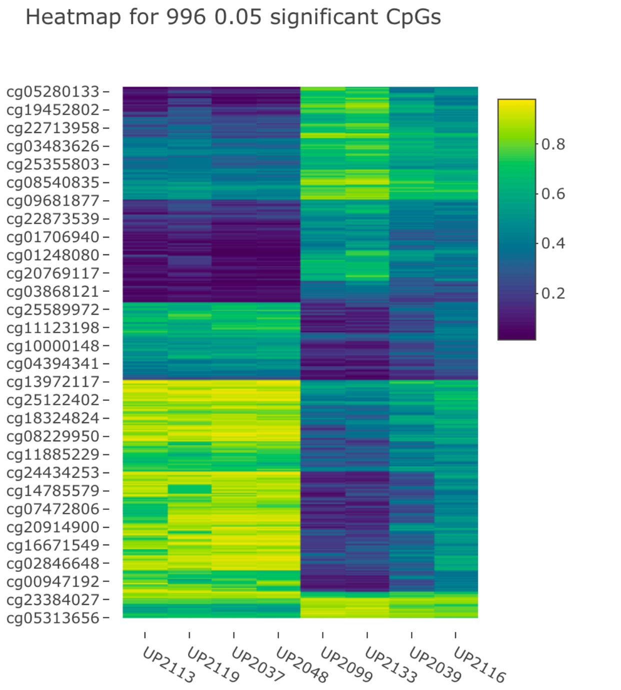
a



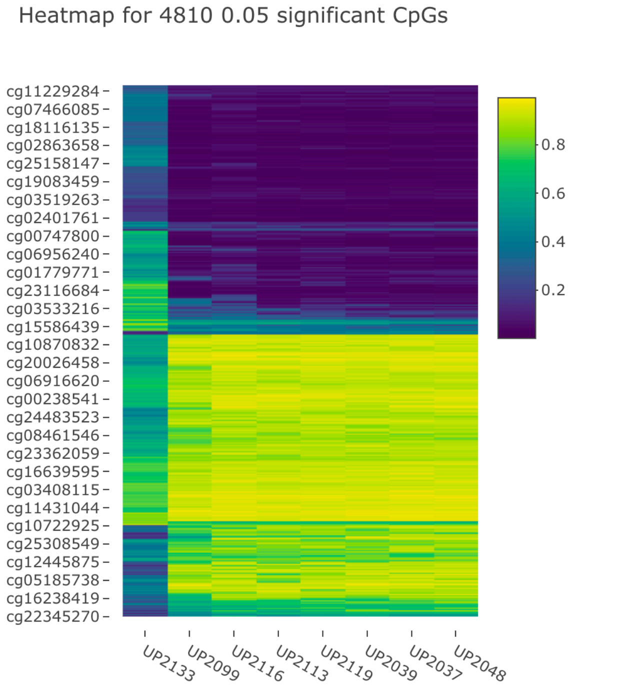
b



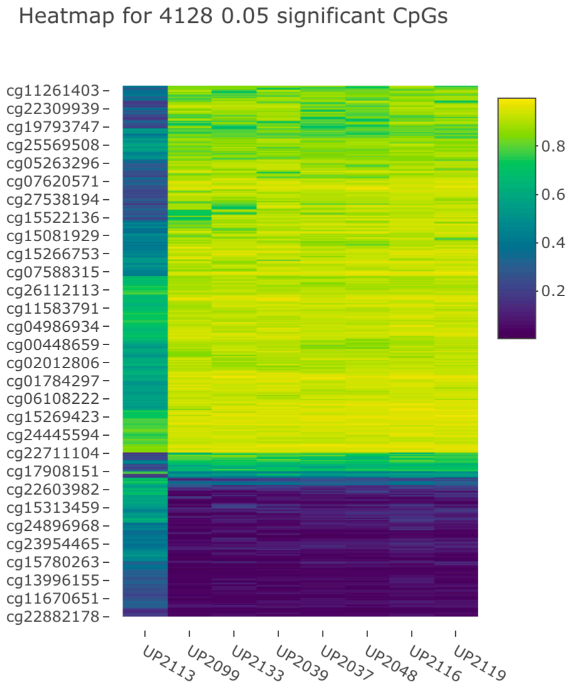
c


Heatmap for 996 0.05 significant CpGs
cg05280133
cg19452802
cg22713958
cg03483626
cg25355803
cg08540835
cg09681877
cg22873539
cg01706940
cg01248080
cg20769117
cg03868121
cg25589972
cg11123198
cg10000148
cg04394341
cg13972117
cg25122402
cg18324824
cg08229950
cg11885229
cg24434253
cg14785579
cg07472806
cg20914900
cg16671549
cg02846648
cg00947192
cg23384027
cg05313656
UP2113
UP2119
UP2037
UP2048
UP2099
UP2133
UP2039
UP2116
0.8
0.6
0.4
0.2

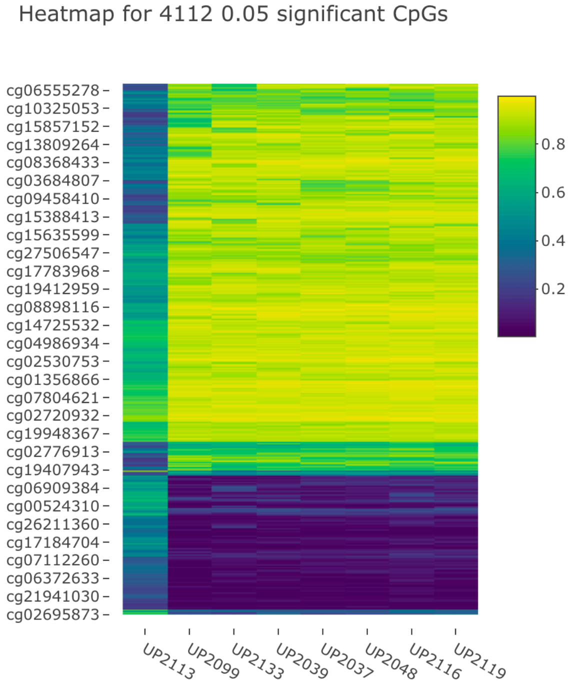
d


Heatmap for 4810 0.05 significant CpGs
cg11229284
cg07466085
cg18116135
cg02863658
cg25158147
cg19083459
cg03519263
cg02401761
cg00747800
cg06956240
cg01779771
cg23116684
cg03533216
cg15586439
cg10870832
cg20026458
cg06916620
cg00238541
cg24483523
cg08461546
cg23362059
cg16639595
cg03408115
cg11431044
cg10722925
cg25308549
cg12445875
cg05185738
cg16238419
cg22345270
UP2133
UP2099
UP2116
UP2113
UP2119
UP2039
UP2037
UP2048
0.8
0.6
0.4
0.2

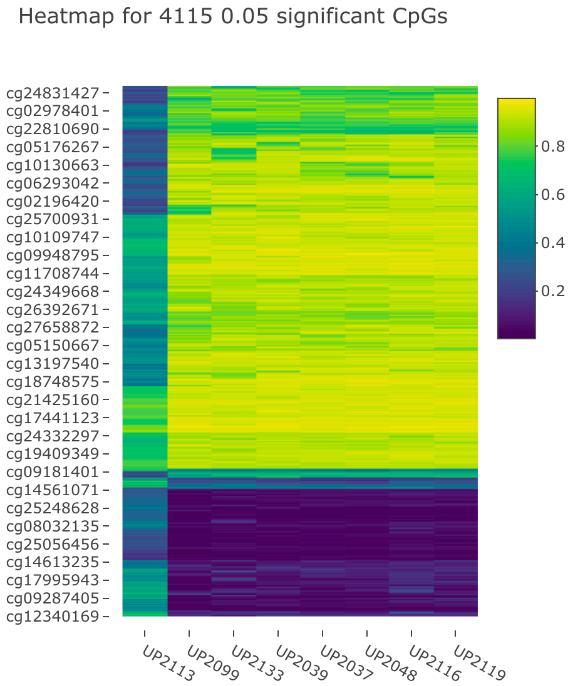
e


Heatmap for 4128 0.05 significant CpGs
cg11261403
cg22309939
cg19793747
cg25569508
cg05263296
cg07620571
cg27538194
cg15522136
cg15081929
cg15266753
cg07588315
cg26112113
cg11583791
cg04986934
cg00448659
cg02012806
cg01784297
cg06108222
cg15269423
cg24445594
cg22711104
cg17908151
cg22603982
cg15313459
cg24896968
cg23954465
cg15780263
cg13996155
cg11670651
cg22882178
UP2113
UP2099
UP2133
UP2039
UP2037
UP2048
UP2116
UP2119
0.8
0.6
0.4
0.2

f


Heatmap for 4112 0.05 significant CpGs
cg06555278
cg10325053
cg15857152
cg13809264
cg08368433
cg03684807
cg09458410
cg15388413
cg15635599
cg27506547
cg17783968
cg19412959
cg08898116
cg14725532
cg04986934
cg02530753
cg01356866
cg07804621
cg02720932
cg19948367
cg02776913
cg19407943
cg06909384
cg00524310
cg26211360
cg17184704
cg07112260
cg06372633
cg21941030
cg02695873
UP2113
UP2099
UP2133
UP2039
UP2037
UP2048
UP2116
UP2119
0.8
0.6
0.4
0.2

g


Heatmap for 4115 0.05 significant CpGs
cg24831427
cg02978401
cg22810690
cg05176267
cg10130663
cg06293042
cg02196420
cg25700931
cg10109747
cg09948795
cg11708744
cg24349668
cg26392671
cg27658872
cg05150667
cg13197540
cg18748575
cg21425160
cg17441123
cg24332297
cg19409349
cg09181401
cg14561071
cg25248628
cg08032135
cg25056456
cg14613235
cg17995943
cg09287405
cg12340169
0.8
0.6
0.4
0.2
UP2113
UP2099
UP2133
UP2039
UP2037
UP2048
UP2116
UP2119

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

Faculty of Health Sciences **Research Ethics Committee**

20 January 2022

Approval Certificate
Amendment

Dear Ms J Craddock,

Ethics Reference No.: 37/2021 – Line 1
Title: Differential genome-wide DNA methylation in prostate tumours from South African men

The **Amendment** as supported by documents received between 2021-11-23 and 2022-01-19 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2022-01-19 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Please remember to use your protocol number (37/2021) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

On behalf of the FHS REC, Dr R Sommers
MBChB, MMed (Int), MPharmMed, PhD
***Deputy Chairperson** of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).