

Imperial College
London

Statistical analysis of ‘-omics’ data

DEVELOPMENTS AND APPLICATIONS



A thesis submitted for the degree of Doctor of Philosophy by

Gianluca Campanella

Department of Epidemiology and Biostatistics, Imperial College London

May 2015

Declaration of originality

I hereby declare that this thesis represents my original work, and that any work (including data, tables, figures, and text citations) reproduced or adapted from any other sources is duly cited and referenced as such.

Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution-NonCommercial-NoDerivatives licence.

Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes, and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

In recent years, increasingly efficient molecular biology techniques created new opportunities to harness large-scale repositories of biological material collected in epidemiological studies; however, methods to manipulate and analyse the wealth of information thus generated have lagged behind.

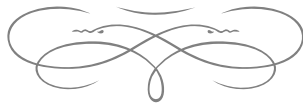
The introductory chapter of this thesis presents the multifaceted field of ‘computational epidemiology’ from the perspectives of molecular biology, measurement theory, and statistical modelling. Focusing on measurement of DNA methylation levels, the author also reviews the state of the art, proposes novel pre-processing methods and evaluation frameworks, and provides recommendations for genome-wide studies of DNA methylation levels using Illumina Infinium® HumanMethylation450 BeadChips.

The remaining chapters, in the form of three self-contained scientific articles, cover applications on the following topics: (i) DNA methylation differences associated with internal migration patterns within Italy; (ii) associations of DNA methylation profiles with adiposity measures, targeted gene expression, biomarkers of lipid and glucose metabolism, and risk of developing three obesity-associated diseases; (iii) associations of a dietary score with blood pressure, and with urinary metabolites as characterised by NMR spectroscopy.

The thesis is concluded with general remarks and the presentation of some open problems that offer potential for future research.



This page intentionally left blank.



Acknowledgements

I am indebted to my supervisors for their support and encouragement over the past three years. In particular, I wish to express my sincere gratitude to Prof. Paul Elliott, who created the necessary conditions for my PhD, and to Dr Marc Chadeau-Hyam. I am equally grateful to the Engineering and Physical Sciences Research Council (EPSRC) for their funding of this research project.

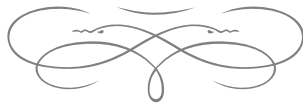
I would also like to thank my collaborators, who have provided me with much knowledge and inspiration, as well as supported my participation in various projects. My sincere appreciation is extended to Prof Paolo Vineis and Dr Silvia Polidoro at the Human Genetics Foundation (Turin, Italy), and to Dr Marc Gunter and the High Performance Computing support team, particularly Simon Burbidge, at Imperial College London.

Thanks to Raphaële and Florence of the ‘dragon team’, to my *amigos portugueses* Nina, Sonia and Filipe, and to my office mates Αρετή, Karin, Li and Pauline for the enjoyable environment they created. Thanks to Hayley of Energia Gym for putting me on a new path, to Sue and Hans for keeping me on it, and to Rad for setting me on yet another.

Un ringraziamento speciale ai miei genitori Anna e Vincenzo, senza i quali tutto questo non sarebbe stato possibile.



This page intentionally left blank.



Contents

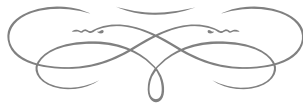
1	Introduction	1
1.1	'-omes' and '-omics'	3
1.1.1	Genomics	4
1.1.2	Epigenomics	7
1.1.3	Transcriptomics	9
1.1.4	Proteomics	10
1.1.5	Metabolomics and metabonomics	10
1.1.6	Tissue heterogeneity and temporal variability	11
1.2	Pre-processing	12
1.2.1	Illumina Infinium [®] assay	13
1.2.2	Infinium [®] HumanMethylation450 BeadChip	17
1.3	Statistical modelling	36
1.3.1	Univariate approaches	37
1.3.2	Multivariate approaches	50
	References	66
2	Analytical strategies for DNA methylation data	77
2.1	Pre-processing methods	78
2.1.1	Literature review	78
2.1.2	Novel methods	82

2.1.3	Evaluation framework	83
2.1.4	Comparison of pre-processing methods	86
2.2	Statistical models	98
2.2.1	Simulation study	98
2.2.2	Comparison of statistical models	102
2.3	Recommendations	107
2.3.1	Pre-processing method	107
2.3.2	Statistical model	107
2.4	Limitations	108
	References	109
3	Epigenetic signatures of internal migration in Italy	113
3.1	Introduction	113
3.2	Methods	115
3.2.1	Study population and sample selection	115
3.2.2	Laboratory analyses	118
3.2.3	Statistical analyses	118
3.3	Results	120
3.4	Discussion	121
	References	125
4	Epigenome-wide study of adiposity and future risk of obesity-related diseases	131
4.1	Introduction	132
4.2	Methods	133
4.2.1	Study population and sample selection	133
4.2.2	Laboratory analyses and data pre-processing	136
4.2.3	Statistical analyses	139

4.3	Results	143
4.3.1	Associations with adiposity measures	143
4.3.2	Associations with transcriptional activity	154
4.3.3	Associations with biomarkers of lipid and glucose meta- bolism	154
4.3.4	Associations with differential blood counts	157
4.3.5	Associations with obesity-related diseases	157
4.4	Discussion	159
	References	167
5	Effects of diet-induced metabolic acidosis on blood pressure	177
5.1	Introduction	177
5.2	Methods	179
5.2.1	Study population	179
5.2.2	Sample selection	179
5.2.3	DIMA indices	181
5.2.4	Laboratory analyses	182
5.2.5	Statistical analyses	182
5.3	Results	184
5.3.1	Dietary diversity	184
5.3.2	Associations with blood pressure and BMI	186
5.3.3	Associations with urinary amino acid concentrations . . .	187
5.3.4	Metabolome-wide association studies	190
5.4	Discussion	190
	References	195
6	Conclusions	201



This page intentionally left blank.



List of Figures

1 Introduction

1.1	Schematic overview of the main ‘-omics’	5
1.2	‘Address’ and ‘probe’ sequences	14
1.3	Infinium [®] I and Infinium [®] II assays	19
1.4	Intensities of hybridisation control beads	26
1.5	Intensities of normalisation control beads associated with sex . .	28
1.6	Intensities of ‘negative’ and ‘out-of-band’ beads	32
1.7	Ratio of ‘in-band’ to ‘out-of-band’ intensities	33
1.8	Scatter plot of intensities for rs2208123	35

2 Analytical strategies for DNA methylation data

2.1	Novel methods for dye bias correction	84
2.2	Correlations among pre-processing methods	90
2.3	Ternary plot of variance components using EPIC-Italy data	93
2.4	Ternary plot of variance components using EpiMigrant data . . .	94
2.5	Nuisance split and residual variance components across pre-pro- cessing methods	97
2.6	Performance metrics averaged across simulations	106

3	Epigenetic signatures of internal migration in Italy	
3.1	Signed Manhattan plot	122
3.2	Box-and-whisker plot across groups	123
3.3	PCA of DNA methylation levels assayed in the pericentric region on the long arm of chromosome 7	124
4	Epigenome-wide study of adiposity and future risk of obesity-related dis- eases	
4.1	Dose-response relationships between DNA methylation levels at cg06500161 (<i>ABCG1</i>) in the replication set and adiposity measures.	153
4.2	Receiver operating characteristic (ROC) curves for case-control lo- gistic regression models	161
5	Effects of diet-induced metabolic acidosis on blood pressure	
5.1	Partial correlation coefficients and 95% CIs for PRAL and its con- stituents with seven broadly defined food groups	185
5.2	Metabolome-wide association study based on samples from the first visit	191
5.3	Metabolome-wide association study based on samples from the second visit	192

List of Tables

1 Introduction

1.1	Classification of HM450 BeadChip assays by number of CpG loci in the ‘probe’ sequence and assay design	21
1.2	Normalisation control beads associated with sex	27

2 Analytical strategies for DNA methylation data

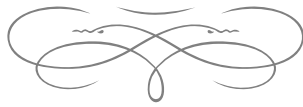
2.1	Datasets used in the evaluation of pre-processing methods	86
2.2	Correlations among pre-processing methods using EPIC-Italy data	88
2.3	Correlations among pre-processing methods using EpiMigrant data	89
2.4	Variance components estimated using EPIC-Italy data	95
2.5	Variance components estimated using EpiMigrant data	96
2.6	Values of ρ and r used in the simulation study and corresponding p -values	100
2.7	Percentages of simulations where the association with the ‘target’ CpG locus was declared significant	103
2.8	Median ranks of p -values for ‘target’ CpG loci across simulations	104
2.9	Average sensitivity across simulations	104
2.10	Percentages of simulations with at least one false positive finding	105
2.11	False discovery rate across simulations	105

3	Epigenetic signatures of internal migration in Italy	
3.1	Characteristics of participants included in the study	117
4	Epigenome-wide study of adiposity and future risk of obesity-related diseases	
4.1	Sample selection and characteristics of participants in the discovery set (at baseline for EPIC-Netherlands)	135
4.2	Sample selection and characteristics of participants in the replication set	137
4.3	Significant associations with BMI in the discovery set	145
4.4	Significant associations with waist circumference in the discovery set	146
4.5	Significant associations with waist-to-hip ratio in the discovery set	146
4.6	Significant associations with waist-to-height ratio in the discovery set	147
4.7	Sex-specific analyses of associations between DNA methylation levels and BMI identified in the discovery set	148
4.8	Sex-specific analyses of associations between DNA methylation levels and waist circumference identified in the discovery set	149
4.9	Sex-specific analyses of associations between DNA methylation levels and waist-to-hip ratio identified in the discovery set	150
4.10	Sex-specific analyses of associations between DNA methylation levels and waist-to-height ratio identified in the discovery set	151
4.11	Associations of DNA methylation levels at informative CpG loci with <i>cis</i> transcriptional activity in EnviroGenoMarkers data	155
4.12	Associations of DNA methylation levels at informative CpG loci with <i>cis</i> transcriptional activity in NOWAC data	156

4.13	Associations of DNA methylation levels at informative CpG loci with biomarkers of lipid and glucose metabolism	158
4.14	Nominally significant associations with post-menopausal breast cancer	159
4.15	Nominally significant associations with colorectal cancer	160
4.16	Nominally significant associations with myocardial infarction . . .	162
5	Effects of diet-induced metabolic acidosis on blood pressure	
5.1	Characteristics of participants included in the study	180
5.2	Associations of PRAL and NEAP with systolic BP	187
5.3	Associations of PRAL and NEAP with diastolic BP	188
5.4	Associations of PRAL and NEAP with BMI	188
5.5	Associations of PRAL and NEAP with urinary amino acid concentrations	189



This page intentionally left blank.



Nomenclature

$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\mathbf{I}_k$	Identity matrix of order k
$\mathbb{E}$	Expected value
$\mathbb{V}$	Variance or variance-covariance matrix
$\mathfrak{I}$	Fisher information matrix
p.d.f.	Probability density function
c.d.f.	Cumulative distribution function
$\mathcal{U}(a, b)$	Continuous uniform distribution on $[a, b]$
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}^2$
χ_k^2	χ^2 -distribution with k degrees of freedom

t_ν	Student's t -distribution with ν degrees of freedom
ϕ	Probability density function of $\mathcal{N}(0, 1)$
Φ	Cumulative distribution function of $\mathcal{N}(0, 1)$
Φ^{-1}	Quantile function of $\mathcal{N}(0, 1)$
i.i.d.	Independent and identically distributed
$\overset{a}{\sim}$	Distributed asymptotically
$\overset{\text{i.i.d.}}{\sim}$	Distributed independently and identically
$\mathcal{H}_0$	Null hypothesis
$\mathcal{H}_a$	Alternative hypothesis
ℓ	Log-likelihood function

Publications

Publications arising from this thesis

1. M Chadeau-Hyam, **G Campanella**, T Jombart, L Bottolo, L Portengen, P Vineis, B Liqueur, and R C H Vermeulen. 'Deciphering the complex: methodological overview of statistical models to derive '-omics'-based biomarkers'. *Environmental and Molecular Mutagenesis* **54.7** (2013), 542–557. Joint first authorship.
2. **G Campanella**, S Polidoro, C Di Gaetano, G Fiorito, S Guarrera, V Krogh, D Palli, S Panico, C Sacerdote, R Tumino, P Elliott, G Matullo, M Chadeau-Hyam, and P Vineis. 'Epigenetic signatures of internal migration in Italy'. *International Journal of Epidemiology* (2014).
3. **G Campanella**, M J Gunter, S Polidoro, C Sacerdote, V Krogh, D Palli, S Panico, R Tumino, S Guarrera, G Matullo, T M Sandanger, E Lund, P Peeters, B Bueno-de-Mesquita, P Georgiadis S A Kyrtopoulos, V Fiano, M Trevisan, P Elliott, M Chadeau-Hyam, and P Vineis. 'Epigenome-wide association study of adiposity and future risk of obesity-related diseases'. *The Journal of the American Medical Association* (2015). Under review.
4. **G Campanella**, B Lehne, S Polidoro, G Fiorito, S Guarrera, V Krogh, D Palli, S Panico, C Sacerdote, R Tumino, G Matullo, P Vineis, J C Chambers, P Elli-

- ott, and M Chadeau-Hyam. 'Pre-processing and statistical analysis of genome-wide DNA methylation data: the CH3 package' (2015). In preparation.
5. **G Campanella**, Q Chan, M L Daviglus, L Van Horn, K Miura, H Ueshima, L Zhao, J Stamler, J K Nicholson, P Elliott, and M Chadeau-Hyam. 'Effects of diet-induced metabolic acidosis on blood pressure' (2015). In preparation.

Other publications

1. S Polidoro, R Broccoletti, **G Campanella**, C Di Gaetano, E Menegatti, M Scoletta, E Lerda, G Matullo, P Vineis, D Berardi, C Scully, and P G Arduino. 'Effects of bisphosphonate treatment on DNA methylation in osteonecrosis of the jaw'. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **757.2** (2013), 104–113.
2. C A Demetriou, J Chen, S Polidoro, K van Veldhoven, C Cuenin, **G Campanella**, K Brennan, F Clavel-Chapelon, L Dossus, M Kvaskoff, D Droган, H Boeing, R Kaaks, A Risch, D Trichopoulos, P Lagiou, G Masala, S Sieri, R Tumino, S Panico, et al. 'Methylome analysis and epigenetic changes associated with menarcheal age'. *PLOS ONE* **8.11** (2013), e79391.
3. M Chadeau-Hyam, P Tubert-Bitter, C Guihenneuc-Jouyaux, **G Campanella**, S Richardson, R C H Vermeulen, M de Iorio, S Galea, and P Vineis. 'Dynamics of the risk of smoking-induced lung cancer: a compartmental hidden Markov model for longitudinal analysis'. *Epidemiology* **25.1** (2014), 28–34.
4. M Chadeau-Hyam, R C H Vermeulen, D G A J HeBELS, R Castagné, **G Campanella**, L Portengen, R S Kelly, I A Bergdahl, B Melin, G Hallmans, D Palli, V Krogh, R Tumino, C Sacerdote, S Panico, T M C M de Kok, M T Smith, J C S Kleinjans, P Vineis, and S A Kyrtopoulos. 'Prediagnostic transcriptomic

- markers of chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis'. *Annals of Oncology* **25.5** (2014), 1065–1072.
5. F Guida, T M. Sandanger, R Castagné, **G Campanella**, S Polidoro, D Palli, V Krogh, R Tumino, C Sacerdote, S Panico, G Severi, S A Kyrtopoulos, P Georgiadis, R C H Vermeulen, E Lund, P Vineis, and M Chadeau-Hyam. 'Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation'. *Human Molecular Genetics* **24.8** (2015), 2349–2359.
 6. S Stringhini, S Polidoro, C Sacerdote, R S Kelly, K van Veldhoven, C Agnoli, S Grioni, R Tumino, M C Giurdanella, S Panico, A Mattiello, D Palli, G Masala, V Gallo, R Castagné, F Paccaud, **G Campanella**, M Chadeau-Hyam, and P Vineis. 'Life-course socioeconomic status and DNA methylation of genes regulating inflammation'. *International Journal of Epidemiology* (2015).
 7. J C Chambers, M Loh, B Lehne, A Drong, J Kriebel, V Motta, S Wahl, H R Elliott, F Rota, W R Scott, W Zhang, S-T Tan, **G Campanella**, M Chadeau-Hyam, L Yengo, R C Richmond, M Adamowicz-Brice, U Afzal, K Bozaoglu, Z Y Mok, et al. 'Epigenome-wide association identifies DNA methylation markers in peripheral blood that predict incident type 2 diabetes amongst Indian Asians and Europeans'. *The Lancet Diabetes & Endocrinology* (2015). In press.
 8. B Liqueur, L Bottolo, **G Campanella**, S Richardson, and M Chadeau-Hyam. 'R2GUESS: a Graphics Processing Unit-based R package for Bayesian variable selection regression of multivariate responses'. *Journal of Statistical Software* (2015). In press.
 9. F Fasanelli, L Baglietto, E Ponzi, F Guida, **G Campanella**, M Johansson, K Grankvist, M Johansson, M Assumma, A Naccarati, M Chadeau-Hyam, B De Stavola, A Hodge, G Giles, M Southey, C Relton, P Haycock, E Lund,

- S Polidoro, T Sandanger, et al. 'Hypomethylation of smoking-related genes is associated with future lung cancer in three prospective cohorts'. *Nature Communications* (2015). Under review.
10. N Kato, M Loh, F Takeuchi, N Verweij, X Wang, W Zhang, T Kelly, D Saleheen, B Lehne, I Mateo Leach, A Drong, J Abbott, S-T Tan, W Scott, **G Campanella**, et al. 'Trans-ethnic genome-wide association study identifies multiple genetic loci influencing blood pressure traits, and implicates a role for DNA methylation: the International Genetics of Blood Pressure (iGEN-BP) study'. *Nature Genetics* (2015). Under review.
 11. **G Campanella**, M De Iorio, A Jasra, and M Chadeau-Hyam. 'The TimeMachine for inference on stochastic trees' (2015). In preparation.
 12. F Cordero, G Ferrero, S Polidoro, G Fiorito, **G Campanella**, C Sacerdote, V Krogh, D Palli, S Panico, R Tumino, P Vineis, and A Naccarati. 'Epigenome-wide study in prediagnostic samples from the European Prospective Investigation into Nutrition and Cancer (EPIC-Italy) cohort: differentially methylated microRNAs in subjects who developed breast cancer' (2015). In preparation.

Chapter 1

Introduction

The completion of the Human Genome Project in 2001 marked the dawn of the ‘high-throughput’ era of molecular biology, and was the harbinger of a new multifaceted field, which can be termed ‘computational epidemiology’, that would emerge from the spread of the high-throughput revolution to traditional epidemiological research.

In the last decade, increasingly efficient standardised laboratory protocols brought about substantial reduction in costs, and created new opportunities to harness large-scale repositories of biological material collected in epidemiological studies. The myriad studies conducted in recent years have yielded a wealth of information at various biological scales. Nonetheless, rigorous methods to ma-

nipulate and analyse these data have lagged relatively behind: classic statistical theory arose to a large extent in settings with few predictors and a comparatively larger number of observations, not in the ‘small n , large p ’ scenario in which ‘-omics’ data are generated. These datasets are also almost invariably characterised by complex correlation structures, sensitivity to experimental conditions, and varying degrees of plasticity in response to endogenous and exogenous exposures including ageing.

Computational epidemiology lies at the intersection not only of molecular biology and statistical modelling, but also of measurement theory and computer science. This is reflected in the organisation of this thesis.

The main theoretical concepts applied in chapters 2 to 5 are introduced in the rest of this chapter. In particular, section 1.1 defines ‘-omes’ and ‘-omics’, and presents the main ‘-omes’ that are currently amenable to agnostic study by high-throughput laboratory techniques. Section 1.2 introduces the measurement process for ‘-omics’, with a particular emphasis on the Infinium[®] assay by Illumina, Inc. Applications of this technology to measurement of DNA methylation levels at a genome-wide scale are also discussed. Finally, section 1.3 is a state-of-the-art review of statistical models for the analysis of ‘-omics’ data: commonly used univariate and multivariate methods are presented, and their strengths and weaknesses are discussed.

Chapter 2 provides a comprehensive review and reasoned comparison of analytical strategies for DNA methylation data generated using Illumina Infinium[®] HumanMethylation450 BeadChips. Section 2.1 illustrates a set of pre-processing methods that have been proposed to deal with technically-induced nuisance variation, presents two novel pre-processing methods, and proposes a general evaluation framework based on variance partitioning; this is also used to assess the performance of the pre-processing methods introduced previously. Section 2.2

compares several model choices (as presented in section 1.3) for DNA methylation levels by means of a simulation study. Finally, section 2.3 provides recommendations for genome-wide investigations of DNA methylation levels in terms of both pre-processing method and choice of statistical model, and section 2.4 highlights some limitations of these analyses and conclusions.

Chapters 3 to 5 cover practical applications of the concepts presented in this chapter; these take the form of self-contained scientific articles. Chapter 3 presents a study of DNA methylation differences in southern Italian participants of the EPIC-Italy cohort that migrated to north-Western Italy. Chapter 4 describes a genome-wide association study of DNA methylation levels with adiposity measures, targeted gene expression, biomarkers of lipid and glucose metabolism, and risk of developing three obesity-associated diseases. Finally, chapter 5 reports associations of two related dietary scores with blood pressure, and with urinary metabolites as characterised by proton NMR spectroscopy.

Chapter 6 concludes with general remarks and the presentation of some open problems that offer potential for future research.

§ 1.1 ‘-omes’ and ‘-omics’

The suffix ‘-ome’ is used in biology to form nouns with the sense ‘all of the specified constituents of a cell, considered collectively or in total’ [1]. The suffix ‘-omics’ denotes the study of specific ‘-omes’, but also carries additional ‘holistic’ connotations: ‘-omics’ aim at characterising the constituents of an ‘-ome’ collectively, endeavouring to explain how dynamic interactions among these entities bring about higher-order structure and function. The ‘-omics’ paradigm is also understood to encompass a certain ‘agnostic’ *modus operandi* of data analysis, which relies on large amounts of data and general ideas and computational

tools, rather than trying to lay a mathematical foundation upon which biological phenomena may be understood [2].

The following sections cover ‘-omes’ for which established high-throughput laboratory techniques are available, and that can thus be studied using the ‘agnostic’ approach described above (fig. 1.1). Many other ‘-omes’ exist, but large-scale applications of the associated ‘-omics’ techniques are often thwarted by financial considerations and time constraints.

1.1.1 Genomics

Macroscopic chromosomal rearrangements (including deletions, duplications, inversions, and translocations) that affect millions to hundreds of millions of base pairs are discernible by optical microscopy, and were the first genetic variations to be reported [4, 5]. They are almost always clinically evident if not outright pathogenic; a prototypical example is partial or complete duplication of chromosome 21, which causes Down’s syndrome. These cytogenetic aberrations occur only rarely, and thus cannot underlie inter-individual differences in phenotypic characters such as physical traits, disease susceptibility, and pharmacokinetics.

The first breakthrough in understanding the genetic basis of phenotypic variation was the completion of the first draft of the sequence of the human genome in 2001 [6, 7]. Automated analysis led to the discovery of millions of point mutations, with distinct patterns of variation among populations. Variations with a minor allele frequency (MAF) greater than or equal to 1% in at least one population came to be known as single nucleotide polymorphisms (SNPs). SNPs are the most common type of genetic variation: it is estimated that they occur at a rate of one every 300 base pairs, which translates to approximately 10 million SNPs in the entire human genome, or about 0.3% by length. SNPs nearer to each other are less likely to be separated during meiotic chromosomal crossover, and are

	Supporting Structure	Platforms (log ₁₀ order of magnitude)	Features
	Genome	DNA Microarrays (6) Sequencing (9)	Categorical data Distance-driven correlation Extremely stable over time
	Epigenome	DNA methylation Histone modifications Non-coding RNA Microarrays (5) Bisulfite sequencing (1)	Continuous data Affected by time and exposures (with reduced plasticity)
	Transcriptome	mRNA Microarrays (5) RNA sequencing (9)	Continuous data Affected by time and exposures Strong measurement noise
	Proteome	Proteins Microarrays (5) Mass spectrometry (5)	Continuous data Affected by time and exposures
	Metabolome	Small molecules Mass spectrometry (5) NMR spectroscopy (4)	Continuous data Structured correlation Strongly affected by exposures

Figure 1.1: Schematic overview of the main ‘-omics’. These ‘-omics’ are amenable to high-throughput laboratory analysis, and can thus be studied in an ‘agnostic’ manner (from Chadeau-Hyam et al. [3]).

thus more likely to be inherited as a group; this genetic linkage gives rise to linkage disequilibrium at the population level, which results in the archetypal distance-driven correlation pattern observed among neighbouring SNPs [8]. Most SNPs are located in intergenic regions, where they cannot alter protein function, and thus appear to have no effect. Others are found within coding or regulatory regions, and may therefore play a more direct role in gene expression, and ultimately phenotypic variation [9]. SNPs in exons are classified as synonymous and non-synonymous, depending on whether the different alleles encode the same or different amino acids (synonymy is afforded by the redundancy of the genetic code). Synonymous SNPs, as well as SNPs in introns or regulatory regions, may still influence the amount and function of gene products, for example by altering mRNA splicing and structure, rate of translation, binding of regulatory miRNAs, and even protein folding [10].

In addition to SNPs, in 2006 Redon et al. [11] reported the existence of almost 1,500 relatively common submicroscopic chromosomal changes in the genomes of 270 healthy individuals. These changes, which cover approximately 12% of the genome by length, were named copy-number variations (CNVs), and defined as affecting more than 1,000 to more than half a million base pairs. As for SNPs, the phenotypic manifestations of CNVs also ensue from altered transcriptional and translational activity when these occur in exons or regulatory regions; however, there seems to be little overlap between SNPs and CNVs in terms of their effects on gene expression [12]. In addition, the presence of CNVs in the coding regions of dosage-sensitive genes may also cause haploinsufficiency [13, 14].

The identification of previously unreported SNPs and CNVs can only proceed from sequencing data, which are often costly to obtain. A cheaper, partly untargeted alternative to whole-genome sequencing is exome sequencing, which is restricted to known or putative coding regions. Large databases of known

variations exist to coordinate this international collaborative research effort [15]. Conversely, large-scale genome-wide association studies (GWAS) can rely on the cheaper targeted genotyping technology afforded by micro-arrays, which can be used to analyse up to a few million markers simultaneously [16, 17].

1.1.2 Epigenomics

The initial enthusiasm of the Human Genome Project was curtailed when increasingly larger GWAS failed to demonstrate a prominent role of genetic variations in complex or multifactorial disorders.

The use of genetic information in cells is regulated at multiple interacting levels, with epigenetic modifications as central modulators of transcriptional activity. Among epigenetic mechanisms, only DNA methylation, which involves modification of cytosine bases (in vertebrates) by covalent addition of a methyl group to form 5-methylcytosines, is amenable to quantification on a genome-wide scale at present. Since this modification occurs almost exclusively in the context of CpG dinucleotides (also known as CpG loci), the genomic location of these alterations is well-defined. This is in contrast with histones, whose location is dynamic and depends even on the very chemical modifications one is trying to measure [18, 19], making it necessary to sequence the DNA attached to modified histones after these have been singled out e.g. by immunoprecipitation.

CpG loci are under-represented (around 20% of the expected frequency) in the genome due to the tendency of 5-methylcytosines to spontaneously deaminate to thymine (on evolutionary time scales); however, long runs of CpGs are found in regions known as CpG islands, which are often located close to promoters [20]. Generally, DNA methylation represses gene expression by attracting proteins that block transcription. CpG loci in promoter regions are thus mostly unmethylated, whereas those in non-promoter regions are mostly methylated to

reduce transcriptional noise [21, 22]. The correlation structure of DNA methylation levels is poorly understood. It is only marginally distance-driven, and is complicated by *cis* associations with genetic variations. It also exhibits a non-local functional component, for example between CpG loci in promoter and transcribed regions, irrespective of genetic distance; this component may also reflect co-regulation patterns akin to the co-expression patterns observed in transcriptomics data.

Three main families of techniques exist to assess DNA methylation levels on a genome-wide scale, based on restriction enzyme digestion, affinity enrichment, and bisulphite conversion [23–25]. Methods based on enzyme digestion exploit the different sensitivities to DNA methylation, and corresponding cutting behaviours, of restriction endonucleases. In approaches based on affinity enrichment such as methylated DNA immunoprecipitation (MeDIP), 5-methylcytosines are recognised by monoclonal antibodies; this is then followed by array-based hybridisation (MeDIP-chip) or sequencing (MeDIP-seq) to determine the proportion of methylated DNA. Finally, in methods based on bisulphite conversion, DNA methylation is measured by querying a purposely created cytosine/thymine polymorphism. First, sodium bisulphite is used to deaminate cytosine into uracil, without affecting 5-methylcytosine. During a subsequent amplification step, uracil is amplified as thymine, whereas 5-methylcytosine is amplified as cytosine, effectively creating the polymorphism. DNA methylation levels can thus be estimated (using either micro-arrays or sequencing techniques) as the proportion of ‘cytosines’ (i.e. 5-methylcytosines that were not affected by bisulphite conversion) at each locus.

There are approximately 30 million CpG loci in the human genome, which can be analysed using untargeted techniques such as whole-genome bisulphite sequencing; however, these methods incur considerable costs. Cheaper tar-

geted alternatives include hybridisation micro-arrays such as the commonly used Infinium[®] HumanMethylation450 BeadChip (described in further detail in section 1.2.2), which measures DNA methylation levels at approximately half a million CpG loci [26].

1.1.3 Transcriptomics

Transcriptomics aims at cataloguing all transcripts produced by the cell, and provides a means to study gene fusions and mutations, alternative splicing, and post-transcriptional regulation by miRNA and other ncRNAs. In addition, mRNAs levels are often used as proxy of gene expression, owing to the predominance of transcriptional regulation over other post-transcriptional regulatory events.

The correlation structure of transcriptomics data is highly non-local, and reflects co-expression patterns among sets of genes with similar or complementary functions [27].

As for genotyping, two main approaches to transcriptome profiling exist. The untargeted and more expensive option is RNA sequencing (RNA-seq). RNA-seq uses deep-sequencing technologies that allow unbiased measurement of all transcripts in a cell, thus providing a more accurate picture of transcriptional activity (including splice variants) [28]. The targeted option, which requires *a priori* characterisation of genes, gene candidates, and splice variants, employs micro-array hybridisation preceded by amplification to complementary DNA (cDNA) using reverse transcriptase, and labelling by incorporation of biotin-labelled nucleotides (biotinylation).

1.1.4 Proteomics

Proteomics is the large-scale study of all proteins produced by the cell, and of other modulating factors such as post-translational modifications and protein-protein interactions.

The correlation structure of proteomics data reflects that of the underlying transcriptome to some extent, and thus the structure of co-expression networks; accordingly, it is akin to that of transcriptomics data.

While targeted micro-arrays that use fluorescent dye labelling of immobilised proteins have been developed, the mainstay of proteomics is untargeted analysis by mass spectrometry (MS) [29, 30]. The typical workflow of MS-based procedures involves isolation of proteins from the biological matrix, followed by separation (usually by gel electrophoresis) and enzymatic digestion (e.g. using trypsin). After these initial steps, protein fragments are separated by liquid chromatography, and injected into the mass spectrometer. Finally, a putative list of proteins (and their concentrations) is constructed by comparing spectra against databases of peptide signatures.

1.1.5 Metabolomics and metabonomics

Metabolomics and metabonomics are two closely related ‘-omics’ focused on the study of chemical fingerprints (in the form of small-molecule metabolites) of specific cellular processes. Metabonomics places a particular emphasis on dynamics and sensitivity, and is defined as ‘the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification’ [31].

Both these ‘-omics’ rely on classic analytical chemistry techniques that have been adapted to allow for high-throughput automated sample processing. These

include mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy, usually in the form of ^1H (proton) and ^{13}C NMR spectroscopy [32–34].

The correlation structure of metabolomics data is particularly complex, and can be understood at three different levels; there exist: (i) a local component emerging from similarities among certain features such as mass-to-charge ratios and chemical shifts in MS- and NMR-based methods, respectively; (ii) a non-local component corresponding to several correlated features arising from one compound; (iii) a biological component related to multiple molecules involved in the process of interest.

As for proteomics, metabolomics experiments can be targeted or untargeted. Quantification of specific compounds is sometimes performed (usually by MS), e.g. for validation purposes; other times, concentrations of a relatively large number of related compounds are measured, as in the case of the MS-based platform developed by Biocrates Life Sciences AG. In untargeted analyses, it has become increasingly common to conduct multiple complementary experiments that allow the untargeted detection of a broad range of molecules across the metabolome [35].

1.1.6 Tissue heterogeneity and temporal variability

With the exception of genomics, which deals with essentially immutable genetic information that is faithfully stored in identical copies across all somatic cells, most other ‘-omics’ are affected by tissue heterogeneity and temporal variability.

Tissue heterogeneity refers to differences in ‘-omics’ profiles (other than genomics) observed across different cell types, especially in mixed tissues such as whole blood [36, 37]. It is mostly a consequence of cellular differentiation, though it is plausible that certain cell types may respond differently to endogenous and exogenous stimuli, thus establishing different ‘-omics’ profiles. Since ‘-omics’

technologies usually average measurements across a large number of possibly heterogeneous cells, the question arises whether phenotypic differences of interest can be linked to the ‘-omics’ profiles under consideration, or whether there exists some higher-order structure (such as the proportion of a specific type of cells) that could explain the observed variation equally well. To be sure, higher-order differences will always be reflected in finer structures such as those that can be appreciated using the ‘magnifying glass’ of ‘-omics’ technologies; however, this distinction may still be relevant for cost-effectiveness considerations. Another issue that stems from tissue heterogeneity is the value of ‘-omics’ measurements in surrogate tissues (such as whole blood) for localised diseases (such as solid tumours). The use of surrogate tissues is motivated mostly by ease of access, and the availability of a large number of stored samples from prospective epidemiological studies.

Temporal variability refers to differences in ‘-omics’ profiles (other than genomics) arising from ultradian, circadian, and infradian rhythms [38]. These oscillations add a dynamic layer on top of a longer-term component associated with chronological ageing. Little is known in this regard for ‘-omics’ other than genomics [39], though the potential for temporal variability is obviously greater as one moves outwards from the genome to the metabolome. Lack of temporal stability limits epidemiological population studies, which can often rely only on stored samples collected at a single time point for each participant.

§ 1.2 Pre-processing

Almost all ‘-omics’ platforms suffer from low signal-to-noise ratio, meaning that biological variability is often hidden behind large amounts of technically-induced variation introduced by complicated laboratory protocols.

Two complementary approaches deal with this problem. First, raw data acquired by laboratory instrumentation can be pre-processed to reduce the amount of systematic variation; these methods are reviewed in this section. Second, statistical models can be adjusted for known sources of technical variability, thus further attenuating their effect; commonly used models, as well as their strengths and weaknesses, are reviewed in section 1.3.

1.2.1 Illumina Infinium[®] assay

The Infinium[®] assay developed by Illumina, Inc. is an integral part of the company's range of micro-arrays (known as BeadChips), and has a multitude of applications, ranging from SNP genotyping and CNV detection, to measurement of DNA methylation levels and abundance of RNA transcripts [40]. Its versatility stems from a general building principle, first described by Michael et al. [41], that allows packing densities many times higher than those obtainable using traditional spotted micro-arrays.

In particular, BeadChips are randomly assembled: a mixture of microscopic silica beads is randomly distributed on an optical substrate (e.g. an etched microarray) containing millions of micrometre-scale wells into which beads assemble. Each bead is covalently attached to multiple copies of an oligonucleotide that defines its type: the first part of the sequence defines the 'address', whose purpose will become clear later; the second part defines the 'probe', and is complementary to some predefined genomic region that constitutes its target (fig. 1.2). Each 'probe' sequence is provided with an 'address' sequence that uniquely identifies it.

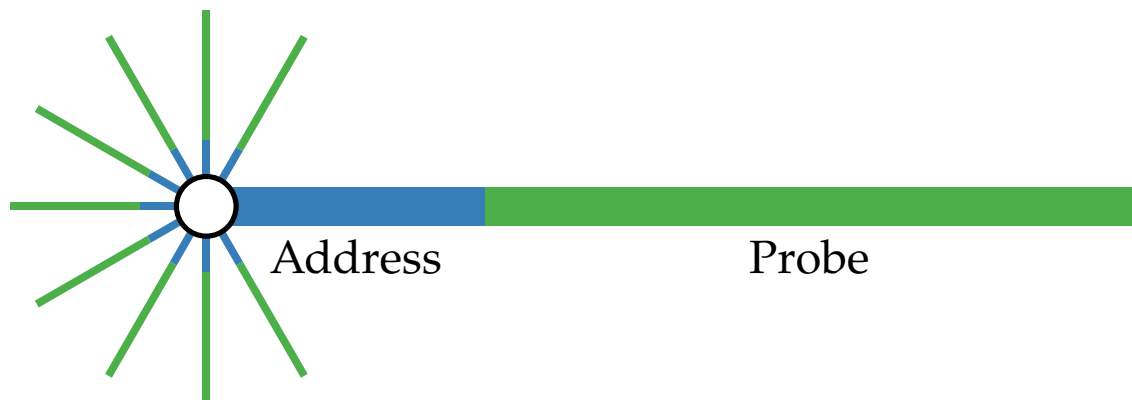


Figure 1.2: ‘Address’ and ‘probe’ sequences. Each bead is covalently attached to multiple copies of an oligonucleotide divided into ‘address’ and ‘probe’ sequences.

Probe sequence

In typical applications of the BeadChip technology, probe sequences span 50 base pairs. While $4^{50} \approx 1.27 \times 10^{30}$ unique 50-bp sequences may be constructed, it is not uncommon for beads to anneal (at least *in silico*) to genomic locations other than the intended; this occurrence is known as cross-hybridisation. Non-specific annealing is due not only to design constraints, but also to the intrinsic redundancy of certain regions of the human genome such as homologous gene families, and regions rich in repeated sequences. Because of the chromosomal imbalance between the two sexes, this issue is particularly relevant whenever cross-hybridisation occurs on the X chromosome.

The potential for cross-hybridisation can be easily predicted *in silico*, and has been repeatedly reported for different applications of the BeadChip technology such as the Infinium[®] HumanMethylation27 BeadChip [42] and its ‘successor’, the Infinium[®] HumanMethylation450 BeadChip [43, 44], which is considered in more detail in the next section. Nevertheless, this phenomenon is poorly under-

stood at the molecular level [45].

Address sequence

Because of the random assembly, the type of bead at each location is not predetermined (as in spotted micro-arrays), but determined stochastically, and must be ascertained in a decoding step. Different encoding schemes have been proposed; the sequential hybridisation method of Gunderson et al. [46] that is currently used to decode BeadChips is presented here.

Instead of relying on dye impregnation of beads, or on particular configurations of optical signalling molecules, beads are recognised by a series of hybridisation reactions. In each step, beads are ‘coloured’ by hybridisation to decoder oligonucleotides (complementary to ‘address’ sequences) labelled with a fluorophore, and the fluorescent signal is read by imaging the entire micro-array; the micro-array is then dehybridised, and the process is repeated. The decoder sequences are unchanged, but the fluorescent labels in each step are chosen so that they give rise to a combinatorial space where each bead type is assigned a unique code. These sequences must adhere to certain design constraints: in particular, they should have minimal cross-complementarity and low similarity to genomic sequences of the target organism; in addition, since they are to be hybridised rapidly and at high concentrations, they must have comparable GC contents and melting temperatures.

This process scales exponentially with the number of bead types. If k different fluorophores are used, at each step beads can be in one of $k + 1$ states: k fluorescent ‘on’ states, and one non-fluorescent ‘off’ state. For example, in the protocol described by Gunderson et al. [46], $k = 2$ (carboxyfluorescein and cyanine 3), and at each step beads can fluoresce in green, red, or not fluoresce. In general, for n bead types, $\lceil \log_{k+1} n \rceil$ steps are required for unambiguous decoding, and

additional steps are typically undertaken to allow for error checking. As noted by Gunderson et al. [46], the complexity of this process can be reduced by assigning multiple ‘address’ sequences to each bead type: for example, if pairs of sequences are used, $2m$ ‘primary’ decoder sequences can be used to identify up to m^2 bead types.

The output of the decoding procedure is a ‘map’ that identifies the type of bead in each well on a specific micro-array, and that is provided by Illumina, Inc. together with it.

Another consequence of the random assembly of BeadChips, and a further fundamental difference with respect to spotted micro-arrays, is that each micro-array will be effectively unique, since the number and arrangement of beads of each type will be randomly determined. In particular, the number of beads of a certain type will be Poisson distributed, with mean generally in the low tens. This increases precision, as measurements can be averaged over different replicates. In addition, random placement of the beads minimises the chance of local effects affecting the overall assay.

Laboratory protocol

The typical laboratory protocol for BeadChips starts with extraction of genomic DNA from biological samples. This is followed by whole-genome amplification and fragmentation, and by subsequent resuspension and annealing to the ‘probes’.

After hybridisation, unhybridised and non-specifically-hybridised oligonucleotides are washed away, and the oligonucleotides on the BeadChip undergo enzymatic single-base extension in the presence of hapten-labelled dideoxynucleotides. In this step, polymerases incorporate biotin-labelled ddCTP and ddGTP, and 2,4-dinitrophenol (DNP)-labelled ddATP and ddUTP, which are then stained

using a dual-colour ‘sandwich’ immunoassay [40, 47].

Finally, fluorescence intensities at each bead location are measured by means of a dual-laser scanner that excites the single-base extension products on the beads, and records high-resolution images of the BeadChip.

Once the BeadChip has been scanned, beads are registered using the ‘decoding map’ (dmap) file provided by Illumina, Inc., intensities are extracted for each bead type, and summary statistics are computed and stored in ‘intensity data’ (idat) files.

1.2.2 Infinium[®] HumanMethylation450 BeadChip

In the previous section, the general building principle of the Infinium[®] technology developed by Illumina, Inc. has been described. The focus is now turned to a specific application, i.e. measurement of DNA methylation levels.

The Infinium[®] HumanMethylation450 (HM450) BeadChip allows interrogation of DNA methylation levels based on bisulphite conversion of DNA (described in section 1.1.2) at 485,512 genomic locations at near-single-nucleotide resolution [26]. It covers 482,421 CpG loci (of which 176,112 outside of CpG islands) and 3,091 non-CpG methylated loci identified in human stem cells. It also includes 65 highly-polymorphic SNPs that were included to confirm the identity of samples from the same individual, as an added level of quality control. Each micro-array can be used to analyse 12 samples (arranged in six rows and two columns) simultaneously.

Assay types

Two different assay types coexist on the HM450 BeadChip, termed Infinium[®] I and Infinium[®] II (fig. 1.3).

Infinium® I

Infinium® I is used in genomic regions of high CpG density, and relies on the assumption that DNA methylation levels are regionally correlated within the 50 bp spanned by the ‘probes’, and thus that all underlying CpG loci are in phase with the target locus.

Infinium® I assays require two bead types per CpG locus, a ‘methylated’ bead type whose 3’ terminus is designed to match the (methylated) cytosine, and an ‘unmethylated’ bead type whose 3’ terminus is designed to match the thymine corresponding to an unmethylated cytosine before bisulphite conversion and amplification.

Following single-base extension and staining, beads of the two types fluoresce in the same colour (which depends on the base adjacent to the target CpG locus), and DNA methylation levels are calculated as the ratio of these fluorescent signals.

Infinium® II

Infinium® II assays do not assume that DNA methylation is in phase among CpG loci, and thus offer actual single-nucleotide resolution; however, they can only be used when up to three CpG loci (excluding the target locus) are found within the 50-bp ‘probe’ sequence.

They use a single bead per CpG locus, whose 3’ terminus complements the base directly upstream of the target locus.

Following single-base extension, the addition of a labelled ddGTP or ddATP signals a ‘methylated’ cytosine or ‘unmethylated’ thymine at the target CpG locus, respectively. Thus, when stained, beads fluoresce simultaneously in two colours, and the ratio of these fluorescent signals can be used to estimate DNA methylation levels.

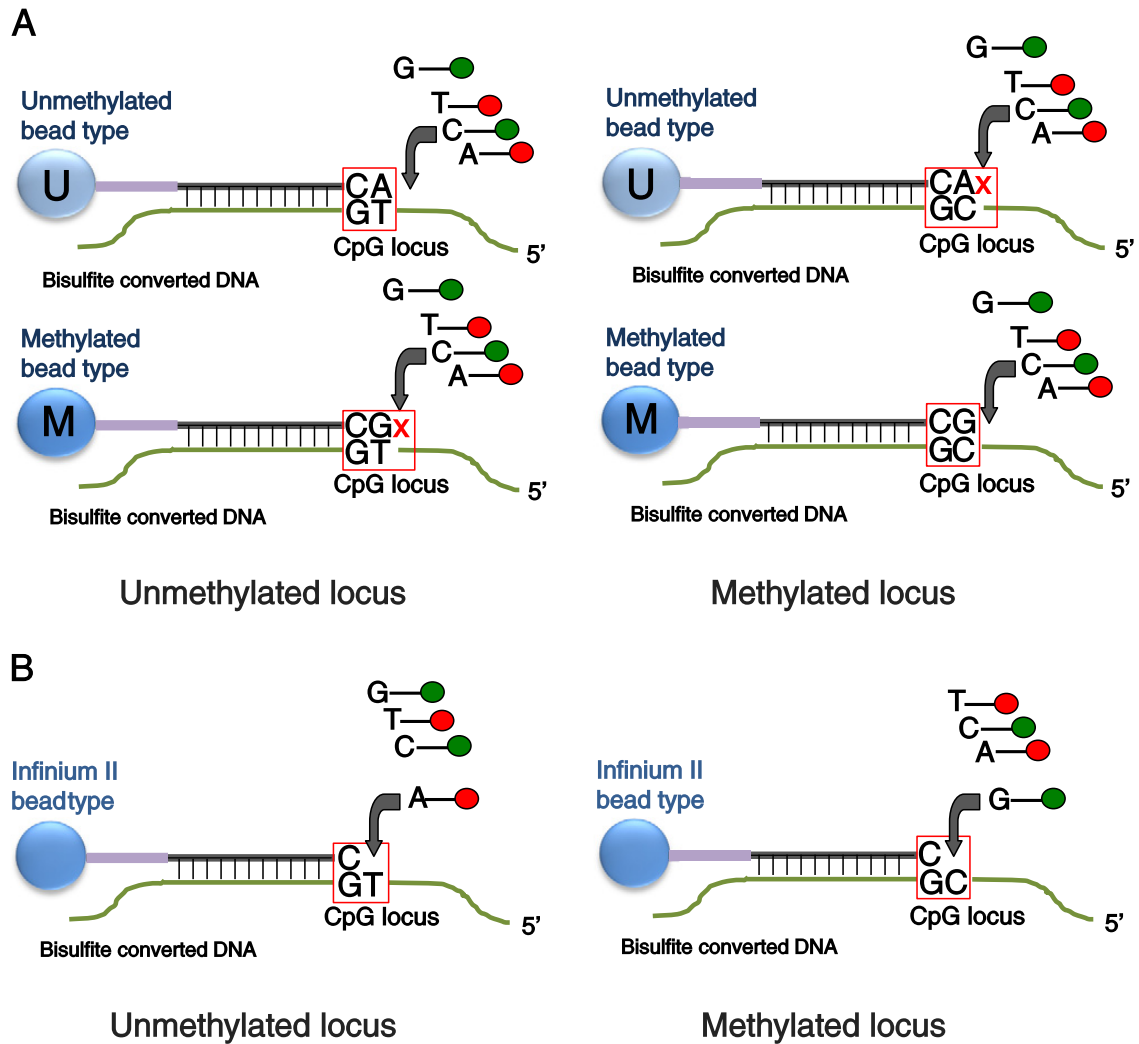


Figure 1.3: Infinium® I and Infinium® II assays. Infinium® I assays (**A**) require two bead types per CpG locus, whereas Infinium® II assays (**B**) use a single bead type that fluoresces in two colours (from Bibikova et al. [26]).

Differences between Infinium[®] I and Infinium[®] II

As described above, Infinium[®] I assays do not measure DNA methylation levels at single-nucleotide resolution, and the choice of assay is often (but not always) driven by the number of CpGs in the 50-bp ‘probe’ sequence, as shown in table 1.1.

Dedeurwaerder et al. [48] were the first to note that DNA methylation levels estimated using different assay designs had different distributions, and proposed a method to correct this ‘bias’. Maksimovic et al. [49], and more recently Teschendorff et al. [50], followed suit, and proposed different methods to address the same ‘problem’.

Nevertheless, it is not to be expected that assays using different designs should provide comparable results.

First, as shown in table 1.1, the choice of design depends on the CpG density around the target locus, which in turn is associated e.g. with genomic context: for instance, higher CpG density is more likely to occur within CpG islands, which are often found near promoters.

Second, since Infinium[®] I assays cannot detect DNA methylation at the target CpG locus irrespective of the status of surrounding loci, they will in general estimate values closer to the extremes. To understand why, consider the case of a ‘probe’ sequence encompassing one CpG locus in addition to the target locus; the sample space of assayed DNA strands can be partitioned as follows, where all counts are assumed positive:

		Target		
		C	T	Total
Non-target	C	n_{CC}	n_{CT}	$n_{C.}$
	T	n_{TC}	n_{TT}	$n_{T.}$
Total		$n_{.C}$	$n_{.T}$	n

CpGs	Assay design		
	Infinium [®] I	Infinium [®] II	Total
1	0	155,016	155,016
2	11,303	112,634	123,937
3	26,201	59,297	85,498
4	36,401	19,998	56,399
5	33,398	0	33,398
6	17,519	0	17,519
7	7,636	0	7,636
8	2,482	0	2,482
9	476	0	476
10	55	0	55
11	5	0	5
Total	135,476	346,945	482,421

Table 1.1: Classification of HM450 BeadChip assays by number of CpG loci in the ‘probe’ sequence and assay design. Infinium[®] I assays are used in genomic regions of high CpG density, and in particular whenever the ‘probe’ sequences encompass four or more CpG loci in addition to the target locus.

The underlying assumption of Infinium[®] assays is that fluorescence intensities are directly proportional to these counts, and can thus be used in their place. An Infinium[®] II assay can assess the methylation status of the target locus irrespective of that of the non-target locus, and so will estimate the DNA methylation level at the target as follows,

$$m_{II} = \frac{n_{.C}}{n} = \frac{n_{.C}}{n_{.C} + n_{.T}} = \frac{n_{CC} + n_{TC}}{n_{CC} + n_{CT} + n_{TC} + n_{TT}}. \quad (1.1)$$

On the other hand, the Infinium[®] I ‘methylated’ bead will only anneal to DNA fragments where both CpG loci are methylated (and similarly for the ‘unmethylated’ bead); thus, an Infinium[®] I assay will estimate the DNA methylation level at the target as follows,

$$m_I = \frac{n_{CC}}{n_{CC} + n_{TT}}, \quad (1.2)$$

which can be understood as conditioning on the non-target locus being in phase. The following relationships hold between eqs. (1.1) and (1.2),

$$m_I \leq m_{II} \iff \frac{n_{TC}}{n_{.C}} \geq \frac{n_{CT}}{n_{.T}}, \quad (1.3)$$

$$m_I \geq m_{II} \iff \frac{n_{TC}}{n_{.C}} \leq \frac{n_{CT}}{n_{.T}}, \quad (1.4)$$

from which it can be appreciated that the Infinium[®] I assay estimates the actual DNA methylation level at the target CpG locus if and only if the proportion of out-of-phase non-target loci is independent of the methylation status of the target locus,

$$m_I = m_{II} \iff \frac{n_{TC}}{n_{.C}} = \frac{n_{CT}}{n_{.T}}. \quad (1.5)$$

Control beads

In addition to beads used to measure DNA methylation or for genotyping, the HM450 BeadChip also includes beads of 835 distinct types that are used for a

variety of quality control purposes. They target sequences that do not contain CpG dinucleotides, and thus do not depend on the methylation status.

Control beads are classified by Illumina, Inc. as 'sample-independent' and 'sample-dependent'.

'Sample-independent' control beads

'Sample-independent' control beads are used to evaluate the performance of specific steps in the laboratory protocol, and are further divided into four categories.

Hybridisation

Hybridisation is monitored by three different beads whose 'probe' sequences perfectly complement three synthetic targets present in the hybridisation buffer at different concentrations (0.2, 1, and 5 pM). Only the green fluorescence is monitored, with an approximately linear dose-response curve (on the log-log scale), as shown in fig. 1.4.

Extension

Single-base extension is monitored by four different beads (one per dideoxynucleotide). The green fluorescence is monitored for ddCTP and ddGTP, and the red fluorescence for ddATP and ddUTP.

Target removal

Two beads test the efficiency of the post-extension stripping step using labelled targets that complement non-extendible probes. These targets are present in the hybridisation buffer.

Staining

Two beads, corresponding to background and high levels of hapten-labelled dideoxynucleotides, are used to examine the efficiency of the staining step for biotin (green fluorescence) and DNP (red fluorescence) separately.

‘Sample-dependent’ control beads

‘Sample-dependent’ control beads are used to evaluate performance across samples, and are divided into five categories.

Bisulphite conversion

These beads assess the efficiency of bisulphite conversion of genomic DNA. Six pairs of beads use the Infinium[®] I design (three pairs per colour), and four beads use the Infinium[®] II design.

Negative

These 600 beads target randomly permuted DNA sequences that should not hybridise, and can thus be used to define the background noise; however, this assumption is unverifiable as Illumina, Inc. does not provide ‘probe’ sequences for any control beads.

Non-polymorphic

Four beads test the overall performance, from amplification to detection, by querying a particular base in a non-polymorphic region of the bisulphite-converted genome.

Specificity

These beads are designed against non-polymorphic T sites, and monitor potential non-specific detection of a signal corresponding to DNA methylation over 'non-methylated' background noise. Six pairs of beads use the Infinium[®] I design, and three beads use the Infinium[®] II design.

Normalisation

These beads are designed for normalisation purposes, but Illumina, Inc. does not provide information on their design. There are a total of 186 beads, divided into 32 pairs of A/G beads, and 61 pairs of C/T beads.

Three A/G pairs and six C/T pairs, reported in table 1.2, exhibit significantly higher fluorescence intensities in women than men, as shown in fig. 1.5; this leads one to believe that they hybridise to the X chromosome.

Pre-processing steps

The pre-processing steps leading to the final set of DNA methylation measurements that can be subjected to statistical analysis are illustrated here on publicly available data generated by Liu et al. [51].

At the end of the Infinium[®] laboratory protocol described in section 1.2.1, each sample yields two idat files that store (in a proprietary binary format) summary statistics of fluorescence intensities in the two colours. These files list the actual number of beads of each type on the micro-array, and the sample means and standard deviations of green and red intensities across replicates. Bead types with less than three replicates are typically disregarded.

In a subsequent step, mean intensities from non-control beads are rearranged into two matrices (named **A** and **B**) with intensity data for each of the two alleles

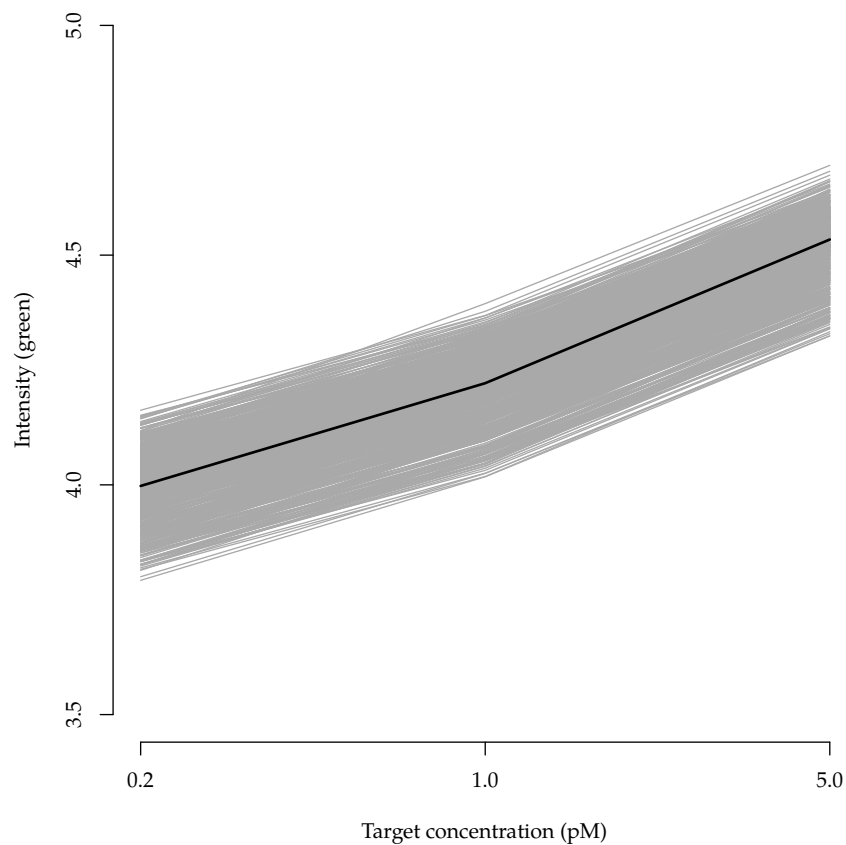
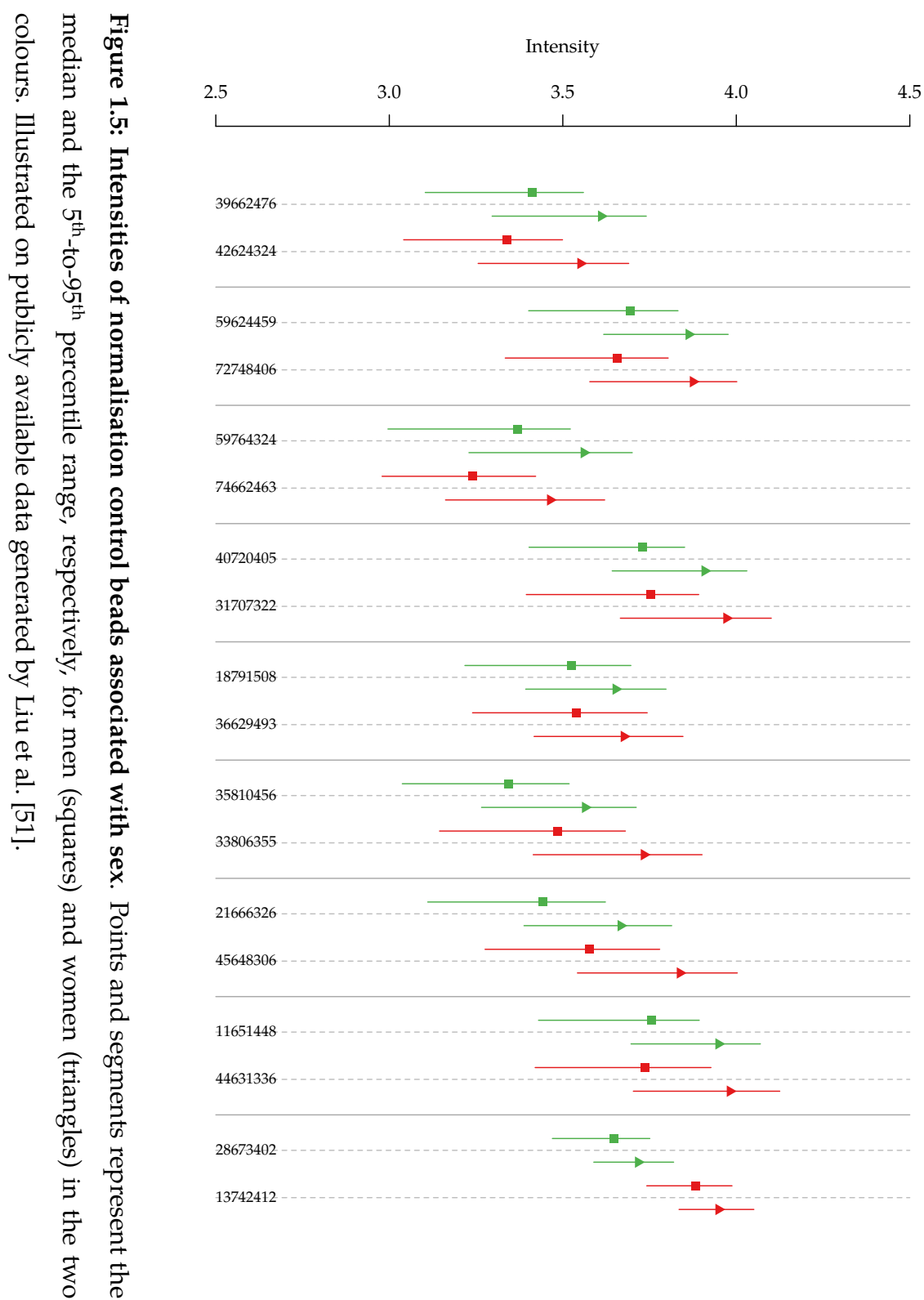


Figure 1.4: Intensities of hybridisation control beads. Each grey line represents a different sample, and the black line represents the median across samples. Target concentrations and intensities are plotted on the log-log scale. Illustrated on publicly available data generated by Liu et al. [51].

Type	Pair	Green bead ID	Red bead ID
A/G	11	39662476	42624324
A/G	23	59624459	72748406
A/G	28	59764324	74662463
C/T	4	40720405	31707322
C/T	21	18791508	36629493
C/T	23	35810456	33806355
C/T	30	21666326	45648306
C/T	34	11651448	44631336
C/T	61	28673402	13742412

Table 1.2: Normalisation control beads associated with sex. Fluorescence intensities measured for these pairs of beads are significantly higher in women than men (fig. 1.5).



at a specific genomic location. For C/T polymorphisms created by bisulphite conversion, **A** corresponds to the thymine (and thus to absence of DNA methylation), and **B** corresponds to the cytosine (and thus to presence of DNA methylation). The treatment depends on the assay design.

Infinium® I

As described above, DNA methylation assays using the Infinium® I design require two beads whose fluorescences are measured in the same colour. The intensity of the ‘unmethylated’ bead type is reproduced in **A**, and that of the ‘methylated’ bead type in **B**; intensities of the two beads in the other colour are disregarded.

Infinium® II

DNA methylation assays using the Infinium® II design rely on a single bead whose fluorescence is measured in the two colours. In particular, the thymine corresponding to absence of DNA methylation anneals to DNP-labelled ddATP, which fluoresces in red; on the other hand, the cytosine corresponding to presence of DNA methylation anneals to biotin-labelled ddGTP, which fluoresces in green. Therefore, the fluorescence intensity in red is reproduced in **A**, and that in green in **B**.

Censoring of values below detection limits

Elements of **A** and **B** are censored if the total intensity $\mathbf{I} = \mathbf{A} + \mathbf{B}$ is below detection thresholds estimated from the 600 ‘negative’ control beads.

First, mean intensities from these beads are used to compute sample means ($\bar{x}_G$ and $\bar{x}_R$) and standard deviations (s_G and s_R) of the background noise in each colour separately.

For Infinium[®] I assays, both **A** and **B** are measured in the same colour, and thus the mean of the background noise of **I** is given by $2\bar{x}_G$ (or $2\bar{x}_R$), and its standard deviation by $\sqrt{2}s_G$ (or $\sqrt{2}s_R$).

For Infinium[®] II assays, **A** is measured in red, whereas **B** is measured in green; therefore, the mean of the background noise of **I** is given by $\bar{x}_G + \bar{x}_R$, and its standard deviation by $\sqrt{s_G^2 + s_R^2}$.

Illumina, Inc. assumes that the background noise in each of these three cases is normally distributed with the given means and standard deviations, and derives the corresponding one-tailed ‘detection p -values’. For a given detection p -value threshold α , the corresponding detection thresholds for **I** are as follows,

$$2\bar{x}_G + z_\alpha \sqrt{2}s_G \quad \text{for Infinium[®] I measured in green,} \quad (1.6)$$

$$2\bar{x}_R + z_\alpha \sqrt{2}s_R \quad \text{for Infinium[®] I measured in red,} \quad (1.7)$$

$$\bar{x}_G + \bar{x}_R + z_\alpha \sqrt{s_G^2 + s_R^2} \quad \text{for Infinium[®] II,} \quad (1.8)$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$.

Background subtraction

In addition to censoring, Illumina, Inc. also recommends performing ‘background subtraction’. This is done by subtracting $\bar{x}_G$ or $\bar{x}_R$ from elements of **A** and **B** (depending on the colour each is measured in), and censoring negative values.

Alternative estimation of background noise

As described above, censoring of values below detection limits, as well as background subtraction, depend on measurements obtained from 600 ‘negative’ control beads.

Triche Jr et al. [52] noted that ‘out-of-band’ intensities, i.e. green intensities from Infinium[®] I beads normally measured in red, and, similarly, red intensit-

ies from Infinium[®] I beads normally measured in green, could alternatively be used to define the background noise. Since Infinium[®] I beads are far more numerous than ‘negative’ control beads, this method would allow a more precise characterisation of the noise distribution.

As shown in fig. 1.6, the distribution of ‘out-of-band’ intensities is generally more skewed and platykurtic than that of intensities from ‘negative’ control beads. This is due to the presence of many Infinium[®] I beads carrying ‘probe’ sequences that cross-hybridise to genomic locations other than the intended. As described above, these beads fluoresce in the two colours depending on the base adjacent to the target CpG locus; in some cases, the ‘out-of-band’ intensity may even be greater than the ‘in-band’ intensity (fig. 1.7).

Dye bias correction

Differences in fluorescence between the two colours, which may be termed ‘dye bias’, may arise e.g. from different rates of incorporation of labelled nucleotides during the extension step, or from variations in the staining procedure.

DNA methylation measurements from Infinium[®] I assays are largely unaffected by these differences, since the two bead types fluoresce in the same colour.

On the other hand, Infinium[®] II assays use a single bead type that fluoresces in two colours to signal methylated and unmethylated cytosines, and thus DNA methylation measurements from these assays will be affected: for example, if the colour corresponding to methylated cytosines fluoresces less intensely than the one corresponding to unmethylated cytosines, DNA methylation levels will be biased towards zero.

Many pre-processing methods have been proposed to ‘equalise’ signals arising from Infinium[®] II assays, and to assimilate them to those from Infinium[®] I assays. Two novel pre-processing methods, as well as a formal comparison of new and

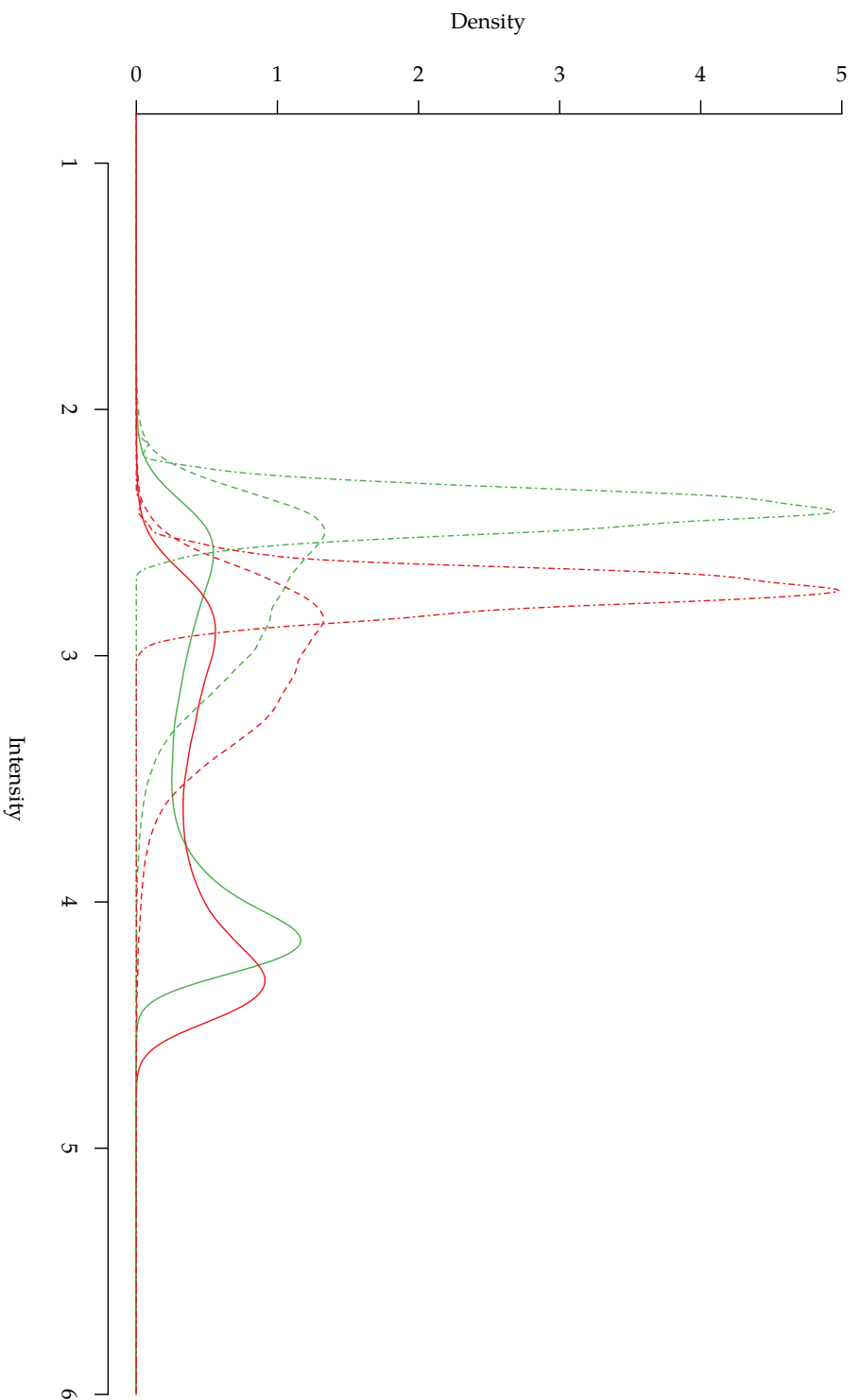


Figure 1.6: Intensities of ‘negative’ and ‘out-of-band’ beads. Kernel density estimates ($\log_{10}$ scale) of Infinium[®] I ‘in-band’ intensities in the two colours (solid lines), ‘out-of-band’ intensities (dashed lines), and ‘negative’ control beads (dash-dotted lines). Illustrated on publicly available data generated by Liu et al. [51].

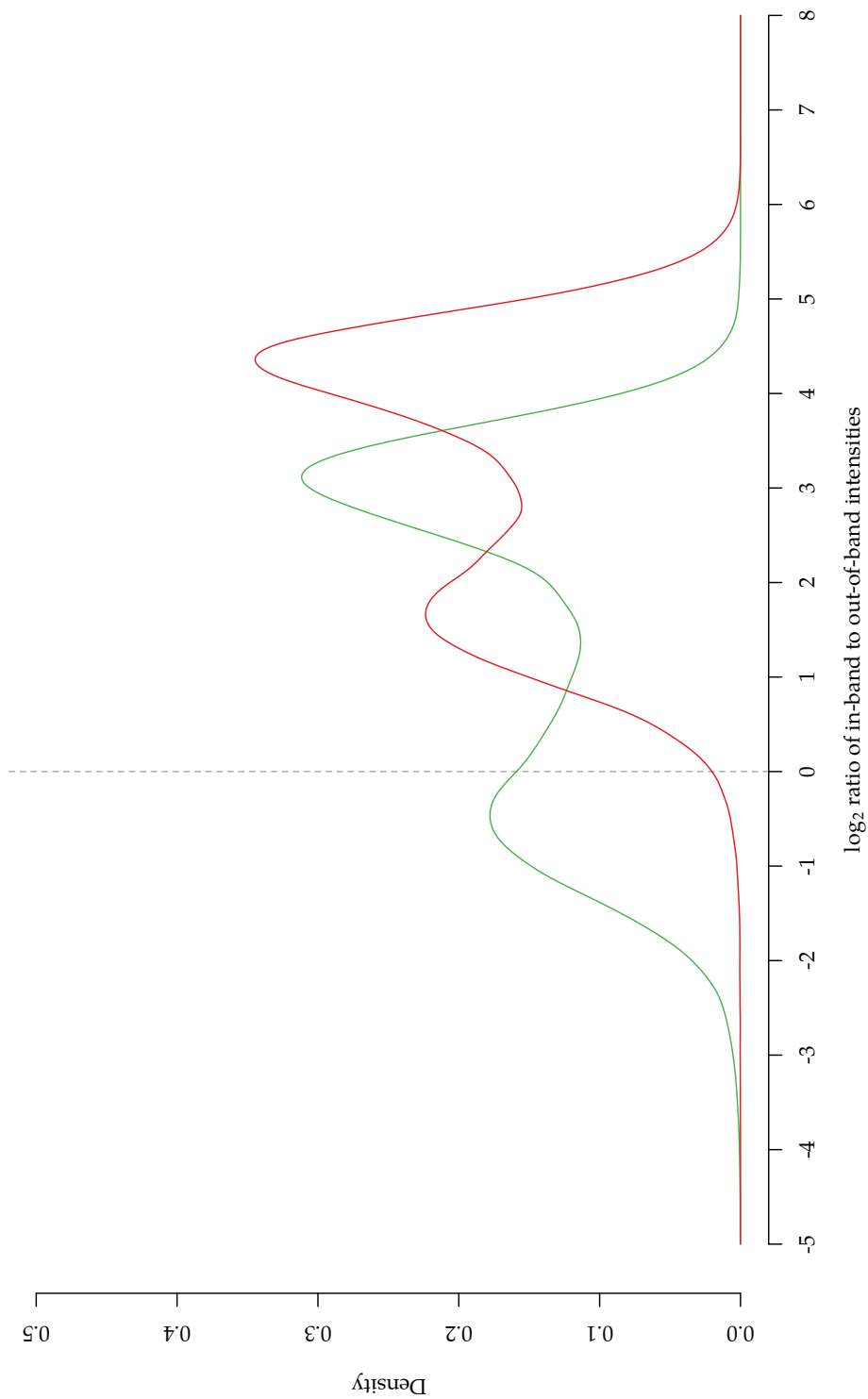


Figure 1.7: Ratio of ‘in-band’ to ‘out-of-band’ intensities. Kernel density estimates ($\log_2$ scale) of ratios of ‘in-band’ to ‘out-of-band’ intensities from Infinium® I beads in the two colours. The dashed grey line corresponds to equal intensities for ‘in-band’ and ‘out-of-band’ signals. Illustrated on publicly available data generated by Liu et al. [51].

existing methods, are presented in section 2.1.

Estimation of DNA methylation levels

Finally, DNA methylation levels are estimated as follows,

$$\mathbf{M} = \frac{\mathbf{B}}{\mathbf{A} + \mathbf{B} + \epsilon}, \quad (1.9)$$

where matrix operations are element-wise, and ϵ is a ‘fudge factor’ that ensures a positive denominator. Illumina, Inc. recommends $\epsilon = 100$, which is approximately two orders of magnitude smaller than observed intensities.

Clearly, elements of $\mathbf{M}$ are restricted to lie on the unit interval, which may violate assumptions of statistical models, as described in section 1.3. For this reason, Du et al. [53] proposed a simple element-wise transformation of $\mathbf{M}$,

$$\tilde{\mathbf{M}} = \text{logit} \frac{\mathbf{B}}{\mathbf{A} + \mathbf{B} + \epsilon} = \log \frac{\mathbf{B}}{\mathbf{A} + \epsilon}, \quad (1.10)$$

since $\text{logit}(x) = \log[x/(1 - x)]$.

In genotyping applications of the BeadChip technology, Illumina, Inc. considers $\mathbf{A}$ and $\mathbf{B}$ as Cartesian coordinates in the first quadrant (since $\mathbf{A} > 0$ and $\mathbf{B} > 0$), and computes the following transformation to polar coordinates,

$$\mathbf{R} = \sqrt{\mathbf{A}^2 + \mathbf{B}^2}, \quad (1.11)$$

$$\mathbf{\Theta} = \arctan \frac{\mathbf{B}}{\mathbf{A}}, \quad (1.12)$$

where matrix operations are element-wise, and it is customary to normalise $\mathbf{\Theta}$ by dividing its elements by the maximum $\pi/2$. As shown in fig. 1.8, genotypes are more easily separated by inspection of the scatter plot of $(\mathbf{\Theta}, \mathbf{R})$ rather than that of $(\mathbf{A}, \mathbf{B})$.

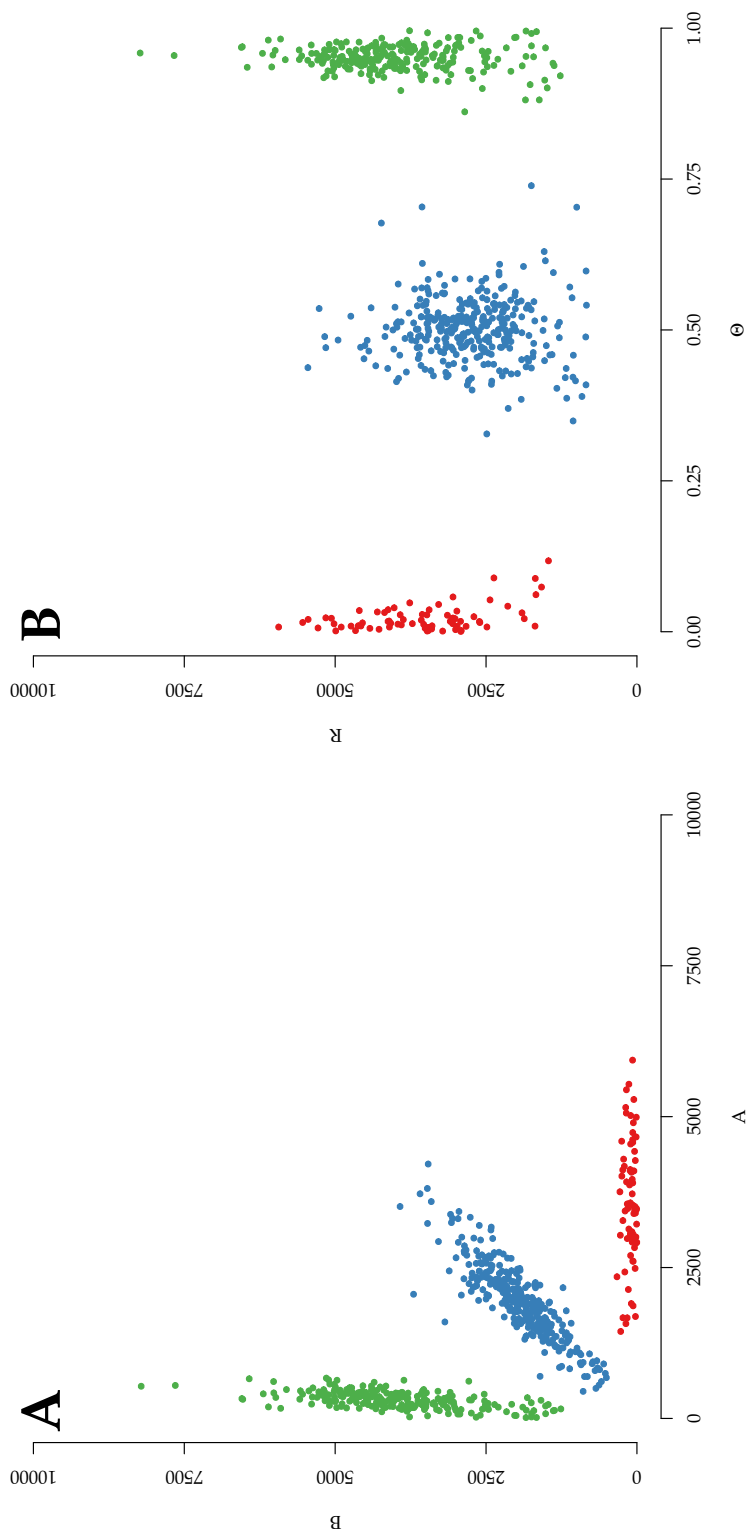


Figure 1.8: Scatter plot of intensities for rs2208123. (A) intensities on the (A, B) plane; (B) transformed intensities on the (Θ , R) plane. Separation of genotypes is often easier after transformation, especially at low fluorescence intensities. Illustrated on publicly available data generated by Liu et al. [51].

§ 1.3 Statistical modelling

For ‘-omics’ data, the number of measurements virtually always exceeds the number of observations, usually by some orders of magnitude. This situation is known as the ‘small n , large p ’ problem, and is intractable using classic statistical theory [54].

Although measurements may define a high-dimensional space, they are often also strongly correlated. These correlations arise from the complex interaction patterns that characterise biological regulatory processes, as described in section 1.1.

In typical cases, this means that a smaller set of variables may be sufficient to accurately represent the biological process of interest. Under this key assumption, and presupposing the existence of a principled approach to achieve this reduction, the ‘small n , large p ’ problem can be turned into a ‘moderate n , small p ’ scenario, and dealt with using classic statistical theory.

The reduction procedure should not only accommodate the presence of many ‘irrelevant’ variables, but also reflect the biological mechanisms underlying the observed interdependence among ‘relevant’ variables. From a computational point of view, it should be scalable and able to cope with an ever-increasing number of measurements.

Two main classes of reduction procedures will be considered in the rest of this section. In the univariate framework (section 1.3.1), measurements are initially considered independently from each other. Associations between each measurement and the phenotype are formally tested using some statistical model, and results (usually in the form of test statistics or p -values) are then aggregated. On the other hand, in the multivariate framework (section 1.3.2), all measurements are considered simultaneously, and the procedure seeks to identify ‘relevant’ vari-

ables that jointly explain the variation in the data, with (supervised methods) or without (unsupervised methods) respect to the phenotype.

1.3.1 Univariate approaches

As outlined above, in the univariate framework measurements are first considered independently, and the same statistical model is repeatedly applied to obtain values for some test statistic of interest, or the corresponding p -values.

Linear models and extensions

Consider the classic linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1.13)$$

where $\mathbf{y}$ is a vector of (continuous) responses, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is a vector of model parameters to be estimated, and $\boldsymbol{\epsilon}$ is a random vector of i.i.d. error terms. Multiple regression and analysis of variance (ANOVA) can also be formulated in this manner.

Note that eq. (1.13) can be equivalently expressed as follows,

$$\mathbb{E}(Y_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad (1.14)$$

where the random variables Y_i are independent, $\mathbf{x}_i^\top$ represents the i^{th} row of $\mathbf{X}$, and $\boldsymbol{\beta}$ is as above.

Nelder et al. [55] introduced the following generalisation of eq. (1.14),

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad (1.15)$$

where the Y_i are independent and distributed in the exponential family, and μ_i and the linear predictor η_i are related by some link function g that is monotone and differentiable. These models are known as generalised linear models (GLMs). The classic linear model of eq. (1.14) is obtained when the Y_i are

normally distributed, and the link function is the identity function. In general, eq. (1.15) will *not* be equivalent to eq. (1.13) after application of g to the values in $\mathbf{y}$.

Hastie et al. [56, 57] introduced a further generalisation of eq. (1.15),

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + \dots, \quad (1.16)$$

where the f_j are arbitrary smooth functions. These models are known as generalised additive models (GAMs), and have been extended to the case of multivariate responses by Yee et al. [58] (vector generalised additive models, VGAMs).

Linear mixed models and extensions

In classic linear models, observations are assumed to be independent and to have equal variances, which leads to a diagonal residual variance matrix $\sigma^2 \mathbf{I}_n$. Linear mixed models are an extension of classic (fixed effects) linear models that include random effects (intercepts and coefficients), and possibly covariance terms. Equation (1.13) is modified as follows,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad (1.17)$$

where $\mathbf{Z}$ is an additional design matrix for random effects, $\mathbf{G}$ is a diagonal matrix (since random effects are assumed uncorrelated) of variance components, and $\mathbf{R}$ is the residual covariance matrix, which equals $\sigma^2 \mathbf{I}_n$ in the absence of additional covariance terms. The variance matrix of the resulting model is given by $\mathbf{Z}\mathbf{G}\mathbf{Z}^\top + \mathbf{R}$, and has block diagonal form.

Generalised linear mixed models (GLMMs) are a natural extension of eq. (1.17) to accommodate a link function as in eq. (1.15),

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad (1.18)$$

where $\mathbf{z}_i^\top$ represents the i^{th} row of $\mathbf{Z}$. The variance matrix of the resulting model can be written as $\mathbb{V}(\boldsymbol{\mu}) + \mathbf{R}$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, but its specification is not as straightforward as in the case of normal data, since $\boldsymbol{\mu}$ is generally not a linear function of $\mathbf{u}$. A first-order approximation is given by $\mathbf{B} \mathbf{Z} \mathbf{G} \mathbf{Z}^\top \mathbf{B} + \mathbf{R}$, where $\mathbf{B}$ depends on the distributional assumptions of the model.

Maximum-likelihood estimation and hypothesis testing

Model parameters in eqs. (1.13) to (1.18) are commonly estimated by the maximum-likelihood principle, which produces estimators with desirable asymptotic properties that hold under fairly general conditions, including consistency, invariance, and asymptotic normality and efficiency [59, 60].

Given the likelihood function $L(\boldsymbol{\theta} | \mathbf{y})$, where $\boldsymbol{\theta}$ is a vector of model parameters (such as $\boldsymbol{\beta}$ above), consider the log-likelihood function

$$\ell(\boldsymbol{\theta} | \mathbf{y}) = \ln L(\boldsymbol{\theta} | \mathbf{y}). \quad (1.19)$$

Clearly, the value $\hat{\boldsymbol{\theta}}$ that maximises L also maximises its monotonic transformation ℓ .

Define the gradient of ℓ with respect to $\boldsymbol{\theta}$ (also known as score vector),

$$\mathbf{s}(\boldsymbol{\theta} | \mathbf{y}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{\partial \ln L}{\partial \boldsymbol{\theta}} = \frac{1}{L} \frac{\partial L}{\partial \boldsymbol{\theta}}. \quad (1.20)$$

The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained by solving $\mathbf{s}(\boldsymbol{\theta} | \mathbf{y}) = \mathbf{0}$. Moreover, $\mathbb{E}(\mathbf{s}) = \mathbf{0}$, and

$$\mathbb{V}(\mathbf{s}) = \mathbb{E}(\mathbf{s}\mathbf{s}^\top) = \mathbb{E} \left[\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}} \right)^\top \right] = \mathfrak{I}(\boldsymbol{\theta}). \quad (1.21)$$

Typically, one is interested in formally testing certain linear hypotheses about the vector of model parameters $\boldsymbol{\theta}$, which in general take the form

$$\mathcal{H}_0 : \mathbf{R} \boldsymbol{\theta} = \mathbf{r}, \quad (1.22)$$

where $\mathbf{R}$ is a $q \times p$, $q < p$, matrix of constants, and $\mathbf{r}$ is a column vector with q known elements. For example, suppose that one wishes to test associations between some exposure and ‘-omics’ measurements. In statistical terms, this research question is tantamount to asking whether constraining some parameters of interest to zero (hence setting $\mathbf{r} = \mathbf{0}$), or equivalently removing the corresponding predictor variables from the model, leads to a significant worsening of the model fit.

Three classic statistical tests for eq. (1.22) are now presented. For the derivation of the Wald and Lagrange multiplier (or score) tests, the following result will also prove useful. Suppose $\mathbf{t} = (t_1, \dots, t_q)$ is a vector of statistics of interest with asymptotic expectation $\mathbb{E}(\mathbf{t})$ and asymptotic variance-covariance matrix $\mathbb{V}(\mathbf{t})$. Then, it holds that

$$[\mathbf{t} - \mathbb{E}(\mathbf{t})]^\top \mathbb{V}(\mathbf{t})^{-1} [\mathbf{t} - \mathbb{E}(\mathbf{t})] \stackrel{a}{\sim} \chi_q^2, \quad (1.23)$$

provided that $\mathbb{V}(\mathbf{t})$ is non-singular.

Likelihood-ratio test

The likelihood ratio is defined as

$$\lambda = \frac{L(\tilde{\boldsymbol{\theta}}|\mathbf{y})}{L(\hat{\boldsymbol{\theta}}|\mathbf{y})}, \quad (1.24)$$

where $L(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is the maximum of the likelihood function attained at the unrestricted estimate $\hat{\boldsymbol{\theta}}$, and $L(\tilde{\boldsymbol{\theta}}|\mathbf{y})$ is obtained by maximising the likelihood function subject to $\mathbf{R}\boldsymbol{\theta} = \mathbf{r}$, yielding the restricted estimate $\tilde{\boldsymbol{\theta}}$. Thus, the likelihood-ratio test requires fitting of two models. A large-sample test for eq. (1.22) can be constructed based on the following result,

$$-2 \ln \lambda = 2[\ell(\hat{\boldsymbol{\theta}}|\mathbf{y}) - \ell(\tilde{\boldsymbol{\theta}}|\mathbf{y})] \stackrel{a}{\sim} \chi_q^2. \quad (1.25)$$

Wald test

Unlike the likelihood-ratio test, the Wald test requires computation of $\hat{\boldsymbol{\theta}}$ only. Since the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal,

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta})), \quad (1.26)$$

it follows that

$$(\mathbf{R} \hat{\boldsymbol{\theta}} - \mathbf{r}) \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{R} \mathcal{I}^{-1}(\boldsymbol{\theta}) \mathbf{R}^\top) \quad (1.27)$$

under $\mathcal{H}_0$, and thus, by eq. (1.23), that

$$[\mathbf{R} \hat{\boldsymbol{\theta}} - \mathbf{r}]^\top (\mathbf{R} \mathcal{I}^{-1}(\boldsymbol{\theta}) \mathbf{R}^\top)^{-1} [\mathbf{R} \hat{\boldsymbol{\theta}} - \mathbf{r}] \overset{a}{\sim} \chi_q^2, \quad (1.28)$$

which can be used to construct a test for eq. (1.22).

Lagrange multiplier (or score) test

While the unrestricted estimate $\hat{\boldsymbol{\theta}}$ satisfies $\mathbf{s}(\hat{\boldsymbol{\theta}}|\mathbf{y}) = 0$, this is in general not the case for the restricted estimate $\tilde{\boldsymbol{\theta}}$; however, under $\mathcal{H}_0$ the difference between the restricted and unrestricted maxima of the likelihood function is close to zero, and $\mathbf{s}(\tilde{\boldsymbol{\theta}}|\mathbf{y}) \approx 0$. Because of the properties of $\mathbf{s}$ summarised above, and as a consequence of eq. (1.23), it holds that

$$\mathbf{s}(\tilde{\boldsymbol{\theta}})^\top \mathcal{I}^{-1}(\tilde{\boldsymbol{\theta}}) \mathbf{s}(\tilde{\boldsymbol{\theta}}) \overset{a}{\sim} \chi_q^2, \quad (1.29)$$

which can be used to construct a test for eq. (1.22).

The three tests presented are asymptotically equivalent, but generally give different numerical results in finite samples. For linear models, it holds that

$$(\mathbf{R} \hat{\boldsymbol{\theta}} - \mathbf{r})^\top (\mathbf{R} \mathcal{I}^{-1}(\boldsymbol{\theta}) \mathbf{R}^\top)^{-1} (\mathbf{R} \hat{\boldsymbol{\theta}} - \mathbf{r}) \geq -2 \ln \lambda \geq \mathbf{s}(\tilde{\boldsymbol{\theta}})^\top \mathcal{I}^{-1}(\tilde{\boldsymbol{\theta}}) \mathbf{s}(\tilde{\boldsymbol{\theta}}), \quad (1.30)$$

i.e. the Wald test statistic will always be greater than or equal to the likelihood-ratio test statistic, which in turn will always be greater than or equal to the Lagrange multiplier (or score) test statistic [59].

Multiple comparisons

Hypothesis testing is subject to error, which occurs when the outcome of the test (i.e. reject, or fail to reject, $\mathcal{H}_0$) does not correspond to the true ‘state of nature’. A type I error occurs when one rejects a true $\mathcal{H}_0$, leading to a so-called ‘false positive’ finding; on the other hand, a type II error occurs when one fails to reject a false $\mathcal{H}_0$, leading to a so-called ‘false negative’ finding. The probability of committing a type I error is the significance level (or size) of the test α , whereas the probability of committing a type II error is the false negative rate β , and is related to the power of the test $1 - \beta$.

When N tests are conducted simultaneously (as is virtually always the case for ‘-omics’ data), outcomes fall into one of four possible situations, and can be tabulated as follows.

	$\mathcal{H}_0$ is true	$\mathcal{H}_a$ is true	
Reject $\mathcal{H}_0$	F_P	T_P	N_P
Do not reject $\mathcal{H}_0$	T_N	F_N	N_N
	N_0	N_a	

The number of ‘false positive’ findings is represented by F_P , and that of ‘true positive’ findings by T_P ; the total number of findings (i.e. the number of rejections of $\mathcal{H}_0$) is N_P . Correspondingly, the number of ‘true negative’ findings is represented by T_N , and that of ‘false positive’ findings by F_N ; their sum is N_N . Irrespective of whether the $\mathcal{H}_0$ are rejected or not, the number of true ‘states of nature’ where $\mathcal{H}_0$ is true is given by N_0 . Similarly, the number of cases where $\mathcal{H}_a$ is true is N_a . Several ratios of interest can be defined from these quantities.

The specificity (or true negative rate, TNR) of the test is given by

$$\text{Specificity} = \frac{T_N}{F_P + T_N} = \frac{T_N}{N_0} \in [0, 1], \quad (1.31)$$

and can be increased by reducing the number of false positive findings F_P . The complement of the specificity is known as false positive rate (FPR).

The sensitivity (or true positive rate, TPR) of the test is defined as

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N} = \frac{T_P}{N_a} \in [0, 1], \quad (1.32)$$

and can be increased by reducing the number of false negative findings F_N . The complement of the sensitivity is known as false negative rate (FNR).

While it would be desirable to maximise both specificity and sensitivity, and thus reduce the number of false (positive or negative) findings, there exists a trade-off between these two ratios. In fact, if one were to always reject $\mathcal{H}_0$, one would achieve maximum specificity, but no sensitivity; on the other hand, if one were to always fail to reject $\mathcal{H}_0$, one would achieve maximum sensitivity, but no specificity.

Other related quantities include the positive predictive value (PPV), also known as precision,

$$\text{PPV} = \frac{T_P}{F_P + T_P} = \frac{T_P}{N_P} \in [0, 1], \quad (1.33)$$

the negative predictive value,

$$\text{NPV} = \frac{T_N}{F_N + T_N} = \frac{T_N}{N_N} \in [0, 1], \quad (1.34)$$

and the accuracy,

$$\text{Accuracy} = \frac{T_P + T_N}{N} \in [0, 1]. \quad (1.35)$$

Two additional probabilistic quantities of interest may be defined.

The first is the family-wise error rate (FWER), which is defined (in the ‘strong’ sense) as the probability of committing at least one type I error, irrespective of the configuration of true and false $\mathcal{H}_0$,

$$\text{FWER} = \Pr(F_P \geq 1) = 1 - \Pr(F_P = 0). \quad (1.36)$$

Alternatively, the FWER may be defined (in the ‘weak’ sense) under the complete null hypothesis that all $\mathcal{H}_0$ are true,

$$\text{FWER}_C = \Pr(F_P \geq 1 | \text{all } \mathcal{H}_0 \text{ are true}). \quad (1.37)$$

Obviously, controlling the FWER in the strong sense also ensures control of the FWER in the weak sense.

The second quantity of interest is the false discovery rate (FDR), which is defined by Benjamini et al. [61] as the expected value of the proportion of false positive findings Q ,

$$\text{FDR} = \mathbb{E}(Q), \quad \text{where} \quad Q = \begin{cases} F_P/N_P & \text{if } N_P > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (1.38)$$

Control of the FDR implies control of the FWER in the weak sense, since $T_P = 0$ under the complete $\mathcal{H}_0$, and thus $Q = 0$ if $F_P = 0$, and $Q = 1$ otherwise, from which it follows that

$$\text{FDR} = \mathbb{E}(Q) = \Pr(F_P \geq 1 | \text{all } \mathcal{H}_0 \text{ are true}) = \text{FWER}_C. \quad (1.39)$$

Note that the definition of Benjamini et al. [61] is not the only possibility. In fact, a natural definition for the FDR would be $\mathbb{E}(F_P/N_P)$; however, this definition is not viable if $\Pr(N_P = 0) > 0$, as is often the case. Storey [62, 63] noted that eq. (1.38) can be equivalently written as follows,

$$\mathbb{E}\left(\frac{F_P}{N_P} \middle| N_P > 0\right) \Pr(N_P > 0), \quad (1.40)$$

and proposed the following definition for the positive FDR,

$$\text{pFDR} = \mathbb{E}\left(\frac{F_P}{N_P} \middle| N_P > 0\right), \quad (1.41)$$

where the term ‘positive’ refers to the conditioning on $N_P > 0$. Equation (1.41) has an interesting Bayesian interpretation as posterior probability if test statistics are assumed to come from a mixture of the distributions under $\mathcal{H}_0$ and $\mathcal{H}_a$ [63].

When a single statistical test is performed, the probability of a type I error α can be controlled by rejecting $\mathcal{H}_0$ when the p -value is smaller than α . On the other hand, if N tests of size α are conducted simultaneously, the FWER (or the FDR) will not generally be controlled at the same level; this is known as the multiple comparisons problem.

Control of the family-wise error rate

Many procedures have been proposed to control the FWER, either in the strong or weak sense.

A commonly adopted procedure is the Bonferroni correction, which rejects $\mathcal{H}_0$ at a corrected significance level α/N . The Bonferroni correction controls the FWER in the strong sense, and

$$\text{FWER} = \Pr(F_P \geq 1) \leq \alpha \frac{N_0}{N} \leq \alpha. \quad (1.42)$$

Introduce N random variables,

$$E_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \mathcal{H}_0 \text{ is true but rejected,} \\ 0 & \text{otherwise.} \end{cases} \quad (1.43)$$

Clearly, $F_P = \sum_i E_i$, and

$$\mathbb{E}(F_P) = \sum_i \mathbb{E}(E_i) = \sum_{i: \mathcal{H}_0 \text{ is true}} \Pr(E_i = 1) = \sum_{i: \mathcal{H}_0 \text{ is true}} \frac{\alpha}{N} = \alpha \frac{N_0}{N}. \quad (1.44)$$

Finally, from Bonferroni's inequality on the probability of finite unions of events,

$$\Pr\left(\bigcup_i X_i\right) \leq \sum_i \Pr(X_i), \quad (1.45)$$

it follows that

$$\text{FWER} = \Pr(F_P \geq 1) \leq \sum_i \Pr(F_P \geq i) = \mathbb{E}(F_P) = \alpha \frac{N_0}{N} \leq \alpha. \quad (1.46)$$

Note that Bonferroni correction does not postulate independence among tests. If this can be assumed, it is possible to obtain a tighter bound on the FWER by means of the Šidák correction, which rejects $\mathcal{H}_0$ at a corrected significance level $1 - (1 - \alpha)^{1/N}$. Consider the same random variables E_i as above; it holds that

$$\text{FWER} = \Pr(F_p \geq 1) = 1 - \Pr(F_p = 0) = 1 - \Pr\left(\bigcap_i E_i = 0\right) \quad (1.47)$$

$$= 1 - \prod_i \Pr(E_i = 0) = 1 - \prod_i (1 - \alpha)^{1/N} = \alpha, \quad (1.48)$$

where eq. (1.48) is a consequence of the hypothesised independence among tests.

The Šidák correction provides a tighter bound than the Bonferroni correction, but at the cost of requiring independence among tests. It is easy to see that

$$\frac{\alpha}{N} \leq 1 - (1 - \alpha)^{1/N}, \quad N \geq 1. \quad (1.49)$$

Furthermore, the corrected significance level of the Bonferroni procedure can be considered as the first-order Taylor approximation around $\alpha = 0$ of that of the Šidák procedure.

Control of the false discovery rate

While the Bonferroni correction does not assume independence among tests, it may be overly conservative when test statistics are highly dependent: in the extreme case of N perfectly dependent tests, the prescribed significance level α/N is much smaller than the actual FWER α . To overcome this issue, and motivated by advances in ‘high-throughput’ sciences including early ‘-omics’ [64], Benjamini et al. [61] proposed controlling the FDR as defined in eq. (1.38).

Controlling the FDR at some level α exerts a less stringent control over false positive findings (and thus affords greater statistical power) than controlling the FWER in the strong sense at the same level, since the latter is tantamount to controlling the probability of even one false positive finding irrespective of the configuration of true and false $\mathcal{H}_0$.

The procedure of Benjamini et al. [61] is now described. Given N tests, let $p_{(1)} \leq \dots \leq p_{(N)}$ be the ordered observed p -values. Define

$$k = \max_i \left\{ p_{(i)} \leq \frac{k}{N} \alpha \right\}, \quad (1.50)$$

and reject the null hypotheses associated with $p_{(1)}, \dots, p_{(k)}$, or do not reject any null hypothesis if no such i exists. As proved by Benjamini et al. [61], this procedure controls the FDR in the strong sense, and

$$\text{FDR} = \mathbb{E}(Q) \leq \alpha \frac{N_0}{N} \leq \alpha. \quad (1.51)$$

Resampling-based multiple comparisons

Consider a vector of p -values $\mathbf{p} = (p_1, \dots, p_N)$ obtained, for example, from separate regression models for some parameter of interest. This vector may be considered a realisation of the corresponding random vector $\mathbf{P} = (P_1, \dots, P_N)$. As described above, Bonferroni-style methods do not take into account distributional characteristics or interdependence in $\mathbf{P}$; however, if the joint distribution of $\mathbf{P}$ under $\mathcal{H}_0$ were known, more powerful single-step adjusted p -values could be computed by comparing the p_i against the theoretical minimum under $\mathcal{H}_0$ [65].

Sometimes the distribution of the random variable representing the minimum p -value $P_{(1)} = \min_i \{P_i\}$ under the complete $\mathcal{H}_0$ can be computed analytically. Assume that tests are independent; if the sampling distribution of the test statistic has been correctly specified, under the complete $\mathcal{H}_0$ it holds that $P_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$ by definition. The c.d.f. of $P_{(1)}$ under the complete $\mathcal{H}_0$ is then given by

$$F_{(1)}(p) = \Pr(P_{(1)} < p) = 1 - \Pr(P_{(1)} > p) = 1 - \Pr\left(\bigcap_i P_i > p\right) \quad (1.52)$$

$$= 1 - \prod_i [1 - F_i(p)] = 1 - [1 - F(p)]^N \quad (1.53)$$

$$= 1 - (1 - p)^N, \quad p \in [0, 1], \quad (1.54)$$

where eqs. (1.53) and (1.54) follow from the fact that $P_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ under the complete $\mathcal{H}_0$. The correspondence with the Šidák corrected significance level is clear.

When tests are not independent, the distribution of $P_{(1)}$ under the complete $\mathcal{H}_0$ is usually intractable. Nevertheless, in many cases it is feasible to simulate vectors $\mathbf{P}^*$ with the same distribution as the original $\mathbf{P}$, from which adjusted p -values may then be approximated. Algorithm 1.1 provides a simple implementation [65].

It is imperative that generation of the simulated data $\mathbf{Y}^*$ be done under the complete $\mathcal{H}_0$, for otherwise p -values will be simulated from non-null distributions, skewing adjustments towards non-significance. Computation of the simulated p -values $\mathbf{p}^*$ should proceed using the same method that produced the original p -values $\mathbf{p}$ from the original data $\mathbf{Y}$.

The variance of the estimates $\tilde{p}_i$ in algorithm 1.1 comes from the binomial distribution,

$$\mathbb{V}(\tilde{p}_i) = \frac{\tilde{p}_i (1 - \tilde{p}_i)}{m}, \quad (1.55)$$

where m is the number of simulations, and can be made arbitrarily small with larger m (and more accurate pseudo-random number generators).

Note that algorithm 1.1 can be adapted to accommodate test statistics rather than p -values, which may reduce the computational burden. Nevertheless, in this case, one must ensure that all tests have similar weight in determining the distribution of the minimum (or maximum, for most test statistics), which is the case e.g. when all tests use the same test statistic. Aside from computational considerations, p -values are natural candidates since they are (at least approximately) marginally uniformly distributed on the unit interval under $\mathcal{H}_0$.

Algorithm 1.1: Resampling-based multiple comparisons. Simulation of p -value vectors $\mathbf{p}^*$ is used to approximate the joint distribution of $\mathbf{P}$ under the complete $\mathcal{H}_0$, and used to compute adjusted p -values [65].

Input: observed p -values $\mathbf{p} = (p_1, \dots, p_N)$, number of simulations m

Output: adjusted p -values $\tilde{\mathbf{p}} = (\tilde{p}_1, \dots, \tilde{p}_N)$

for $i = 1, \dots, N$ **do**

$k_i \leftarrow 0$

for $j = 1, \dots, m$ **do**

 generate data $\mathbf{Y}^*$ from an estimated complete $\mathcal{H}_0$

 compute $\mathbf{p}^*$ from $\mathbf{Y}^*$

for $i = 1, \dots, N$ **do**

if $p_i \geq \min \mathbf{p}^*$ **then**

$k_i \leftarrow k_i + 1$

return $\tilde{p}_i = k_i/m$

Strengths and weaknesses

Univariate approaches have been widely adopted for several reasons. First, they are generally computationally efficient, although the running time can become prohibitive when more refined models (e.g. including random effects or non-parametric smoothing functions) are used. Second, they are extremely flexible, and can accommodate most types of predictor and outcome variables, and a wide range of parametric and non-parametric relationships. Third, they are readily available as part of most statistical packages including the free software environment for statistical computing R.

Coupled with the various multiple comparisons strategies that have been proposed in the literature, these methods are able to effectively handle extremely

large datasets. Moreover, the inclusion of random effects allows estimation and partitioning of structured nuisance variation induced by experimental conditions that could otherwise hinder statistical inference.

The main drawback of univariate methods is that they can only uncover simple relationships between the outcome and each independent variable, since by construction they only try to model the marginal effect of each predictor. For this reason, they may fail to reveal key features of the data, and may miss complex correlation patterns among predictors, leading e.g. to redundancy in the results. More importantly, these methods disregard potential interactions and effect modifications, and may thus overlook relationships involving combinations of predictors that jointly mediate development of the phenotype of interest.

These limitations have motivated the introduction of multivariate approaches that are not only able to account for the correlation among predictors, but can also identify subsets of ‘relevant’ independent variables; these are reviewed in the next section.

1.3.2 Multivariate approaches

Principal component analysis (PCA)

Principal component analysis (PCA) is possibly the best-known multivariate analytical tool. In approximate terms, PCA identifies a rotation of a data matrix along directions of maximum variability (or, equivalently, high information), and projects it onto a lower-dimensional space while preserving as much variability as possible. The rotation is specified in terms of a column matrix of loading vectors; these are chosen in directions of maximum variability, subject to orthogonality constraints.

More formally, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a real matrix of size $n \times p$ whose columns have

been centred (and possibly scaled). To derive the form of the PCs, consider the first loading vector $\mathbf{w}_1$, defined to maximise

$$\mathbb{V}(\mathbf{w}_1^\top \mathbf{X}) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1, \quad (1.56)$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix (or, more commonly, its empirical estimate) of $\mathbf{X}$. Equation (1.56) is unbounded for unconstrained $\mathbf{w}_1$, so a normalisation constraint must be imposed, for example $\mathbf{w}_1^\top \mathbf{w}_1 = 1$. Using the technique of Lagrange multipliers,

$$\max_{\mathbf{w}_1} (\mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1) \quad \text{s.t. } \mathbf{w}_1^\top \mathbf{w}_1 = 1 \quad (1.57)$$

can be turned into the unconstrained maximisation problem

$$\max_{\mathbf{w}_1} [\mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 - \lambda (\mathbf{w}_1^\top \mathbf{w}_1 - 1)], \quad (1.58)$$

where λ is the Lagrange multiplier. Differentiation with respect to $\mathbf{w}_1$ yields

$$(\boldsymbol{\Sigma} - \lambda \mathbf{I}_p) \mathbf{w}_1 = \mathbf{0}, \quad (1.59)$$

and thus λ is an eigenvalue of $\boldsymbol{\Sigma}$ with corresponding eigenvector $\mathbf{w}_1$. Note that

$$\mathbb{V}(\mathbf{w}_1^\top \mathbf{X}) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 = \mathbf{w}_1^\top (\boldsymbol{\Sigma} \mathbf{w}_1) = \mathbf{w}_1^\top \lambda \mathbf{w}_1 = \lambda \mathbf{w}_1^\top \mathbf{w}_1 = \lambda, \quad (1.60)$$

since $\mathbf{w}_1^\top \mathbf{w}_1 = 1$ by construction, and so the $\mathbf{w}_1$ that maximises eq. (1.56) is the eigenvector associated with the largest eigenvalue λ_1 , and $\mathbb{V}(\mathbf{w}_1^\top \mathbf{X}) = \lambda_1$.

Consider now the second loading vector $\mathbf{w}_2$, which maximises

$$\mathbb{V}(\mathbf{w}_2^\top \mathbf{X}) = \mathbf{w}_2^\top \boldsymbol{\Sigma} \mathbf{w}_2 \quad \text{s.t. } \mathbf{w}_2^\top \mathbf{w}_2 = 1, \mathbb{V}(\mathbf{w}_1^\top \mathbf{X}, \mathbf{w}_2^\top \mathbf{X}) = 0, \quad (1.61)$$

and note that

$$\mathbb{V}(\mathbf{w}_1^\top \mathbf{X}, \mathbf{w}_2^\top \mathbf{X}) = \mathbb{V}(\mathbf{w}_2^\top \mathbf{X}, \mathbf{w}_1^\top \mathbf{X}) = \mathbf{w}_2^\top \boldsymbol{\Sigma} \mathbf{w}_1 \quad (1.62)$$

$$= \mathbf{w}_2^\top (\boldsymbol{\Sigma} \mathbf{w}_1) = \mathbf{w}_2^\top \lambda_1 \mathbf{w}_1 = \lambda_1 \mathbf{w}_2^\top \mathbf{w}_1 = \lambda_1 \mathbf{w}_1^\top \mathbf{w}_2, \quad (1.63)$$

so that the second constraint in eq. (1.61) effectively reduces to $\mathbf{w}_1^\top \mathbf{w}_2 = \mathbf{w}_2^\top \mathbf{w}_1 = 0$. Applying again the technique of Lagrange multipliers, the following unconstrained maximisation problem is obtained,

$$\max_{\mathbf{w}_2} [\mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda (\mathbf{w}_2^\top \mathbf{w}_2 - 1) - \mu \mathbf{w}_2^\top \mathbf{w}_1], \quad (1.64)$$

where λ and μ are the Lagrange multipliers. Differentiation with respect to $\mathbf{w}_2$ yields

$$\mathbf{\Sigma} \mathbf{w}_2 - \lambda \mathbf{w}_2 - \mu \mathbf{w}_1 = \mathbf{0}, \quad (1.65)$$

and multiplication on the left by $\mathbf{w}_1^\top$ leads to

$$\mathbf{w}_1^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda \mathbf{w}_1^\top \mathbf{w}_2 - \mu \mathbf{w}_1^\top \mathbf{w}_1 = \mu = 0, \quad (1.66)$$

since the first two terms are zero, and $\mathbf{w}_1^\top \mathbf{w}_1 = 1$ by construction. Therefore,

$$(\mathbf{\Sigma} - \lambda \mathbf{I}_p) \mathbf{w}_2 = \mathbf{0}, \quad (1.67)$$

and λ is again an eigenvalue of $\mathbf{\Sigma}$ with corresponding eigenvector $\mathbf{w}_2$. As before, maximising $\mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2$ is tantamount to choosing the largest possible value for λ . Assuming that the eigenvalues of $\mathbf{\Sigma}$ have unit algebraic multiplicity, it must be $\lambda \neq \lambda_1$, for otherwise $\mathbf{w}_2 = \mathbf{w}_1$, violating $\mathbf{w}_2^\top \mathbf{w}_1 = 0$. Thus, λ is the second largest eigenvalue λ_2 of $\mathbf{\Sigma}$, $\mathbf{w}_2$ is the associated eigenvector, and $\mathbb{V}(\mathbf{w}_2^\top \mathbf{X}) = \lambda_2$.

By the same token, it can be shown that the i^{th} loading vector $\mathbf{w}_i$ is the eigenvector of $\mathbf{\Sigma}$ corresponding to the eigenvalue λ_i , and $\mathbb{V}(\mathbf{w}_i^\top \mathbf{X}) = \lambda_i$. Thus, the eigenvalues of $\mathbf{\Sigma}$ have a natural interpretation, and the ratio

$$\tilde{\mathbb{V}}_i = \frac{\mathbb{V}(\mathbf{w}_i^\top \mathbf{X})}{\sum_j \mathbb{V}(\mathbf{w}_j^\top \mathbf{X})} = \frac{\lambda_i}{\sum_j \lambda_j} \quad (1.68)$$

represents the percentage of variability in $\mathbf{X}$ accounted for by each PC.

Given the real symmetric matrix $\mathbf{\Sigma}$ of order n , PCs can therefore be computed by considering its eigendecomposition,

$$\mathbf{\Sigma} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top, \quad (1.69)$$

where $\mathbf{Q}$ is an orthonormal matrix of order n whose columns are the eigenvectors of $\mathbf{\Sigma}$ (or, equivalently, the loading vectors), and $\mathbf{\Lambda}$ is a diagonal matrix of the same order whose entries are the associated eigenvalues (or, equivalently, the variation explained by each PC).

For large p , it is generally not feasible to compute the eigendecomposition of $\mathbf{\Sigma}$ directly; instead, PCA is usually performed by means of singular value decomposition (SVD) of $\mathbf{X}$,

$$\mathbf{X} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top, \quad (1.70)$$

where $\mathbf{U}$ is an orthogonal matrix of order n , $\mathbf{\Lambda}$ is a non-negative $n \times p$ diagonal matrix, and $\mathbf{V}$ is an orthogonal matrix of order p . It holds that

$$\mathbf{\Sigma} = \mathbf{X}^\top \mathbf{X} = (\mathbf{V} \mathbf{\Lambda} \mathbf{U}^\top) (\mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top) = \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^\top, \quad (1.71)$$

since $\mathbf{U}$ is orthogonal, $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_n$. Note that, in eq. (1.71), the right-singular vectors in $\mathbf{V}$ are the eigenvector of $\mathbf{\Sigma}$, and the entries of the diagonal matrix $\mathbf{\Lambda}^2$ are the associated eigenvalues.

The full PC decomposition of $\mathbf{X}$ can be given as

$$\mathbf{T} = \mathbf{X} \mathbf{W}, \quad (1.72)$$

where $\mathbf{W} = (\mathbf{w}_1 | \dots | \mathbf{w}_p)$ is an orthonormal matrix of order p whose columns are the eigenvectors of $\mathbf{\Sigma}$ (or, equivalently, the loading vectors as defined above). Without loss of generality, it will be assumed in the following that the eigenvectors in $\mathbf{W}$ are sorted in non-increasing order of explained variability, or equivalently that the associated eigenvalues satisfy $\lambda_1 \geq \dots \geq \lambda_p$. Given eq. (1.72), it is useful to consider the lower-order approximation

$$\mathbf{T}_k = \mathbf{X} \mathbf{W}_k, \quad k < p, \quad (1.73)$$

where $\mathbf{W}_k = (\mathbf{w}_1 | \dots | \mathbf{w}_k)$, and $\mathbf{T}_k$ is the projection of $\mathbf{X}$ on the k -dimensional space defined by the first k PCs. The number of PCs k is commonly determined

according to

$$k = \arg \min_i \left(\sum_{j=1}^i \tilde{V}_j \geq \alpha \right), \quad (1.74)$$

which corresponds to the lowest-dimensional projection that retains at least a fraction α of the total variability in $\mathbf{X}$. For $k \in \{1, \dots, 3\}$, observations may be easily plotted to detect possible patterns. Once a lower-dimensional projection $\mathbf{T}_k$ has been selected, it may also be used in regression models in place of $\mathbf{X}$; this is known as PC regression. Note that, if $n < p$, at most $n - 1$ components may be defined, since $\lambda_n = \dots = \lambda_p = 0$.

Partial least squares (PLS)

While PCs may be used in regression models, in general the decomposition that maximises the explained variability in $\mathbf{X}$ will not necessarily be predictive of $\mathbf{Y}$. Partial least squares (PLS) redefines the loading vectors $\mathbf{w}_i$ so that they maximise the covariance between $\mathbf{X}$ and $\mathbf{Y}$ (rather than the variance in $\mathbf{X}$ alone).

Univariate PLS (PLS1)

More formally, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a real matrix of size $n \times p$ whose columns have been centred (and possibly scaled), and $\mathbf{y} \in \mathbb{R}^n$ a centred (and possibly scaled) real vector. The PLS procedure originally introduced by Wold et al. [66] is presented in algorithm 1.2, and will now be described in further detail. Note that alternative approaches to PLS regression have been introduced [67], some with improved numerical stability.

As described above, PLS seeks directions in the space of $\mathbf{X}$ that are relevant to explaining the variation in $\mathbf{y}$; in fact, the first loading vector $\mathbf{w}_1$ is formed by standardising the covariance matrix $\mathbf{X}^\top \mathbf{y}$, and its transpose $\mathbf{w}_1^\top$ is proportional to the ordinary least squares regression coefficient estimates $\mathbf{y}^\top \mathbf{X} / \mathbf{y}^\top \mathbf{y}$. The score

vector $\mathbf{t}_1 \in \mathbb{R}^n$ is a linear combination of the columns of $\mathbf{X}$ with weights $\mathbf{w}_1$, and may thus be considered the ‘optimal’ linear combination to predict $\mathbf{y}$. The vector $\mathbf{p}_1 \in \mathbb{R}^p$ and the regression coefficient c_1 are computed by linear regression of $\mathbf{X}$ and $\mathbf{y}$ on $\mathbf{t}_1$, respectively. They are used in the two ‘deflation’ steps,

$$\mathbf{X}_2 = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^\top, \quad (1.75)$$

$$\mathbf{y}_2 = \mathbf{y} - c_1 \mathbf{t}_1, \quad (1.76)$$

from which it can be seen that the procedure is then repeated on the residuals $\mathbf{X}_2$ and $\mathbf{y}_2$, effectively ensuring orthogonality of the $\mathbf{t}_i$. Each iteration of algorithm 1.2 reveals additional information on the relationship between $\mathbf{X}$ and $\mathbf{y}$. At the i^{th} iteration, it holds that

$$\mathbf{X} = \mathbf{T} \mathbf{P}^\top + \mathbf{X}_{i+1}, \quad (1.77)$$

$$\mathbf{y} = \mathbf{T} \mathbf{c}^\top + \mathbf{y}_{i+1}, \quad (1.78)$$

and so the number of components k should be selected such that $\mathbf{X}_{i+1}$ and $\mathbf{y}_{i+1}$ are approximately uncorrelated. The PLS regression coefficients $\hat{\boldsymbol{\beta}}$ computed in algorithm 1.2 may be used for prediction: given a new sample $\mathbf{x}^* \in \mathbb{R}^p$, the associated prediction is given by

$$y^* = \bar{y} + (\mathbf{x}^* - \bar{\mathbf{x}}) \hat{\boldsymbol{\beta}}, \quad (1.79)$$

where $\bar{\mathbf{x}} \in \mathbb{R}^p$ and $\bar{y} \in \mathbb{R}$ are the empirical means of $\mathbf{X}$ and $\mathbf{y}$, respectively.

Multivariate PLS (PLS2)

Algorithm 1.2 applies to the case of a single response vector $\mathbf{y} \in \mathbb{R}^n$. If one is interested in modelling a multivariate response $\mathbf{Y} = (\mathbf{y}_1 | \dots | \mathbf{y}_q)$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$, one may apply algorithm 1.2 separately to each $\mathbf{y}_i$. For ordinary least squares, this is tantamount to estimating $\mathbf{B}$ for the regression model

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}, \quad (1.80)$$

where $\mathbf{B} \in \mathbb{R}^{p \times q}$, and $\mathbf{E} = (\epsilon_1 | \dots | \epsilon_q)$ is a random matrix of size $n \times q$; however, this is not the case for PLS.

A simple modification, presented in algorithm 1.3, takes into account correlations in $\mathbf{Y}$; this is done by substituting eq. (1.78) with a model for $\mathbf{Y}$ similar to that of eq. (1.77),

$$\mathbf{X} = \mathbf{T} \mathbf{P}^\top + \mathbf{X}_{i+1}, \quad (1.81)$$

$$\mathbf{Y} = \mathbf{U} \mathbf{Q}^\top + \mathbf{Y}_{i+1}, \quad (1.82)$$

which includes a second set of scores $\mathbf{U} = (\mathbf{u}_1 | \dots | \mathbf{u}_k)$ for $\mathbf{Y}$. Equations (1.81) and (1.82) are linked by the inner relationship

$$\mathbf{U} = \mathbf{T} \mathbf{C} + \mathbf{U}_{i+1}. \quad (1.83)$$

In the special case $q = 1$, algorithm 1.3 reduces to algorithm 1.2, for then $\mathbf{u}_i = \mathbf{y}_i$, and $\mathbf{q}_i = 1$.

While algorithms 1.2 and 1.3 are mostly similar, there are a few differences that should be noted. In particular, the definition of the $\mathbf{w}_i$ in algorithm 1.3 is based on the covariance between $\mathbf{X}_i$ and an arbitrarily selected column $\mathbf{u}_i$ of $\mathbf{Y}_i$, assumed to be representative of the response matrix; the score vector $\mathbf{t}_i$ is again a linear combination of the columns of $\mathbf{X}_i$ with weights $\mathbf{w}_i$. Similarly, the weights $\mathbf{q}_i$ depend on the covariance between $\mathbf{Y}_i$ and $\mathbf{t}_i$, and the corresponding score vector $\mathbf{u}_i$ is a linear combination of the columns of $\mathbf{Y}_i$ weighted by $\mathbf{q}_i$. The vector $\mathbf{p}_i$ and the regression coefficient c_i are computed by linear regression of $\mathbf{X}_i$ and $\mathbf{u}_i$ on $\mathbf{t}_i$, respectively. As in algorithm 1.2, they are used in the two ‘deflation’ steps,

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{t}_i \mathbf{p}_i^\top, \quad (1.84)$$

$$\mathbf{Y}_{i+1} = \mathbf{Y}_i - c_i \mathbf{t}_i \mathbf{q}_i^\top, \quad (1.85)$$

from which it can be seen (by application of eq. (1.83)) that the procedure is then repeated on the residuals $\mathbf{X}_{i+1}$ of $\mathbf{X}_i$ on $\mathbf{t}_i$, and $\mathbf{Y}_{i+1}$ of $\mathbf{Y}_i$ on $\mathbf{u}_i$; this ensures orthogonality of the $\mathbf{t}_i$.

Furthermore, at convergence it holds that

$$\mathbf{X}_i^\top \mathbf{Y}_i \mathbf{Y}_i^\top \mathbf{X}_i \mathbf{w}_i = \mathbf{X}_i^\top \mathbf{Y}_i \mathbf{Y}_i^\top \mathbf{t}_i \quad \text{since } \mathbf{t}_i = \mathbf{X}_i \mathbf{w}_i \quad (1.86)$$

$$= \mathbf{X}_i^\top \mathbf{Y}_i \mathbf{q}_i \|\mathbf{Y}_i^\top \mathbf{t}_i\| \quad \text{since } \mathbf{q}_i = \mathbf{Y}_i^\top \mathbf{t}_i / \|\mathbf{Y}_i^\top \mathbf{t}_i\| \quad (1.87)$$

$$= \mathbf{X}_i^\top \mathbf{u}_i \|\mathbf{Y}_i^\top \mathbf{t}_i\| \quad \text{since } \mathbf{u}_i = \mathbf{Y}_i \mathbf{q}_i \quad (1.88)$$

$$= \mathbf{w}_i \|\mathbf{X}_i^\top \mathbf{u}_i\| \|\mathbf{Y}_i^\top \mathbf{t}_i\| \quad \text{since } \mathbf{w}_i = \mathbf{X}_i^\top \mathbf{u}_i / \|\mathbf{X}_i^\top \mathbf{u}_i\| \quad (1.89)$$

$$= \lambda_i \mathbf{w}_i, \quad (1.90)$$

and thus the $\mathbf{w}_i$ are the eigenvectors of $\mathbf{X}_i^\top \mathbf{Y}_i \mathbf{Y}_i^\top \mathbf{X}_i$ with associated eigenvalues

$$\lambda_i = \|\mathbf{X}_i^\top \mathbf{u}_i\| \|\mathbf{Y}_i^\top \mathbf{t}_i\|. \quad (1.91)$$

Similarly, it may be shown that the $\mathbf{q}_i$ are the eigenvectors of $\mathbf{Y}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{Y}_i$, that the $\mathbf{t}_i$ are the eigenvectors of $\mathbf{X}_i \mathbf{X}_i^\top \mathbf{Y}_i \mathbf{Y}_i^\top$, and that the $\mathbf{u}_i$ are the eigenvectors of $\mathbf{Y}_i \mathbf{Y}_i^\top \mathbf{X}_i \mathbf{X}_i^\top$, all with associated eigenvalues as in eq. (1.91).

Orthogonal PLS methods

Standard PLS is highly sensitive to structured noise in the predictor matrix $\mathbf{X}$, which may cause it to include more components to capture more variance; in turn, this might lead to a situation in which the number of components exceeds that of response variables. Moreover, noise (defined as variation in $\mathbf{X}$ that is orthogonal to the response matrix $\mathbf{Y}$) may influence the first PLS component, adversely affecting its correlation with $\mathbf{Y}$, and thus leading to poor interpretability of the model.

To address this problem, Trygg et al. [68] proposed an improved algorithm, called Orthogonal Projections to Latent Structures (O-PLS). It aims at removing noise, defined as systematic variation in the predictor matrix $\mathbf{X}$ that is uncorrelated to the response matrix $\mathbf{Y}$, based on a pre-processing method known as Orthogonal Signal Correction (OSC) [69], an idea that has also been exploited for

data normalization purposes [70]. The predictivity of O-PLS is the same as for standard PLS, but it also separates relevant (predictive) and non-relevant (orthogonal to $\mathbf{Y}$) variation, thus improving the overall interpretability of the model.

Bidirectional Orthogonal Projections to Latent Structures (O2-PLS) [68, 71] is a further improvement of O-PLS that also removes systematic variation in the response matrix $\mathbf{Y}$ that is uncorrelated to the predictor matrix $\mathbf{X}$, thus enabling bidirectional predictions to be made. The systematic variation in $\mathbf{X}$ and $\mathbf{Y}$ is thus separated into three parts: one that is common to both ($\mathbf{X}$ - $\mathbf{Y}$ joint variation), one that is unique to $\mathbf{X}$ ($\mathbf{Y}$ -orthogonal variation), and one that is unique to $\mathbf{Y}$ ($\mathbf{X}$ -orthogonal variation). The O2-PLS decomposition takes the following form,

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{T}_\perp\mathbf{P}_\perp^\top + \mathbf{E}, \quad (1.92)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\top + \mathbf{U}_\perp\mathbf{Q}_\perp^\top + \mathbf{F}, \quad (1.93)$$

where $\mathbf{T}_\perp$ and $\mathbf{P}_\perp$ are the score and loading matrices, respectively, of the part of $\mathbf{X}$ that is orthogonal, and thus uncorrelated, to $\mathbf{Y}$, and *vice versa* for $\mathbf{U}_\perp$ and $\mathbf{Q}_\perp$.

OnPLS [72] generalises O2-PLS to the case of multiple matrices, building bidirectional pairwise prediction models between all different pairs. OnPLS has been further extended to allow for extraction of orthogonal variation in both columns and rows [73].

Regularisation and variable selection

While methods described in sections 1.3.2 and 1.3.2 can be used in the ‘small n , large p ’ scenario, the corresponding models often lack interpretability.

As discussed in the introduction to this section, when $n \ll p$, it is assumed that only a small number of predictors contribute to the response (or responses), which is tantamount to assuming that the vector of regression coefficients $\boldsymbol{\beta}$ (or the matrix $\mathbf{B}$ for multivariate $\mathbf{Y}$) is sparse.

The parsimonious representation afforded by sparse models lends greater interpretability by identifying the subset of ‘relevant’ predictors among many ‘irrelevant’ variables. In addition, sparse models circumvent the multiple comparisons problem expounded in section 1.3.1, for they consider all p variables at the same time while trying to identify a small subset of ‘relevant’ predictors.

Two approaches to incorporate sparsity are now presented.

Regularisation (or penalisation)

Regularised (or penalised) methods penalise models with ‘unrealistic’ values for the model parameters θ , thus introducing additional information to prevent overfitting. From a Bayesian perspective, this may be thought of as imposing certain prior distributions on θ , and finding the mode of the posterior distribution of θ given the data; in many instances, this may be proved analytically.

More formally, instead of maximising $\ell(\theta|\mathbf{y})$ directly, a penalty function P that depends on θ only is introduced, and the following function is maximised,

$$R(\theta|\mathbf{y}) = \ell(\theta|\mathbf{y}) - \lambda P(\theta), \quad (1.94)$$

where the regularisation parameter $\lambda \geq 0$ controls the trade-off between ℓ and P . It is directly related to the bias-variance dilemma: small values of λ will give more weight to ℓ , and the resulting model will tend to overfit the data (and thus have high variance); on the other hand, large values of λ will lead to models that underfit the data, and may thus be overly simplistic and potentially biased. Tuning of λ is therefore of primary importance, and typically proceeds by cross-validation.

Note also that the objective function of eq. (1.94) is equivalent to that obtained by applying the technique of Lagrange multipliers to the constrained maximisa-

tion problem

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{y}) \quad \text{s.t. } P(\boldsymbol{\theta}) \leq t, \quad (1.95)$$

for some tuning parameter t , from which it can be seen that λ in eq. (1.94) is in fact a Lagrange multiplier.

In place of maximising eq. (1.94), the following function may equivalently be minimised,

$$r(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|\mathbf{y}) + \lambda P(\boldsymbol{\theta}), \quad (1.96)$$

where L here is a loss function that is usually proportional to the negative log-likelihood of the model, e.g. the residual sum of squares in classic linear models.

As described above, the penalty function P should penalise ‘unrealistic’ values for $\boldsymbol{\theta}$, which in the context of sparse models is taken to mean values away from zero.

Two commonly used penalty functions are based on L^p norms,

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p}, \quad (1.97)$$

of $\boldsymbol{\theta}$ for $p \in \{1, 2\}$, and are as follows,

$$P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i| \quad (\text{lasso penalty}), \quad (1.98)$$

$$P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2 = \sum_i \theta_i^2 \quad (\text{Tikhonov or ridge penalty}). \quad (1.99)$$

While the lasso penalty of eq. (1.98) ensures sparsity, it has some limitations: first, for $p > n$, it will select at most n variables among p . Second, if one predictor belonging to a ‘group’ of highly correlated variables is selected, it will tend to ignore the others. These limitations are overcome by the Tikhonov or ridge penalty of eq. (1.99), which however does not favour sparse solutions as strongly as the lasso penalty.

Their convex combination,

$$P(\boldsymbol{\theta}) = \alpha \|\boldsymbol{\theta}\|_2 + (1 - \alpha)\|\boldsymbol{\theta}\|_1, \quad \alpha \in [0, 1], \quad (1.100)$$

is known as the ‘elastic net’ penalty. It leverages their strengths while diminishing their weaknesses, at the cost of one additional parameter α that is again typically tuned by cross-validation. Note that eqs. (1.98) and (1.99) are special cases of eq. (1.100), and are recovered for $\alpha = 0$ and $\alpha = 1$, respectively.

The elastic net penalty of eq. (1.100) may be introduced in different models. For example, in the context of eq. (1.96), the estimate $\hat{\boldsymbol{\beta}}$ of the regression coefficients of the classic linear model of eq. (1.13) becomes

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2, \quad (1.101)$$

where two separate regularisation parameters λ_1, λ_2 were introduced in place of combinations of λ and α . Jolliffe et al. [74] introduced the lasso penalty of eq. (1.98) in PCA, obtaining the series of maximisation problems

$$\max_{\mathbf{w}_i} (\mathbf{w}_i^\top \boldsymbol{\Sigma} \mathbf{w}_i) \quad \text{s.t.} \quad \mathbf{w}_i^\top \mathbf{w}_i = 1, \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall j < i \leq 2, \|\mathbf{w}_i\|_1 \leq t, \quad (1.102)$$

where t is again a tuning parameter. Due to lack of computational efficiency that makes it infeasible to tune t e.g. by cross-validation, Zou et al. [75] introduced the elastic net penalisation in PCA, and considered the maximisation problem

$$\max_{\mathbf{A}, \mathbf{B}} \|\mathbf{X} - \mathbf{XBA}^\top\|_2 + \sum_{i=1}^k \lambda_{1i} \|\boldsymbol{\beta}_i\|_1 + \lambda_2 \sum_{i=1}^k \|\boldsymbol{\beta}_i\|_2 \quad \text{s.t.} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I}_k, \quad (1.103)$$

which is based on a transformation of the PCA problem into a regression-type problem. Similar modifications are possible for PLS [76].

Variable selection

The (binary) variable selection problem is formulated as follows.

Let $\gamma \in [0, 1]^p$ be a latent binary vector, and define $\mathbf{X}_\gamma$ as the design matrix of size $n \times p_\gamma$, where $p_\gamma = \gamma^\top \mathbf{1}_p$, with columns of $\mathbf{X}$ corresponding to $\gamma_i = 1$; in this setting, it is customary to separate the intercept term so that it is not affected by the variable selection procedure. Similarly, define $\boldsymbol{\theta}_\gamma$ as the corresponding vector of (non-zero) model parameters. The variable selection problem is the minimisation problem

$$\arg \min_{\gamma} C(\boldsymbol{\theta}_\gamma | \mathbf{y}), \quad (1.104)$$

where the form of the cost function C may depend on the model.

Common choices for C are presented in the following.

Akaike information criterion (AIC)

Defined as

$$\text{AIC}(\boldsymbol{\theta}_\gamma | \mathbf{y}) = -2\ell(\hat{\boldsymbol{\theta}}_\gamma | \mathbf{y}) + 2p_\gamma, \quad (1.105)$$

where ℓ is evaluated at the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_\gamma$.

Bayesian information criterion (BIC)

Defined as

$$\text{BIC}(\boldsymbol{\theta}_\gamma | \mathbf{y}) = -2\ell(\hat{\boldsymbol{\theta}}_\gamma | \mathbf{y}) + p_\gamma \ln n, \quad (1.106)$$

where ℓ is evaluated at the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_\gamma$.

Adjusted coefficient of determination

Given the coefficient of determination

$$R^2 = 1 - \frac{\text{SS}_M}{\text{SS}_\emptyset}, \quad \text{SS}_M = \sum_i (y_i - \hat{y}_i)^2, \quad \text{SS}_\emptyset = \sum_i (y_i - \bar{y})^2, \quad (1.107)$$

where $\hat{\mathbf{y}}$ is the vector of predicted values from the model (that thus depends on γ), and $\bar{y} = \sum_i y_i / n$, the adjusted coefficient of determination is defined as follows,

$$\bar{R}^2 = 1 - \frac{SS_M / (n - p_\gamma - 1)}{SS_\emptyset / (n - 1)} = 1 - \frac{n - 1}{n - p_\gamma - 1} (1 - R^2) \leq R^2, \quad (1.108)$$

where SS_M and $SS_\emptyset$ are now divided by the corresponding degrees of freedom of the variance estimates.

Predicted residual sum of squares (PRESS)

Defined as

$$PRESS = \sum_i (y_i - \hat{y}_i^{-i})^2, \quad (1.109)$$

where $\hat{y}_i^{-i}$ is the predicted value for the i^{th} observation estimated from the model (again depending on γ) fitted to all observations but the i^{th} . As such, its computation is effectively a ‘leave-one-out’ cross-validation procedure.

Mallows’ C_p

Defined as

$$C_p = \frac{SS_M}{SS_F / (n - p - 1)} + 2(p_\gamma + 1) - n, \quad (1.110)$$

where SS_M is as in eq. (1.107), and thus depends on the model selected through γ , and SS_F is similarly obtained from the full model including all p predictors (i.e. the model associated with $\gamma = \mathbf{1}_p$) as well as the intercept term.

As eqs. (1.105) to (1.110) exemplify, the prototypical cost function C tries to strike a balance between two competing aims, model fit and model size. If model complexity was not controlled, it would always be possible to improve model fit by including more parameters: for example, the unadjusted coefficient of determination R^2 of eq. (1.107) can always be brought arbitrarily close to 1. The same balancing principle underlies regularisation, and the parallel with eq. (1.96)

is clear; in fact, the penalty function $P(\theta_\gamma) = p_\gamma$ corresponds to the L^p norm of eq. (1.97) for $p \rightarrow 0$.

Nevertheless, variable selection approaches are radically different in that the minimisation problem of eq. (1.104) is discrete. The search space defined by all possible choices of γ has size exponential in p , and selection of the ‘best’ model (according to some cost function C) requires evaluation of all 2^p possibilities. Of course, complete enumeration is infeasible for all but the smallest values of p , and definitely not in the ‘small n , large p ’ scenario.

The main issue of variable selection then becomes the construction of clever algorithms to navigate the search space in some principled way that avoids exploration of ‘unpromising’ models. Typical approaches include greedy algorithms such as stepwise variable selection, and classic global optimisation solutions such as ‘branch-and-bound’ methods. Stochastic methods have found wide application in a Bayesian context. Bayesian variable selection techniques provide easily interpretable measures, such as the posterior model probability and the marginal inclusion probability of each predictor, that have fostered their adoption. Moreover, they allow incorporation of prior information, e.g. on model complexity by means of distributional assumptions on γ .

Strengths and weaknesses

Dimensionality reduction techniques such as PCA, PLS, and derived methods facilitate joint statistical analysis of the full set of independent variables by reducing the dimension of the space of predictors, effectively compressing the data. They can efficiently identify the main trends and sources of variation in the data, and compressing them into a small number of latent variables.

A peculiar issue worth noting is the limited ability to directly include known confounders in these models. A two-stage procedure is commonly employed:

first, univariate models are used to estimate the effect of confounders, and a new data matrix is built from the residuals; second, dimensionality reduction algorithms are applied. Nonetheless, this procedure assumes that confounders have independent effects, which may lead to biased adjustments.

More importantly, in high-dimensional settings these techniques are prone to select features that are irrelevant with respect to the outcome of interest. Furthermore, unless the process under examination is mainly driven by a few independent factors measured with little to no error, latent variables may be defined in such a way that they capture mostly noise (rather than biologically relevant variation); in this situation, identifying and accounting for technically-induced variation becomes impractical, and the resulting model lacks any biological interpretability.

Several approaches have been proposed to overcome these issues. The overarching principle of regularised methods and variable selection strategies is the inclusion of prior information in the model to prevent overfitting. In addition, these methods are able to take into account the effect of confounders, which may be excluded from penalisation or the selection procedure, and provide a small interpretable set of ‘relevant’ variables.

Nevertheless, they may be computationally very intensive: this is especially true of variable selection techniques, although recent improvements have enabled their application to large-scale datasets. Another peculiarity of these methods is that models are defined by maximising some measure of prediction accuracy or explained variation. For this reason, aetiological interpretation will usually require additional targeted analyses by means of validation studies.

References

1. A Stevenson. *Oxford Dictionary of English*. 3rd ed. Oxford University Press, 2010.
2. D Napoletani, M Panza, and D C Struppa. 'Agnostic science: towards a philosophy of data analysis'. *Foundations of Science* **16.1** (2011), 1–20.
3. M Chadeau-Hyam, G Campanella, T Jombart, L Bottolo, L Portengen, P Vineis, B Liqueur, and R C H Vermeulen. 'Deciphering the complex: methodological overview of statistical models to derive '–omics'–based biomarkers'. *Environmental and Molecular Mutagenesis* **54.7** (2013), 542–557.
4. B Alberts, D Bray, K Hopkin, A Johnson, K Roberts, J Lewis, M Raff, and P Walter. *Essential cell biology*. 4th ed. Garland Science, 2013.
5. B Alberts, A Johnson, J Lewis, D Morgan, M Raff, K Roberts, and P Walter. *Molecular biology of the cell*. 6th ed. Garland Science, 2014.
6. International Human Genome Sequencing Consortium. 'Initial sequencing and analysis of the human genome'. *Nature* **409** (2001), 860–921.
7. J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J Russo Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, et al. 'The sequence of the human genome'. *Science* **291.5507** (2001), 1304–1351.
8. D E Reich, M Cargill, S Bolck, J Ireland, P C Sabeti, D J Richter, T Lavery, R Kouyoumjian, S F Farhadian, R Ward, and E S Lander. 'Linkage disequilibrium in the human genome'. *Nature* **411** (2001), 199–204.

9. M Cargill, D Altshuler, J Ireland, P Sklar, K Ardlie, N Patil, C R Lane, E P Lim, N Kalyanaraman, J Nemesh, L Ziaugra, L Friedland, A Rolfe, J Warrington, R Lipshutz, G Q Daley, and E S Lander. 'Characterization of single-nucleotide polymorphisms in coding regions of human genes'. *Nature Genetics* **22** (1999), 231–238.
10. R Hunt, Z E Sauna, S V Ambudkar, M M Gottesman, and C Kimchi-Sarfaty. 'Silent (synonymous) SNPs: should we care about them?' *Single Nucleotide Polymorphisms*. Ed. by A A Komar. **578**. Methods in Molecular Biology. Humana Press, 2009, 23–39.
11. R Redon, S Ishikawa, K R Fitch, L Feuk, G H Perry, T D Andrews, H Fiegler, M H Shapero, A R Carson, W Chen, E K Cho, S Dallaire, J L Freeman, J R González, M Gratacòs, J Huang, D Kalaitzopoulos, D Komura, J R MacDonald, C R Marshall, et al. 'Global variation in copy number in the human genome'. *Nature* **444** (2006), 444–454.
12. B E Stranger, M S Forrest, M Dunning, C E Ingle, C Beazley, N Thorne, R Redon, C P Bird, A de Grassi, C Lee, C Tyler-Smith, N Carter, S W Scherer, S Tavaré, P Deloukas, M E Hurles, and E T Dermitzakis. 'Relative impact of nucleotide and copy number variation on gene expression phenotypes'. *Science* **315**.5813 (2007), 848–853.
13. M O'Driscoll. 'Haploinsufficiency of DNA damage response genes and their potential influence in human genomic disorders'. *Current Genomics* **9**.3 (2008), 137–146.
14. V T Dang, K S Kassahn, A E Marcos, and M A Ragan. 'Identification of human haploinsufficient genes and their genomic proximity to segmental duplications'. *European Journal of Human Genetics* **16** (2008), 1350–1357.

15. S T Sherry, M-H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. 'dbSNP: the NCBI database of genetic variation'. *Nucleic Acids Research* **29.1** (2001), 308–311.
16. J N Hirschhorn and M J Daly. 'Genome-wide association studies for common diseases and complex traits'. *Nature Reviews Genetics* **6** (2005), 95–108.
17. M I McCarthy, G R Abecasis, L R Cardon, D B Goldstein, J Little, J P A Ioannidis, and J N Hirschhorn. 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges'. *Nature Reviews Genetics* **9** (2008), 356–369.
18. B D Strahl and C D Allis. 'The language of covalent histone modifications'. *Nature* **403** (2000), 41–45.
19. T Jenuwein and C D Allis. 'Translating the histone code'. *Science* **293**.5532 (2001), 1074–1080.
20. F Antequera. 'Structure, function and evolution of CpG island promoters'. *Cellular and Molecular Life Sciences* **60.8** (2003), 1647–1658.
21. A P Bird and A P Wolffe. 'Methylation-induced repression – belts, braces, and chromatin'. *Cell* **99.5** (1999), 451–454.
22. R J Klose and A P Bird. 'Genomic DNA methylation: the mark and its mediators'. *Trends in Biochemical Sciences* **31.2** (2006), 89–97.
23. S J Clark, J Harrison, C L Paul, and M Frommer. 'High sensitivity mapping of methylated cytosines'. *Nucleic Acids Research* **22.15** (1994), 2990–2997.
24. P W Laird. 'Principles and challenges of genome-wide DNA methylation analysis'. *Nature Reviews Genetics* **11** (2010), 191–203.

25. C Bock, E M Tomazou, A B Brinkman, F Müller, F Simmer, H Gu, N Jäger, A Gnirke, H G Stunnenberg, and A Meissner. 'Quantitative comparison of genome-wide DNA methylation mapping technologies'. *Nature Biotechnology* **28** (2010), 1106–1114.
26. M Bibikova, B Barnes, C Tsan, V Ho, B Klotzle, J M Le, D Delano, L Zhang, G P Schroth, K L Gunderson, J-B Fan, and R Shen. 'High density DNA methylation array with single CpG site resolution'. *Genomics* **98.4** (2011), 288–295.
27. M B Eisen, P T Spellman, P O Brown, and D Botstein. 'Cluster analysis and display of genome-wide expression patterns'. *Proceedings of the National Academy of Sciences* **95.25** (1998), 14863–14868.
28. Z Wang, M Gerstein, and M Snyder. 'RNA-seq: a revolutionary tool for transcriptomics'. *Nature Reviews Genetics* **10** (2009), 57–63.
29. R Aebersold and M Mann. 'Mass spectrometry-based proteomics'. *Nature* **422** (2003), 198–207.
30. T R Burkard, M Planyavsky, I Kaupe, F P Breitwieser, T Bürckstümmer, K L Bennett, G Superti-Furga, and J Colinge. 'Initial characterization of the human central proteome'. *BMC Systems Biology* **5** (2011), 17.
31. J K Nicholson, J C Lindon, and E Holmes. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data'. *Xenobiotica* **29.11** (1999), 1181–1189.
32. R Goodacre, S Vaidyanathan, W B Dunn, G G Harrigan, and D B Kell. 'Metabolomics by numbers: acquiring and understanding global metabolite data'. *Trends in Biotechnology* **22.5** (2004), 245–252.
33. W B Dunn, N J C Bailey, and H E Johnson. 'Measuring the metabolome: current analytical technologies'. *Analyst* **130.5** (2005), 606–625.

34. K Dettmer, P A Aronov, and B D Hammock. 'Mass spectrometry-based metabolomics'. *Mass Spectrometry Reviews* **26.1** (2007), 51–78.
35. A M Evans, C D DeHaven, T Barrett, M Mitchell, and E Milgram. 'Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems'. *Analytical Chemistry* **81.16** (2009), 6656–6667.
36. A R Whitney, M Diehn, S J Popper, A A Alizadeh, J C Boldrick, D A Relman, and P O Brown. 'Individuality and variation in gene expression patterns in human blood'. *Proceedings of the National Academy of Sciences* **100.4** (2003), 1896–1901.
37. A E Jaffe and R A Irizarry. 'Accounting for cellular heterogeneity is critical in epigenome-wide association studies'. *Genome Biology* **15.2** (2014), R31.
38. S M Reppert and D R Weaver. 'Coordination of circadian timing in mammals'. *Nature* **418** (2002), 935–941.
39. K-F Storch, O Lipan, I Leykin, N Viswanathan, F C Davis, W H Wong, and C J Weitz. 'Extensive and divergent circadian gene expression in liver and heart'. *Nature* **417** (2002), 78–83.
40. F J Steemers, W Chang, G Lee, D L Barker, R Shen, and K L Gunderson. 'Whole-genome genotyping with the single-base extension assay'. *Nature Methods* **3** (2006), 31–33.
41. K L Michael, L C Taylor, S L Schultz, and D R Walt. 'Randomly ordered addressable high-density optical sensor arrays'. *Analytical Chemistry* **70.7** (1998), 1242–1248.

42. Y Chen, S Choufani, J Carlos Ferreira, D Grafodatskaya, D T Butcher, and R Weksberg. 'Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray'. *Genomics* **97.4** (2011), 214–222.
43. Y Chen, M Lemire, S Choufani, D T Butcher, D Grafodatskaya, B W Zanke, S Gallinger, T J Hudson, and R Weksberg. 'Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray'. *Epigenetics* **8.2** (2013), 203–209.
44. M E Price, A Cotton, L Lam, P Farre, E Emberly, C Brown, W Robinson, and M Kobor. 'Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array'. *Epigenetics & Chromatin* **6.1** (2013), 4.
45. C Wu, R Carta, and L Zhang. 'Sequence dependence of cross-hybridization on short oligo microarrays'. *Nucleic Acids Research* **33.9** (2005), e84.
46. K L Gunderson, S Kruglyak, M S Graige, F Garcia, B G Kermani, C Zhao, D Che, T Dickinson, E Wickham, J Bierle, D Doucet, M Milewski, R Yang, C Siegmund, J Haas, L Zhou, A Oliphant, J-B Fan, S Barnard, and M S Chee. 'Decoding randomly ordered DNA arrays'. *Genome Research* **14.5** (2004), 870–877.
47. D Pinkel, T Straume, and J W Gray. 'Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization'. *Proceedings of the National Academy of Sciences* **83.9** (1986), 2934–2938.
48. S Dedeurwaerder, M Defrance, E Calonne, H Denis, C Sotiriou, and F Fuks. 'Evaluation of the Infinium Methylation 450K technology'. *Epigenomics* **3.6** (2011), 771–784.

49. J Maksimovic, L Gordon, and A Oshlack. 'SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 Bead-Chips'. *Genome Biology* **13.6** (2012), R44.
50. A E Teschendorff, F Marabita, M Lechner, T Bartlett, J Tegner, D Gomez-Cabrero, and S Beck. 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data'. *Bioinformatics* **29.2** (2013), 189–196.
51. Y Liu, M J Aryee, L Padyukov, M D Fallin, E Hesselberg, A Runarsson, L Reinius, N Acevedo, M Taub, M Ronninger, K Shchetynsky, A Scheynius, J Kere, L Alfredsson, L Klareskog, T J Ekström, and A P Feinberg. 'Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis'. *Nature Biotechnology* **31** (2013), 142–147.
52. T J Triche Jr, D J Weisenberger, D Van Den Berg, P W Laird, and K D Siegmund. 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays'. *Nucleic Acids Research* **41.7** (2013), e90.
53. P Du, X Zhang, C-C Huang, N Jafari, W A Kibbe, L Hou, and S M Lin. 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis'. *BMC Bioinformatics* **11** (2010), 587.
54. I M Johnstone and D M Titterton. 'Statistical challenges of high-dimensional data'. *Philosophical Transactions of the Royal Society A* **367.1906** (2002), 4237–4253.
55. J A Nelder and R W M Wedderburn. 'Generalized linear models'. *Journal of the Royal Statistical Society, Series A* **135.3** (1972), 370–384.
56. T J Hastie and R J Tibshirani. 'Generalized additive models'. *Statistical Science* **1.3** (1986), 297–310.

57. T J Hastie and R J Tibshirani. *Generalized additive models*. Vol. 43. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 1990.
58. T W Yee and C J Wild. 'Vector generalized additive models'. *Journal of the Royal Statistical Society, Series B* **58.3** (1996), 481–493.
59. J Johnston and J DiNardo. *Econometric methods*. 4th ed. McGraw-Hill, 1997.
60. R B Millar. *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*. John Wiley & Sons, Inc., 2011.
61. Y Benjamini and Y Hochberg. 'Controlling the false discovery rate: a practical and powerful approach to multiple testing'. *Journal of the Royal Statistical Society, Series B* **57.1** (1995), 289–300.
62. J D Storey. 'A direct approach to false discovery rates'. *Journal of the Royal Statistical Society, Series B* **64.3** (2002), 479–498.
63. J D Storey. 'The positive false discovery rate: a Bayesian interpretation and the q -value'. *The Annals of Statistics* **31.6** (2003), 2013–2035.
64. Y Benjamini. 'Discovering the false discovery rate'. *Journal of the Royal Statistical Society, Series B* **72.4** (2010), 405–416.
65. P H Westfall and S S Young. *Resampling-based multiple testing: examples and methods for p -value adjustment*. John Wiley & Sons, Inc., 1993.
66. S Wold, H Martens, and H Wold. 'The multivariate calibration problem in chemistry solved by the PLS method'. *Matrix pencils*. Ed. by B Kågström and A Ruhe. **973**. Lecture Notes in Mathematics. Springer, 1983, 286–293.
67. M Andersson. 'A comparison of nine PLS1 algorithms'. *Journal of Chemometrics* **23.10** (2009), 518–529.
68. J Trygg and S Wold. 'Orthogonal projections to latent structures (O-PLS)'. *Journal of Chemometrics* **16.3** (2002), 119–128.

69. O Svensson, T Kourti, and J F MacGregor. 'An investigation of orthogonal signal correction algorithms and their characteristics'. *Journal of Chemometrics* **16.4** (2002), 176–188.
70. M Bylesjö, D Eriksson, A Sjödin, S Jansson, T Moritz, and J Trygg. 'Orthogonal projections to latent structures as a strategy for microarray data normalization'. *BMC Bioinformatics* **8.1** (2007), 207.
71. J Trygg and S Wold. 'O2-PLS, a two-block (X – Y) latent variable regression (LVR) method with an integral OSC filter'. *Journal of Chemometrics* **17.1** (2003), 53–64.
72. T Löfstedt and J Trygg. 'OnPLS – a novel multiblock method for the modelling of predictive and orthogonal variation'. *Journal of Chemometrics* **25.8** (2011), 441–455.
73. T Löfstedt, L Eriksson, G Wormbs, and J Trygg. 'Bi-modal OnPLS'. *Journal of Chemometrics* **26.6** (2012), 236–245.
74. I T Jolliffe, N T Trendafilov, and M Uddin. 'A modified principal component technique based on the LASSO'. *Journal of Computational and Graphical Statistics* **12.3** (2003), 531–547.
75. H Zou, T Hastie, and R Tibshirani. 'Sparse principal component analysis'. *Journal of Computational and Graphical Statistics* **15.2** (2006), 265–286.
76. H Chun and S Keleş. 'Sparse partial least squares regression for simultaneous dimension reduction and variable selection'. *Journal of the Royal Statistical Society: Series B* **72.1** (2010), 3–25.

Algorithm 1.2: Univariate partial least squares (PLS1). This procedure was originally introduced by Wold et al. [66]; alternative approaches have been reviewed by [67].

Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, $k \in \{1, \dots, n\}$

Output: $\hat{\mathbf{X}} = \mathbf{T} \mathbf{P}^\top$, $\hat{\mathbf{y}} = \mathbf{T} \mathbf{c}^\top$, $\hat{\boldsymbol{\beta}}$

$\mathbf{X}_1 \leftarrow \mathbf{X}$

$\mathbf{y}_1 \leftarrow \mathbf{y}$

for $i = 1, \dots, k$ **do**

$\mathbf{w}_i \leftarrow \mathbf{X}_i^\top \mathbf{y}_i / \|\mathbf{X}_i^\top \mathbf{y}_i\|$

$\mathbf{t}_i \leftarrow \mathbf{X}_i \mathbf{w}_i$

$\mathbf{p}_i \leftarrow \mathbf{X}_i^\top \mathbf{t}_i / \mathbf{t}_i^\top \mathbf{t}_i$

$c_i \leftarrow \mathbf{t}_i^\top \mathbf{y}_i / \mathbf{t}_i^\top \mathbf{t}_i$

$\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i - \mathbf{t}_i \mathbf{p}_i^\top$

$\mathbf{y}_{i+1} \leftarrow \mathbf{y}_i - c_i \mathbf{t}_i$

$\mathbf{W} \leftarrow (\mathbf{w}_1 | \dots | \mathbf{w}_k)$

$\mathbf{T} \leftarrow (\mathbf{t}_1 | \dots | \mathbf{t}_k)$

$\mathbf{P} \leftarrow (\mathbf{p}_1 | \dots | \mathbf{p}_k)$

$\mathbf{c} \leftarrow (c_1, \dots, c_k)$

$\hat{\boldsymbol{\beta}} \leftarrow \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1} \mathbf{c}$

Algorithm 1.3: Multivariate partial least squares (PLS2). This procedure was originally introduced by Wold et al. [66].

Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$, $k \in \{1, \dots, n\}$

Output: $\hat{\mathbf{X}} = \mathbf{T} \mathbf{P}^\top$, $\hat{\mathbf{Y}} = \mathbf{U} \mathbf{Q}^\top = (\mathbf{T} \mathbf{C}) \mathbf{Q}^\top$, $\hat{\mathbf{B}}$

$\mathbf{X}_1 \leftarrow \mathbf{X}$

$\mathbf{Y}_1 \leftarrow \mathbf{Y}$

for $i = 1, \dots, k$ **do**

$\mathbf{u}_i \leftarrow$ arbitrary column of $\mathbf{Y}_i$

repeat

$\mathbf{u}_i^* \leftarrow \mathbf{u}$

$\mathbf{w}_i \leftarrow \mathbf{X}_i^\top \mathbf{u}_i / \|\mathbf{X}_i^\top \mathbf{u}_i\|$

$\mathbf{t}_i \leftarrow \mathbf{X}_i \mathbf{w}_i$

$\mathbf{q}_i \leftarrow \mathbf{Y}_i^\top \mathbf{t}_i / \|\mathbf{Y}_i^\top \mathbf{t}_i\|$

$\mathbf{u}_i \leftarrow \mathbf{Y}_i \mathbf{q}_i$

until $\|\mathbf{u}_i - \mathbf{u}_i^*\| < \epsilon$

$\mathbf{p}_i \leftarrow \mathbf{X}_i^\top \mathbf{t}_i / \mathbf{t}_i^\top \mathbf{t}_i$

$c_i \leftarrow \mathbf{t}_i^\top \mathbf{u}_i / \mathbf{t}_i^\top \mathbf{t}_i$

$\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i - \mathbf{t}_i \mathbf{p}_i^\top$

$\mathbf{Y}_{i+1} \leftarrow \mathbf{Y}_i - c_i \mathbf{t}_i \mathbf{q}_i^\top$

$\mathbf{U} \leftarrow (\mathbf{u}_1 | \dots | \mathbf{u}_k)$

$\mathbf{W} \leftarrow (\mathbf{w}_1 | \dots | \mathbf{w}_k)$

$\mathbf{T} \leftarrow (\mathbf{t}_1 | \dots | \mathbf{t}_k)$

$\mathbf{Q} \leftarrow (\mathbf{q}_1 | \dots | \mathbf{q}_k)$

$\mathbf{P} \leftarrow (\mathbf{p}_1 | \dots | \mathbf{p}_k)$

$\mathbf{C} \leftarrow \text{diag}(c_1, \dots, c_k)$

$\hat{\mathbf{B}} \leftarrow \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1} \mathbf{C} \mathbf{Q}^\top$

Chapter 2

Analytical strategies for DNA methylation data

While genome-wide DNA methylation data generated using Infinium[®] assays are becoming increasingly available, no formal assessments of pre-processing methods or modelling strategies have been carried out to date.

This chapter endeavours to provide reasoned comparisons of different approaches based on sound statistical frameworks; it is organised as follows. The first section of section 2.1 reviews three three comparisons of pre-processing methods that have been attempted in the literature, and points out their shortcomings. Section 2.1.2 presents two novel pre-processing methods with desirable compu-

tational properties. Section 2.1.3 proposes a new evaluation framework based on variance partitioning in the presence of technical replicates; this is then employed, in section 2.1.4, to appraise 15 different pre-processing methods including those introduced in section 2.1.2. Section 2.2 compares six different modeling choices for DNA methylation data by means of a simulation study. Finally, results from the comparisons of sections 2.1 and 2.2.2 are condensed in the recommendations of section 2.3; some limitations of the analyses and conclusions presented herein are also discussed in section 2.4.

§ 2.1 Pre-processing methods

As outlined in section 1.2.2, many pre-processing methods have been proposed in recent years for DNA methylation data generated using the Illumina Infinium® HumanMethylation450 BeadChip, and a few comparisons have been attempted.

2.1.1 Literature review

Marabita et al. [1] considered four datasets of DNA methylation profiles that had been obtained from:

- (i) peripheral blood leukocytes ($N = 50$) or CD14+ monocytes sorted from peripheral blood ($N = 36$), including eight samples analysed in duplicate, and one sample analysed in triplicate;
- (ii) paired muscle biopsies ($N = 2 \times 17$) before and after a three-month period of supervised exercise training;
- (iii) peripheral blood leukocytes or CD14+ monocytes (paired subset of data from Reinius et al. [2], $N = 2 \times 17$);

- (iv) cell lines ($N = 6$) with matched DNA methylation measurements obtained by bisulphite pyrosequencing for 15 CpG loci (from Dedeurwaerder et al. [3]).

They evaluated six pipelines based on the following criteria:

- (i) reduction of the ‘bias’ between Infinium[®] I and Infinium[®] II assays, which was assessed by comparing DNA methylation levels at pairs of CpG loci less than 200 bp away discordant for assay design;
- (ii) reduction of the variance between technical replicates, which was evaluated by computing average absolute differences between technical replicates;
- (iii) ability to identify differentially methylated CpG loci when comparing DNA methylation levels in peripheral blood leukocytes to those in CD14+ monocytes;
- (iv) comparison of DNA methylation levels measured using HM450 BeadChips with matched bisulphite pyrosequencing measurements.

The first criterion rests on the idea that DNA methylation levels at adjacent CpG loci are generally correlated, and their measurement should not be affected by assay design. It was initially proposed by Teschendorff et al. [4] to evaluate their ‘bias’ reduction method, and unsurprisingly Marabita et al. [1] found that the ‘optimal’ pipeline included this method. The second criterion encourages a distorted view, not only because of the limited sample size, but more importantly because it does not take into account between-sample variance. The trivial method that averages DNA methylation levels at each CpG locus across samples minimises the variance between technical replicates, but it is hardly a viable method. In the absence of a reference to define ‘true’ findings, evaluation of the third criterion between different pre-processing methods is inconsequential. Finally, the

fourth criterion is limited in terms of generalisability not only because of the small sample size, but also because of the restricted number of CpG loci considered.

Pidsley et al. [5] considered 11 unpublished datasets of DNA methylation profiles, for a total of 696 samples with no technical replicates. They evaluated the performance of 14 pre-processing methods using three metrics based on:

- (i) DNA methylation levels at CpG loci located in imprinted regions, which are characterised by allele-specific parent-of-origin DNA methylation;
- (ii) DNA methylation levels at CpG loci located on the X chromosome, which differ significantly between men and women;
- (iii) separability of genotypes at the 65 highly-polymorphic SNPs included on the HM450 BeadChip.

The first two metrics are based on biological phenomena, namely genomic imprinting and X chromosome inactivation. Pidsley et al. [5] observe that DNA methylation levels at imprinted regions have an expectation of 50%, and define their first metric as the standard deviation of the full set of DNA methylation levels at 237 assayed CpG loci located in imprinted regions over the square root of the number of samples. Once again, the trivial method that averages DNA methylation levels at each CpG locus across samples minimises this metric, and would thus be considered ‘optimal’. Their second metric is based on differences among sexes in DNA methylation levels at CpG loci located almost exclusively on the X chromosome. In particular, they use p -values obtained from a genome-wide analysis of sex differences as predictors of whether a CpG locus is located on the X chromosome, and estimate their performance by computing the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Finally, their third metric is based on separability of genotypes. Instead of considering points in the $(\mathbf{A}, \mathbf{B})$ or in the $(\mathbf{\Theta}, \mathbf{R})$ plane of eq. (1.11), Pidsley et al. [5] use ratios

of intensities as in eq. (1.9), and partition observations into three groups (corresponding to the degree of homozygosity) using the k -means algorithm. The metric is then derived by averaging quantities directly related to the within-group sum of squares and to group sizes. While the authors claim that this metric could be used as a general quality score for samples, it is inexpedient to generalise results from 65 genotyping assays to the far larger number of DNA methylation assays on the micro-array.

Once the three metrics have been calculated for each dataset and each pre-processing method, the latter are evaluated by the average rank across metrics and across datasets; this process is repeated for Infinium[®] I and Infinium[®] II assays separately, and ranks are again averaged across assay designs. A first justifiable assumption underlying this ranking system is that the three metrics should be assigned equal weight. The system also assigns equal weight to each dataset, which is debatable given the different sample sizes and sex ratios. Nevertheless, the main drawback and source of bias of this ranking system is the lack of proper weighting for Infinium[®] I and Infinium[®] II assays, which are assigned equal weight despite not being equally represented on the micro-array. Since the proportion of Infinium[®] I assays is approximately 28% (table 1.1), methods that perform better for this subset of assays are favoured, and correspondingly those that perform better for Infinium[®] II assays are put at a disadvantage.

More recently, Wu et al. [6] evaluated four pre-processing methods on a large dataset consisting of 1,204 DNA methylation profiles including 13 samples analysed in duplicate, and two samples analysed 13 times. They assessed performances based on two metrics computed between pairs of technical replicates: Pearson's correlation coefficient, and the 99th percentile of absolute differences. Differently from Marabita et al. [1] and Pidsley et al. [5], Wu et al. [6] acknowledge the dangers of over-normalisation, and examine the same metrics within non-

replicate samples by random pairing; however, they consider only one such pairing, and fail to see the opportunity of approximating the null distribution of the two metrics, e.g. by considering all possible pairings, or a random subset thereof.

2.1.2 Novel methods

Before introducing and applying the evaluation framework in the next section, two novel methods for dye bias correction are presented. These use information from the normalisation control beads, which consists of 93 pairs of intensities in the two colours, $\{(g_i, r_i), i = 1, \dots, 93\}$, as outlined in section 1.2.2.

Both methods can be applied to each sample independently. This greatly improves computational efficiency and reduces memory requirements, while at the same time ensuring procedural reproducibility, which is paramount when pooling data from multiple studies, e.g. in meta-analyses.

First method

The first method is similar to the ‘normalisation’ option implemented by Illumina, Inc. in the data analysis software GenomeStudio, but computes two multiplicative correction factors independently for each sample rather than with respect to an arbitrary reference. These are defined as follows,

$$k/\bar{g} \quad \text{for green intensities,} \quad (2.1)$$

$$k/\bar{r} \quad \text{for red intensities,} \quad (2.2)$$

where $\bar{g} = \sum_i g_i/93$ and $\bar{r} = \sum_i r_i/93$ are the mean intensities from normalisation control beads in green and red, respectively, and $k = (\bar{g} + \bar{r})/2$. The correction is

applied as follows,

$$\tilde{\mathbf{A}}_{\text{II}} = k/\bar{r} \mathbf{A}_{\text{II}}, \quad (2.3)$$

$$\tilde{\mathbf{B}}_{\text{II}} = k/\bar{g} \mathbf{B}_{\text{II}}, \quad (2.4)$$

where $\mathbf{A}_{\text{II}}$ and $\mathbf{B}_{\text{II}}$ are the matrices of intensities $\mathbf{A}$ and $\mathbf{B}$ restricted to Infinium[®] II assays, and matrix operations are element-wise.

Second method

The second method fits the linear model

$$\log g_i = \beta_0 + \beta_R \log r_i + \epsilon_i, \quad (2.5)$$

where β_0 and β_R are parameters to be estimated, and the ϵ_i are i.i.d. error terms. Once β_0 and β_R have been estimated for a given sample, intensities in red are ‘greenified’ as follows,

$$\tilde{\mathbf{A}}_{\text{II}} = e^{\beta_0} \mathbf{A}_{\text{II}}^{\beta_R}, \quad (2.6)$$

where $\mathbf{A}_{\text{II}}$ is as above, and matrix operations are again element-wise.

The effect of both corrections on intensities in green and red, as well as the resulting estimated DNA methylation levels, are shown in fig. 2.1. In $\log_{10}$ scale, the first method shifts the distributions of green and red intensities, whereas the second method (which acts on red intensities only) additionally scales them.

2.1.3 Evaluation framework

The efforts presented in section 2.1.1 have mostly been thwarted by lack of available data including technical replicates. Furthermore, they have often resorted to metrics that do not take into account between-sample variance, leading to erroneous conclusions.

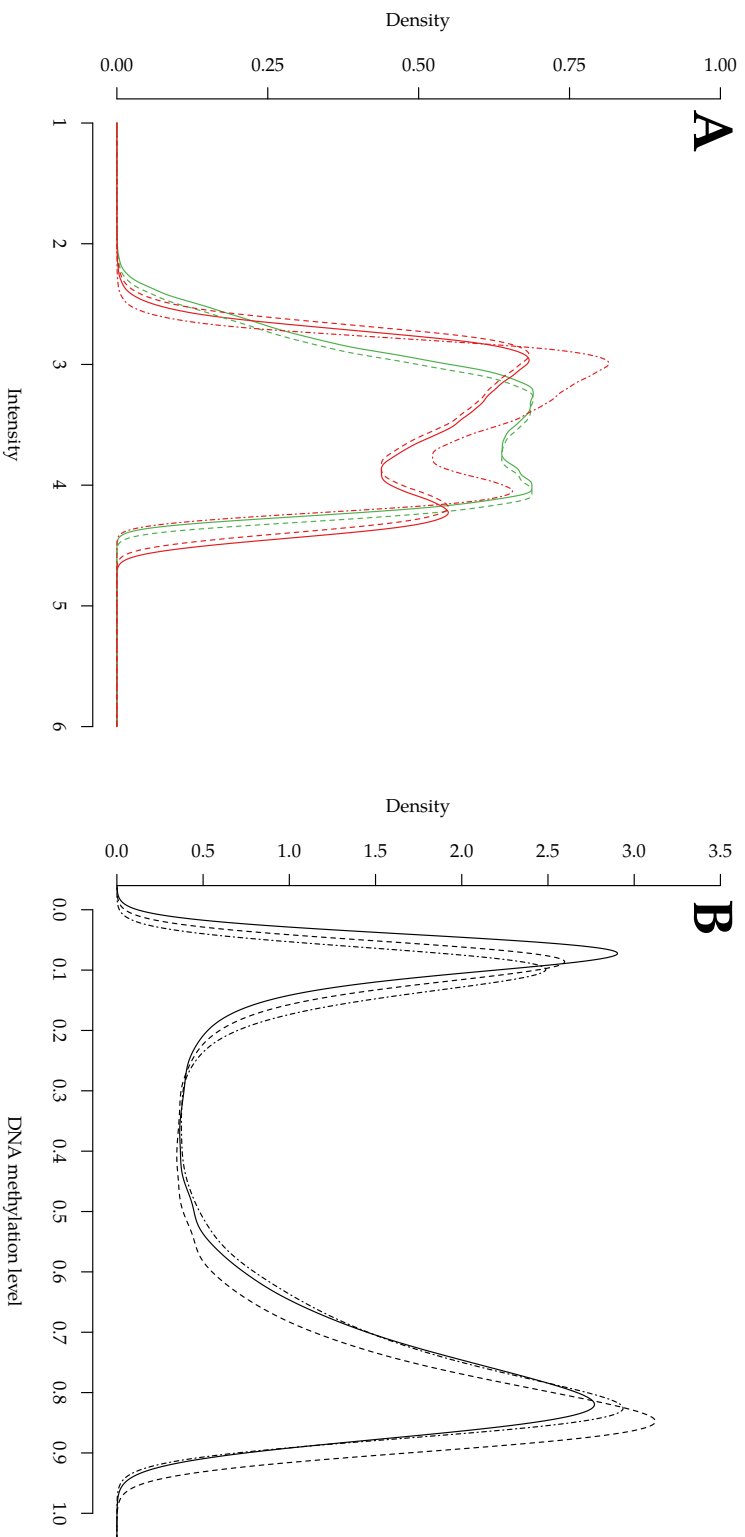


Figure 2.1: Novel methods for dye bias correction. (A) Kernel density estimates (log₁₀ scale) of intensities from Infinium® II beads in the two colours before dye bias correction (solid lines), using the first method (dashed lines), or using the second method (dash-dotted line); (B) kernel density estimates of the resulting DNA methylation levels computed according to eq. (1.9), setting $\epsilon = 0$. Illustrated on publicly available data generated by Liu et al. [7].

The author proposes the following linear mixed model to evaluate pre-processing methods. For each DNA methylation measurement $\mathbf{y}$, the following model (as in eq. (1.17)) is estimated,

$$\mathbf{y} = \beta_0 + \mathbf{u}_{\text{technical}} + \mathbf{u}_{\text{within-replicate}} + \boldsymbol{\epsilon}, \quad (2.7)$$

where β_0 is the intercept term corresponding to the mean DNA methylation level, $\mathbf{u}_{\text{technical}}$ and $\mathbf{u}_{\text{within-replicate}}$ are random intercepts corresponding to technical and within-replicate variation, respectively, and $\boldsymbol{\epsilon}$ is a random vector of i.i.d. error terms that represents residual variation. Equation (2.7) effectively partitions the variance of $\mathbf{y}$ into three components:

- (i) the technical variance component corresponds to known sources of technically-induced nuisance variation, and is typically defined in terms of batches of samples that have undergone laboratory procedures (partly or completely) as a group;
- (ii) the within-replicate variance component corresponds to technically-induced nuisance variation that is left unaccounted for by the technical component, and can thus be interpreted as ‘intrinsic’ to the measurement process;
- (iii) the residual variance component corresponds to non-technically-induced variation, including biologically-relevant variation.

Equation (2.7) also explicitly incorporates the two competing aims that pre-processing methods ought to achieve: on the one hand, nuisance variation (technical and within-replicate variance) should be minimised; on the other hand, biologically-relevant variation (represented here, at least partly, by the residual variance) should be preserved as much as possible. Because of the inherent trade-off between these objectives, eq. (2.7) provides the ideal framework in which to evaluate them simultaneously.

	EPIC-Italy	EpiMigrant
Samples	1,217	3,187
No replicates	1,131	3,115
Two replicates	40	36
Three replicates	2	—
Assays	483,503	483,802
CpG loci	480,349	480,648
CpH loci	3,089	3,089
SNPs	65	65
Micro-arrays	103	280
Positions	12	12

Table 2.1: Datasets used in the evaluation of pre-processing methods. The number of assays is computed after quality control procedures and across pre-processing methods.

2.1.4 Comparison of pre-processing methods

Two datasets of DNA methylation profiles in peripheral blood leukocytes were considered. The first dataset was generated as part of a series of case-control studies nested within the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy) [8]. It comprises 1,217 samples including more than 40 pairs of ‘accidental’ technical replicates whose identity was confirmed using the 65 highly-polymorphic SNPs present on the HM450 Bead-Chip. The second dataset was generated by the EpiMigrant project. It comprises 3,187 samples including 36 pairs of technical replicates with confirmed identity. Information on the two studies is summarised in table 2.1.

For each dataset separately, a reference set of DNA methylation measurements was created by loading idat files generated by GenomeStudio, censoring values below detection limits estimated from ‘negative’ control beads, and performing background subtraction as described in section 1.2.2. A total of 15 pre-processing methods were considered in addition to the reference set. Two sets were produced using the two dye bias correction methods introduced in section 1.2.2. The remaining 13 sets were generated using the R package watermelon that implements all 14 pre-processing methods considered by Pidsley et al. [5]; of these, method tost systematically failed to run, and was thus excluded from further consideration.

First, DNA methylation measurements obtained using the different pre-processing methods were compared against those in the reference set by computing Spearman’s rank correlation coefficients across assays for each method. These correlation coefficients cannot be used for evaluation purposes, but give an indication of how ‘conservative’ each pre-processing method is with respect to the reference (i.e. to lack of pre-processing). Summary statistics are reported in tables 2.2 and 2.3 for EPIC-Italy and EpiMigrant data, respectively, and depicted in fig. 2.2 for both datasets. In general, most pre-processing methods exhibited very high correlations with the reference set; this was more pronounced using EpiMigrant data. DNA methylation values obtained using the pre-processing method swan correlated very poorly with those in the reference set. Upon further inspection of the data, it became clear that the implementation of this method was not viable, and it was thus excluded from further analyses.

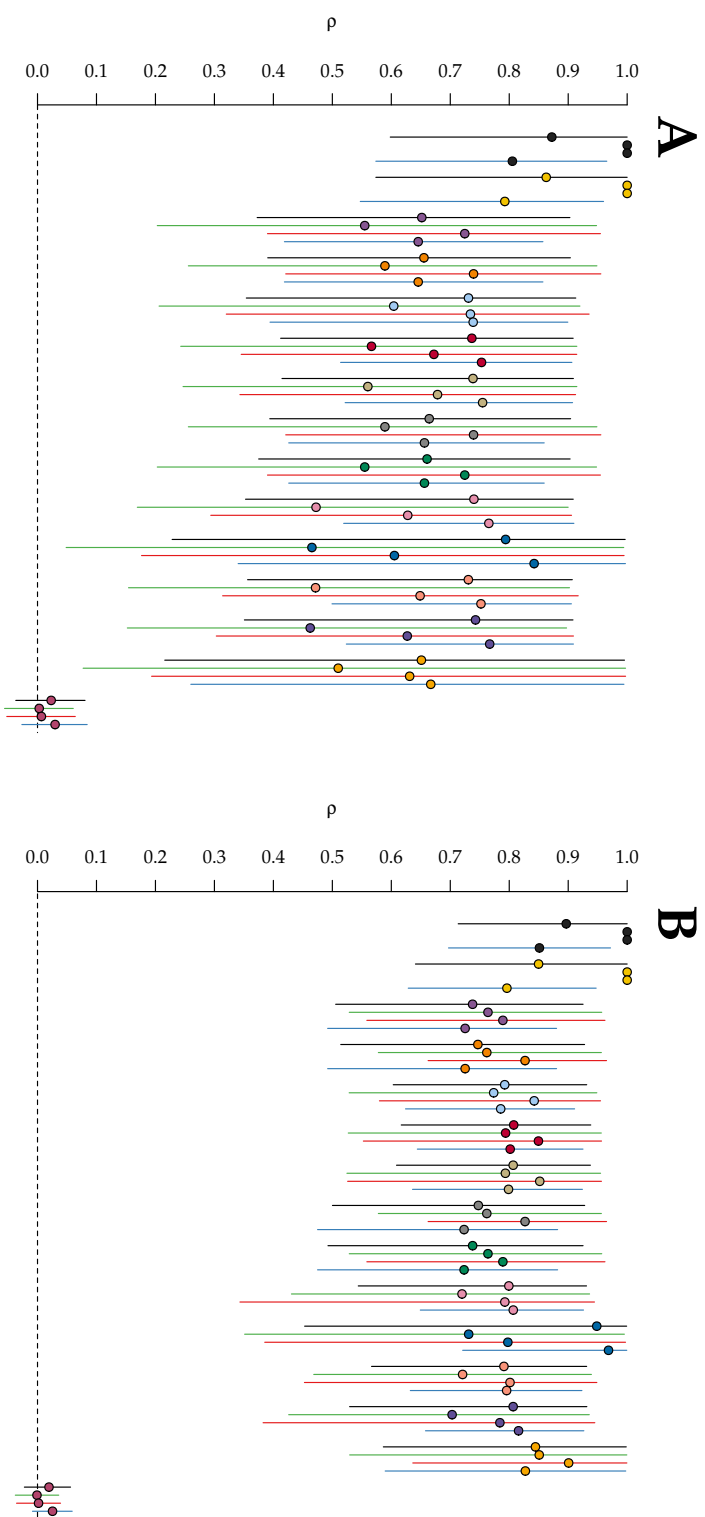
Second, eq. (2.7) was estimated for each DNA methylation measurement and each pre-processing method separately for EPIC-Italy and EpiMigrant data. The estimated technical, within-replicate, and residual variances were stored and subsequently normalised to unit sum to represent contributions of each component

Method	All assays		Infinium® I in green		Infinium® I in red		Infinium® II	
	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile
● Method 1	0.87	0.60 to 1.00	1.00	1.00 to 1.00	1.00	1.00 to 1.00	0.81	0.57 to 0.97
● Method 2	0.86	0.57 to 1.00	1.00	1.00 to 1.00	1.00	1.00 to 1.00	0.79	0.55 to 0.96
● dasen	0.65	0.37 to 0.90	0.55	0.20 to 0.95	0.72	0.39 to 0.95	0.65	0.42 to 0.86
● nasen	0.66	0.39 to 0.90	0.59	0.26 to 0.95	0.74	0.42 to 0.96	0.65	0.42 to 0.86
● betagn	0.73	0.35 to 0.91	0.60	0.21 to 0.92	0.73	0.32 to 0.94	0.74	0.39 to 0.90
● naten	0.74	0.41 to 0.91	0.57	0.24 to 0.91	0.67	0.35 to 0.91	0.75	0.51 to 0.91
● nanet	0.74	0.41 to 0.91	0.56	0.25 to 0.91	0.68	0.34 to 0.91	0.75	0.52 to 0.91
● nanes	0.66	0.39 to 0.90	0.59	0.26 to 0.95	0.74	0.42 to 0.96	0.66	0.43 to 0.86
● danes	0.66	0.38 to 0.90	0.55	0.20 to 0.95	0.72	0.39 to 0.95	0.66	0.43 to 0.86
● danet	0.74	0.35 to 0.91	0.47	0.17 to 0.90	0.63	0.29 to 0.91	0.77	0.52 to 0.91
● danen	0.79	0.23 to 1.00	0.47	0.05 to 0.99	0.61	0.18 to 0.99	0.84	0.34 to 1.00
● daten1	0.73	0.36 to 0.91	0.47	0.15 to 0.90	0.65	0.31 to 0.92	0.75	0.50 to 0.91
● daten2	0.74	0.35 to 0.91	0.46	0.15 to 0.90	0.63	0.30 to 0.91	0.77	0.52 to 0.91
● fuks	0.65	0.22 to 1.00	0.51	0.08 to 1.00	0.63	0.19 to 1.00	0.67	0.26 to 0.99
● swan	0.02	−0.04 to 0.08	0.00	−0.06 to 0.06	0.01	−0.05 to 0.06	0.03	−0.03 to 0.08

Table 2.2: Correlations among pre-processing methods using EPIC-Italy data. Summary statistics of Spearman’s rank correlation coefficients were computed separately for all assays, Infinium® I assays in green and red, and Infinium® II assays.

Method	All assays			Infinium® I in green			Infinium® I in red			Infinium® II		
	Median	5 th -to-95 th percentile		Median	5 th -to-95 th percentile		Median	5 th -to-95 th percentile		Median	5 th -to-95 th percentile	
● Method 1	0.90	0.71 to 1.00		1.00	1.00 to 1.00		1.00	1.00 to 1.00		0.85	0.70 to 0.97	
● Method 2	0.85	0.64 to 1.00		1.00	1.00 to 1.00		1.00	1.00 to 1.00		0.80	0.63 to 0.95	
● dasen	0.74	0.51 to 0.93		0.76	0.53 to 0.96		0.79	0.56 to 0.96		0.73	0.49 to 0.88	
● nasen	0.75	0.51 to 0.93		0.76	0.58 to 0.96		0.83	0.66 to 0.97		0.73	0.49 to 0.88	
● betaqn	0.79	0.60 to 0.93		0.77	0.53 to 0.95		0.84	0.58 to 0.95		0.79	0.62 to 0.91	
● naten	0.81	0.62 to 0.94		0.79	0.53 to 0.96		0.85	0.55 to 0.96		0.80	0.64 to 0.93	
● nanet	0.81	0.61 to 0.94		0.79	0.52 to 0.95		0.85	0.53 to 0.96		0.80	0.64 to 0.92	
● nanes	0.75	0.50 to 0.93		0.76	0.58 to 0.96		0.83	0.66 to 0.97		0.72	0.47 to 0.88	
● danes	0.74	0.49 to 0.93		0.76	0.53 to 0.96		0.79	0.56 to 0.96		0.72	0.47 to 0.88	
● danet	0.80	0.54 to 0.93		0.72	0.43 to 0.94		0.79	0.34 to 0.94		0.81	0.65 to 0.93	
● danen	0.95	0.45 to 1.00		0.73	0.35 to 1.00		0.80	0.39 to 1.00		0.97	0.72 to 1.00	
● daten1	0.79	0.57 to 0.93		0.72	0.47 to 0.94		0.80	0.45 to 0.95		0.80	0.63 to 0.92	
● daten2	0.81	0.53 to 0.93		0.70	0.43 to 0.94		0.78	0.38 to 0.95		0.82	0.66 to 0.93	
● fuks	0.84	0.59 to 1.00		0.85	0.53 to 1.00		0.90	0.64 to 1.00		0.83	0.59 to 1.00	
● swan	0.02	−0.02 to 0.06		−0.00	−0.04 to 0.04		0.00	−0.04 to 0.04		0.03	−0.01 to 0.06	

Table 2.3: Correlations among pre-processing methods using EpiMigrant data. Summary statistics of Spearman’s rank correlation coefficients were computed separately for all assays, Infinium® I assays in green and red, and Infinium® II assays.



to the total variance.

Distributions across assays were summarised in two complementary ways (separately for EPIC-Italy and EpiMigrant).

Averaging of each component across assays yields three values for each pre-processing method that can be interpreted as the average partitioning of the variance of DNA methylation measurements. Since arithmetic means preserve unit sums, this allows ternary representations as in figs. 2.3 and 2.4. It is evident from both plots that most methods yield only modest improvements compared to lack of pre-processing, which achieves a roughly equal split between the technical and within-replicate variance components. With the exception of a few pre-processing methods (fucs using EPIC-Italy data, and danen using both datasets), most methods achieve approximately the same value for the residual variance component (between 40% and 45%), and slightly favour reductions in the within-replicate (rather than the technical) variance component.

Nevertheless, it must be recognised that, owing to the presence of variance components estimated to be zero, arithmetic means are hardly informative. For this reason, the distributions of the different variance components were also summarised by:

- (i) the proportion of assays for which a given variance component was estimated to be zero (within machine precision);
- (ii) the median and 5th-to-95th percentile range restricted to positive variances (within machine precision);
- (iii) the ‘nuisance split’, which is defined as

$$\text{Nuisance split} = \frac{2}{\pi} \arctan \frac{\tilde{\sigma}_{\text{technical}}^2}{\tilde{\sigma}_{\text{within-replicate}}^2} \in (0, 1), \quad (2.8)$$

where $\tilde{\sigma}_{\text{technical}} > 0$ and $\tilde{\sigma}_{\text{within-replicate}} > 0$ are the technical and within-replicate variance components, respectively, and the value is normalised

by dividing by the maximum $\pi/2$.

Equation (2.8) is proportional to the angle from the top corners of figs. 2.3 and 2.4 to each point, and represents the propensity of each pre-processing method to favour reductions in the within-replicate (values close to zero) or technical (values close to one) variance components.

Summary statistics are reported in tables 2.4 and 2.5 for EPIC-Italy and Epi-Migrants data, respectively. The ‘polar’ representation of eq. (2.8) is depicted in fig. 2.5.

Given a fixed value for the residual variance component, it is desirable to achieve a roughly equal or greater reduction in the within-replicate rather than the technical component (corresponding to a nuisance split below 50%). This is because, as detailed above, the technical component corresponds to known sources of nuisance variation that can be further accounted for in statistical models (as described in section 1.3); on the other hand, the within-replicate component is ‘intrinsic’ to the measurement process, and cannot be modelled directly in the absence of technical replicates.

Interestingly, most pre-processing methods were found to favour reductions in the within-replicate variance component. Nevertheless, with the exception of the pre-processing methods *fuks* and *danen* that ought not be used, it is inexpedient to appoint an ‘optimal’ method based on variance partitioning. This gives way to other considerations such as computational efficiency (in terms of memory footprint and running time) and procedural reproducibility, since most methods process entire batches of samples simultaneously, and may produce slightly different results if data from multiple studies are pooled.

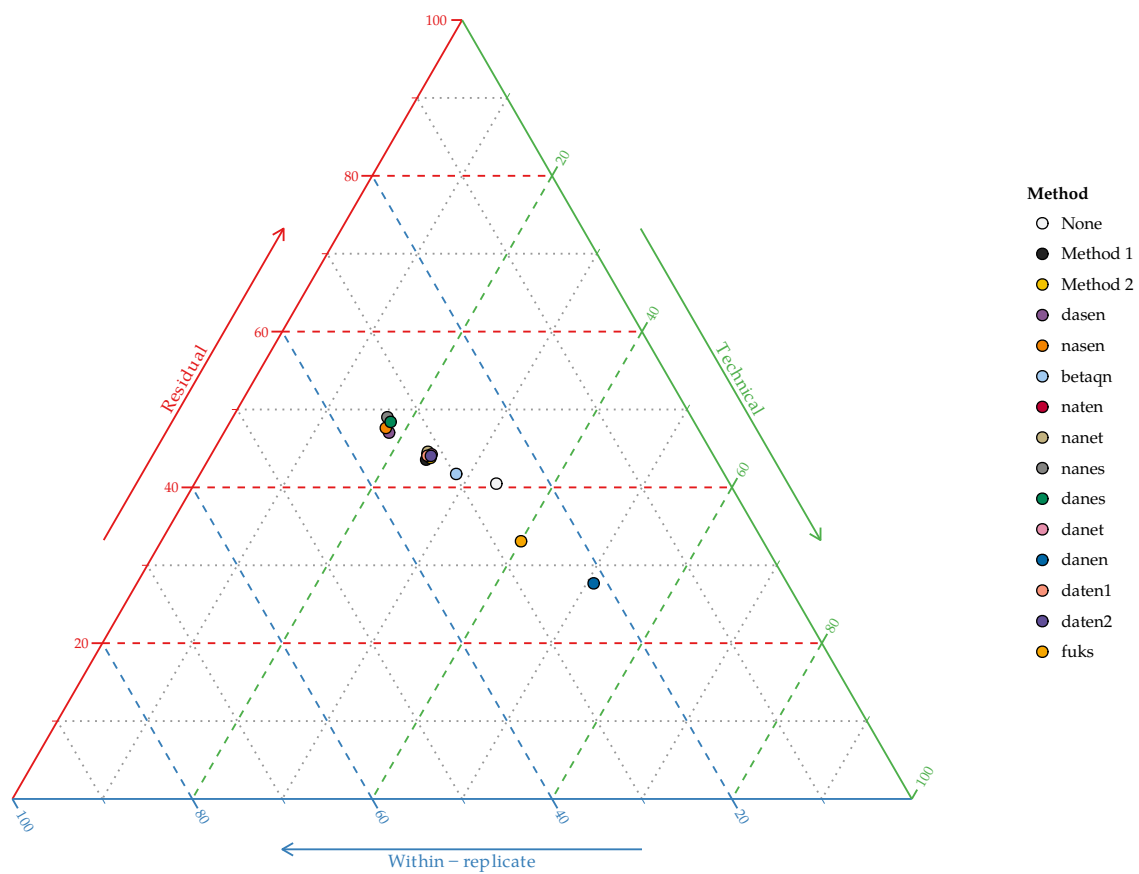


Figure 2.3: Ternary plot of variance components using EPIC-Italy data. Each point represents the average partitioning of the variance of DNA methylation measurements across all assays.

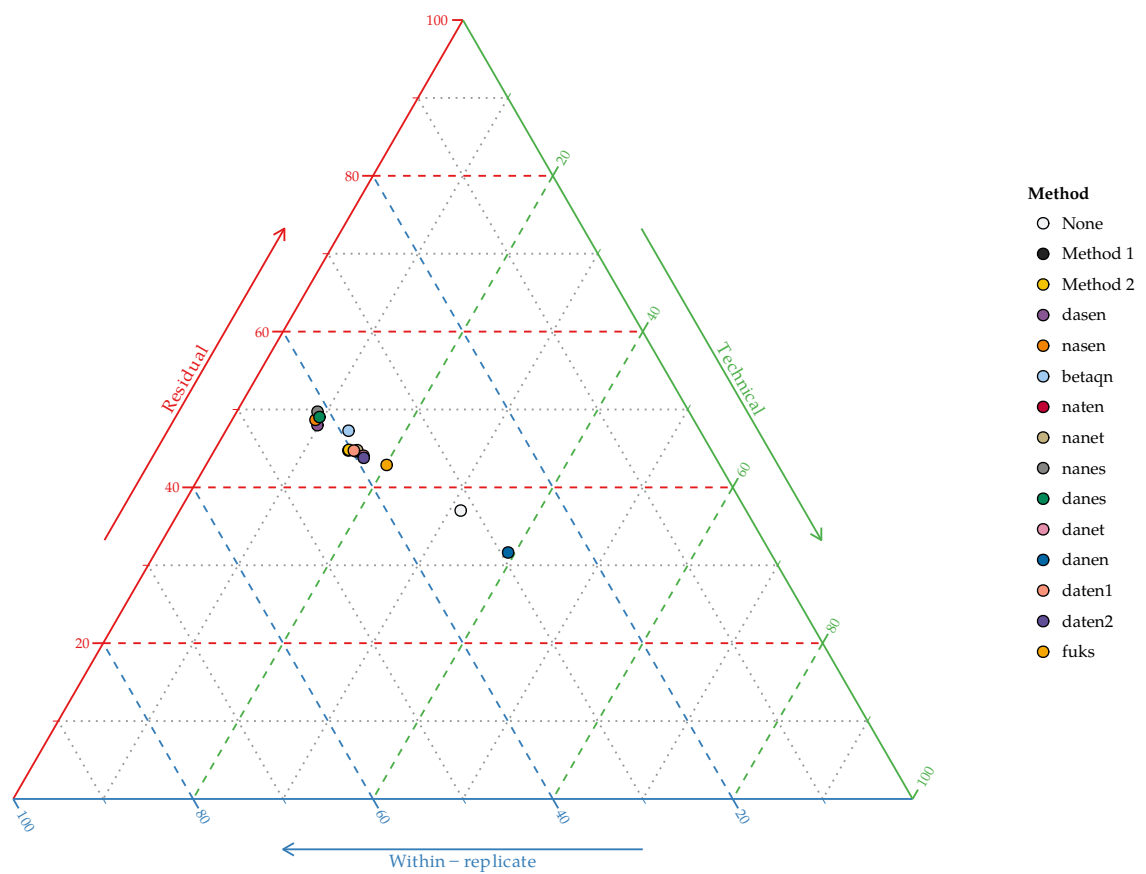


Figure 2.4: Ternary plot of variance components using EpiMigrant data. Each point represents the average partitioning of the variance of DNA methylation measurements across all assays.

Method	Within-replicate			Technical			Nuisance split			Residual	
	Null	Median	5 th -to-95 th percentile	Null	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile	Residual
○ None	17.78%	26.55%	1.53% to 79.67%	0.13%	31.76%	7.60% to 65.06%	57.48%	6.28% to 97.90%	37.47%	9.29% to 81.31%	
● Method 1	17.89%	35.23%	1.88% to 87.38%	0.18%	19.27%	4.23% to 59.29%	33.28%	2.91% to 96.93%	42.21%	7.52% to 84.76%	
● Method 2	17.98%	33.87%	1.72% to 87.29%	0.18%	20.13%	4.14% to 59.37%	35.25%	2.89% to 97.20%	42.67%	7.59% to 84.87%	
● dasen	13.04%	34.89%	2.86% to 88.02%	0.39%	14.39%	2.04% to 48.47%	25.96%	1.52% to 93.62%	47.01%	8.70% to 86.69%	
● nasen	13.15%	35.07%	2.91% to 88.05%	0.40%	13.96%	2.01% to 47.16%	25.35%	1.50% to 93.12%	47.74%	8.70% to 87.16%	
● betaqn	14.02%	28.07%	1.93% to 84.71%	0.12%	25.68%	3.92% to 62.08%	48.41%	2.92% to 97.19%	40.39%	9.44% to 79.17%	
● naten	13.26%	31.57%	2.37% to 84.50%	0.16%	20.36%	3.61% to 56.37%	37.83%	2.78% to 95.94%	43.31%	9.64% to 82.46%	
● nanet	13.34%	31.28%	2.32% to 84.25%	0.16%	20.26%	3.59% to 56.44%	37.96%	2.78% to 95.98%	43.74%	9.81% to 82.56%	
● nanes	13.63%	34.06%	2.65% to 87.56%	0.45%	13.31%	1.93% to 46.66%	25.00%	1.48% to 93.46%	49.69%	9.21% to 88.24%	
● danes	13.51%	33.89%	2.62% to 87.52%	0.44%	13.71%	1.96% to 48.07%	25.59%	1.50% to 93.94%	48.80%	9.21% to 87.89%	
● danet	13.21%	30.71%	2.20% to 84.19%	0.16%	20.49%	3.68% to 58.45%	38.87%	2.82% to 96.38%	43.00%	9.88% to 82.22%	
● danen	11.46%	15.75%	1.04% to 73.38%	0.01%	55.07%	13.73% to 72.28%	82.15%	11.08% to 98.97%	27.29%	9.91% to 47.30%	
● daten1	13.03%	31.43%	2.29% to 84.69%	0.18%	20.05%	3.38% to 58.92%	37.25%	2.63% to 96.27%	42.55%	9.58% to 82.85%	
● daten2	13.24%	31.00%	2.22% to 84.37%	0.17%	20.38%	3.63% to 59.08%	38.36%	2.81% to 96.36%	42.47%	9.72% to 82.77%	
● fuchs	15.04%	25.47%	1.21% to 80.41%	0.03%	39.73%	8.42% to 70.88%	62.02%	6.21% to 98.70%	31.88%	9.49% to 62.23%	

Table 2.4: Variance components estimated using EPIC-Italy data. ‘Null’ reports the proportion of assays for which the corresponding variance component was estimated to be zero (within machine precision).

Method	Within-replicate			Technical			Nuisance split		Residual	
	Null	Median	5 th -to-95 th percentile	Null	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile	Median	5 th -to-95 th percentile
○ None	15.13%	34.47%	0.00% to 85.58%	0.22%	29.59%	3.86% to 65.08%	48.65%	3.40% to 100.00%	31.32%	5.20% to 85.54%
● Method 1	15.13%	49.33%	0.67% to 91.79%	0.37%	11.03%	1.89% to 42.34%	14.89%	1.31% to 98.39%	42.60%	4.99% to 89.97%
● Method 2	14.47%	48.15%	0.56% to 91.70%	0.34%	11.23%	1.85% to 41.66%	15.43%	1.29% to 98.66%	43.16%	5.11% to 89.42%
● dasen	10.94%	46.46%	2.52% to 91.85%	0.50%	6.74%	1.05% to 29.71%	10.02%	0.87% to 88.67%	47.01%	4.87% to 94.57%
● nasen	11.27%	46.61%	2.54% to 91.86%	0.51%	6.12%	1.03% to 29.16%	9.57%	0.85% to 87.55%	47.60%	4.88% to 95.32%
● betagn	12.51%	42.31%	1.50% to 91.02%	0.21%	8.99%	1.65% to 44.28%	15.05%	1.29% to 95.39%	46.32%	5.09% to 93.58%
● naten	10.87%	42.71%	1.74% to 89.54%	0.20%	11.04%	1.93% to 45.14%	17.27%	1.67% to 95.36%	42.45%	5.09% to 92.28%
● nanet	10.83%	42.28%	1.68% to 89.41%	0.19%	11.06%	2.02% to 45.17%	17.60%	1.73% to 95.59%	42.77%	5.16% to 92.22%
● nanes	11.40%	45.29%	2.25% to 91.40%	0.54%	5.94%	0.99% to 28.20%	9.48%	0.84% to 88.56%	49.48%	5.26% to 95.52%
● danes	11.06%	45.16%	2.23% to 91.39%	0.52%	6.52%	1.01% to 28.83%	9.92%	0.86% to 89.64%	48.79%	5.26% to 94.85%
● danet	10.98%	41.82%	1.48% to 89.49%	0.18%	11.58%	2.36% to 50.38%	18.84%	1.84% to 96.48%	41.38%	5.15% to 90.68%
● danen	14.52%	29.67%	0.00% to 84.37%	0.01%	39.28%	7.45% to 69.02%	59.88%	5.18% to 100.00%	28.79%	5.24% to 68.02%
● daten1	11.02%	43.00%	1.77% to 89.88%	0.26%	10.93%	1.71% to 43.42%	16.74%	1.48% to 95.17%	42.94%	5.05% to 91.50%
● daten2	11.12%	42.05%	1.57% to 89.56%	0.19%	11.57%	2.28% to 49.98%	18.54%	1.80% to 96.26%	41.29%	5.11% to 91.07%
● faks	15.61%	45.39%	0.72% to 87.83%	0.05%	16.87%	4.05% to 48.76%	26.35%	2.78% to 98.34%	39.69%	5.17% to 85.34%

Table 2.5: Variance components estimated using EpiMigrant data. ‘Null’ reports the proportion of assays for which the corresponding variance component was estimated to be zero (within machine precision).

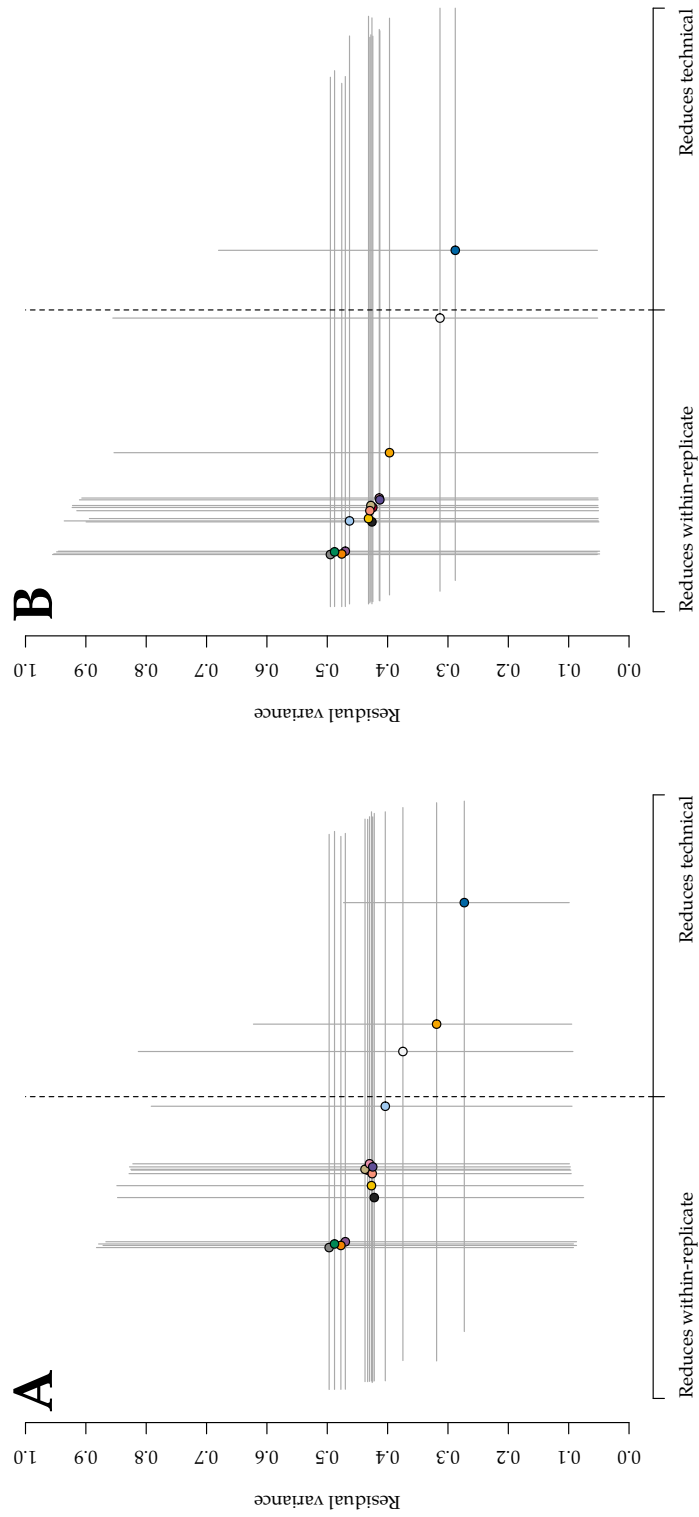


Figure 2.5: Nuisance split and residual variance components across pre-processing methods. Summary statistics were computed using: (A) EPIC-Italy; (B) EpiMigrant data. Points and segments represent the median and the 5th-to-95th percentile range, of nuisance splits (x-axis) and residual variance components (y-axis). The colour legend for pre-processing methods is reported in tables 2.4 and 2.5.

§ 2.2 Statistical models

While a few comparisons of different pre-processing methods have been attempted in the literature, much less attention has been paid to statistical modelling.

DNA methylation measurements lie on the unit interval, which may complicate their analysis, as described in section 1.2.2. For this reason, Du et al. [9] proposed the simple logit transformation of eq. (1.10), and argued that transformed values provide better performance (in terms of detection rate and true positive rate) based on a small titration experiment. Of course, as noted by Piepho [10], the logit is but one of several transformation that have been suggested for proportions. Moreover, other models, e.g. beta regression and fractional logistic regression, can naturally accommodate data restricted to the unit interval.

2.2.1 Simulation study

The author proposes a simulation study of six different model choices based on publicly available data generated by Liu et al. [7] as part of a case-control study of rheumatoid arthritis. The dataset includes 689 DNA methylation profiles in peripheral blood leukocytes, and was pre-processed using the second method described in section 2.1.2.

First, 100 autosomal ‘target’ CpG loci with DNA methylation measurements available in at least 95% of the samples were selected at random. These were chosen to be more than 100 bp away from non-rare (MAF greater than 1%) genetic variations, and without convincing evidence against a unimodal distribution of DNA methylation levels (p -value greater than 0.05 in the ‘dip’ test of Hartigan et al. [11]) that might indicate the presence of unaccounted genetic variations, or other issues including cross-hybridisation to multiple genomic locations.

Second, for each ‘target’ CpG locus, seven exposure variables were generated by inducing Spearman’s rank correlations of 20% to 50% in increments of 5% (using the method of Iman et al. [12] presented in algorithm 2.1) between DNA methylation measurements and $\mathbf{x} = (x_1, \dots, x_n)$, where

$$x_i = \Phi^{-1} \left(\frac{i - 1/2}{n} \right), \quad (2.9)$$

and n is the number of samples.

The lower bound for the Spearman’s rank correlation coefficient ρ was determined as follows. The null hypothesis $\mathcal{H}_0 : \rho = 0$ can be tested using the following approximate result [13],

$$\rho \sqrt{\frac{n-2}{1-\rho^2}} \stackrel{a}{\sim} t_{n-2}, \quad (2.10)$$

where n is the number of samples, from which it follows that the (positive) value of ρ corresponding to a given p -value p is given by

$$\rho = \frac{t_{p/2}}{\sqrt{t_{p/2}^2 + n - 2}}, \quad (2.11)$$

where $t_{p/2}$ is the 100 $(1 - p/2)\%$ quantile of the Student’s t -distribution with $n - 2$ degrees of freedom. Substituting $n = 689$ and $p = 0.05/470870 \approx 1.06 \times 10^{-7}$ gives $\rho \approx 0.20$. To increase accuracy, Spearman’s rank correlation coefficients ρ were transformed to Pearson product-moment correlation coefficients r in algorithm 2.1 by means of the following transformation that holds between normally distributed random variables [14],

$$r = 2 \sin(\rho \pi / 6), \quad (2.12)$$

which gives $r \approx 0.21$. Table 2.6 lists the values of ρ and r used in the simulation study, and the corresponding p -values computed using the approximation of eq. (2.10).

ρ	r	p -value
0.20	0.21	1.20×10^{-7}
0.25	0.26	2.81×10^{-11}
0.30	0.31	8.53×10^{-16}
0.35	0.36	2.76×10^{-21}
0.40	0.42	7.38×10^{-28}
0.45	0.47	1.18×10^{-35}
0.50	0.52	7.35×10^{-45}

Table 2.6: Values of ρ and r used in the simulation study and corresponding p -values. Values of r were computed using eq. (2.12); p -values are based on the approximation of eq. (2.10), assuming $n = 689$.

Third, for each ‘target’ CpG locus and each simulated exposure, a series of genome-wide association studies was conducted for each of the following model choices:

- (i) linear model (LM) of eq. (1.13) adjusted for micro-array and position on the micro-array, before and after logit transformation of DNA methylation levels;
- (ii) linear mixed model (LMM) of eq. (1.17) with random intercepts for each micro-array and position on the micro-array, before and after logit transformation of DNA methylation levels;
- (iii) beta regression (BR) model using the parametrisation of Ferrari et al. [15] and adjusted for micro-array and position on the micro-array;
- (iv) fractional logistic regression (FLR) model of Papke et al. [16] adjusted for micro-array and position on the micro-array.

Algorithm 2.1: Induction of rank correlations between given marginal distributions. The method by Iman et al. [12] determines the appropriate shuffle by ranking variables in $\mathbf{X}$ as in a reference sample $\mathbf{T}$ with the desired rank correlation.

Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$, target rank correlation matrix $\mathbf{S} \in [-1, 1]^{p \times p}$

Output: $\tilde{\mathbf{X}}$ with rank correlation matrix approximately equal to $\mathbf{S}$

$\mathbf{M} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$

$\mathbf{E} \leftarrow n^{-1} \mathbf{M}^\top \mathbf{M}$

$\mathbf{E} \rightarrow \mathbf{F}^\top \mathbf{F}$ // Choleski decomposition

$\mathbf{S} \rightarrow \mathbf{C}^\top \mathbf{C}$ // Choleski decomposition

$\mathbf{T} \leftarrow \mathbf{M} \mathbf{F}^{-1} \mathbf{C}$

Construct $\tilde{\mathbf{X}}$ by rearranging the values in each column of $\mathbf{X}$ to have the same rank ordering as the corresponding column of $\mathbf{T}$

return $\tilde{\mathbf{X}}$

Finally, for each genome-wide association study, the smallest observed p -value, and the p -value and rank for the ‘target’ CpG locus were recorded. The quantities described in section 1.3.1 were also computed as follows. Multiple comparisons were taken into account by adopting a 5% Bonferroni-corrected significance threshold $\alpha = 0.05/470870 \approx 1.06 \times 10^{-7}$. Given a ‘target’ CpG locus t and a simulated exposure variable x , associations with each CpG locus i were considered ‘true’ if:

- (i) the Spearman’s rank correlation coefficient $\rho_{x,i}$ between x and i was declared statistically significant after Bonferroni correction, $|\rho_{x,i}| \geq \rho_1$, where $\rho_1 \approx 0.20$ was computed from eq. (2.11) as described above;
- (ii) the Spearman’s rank correlation coefficient $\rho_{t,i}$ between t and i was declared statistically significant after Bonferroni correction for the total number of

pairwise correlations between DNA methylation levels, $|\rho_{t,i}| \geq \rho_2$, where $\rho_2 \approx 0.27$ was computed from eq. (2.11) with $p = 0.05/(470870 \times 470869/2) \approx 4.51 \times 10^{-13}$;

(iii) $\rho_{x,i}$ and $\rho_{t,i}$ were concordant, $\rho_{x,i}\rho_{t,i} > 0$.

These quantities were then used to compute the performance metrics of eqs. (1.31) to (1.35) and (1.38), which were summarised across simulations as described below.

2.2.2 Comparison of statistical models

The p -value for the ‘target’ CpG locus could not be estimated using BR models in three simulations; these were excluded from further consideration.

The percentages of simulations where the association with the ‘target’ CpG locus was declared significant after Bonferroni correction are reported in table 2.7. For $\rho = 0.35$, all model choices, with the exception of LM after logit transformation, identified the association with the ‘target’ CpG locus in more than 95% of the cases. For smaller values of ρ , the best-performing model choice was LMM, followed by BR and FLR, and finally LM; notably, logit transformation of DNA methylation values worsened the performance of both LM and LMM. Similar conclusions can be reached by examining the median rank of p -values for associations with ‘target’ CpG loci (table 2.8).

The specificity of all model choices exceeded 95% across simulations; on the other hand, the average sensitivity of LMM, BR, and FLR was comparable for all values of ρ considered, and generally lower than that of LM (table 2.9, fig. 2.6A). The sensitivity of LM and LMM was only slightly worsened by logit transformation.

Model	ρ						
	0.20	0.25	0.30	0.35	0.40	0.45	0.50
● LM	9.00%	42.00%	82.00%	95.00%	100.00%	99.00%	100.00%
● + logit	4.00%	39.00%	71.00%	88.00%	99.00%	99.00%	100.00%
● LMM	24.00%	76.00%	95.00%	98.00%	100.00%	100.00%	100.00%
● + logit	21.00%	59.00%	91.00%	98.00%	100.00%	100.00%	100.00%
● BR	19.00%	52.00%	85.86%	96.97%	100.00%	99.00%	100.00%
● FLR	16.00%	51.00%	86.00%	96.00%	98.00%	99.00%	100.00%

Table 2.7: Percentages of simulations where the association with the ‘target’ CpG locus was declared significant.

Negative predictive value and accuracy were virtually indistinguishable for all model choices, and showed a weak inverse association with ρ . On the contrary, the positive predictive value tended to increase with ρ (fig. 2.6B), plateauing between 0.8 and 0.9 for all methods at $\rho = 0.5$.

The percentages of simulations with at least one false positive finding were used as estimates of the family-wise error rate (table 2.10 and fig. 2.6C). The false discovery rate was computed according to eq. (1.38) across simulations (table 2.11 and fig. 2.6D). Irrespective of logit transformation, LM attained the lowest FWER and FDR for all values of ρ considered. LMM, BR, and FLR all performed in a similar fashion for the FWER, but LMM achieved a slightly lower FDR than BR and FLR.

These results reaffirm the inherent trade-off between the probabilities of committing type I and type II errors described in section 1.3.1. For example, LMM achieves a very high sensitivity, but at the cost of being more prone to type I errors than LM; the same is true of BR and FLR.

		ρ						
Model		0.20	0.25	0.30	0.35	0.40	0.45	0.50
●	LM	42.5	3.0	1.0	1.0	1.0	1.0	1.0
●	+ logit	86.0	4.5	1.0	1.0	1.0	1.0	1.0
●	LMM	13.5	1.0	1.0	1.0	1.0	1.0	1.0
●	+ logit	37.5	2.5	1.0	1.0	1.0	1.0	1.0
●	BR	84.5	3.0	1.0	1.0	1.0	1.0	1.0
●	FLR	46.0	3.0	1.0	1.0	1.0	1.0	1.0

Table 2.8: Median ranks of p -values for ‘target’ CpG loci across simulations.

		ρ						
Model		0.20	0.25	0.30	0.35	0.40	0.45	0.50
●	LM	3.81%	22.40%	16.90%	15.23%	15.62%	14.73%	15.16%
●	+ logit	2.01%	20.17%	16.49%	15.11%	15.36%	14.44%	14.45%
●	LMM	12.06%	32.07%	23.51%	21.89%	21.74%	23.35%	23.70%
●	+ logit	11.76%	27.99%	22.75%	21.47%	21.08%	22.32%	22.45%
●	BR	11.68%	26.86%	21.60%	19.13%	18.80%	18.54%	18.89%
●	FLR	10.11%	28.04%	20.92%	18.68%	18.01%	17.15%	17.43%

Table 2.9: Average sensitivity across simulations.

Model	ρ						
	0.20	0.25	0.30	0.35	0.40	0.45	0.50
● LM	20.00%	24.00%	30.00%	40.00%	42.00%	52.00%	64.00%
● + logit	23.00%	24.00%	29.00%	41.00%	50.00%	56.00%	64.00%
● LMM	47.00%	41.00%	58.00%	68.00%	77.00%	86.00%	95.00%
● + logit	46.00%	44.00%	62.00%	75.00%	83.00%	87.00%	95.00%
● BR	53.00%	58.00%	67.68%	68.69%	78.00%	77.00%	85.86%
● FLR	51.00%	48.00%	60.00%	58.00%	67.00%	69.00%	78.00%

Table 2.10: Percentages of simulations with at least one false positive finding.

The values can be interpreted as estimates of the family-wise error rate across simulations.

Model	ρ						
	0.20	0.25	0.30	0.35	0.40	0.45	0.50
● LM	17.63%	10.75%	8.11%	12.48%	7.62%	8.64%	12.04%
● + logit	20.31%	11.78%	8.42%	13.18%	9.66%	10.21%	11.60%
● LMM	39.63%	17.37%	16.23%	17.02%	10.05%	11.57%	12.37%
● + logit	38.80%	22.19%	18.37%	20.34%	14.37%	12.67%	13.13%
● BR	46.18%	36.50%	29.42%	26.17%	21.64%	18.51%	18.73%
● FLR	46.09%	28.02%	24.24%	22.72%	17.96%	13.82%	16.20%

Table 2.11: False discovery rate across simulations.

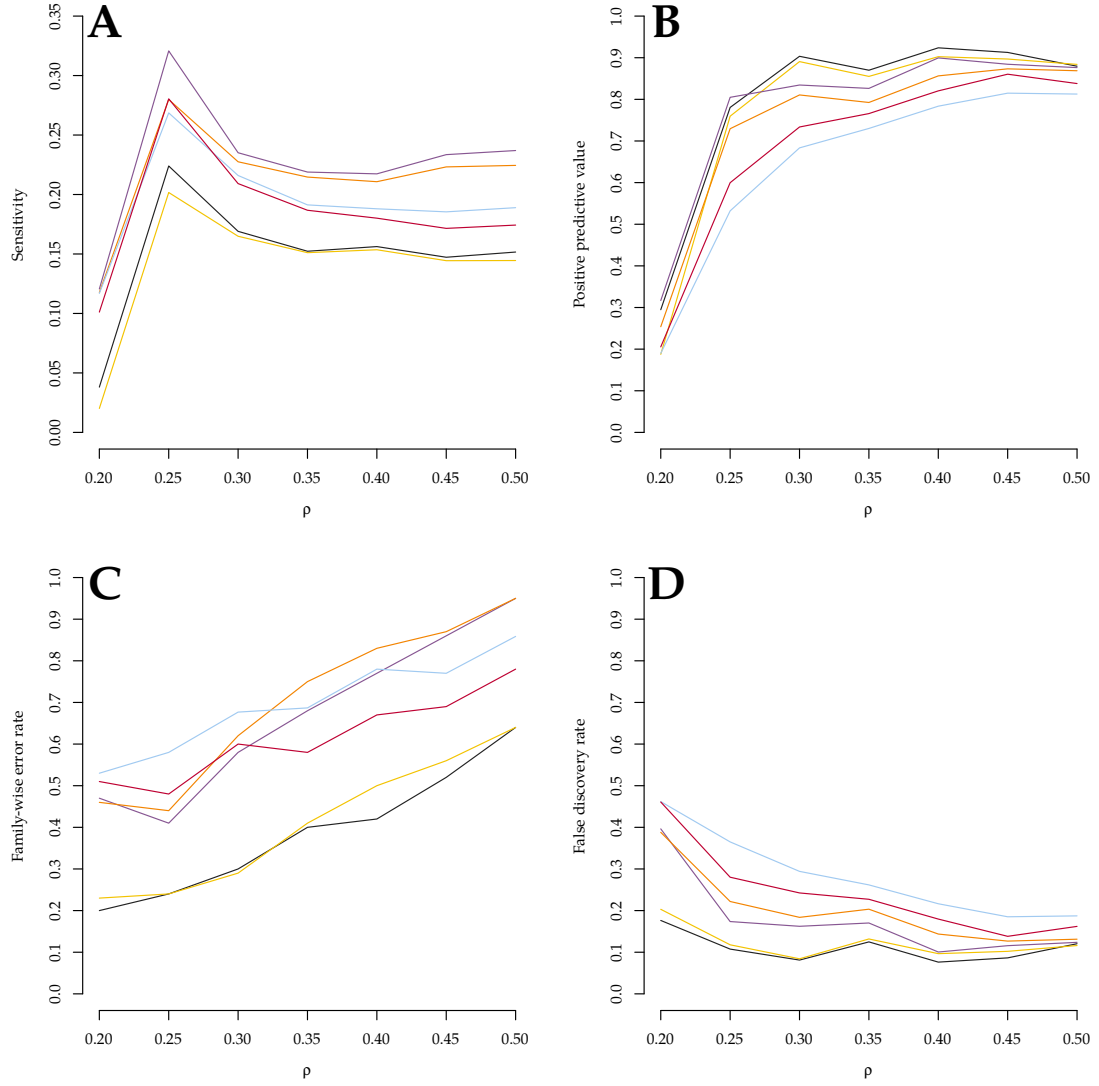


Figure 2.6: Performance metrics averaged across simulations. (A) sensitivity; (B) positive predictive value; (C) family-wise error rate; (D) false discovery rate. The colour legend is reported in tables 2.7 to 2.11.

§ 2.3 Recommendations

2.3.1 Pre-processing method

As outlined in section 2.1.4, most pre-processing methods perform equally well, leading to other considerations. The two dye bias correction methods introduced in section 1.2.2 are appealing for several reasons. First, they can be executed independently for each sample, making them not only easily parallelisable, but also completely reproducible. Furthermore, they alter the data as little as possible, and yet achieve performances comparable with those of more complicated pre-processing methods.

For these reasons, the author recommends that either method is adopted to pre-process DNA methylation data generated using HM450 BeadChips.

2.3.2 Statistical model

Based on the results presented in section 2.2.2, model choices can be roughly divided into two categories. More conservative approaches include LM before and after logit transformation of DNA methylation levels; while they are less susceptible to false positive findings, they may not be sufficiently powered, especially in small studies. The converse is true of the second class of model choices, which may be called ‘exploratory’, that includes LMM before and after logit transformation, BR, and FLR. Naturally, these differences become immaterial as the strength of the association (represented by ρ in the simulations) is increased; it is reasonable to expect a similar behaviour with increased sample sizes.

For these reasons, it is inexpedient to provide a single recommendation. Nonetheless, given the exploratory nature of most ‘-omics’ studies, it is the author’s

opinion that model choices that afford greater statistical power should be favoured, and associations should be subsequently validated (and possibly replicated) using different laboratory techniques; this is potentially more relevant in smaller studies. In addition, other factors may inform the decision of statistical models. For example, phenotypically-relevant differences in the variance of DNA methylation levels have recently received attention [17]; BR offers a natural parametrisation that is perfectly suited to their exploration.

§ 2.4 Limitations

The comparison of pre-processing methods and the simulation study of sections 2.1.4 and 2.2.2, as well as the recommendations condensed in section 2.3, are subject to some limitations.

The comparison of pre-processing methods included more than 4,400 samples and 75 pairs of technical replicates; however, the number of technical replicates, albeit larger than that of any comparisons attempted to date, was relatively small. Nonetheless, the two studies yielded comparable results.

For computational reasons, the comparison of statistical models was based on 100 simulations with fixed sample size for each combination of simulated exposure and model choice. In the absence of reasonable *a priori* assumptions, it was also assumed that exposure variables were normally distributed.

Finally, both studies were conducted using DNA methylation profiles in peripheral blood leukocytes; it is possible that results presented herein do not generalise to other tissues.

References

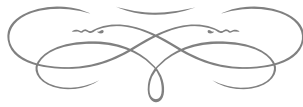
1. F Marabita, M Almgren, M E Lindholm, S Ruhrmann, F Fagerström-Billai, M Jagodic, C J Sundberg, T J Ekström, A E Teschendorff, J Tegnér, and D Gomez-Cabrero. 'An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform'. *Epigenetics* **8.3** (2013), 333–346.
2. L E Reinius, N Acevedo, M Joerink, G Pershagen, S-E Dahlén, D Greco, C Söderhäll, A Scheynius, and J Kere. 'Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility'. *PLOS ONE* **7.7** (2012), e41361.
3. S Dedeurwaerder, M Defrance, E Calonne, H Denis, C Sotiriou, and F Fuks. 'Evaluation of the Infinium Methylation 450K technology'. *Epigenomics* **3.6** (2011), 771–784.
4. A E Teschendorff, F Marabita, M Lechner, T Bartlett, J Tegner, D Gomez-Cabrero, and S Beck. 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data'. *Bioinformatics* **29.2** (2013), 189–196.
5. R Pidsley, C C Y Wong, M Volta, K Lunnon, J Mill, and L C Schalkwyk. 'A data-driven approach to preprocessing Illumina 450K methylation array data'. *BMC Genomics* **14** (2013), 293.
6. M C Wu, B R Joubert, P-F Kuan, S E Håberg, W Nystad, S D Peddada, and S J London. 'A systematic assessment of normalization approaches for the Infinium 450K methylation platform'. *Epigenetics* **9.2** (2014), 318–329.

7. Y Liu, M J Aryee, L Padyukov, M D Fallin, E Hesselberg, A Runarsson, L Reinius, N Acevedo, M Taub, M Ronninger, K Shchetynsky, A Scheynius, J Kere, L Alfredsson, L Klareskog, T J Ekström, and A P Feinberg. 'Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis'. *Nature Biotechnology* **31** (2013), 142–147.
8. D Palli, F Berrino, P Vineis, R Tumino, S Panico, G Masala, C Saieva, S Salvini, M Ceroti, V Pala, S Sieri, G Frasca, M C Giurdanella, C Sacerdote, L Fiorini, E Celentano, R Galasso, A Decarli, and V Krogh. 'A molecular epidemiology project on diet and cancer: the EPIC-Italy prospective study. Design and baseline characteristics of participants'. *Tumori* **89** (2003), 586–593.
9. P Du, X Zhang, C-C Huang, N Jafari, W A Kibbe, L Hou, and S M Lin. 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis'. *BMC Bioinformatics* **11** (2010), 587.
10. H-P Piepho. 'The folded exponential transformation for proportions'. *Journal of the Royal Statistical Society, Series D* **52.4** (2003), 575–589.
11. J A Hartigan and P M Hartigan. 'The dip test of unimodality'. *The Annals of Statistics* **13.1** (1985), 70–84.
12. R L Iman and W J Conover. 'A distribution-free approach to inducing rank correlation among input variables'. *Communications in Statistics – Simulation and Computation* **11.3** (1982), 311–334.
13. W H Press, S A Teukolsky, W T Vetterling, and B P Flannery. *Numerical recipes: the art of scientific computing*. 3rd ed. Cambridge University Press, 2007.
14. K Pearson. 'Mathematical contributions to the theory of evolution. XVI. On further methods of determining correlation'. *Drapers' Company Research Memoirs, Biometric Series IV* (1907).

15. S L P Ferrari and F Cribari-Neto. 'Beta regression for modelling rates and proportions'. *Journal of Applied Statistics* **31.7** (2004), 799–815.
16. L E Papke and J M Wooldridge. 'Econometric methods for fractional response variables with an application to 401(k) plan participation rates'. *Journal of Applied Econometrics* **11** (1996), 619–632.
17. R Jiang, M J Jones, E Chen, S M Neumann, H B Fraser, G E Miller, and M S Kobor. 'Discordance of DNA methylation variance between two accessible human tissues'. *Scientific Reports* **5** (2015), 8257.



This page intentionally left blank.



Chapter 3

Epigenetic signatures of internal migration in Italy

This chapter has been published as a paper in the *International Journal of Epidemiology* [1] under a Creative Commons Attribution licence, and is reproduced here with minimal alterations. © ⓘ

§ 3.1 Introduction

Observational studies have contributed to consolidating the idea that the risks of many non-communicable diseases in voluntary migrants become similar to those in the host (native) population after one or more generations [2–4]. For ex-

ample, seminal studies have revealed a gradient of increasing incidence of coronary heart disease in Japanese men from Japan to Hawaii to California [4, 5]. These observations have been used to support the hypothesis that these diseases have a predominantly environmental origin (rather than inherited). Nonetheless, differences persist between native and migrant populations. For instance, migrants from non-western countries are more prone to cancers related to infections experienced in early life, and less likely to suffer from cancers commonly associated with a Westernized lifestyle [6].

In this paper, we speculate that the observed health differentials might be mediated at the molecular level by changes in DNA methylation. In particular, we hypothesize that these changes are brought about by exposure differentials between the origin and host populations, and that they are instrumental in coping with the ‘mismatch’ between early-life programming (due to perinatal exposures) and changes in those same exposures that occurred later in life as a result of migration. This hypothesis is based on the concept that developmental history leaves its mark primarily through potentially reversible epigenetic changes [7]. It is also supported by the observation that disease risks in migrants tend to increase with duration of residence in the host population, eventually becoming indistinguishable from those in natives [4]. Among epigenetic changes, DNA methylation is thought to be relatively stable due to its heritability across cell generations, and yet flexible enough to allow differentiation into different cell types, as well as adaptation to stress and the external environment [8–10]. In addition, DNA methylation plays a pivotal role in transcriptional repression and suppression of transcriptional noise [11], and is tightly linked to other epigenetic mechanisms such as histone modifications and chromatin remodelling [12, 13]. DNA methylation levels are associated with environmental and lifestyle exposures such as tobacco smoking [14], and altered DNA methylation patterns have

also been implicated in many human diseases [15].

Traditionally, epidemiological studies of migrants endeavoured to elucidate the relative contributions of genetic background, environment and their interaction [16, 17]. Most studies have focused on the effects of international migration, since risk factor differentials tend to be larger across countries. In this case, genetic differences between migrants and the host population may hinder the identification of migration-specific effects. Italy represents an interesting natural experiment, not only for its pronounced economic, social and environmental south-to-north gradient and the mass migration of labour that took place from the mid 1940s to the 1970s [18, 19], but also for its relative genetic homogeneity (with the possible exception of Sardinia [20]).

Using genome-wide DNA methylation profiles obtained from prospectively collected peripheral blood samples from 1,066 participants in the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy) [21], we present the first epigenome-wide association study (EWAS) to identify DNA methylation changes associated with voluntary south-to-north migration that occurred within Italy in the three decades after the end of WWII.

§ 3.2 Methods

3.2.1 Study population and sample selection

All participants were recruited between 1993 and 1998 as part of EPIC-Italy [21]. Detailed lifestyle and dietary information was collected at enrolment using self-administered questionnaires and a validated food frequency questionnaire [22], respectively. Anthropometric measurements were obtained at the inclusion visit, as were peripheral blood samples that were aliquoted and stored in liquid nitro-

gen on the day of collection.

A total of 1,222 genome-wide DNA methylation profiles were acquired as part of three separate prospective case-control studies nested within EPIC-Italy on breast cancer ($N = 332$), colorectal cancer ($N = 338$) and myocardial infarction (EPICOR [23], $N = 552$). Eight profiles were excluded because of unsatisfactory technical quality. A single profile was retained, on the basis of technical quality, for participants included in more than one study, leaving a total of 1,170 unique profiled participants. Within each study, participants who developed the relevant condition less than 1 year after blood draw ($N = 46$), or who developed any kind of haematological malignancy at any time after enrolment ($N = 4$), were excluded; all remaining subjects were considered healthy at inception. A total of 23 participants were excluded because of incomplete dietary or lifestyle information. To minimize confounding by genetic factors, participants born outside Italy ($N = 18$) or in the insular region of Sardinia ($N = 18$) were also excluded. The remaining 1061 participants were categorized as follows:

- (i) south-to-north migrants ($N = 190$), recruited in Turin ($N = 148$) or Varese ($N = 42$), and born in any southern Italian region;
- (ii) southern natives (origin population, $N = 123$), recruited by the two southern Italian EPIC centres of Naples ($N = 40$) and Ragusa ($N = 83$), and born in any southern Italian region;
- (iii) north-western natives (host population, $N = 543$), recruited by the two north-western Italian EPIC centres of Turin ($N = 317$) and Varese ($N = 226$), and born in any north-western Italian region.

A total of 205 participants did not fall into any of the above categories, and were excluded from subsequent analyses. Detailed information on the 856 participants included in the study is summarized in table 3.1.

	South-to-north migrants	Southern natives	North-western natives
<i>N</i>	190	123	543
Men	115 (60.5%)	62 (50.4%)	232 (42.7%)
Women	75 (39.5%)	61 (49.6%)	311 (57.3%)
Age (years)	51.2 (35.8 to 66.0)	52.7 (34.7 to 67.9)	54.4 (35.2 to 72.0)
Smoking status			
Never	64 (33.7%)	41 (33.3%)	248 (45.7%)
Current	65 (34.2%)	40 (32.5%)	151 (27.8%)
Former	61 (32.1%)	42 (34.1%)	144 (26.5%)
Physical activity			
1	30 (15.8%)	28 (22.8%)	116 (21.4%)
2	67 (35.3%)	36 (29.3%)	148 (27.3%)
3	67 (35.3%)	51 (41.5%)	251 (46.2%)
4	26 (13.7%)	8 (6.5%)	28 (5.2%)
5	0 (0%)	0 (0%)	0 (0%)
Energy (kcal)	2143.5 (487.7 to 5225.3)	2375.9 (1100.7 to 5476.9)	2184.0 (722.7 to 5309.2)
Protein (g)	85.5 (26.5 to 189.6)	94.1 (39.5 to 190.8)	89.7 (24.4 to 209.3)
From animal sources	53.9 (17.2 to 132.0)	51.3 (12.5 to 103.3)	58.2 (14.3 to 167.8)
From vegetable sources	29.3 (5.6 to 92.0)	40.0 (13.8 to 124.9)	26.3 (5.5 to 70.3)
Fat (g)	77.6 (21.0 to 199.0)	85.9 (36.3 to 178.4)	83.5 (18.9 to 227.7)
From animal sources	41.6 (14.5 to 108.5)	41.9 (12.1 to 105.1)	47.8 (6.4 to 174.6)
From vegetable sources	36.0 (5.6 to 90.5)	42.0 (12.5 to 91.6)	35.0 (4.4 to 91.0)
Cholesterol (mg)	326.7 (77.4 to 772.9)	288.4 (72.9 to 692.5)	358.0 (72.4 to 1002.5)
Available carbohydrates (g)	259.1 (50.8 to 685.9)	320.0 (120.3 to 780.5)	249.2 (48.9 to 760.6)
Soluble carbohydrates	93.6 (15.8 to 284.1)	95.5 (39.0 to 253.0)	98.5 (28.1 to 384.2)
Starch	158.5 (33.7 to 571.6)	209.2 (55.8 to 655.1)	144.0 (5.1 to 460.3)
Fibre (g)	21.0 (3.6 to 51.5)	33.4 (10.5 to 131.2)	19.5 (5.3 to 70.5)
Alcohol (g)	11.5 (0.0 to 90.9)	1.6 (0.0 to 50.4)	11.2 (0.0 to 105.4)
Vitamins			
Vitamin A (μ g RE)	944.8 (167.1 to 3467.5)	898.5 (260.1 to 5092.5)	1031.7 (175.4 to 6716.7)
Vitamin B ₁ (mg)	0.96 (0.24 to 2.12)	1.22 (0.40 to 4.40)	0.98 (0.33 to 2.18)
Vitamin B ₂ (mg)	1.46 (0.30 to 3.12)	1.58 (0.57 to 3.19)	1.58 (0.43 to 4.19)
Vitamin B ₃ (mg)	17.8 (5.2 to 39.6)	21.9 (7.8 to 63.9)	17.8 (4.9 to 39.5)
Vitamin B ₆ (mg)	1.84 (0.45 to 4.40)	2.28 (0.85 to 5.75)	1.87 (0.46 to 4.12)
Folic acid (μ g)	269.8 (48.3 to 775.7)	335.0 (124.4 to 835.0)	261.0 (50.8 to 673.8)
Vitamin C (mg)	127.5 (16.9 to 413.5)	153.2 (63.4 to 672.7)	121.5 (5.9 to 888.5)
Vitamin D (μ g)	1.24 (0.13 to 5.85)	1.05 (0.04 to 6.30)	1.34 (0.08 to 12.45)
Vitamin E (mg)	7.65 (1.90 to 24.18)	9.08 (3.78 to 20.04)	7.69 (1.50 to 22.60)
Minerals			
Calcium (mg)	919.9 (279.6 to 2956.3)	834.3 (239.0 to 2050.2)	1020.1 (201.7 to 3833.4)
Iron (mg)	13.8 (2.9 to 35.0)	15.1 (5.8 to 42.9)	13.8 (4.8 to 31.9)
Phosphorus (mg)	1342.6 (382.9 to 2945.3)	1558.9 (576.4 to 3816.1)	1413.4 (487.0 to 3393.8)
Potassium (mg)	3207.2 (690.1 to 7023.0)	3564.4 (1414.5 to 8445.0)	3271.4 (982.1 to 8882.6)
Sodium (mg)	2164.9 (359.2 to 9837.6)	2365.8 (729.8 to 8879.0)	2227.3 (728.5 to 8636.5)
Zinc (mg)	11.7 (4.0 to 26.1)	14.2 (4.4 to 42.9)	12.4 (3.6 to 32.8)

Table 3.1: Characteristics of participants included in the study. Counts and percentages are reported for categorical variables, and medians and ranges for continuous variables.

3.2.2 Laboratory analyses

Genome-wide DNA methylation analyses using the Illumina Infinium® HumanMethylation450 (HM450) platform were carried out at the Human Genetics Foundation (Turin, Italy) according to manufacturers' protocols. Buffy coats stored in liquid nitrogen were thawed, and genomic DNA was extracted using the QIAGEN QIAasympyony DNA Midi Kit; 500 ng of DNA were bisulphite-converted using the Zymo Research EZ-96 DNA Methylation-Gold™ Kit, and hybridized to Illumina Infinium® HumanMethylation450 BeadChips. These were subsequently scanned using the Illumina HiScanSQ system, and sample quality was assessed using control probes present on the microarrays. Finally, raw intensity data were exported from Illumina GenomeStudio (version 2011.1). Data pre-processing was carried out using in-house software written for the R statistical computing environment. For each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction and dye bias correction (for probes using the Infinium II design) were also performed. The subset of 470,870 probes targeting autosomal CpG loci was selected for further analyses. DNA methylation levels at each locus were expressed as the ratio of intensities arising from methylated cytosines over total intensities.

3.2.3 Statistical analyses

An EWAS was conducted to compare south-to-north migrants with their origin (southern natives) and host (north-western natives) populations, with the objective of characterizing epigenetic adaptation processes subsequent to migration to north-western Italy. For each probe, DNA methylation levels were modelled as

dependent variable in a generalized linear model with beta-distributed response using the parametrisation of Ferrari et al. [24]. All models were adjusted for micro-array ($N = 102$) and position on the micro-array ($N = 12$), sex, and case-control status (separately for cancers and myocardial infarction). In place of age at recruitment, models were also adjusted for two continuous variables representing the time to birth and to recruitment of each participant (computed from an arbitrary reference date). Since the difference between these two quantities equals age at recruitment for any choice of reference date, this approach grants an additional degree of freedom to account for possible differences in migration behaviour associated with birth date. The effect of dietary and lifestyle factors, which are radically different in southern regions [25], was investigated using a second set of models additionally adjusted for 25 dietary variables (total energy intake, protein from animal and vegetable sources, fat from animal and vegetable sources, cholesterol, soluble carbohydrates, starch, fibre, alcohol and vitamins and minerals as listed in table 3.1), smoking status and level of physical activity (categorical variable). To prevent inclusion of highly correlated variables and reduce the number of estimated regression coefficients, dietary variables were subjected to principal component analysis (PCA), and the first 16 principal components (explaining more than 99% of the variance) were included in the models. Multiple comparisons were accounted for by considering a Bonferroni-corrected significance threshold $\alpha = 0.05/470,870 \approx 1.1 \times 10^{-7}$, ensuring a stringent control of the family-wise error rate at level 5%. Candidate CpG loci were additionally filtered as follows. First, probe sequences were aligned to the reference human genome using Bowtie 2 [26] to assess the potential to cross-hybridise to multiple genomic locations, thus affecting DNA methylation measurements [27]. CpG loci targeted by cross-hybridising probes (defined as those lacking unique genome alignments, with up to three base mismatches) were excluded from fur-

ther consideration. Second, potential sources of genetic confounding and context disruption for DNA methylation (such as polymorphisms at the CpG locus) were identified by retrieving known genetic variations and computing the corresponding minor allele frequencies (MAFs) in the European population, based on publicly available data generated by the 1000 Genomes project [28]. As a precautionary measure, CpG loci found within 100 base pairs (bp) of non-rare variants (MAF greater than 1%) were removed from the list of candidates.

§ 3.3 Results

As illustrated in fig. 3.1A, the EWAS identified 20 differentially methylated CpG loci in south-to-north migrants with respect to the origin population (southern natives). Two probe sequences were ambiguously aligned to the reference human genome, and genetic variations were found in the vicinity of nine candidate CpG loci. A total of nine CpG loci were left for further consideration, of which none survived the adjustment for dietary and lifestyle factors. Comparison of south-to-north migrants with respect to the host population (north-western natives) revealed 91 differentially methylated CpG loci (fig. 3.1B). After removal of 23 candidates whose associated probe sequences could not be uniquely aligned to the reference human genome, and of 33 candidates in the proximity of non-rare genetic variations, 35 CpG loci were left for further consideration, and 22 survived the adjustment for dietary and lifestyle factors. Of these, 17 were found to be relatively hypermethylated in south-to-north migrants, and seven were found in the pericentric region on the long arm of chromosome 7 (from the centromere to 6.37×10^7 bp). These loci exhibited a consistent decreasing gradient from south-to-north migrants to southern natives to north-western natives (fig. 3.2). They were also flanked by several other loci that shared the same direction of associ-

ation. This region was additionally characterized by PCA of DNA methylation measurements at 43 enclosed CpG loci assayed by the HM450 platform (filtered according to the criteria described above), before and after adjustment for dietary and lifestyle factors. Irrespective of adjustment, the first principal component explained approximately 35% of the variance (fig. 3.3A), and was the only component explaining more than 10% of the variance. The association of each principal component with migratory status was formally assessed using Kruskal–Wallis rank sum tests. Results were comparable before and after adjustment; however, the second, third and 36th principal components lost statistical significance after adjustment (fig. 3.3B). The first principal component was consistently associated with migratory status (p -values 1.71×10^{-8} and 6.28×10^{-6} before and after adjustment, respectively), as was the 15th (p -values 0.043 and 0.041, respectively). Scores exhibited a decreasing gradient similar to that observed in fig. 3.2, albeit less markedly for the 15th principal component (fig. 3.3C).

§ 3.4 Discussion

To our knowledge, this is the first EWAS to examine DNA methylation changes in voluntary migrants. The gamut of alterations observed in south-to-north migrants offers evidence that important environmental and lifestyle changes may induce molecular adaptation mechanisms to stressors that are inheritable across cell divisions. Some of the differences are evident even after adjustment for dietary and lifestyle factors, suggesting that these DNA methylation changes are not merely ascribable to behaviour modification following migration. Intriguingly, we found DNA methylation changes in south-to-north migrants compared to the host population at several CpG loci located on a large pericentric region on the long arm of chromosome 7. Pericentric regions have long been thought to be

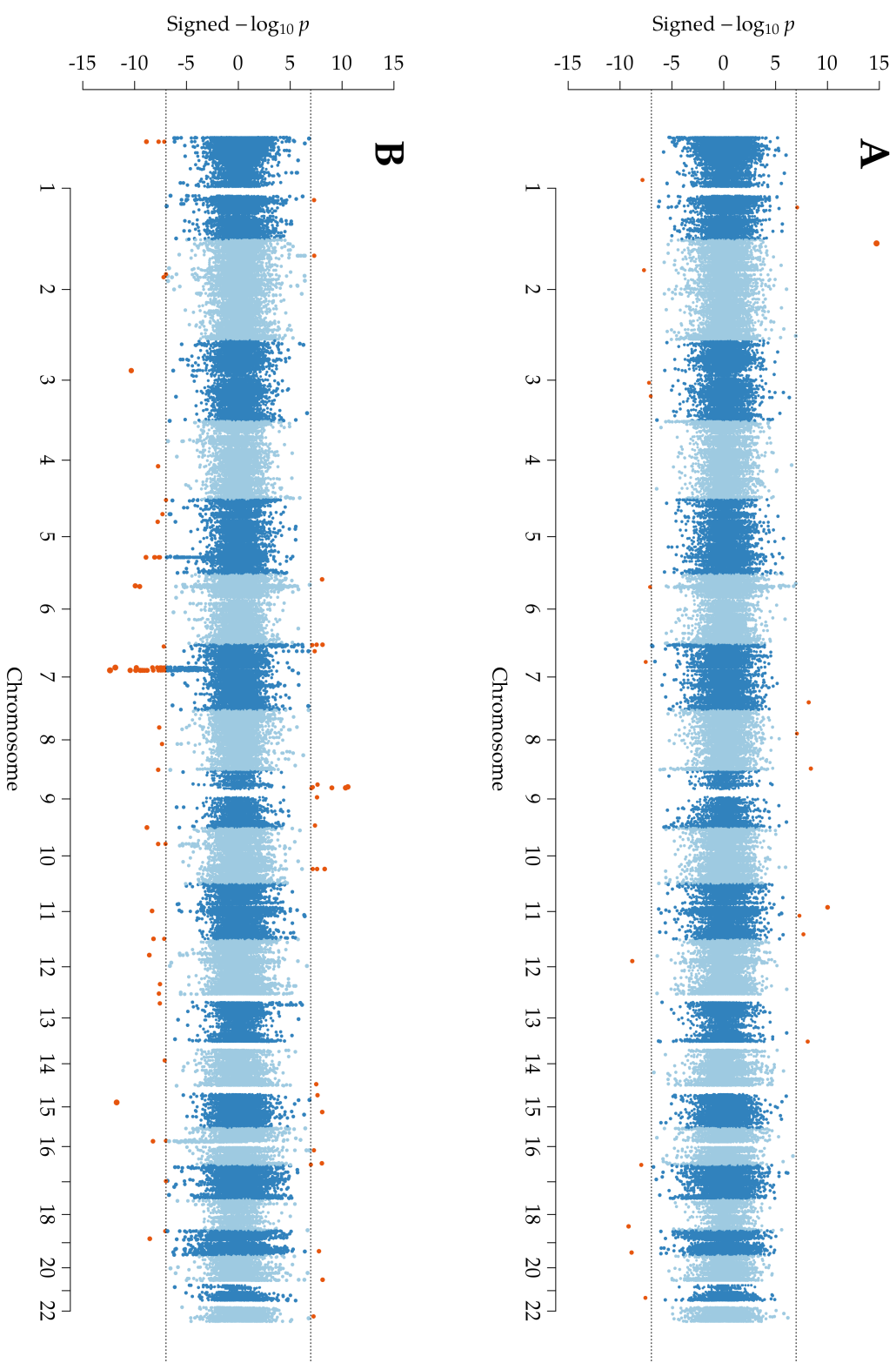


Figure 3.1: Signed Manhattan plot. EWAS comparing south-to-north migrants with: (A) the origin population (southern natives); (B) the host population (north-western natives).

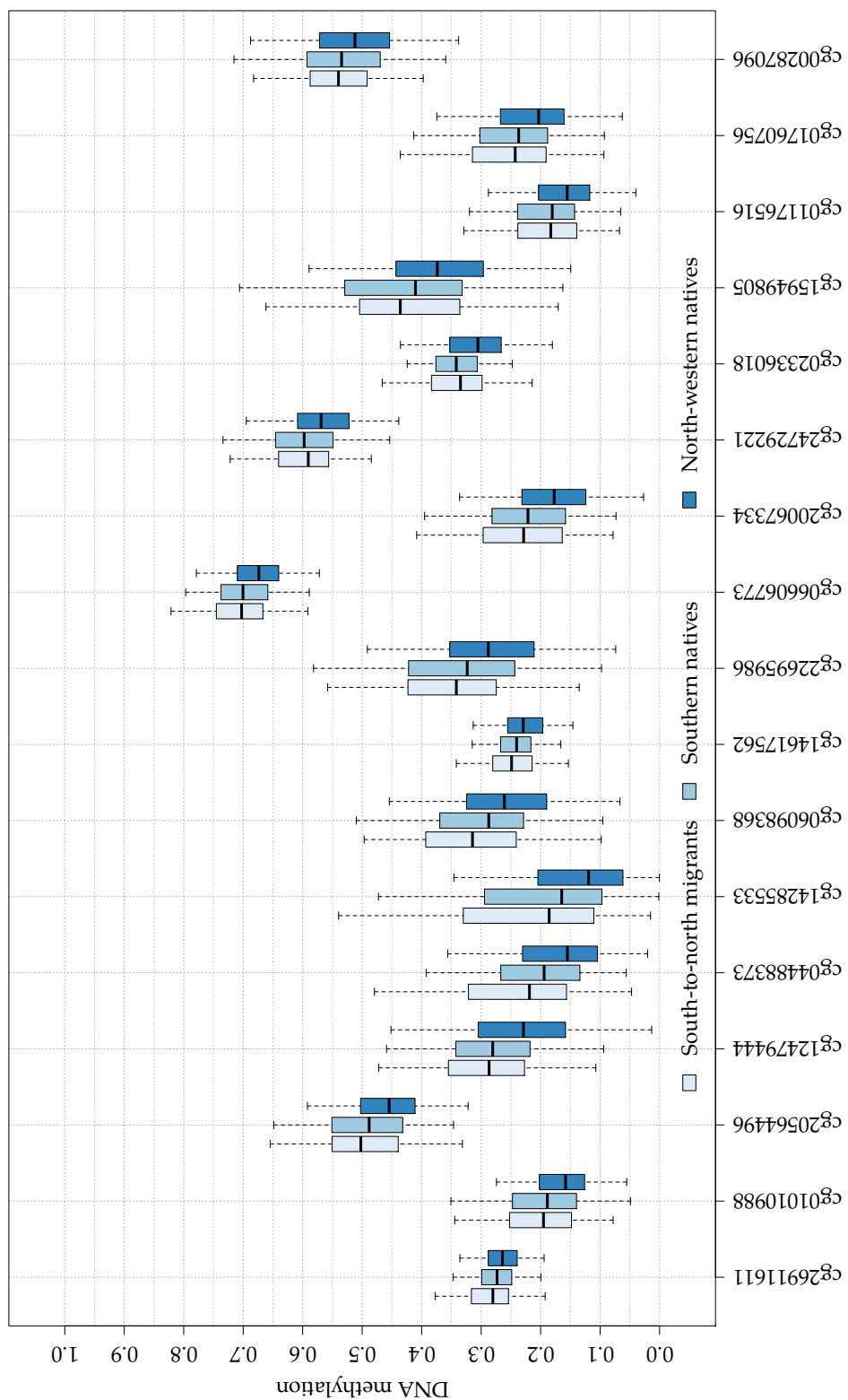


Figure 3.2: Box-and-whisker plot across groups. CpG loci identified by the EWAS in the pericentric region on the long arm of chromosome 7.

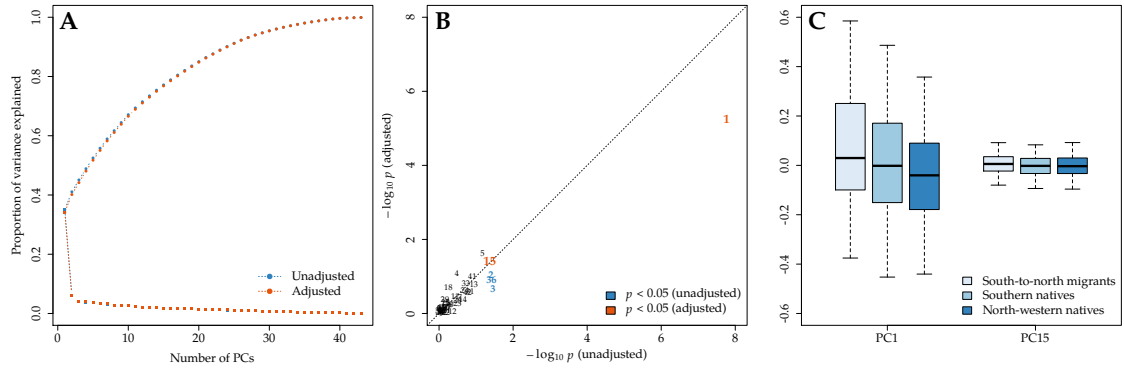


Figure 3.3: PCA of DNA methylation levels assayed in the pericentric region on the long arm of chromosome 7. (A) scree plot; (B) Manhattan plot for the association of PCs with migrant status (Kruskal-Wallis rank sum test); (C) box-and-whisker plot for PCs associated with migrant status ($p < 0.05$) after adjustment for dietary and lifestyle factors.

transcriptionally inert, but recent evidence suggests that pericentric and centromeric transcripts play an important role in preserving genome stability [29]. Additionally, transcription of pericentric satellites appears to be a general cellular response to external stressors including heat shock, ultraviolet radiation and oxidative stress [30]. In this light, it appears that molecular consequences of migration may not be limited to specific genes, but may act at a higher complexity level, for example on gene regulatory networks. The gradient observed in fig. 3.2 may thus epitomise an adaptive mechanism to cope with the ‘mismatch’ between early-life programming and exposure changes in later life: before migration, south-to-north migrants and southern natives share common environmental factors that affect DNA methylation patterns and (possibly) differentiate them from northern natives; the amplified response observed after migration might therefore be a consequence of relative abundance or lack of these factors in the host population. Such factors could include, for example, vitamin D (in re-

lation to more limited sun exposure in northern Italy), other vitamins contained in food, occupational and environmental exposure to pollutants and even exposure to different infectious agents (with some viruses, for example the hepatitis B virus, being more prevalent in southern populations). This would not only explain the observed DNA methylation gradient, but it would also be consistent with the 'developmental origins of disease' hypothesis [31, 32] and with current understanding of the role of perinatal exposures in health and disease.

The main strengths of our study are its sample size and the relative genetic homogeneity of its participants (all born in Italy), which limits the potential for genetic confounding. Its main limitation is the lack of information regarding the time of migration, from which age at migration and duration of stay could be computed and accounted for. Nevertheless, absence of this information is more likely to dilute any observable effect on DNA methylation, rather than lead to false-positive results. The biological interpretation of our results could be enhanced were genome-wide gene expression data available for the same subjects. These would allow us to establish whether the observed DNA methylation changes are associated with gene expression and its regulation, and would thus provide a much deeper understanding of how migration exerts its biological effects at different cellular complexity levels. Despite these limitations, we think this work exemplifies the promising potential of EWAS approaches to elucidate complex and subtle effects of migration at the population level.

References

1. G Campanella, S Polidoro, C Di Gaetano, G Fiorito, S Guarrera, V Krogh, D Palli, S Panico, C Sacerdote, R Tumino, P Elliott, G Matullo, M Chadeau-

- Hyam, and P Vineis. 'Epigenetic signatures of internal migration in Italy'. *International Journal of Epidemiology* (2014).
2. M Geddes, D Balzi, E Buiatti, M Khlat, and D Parkin. 'Cancer in Italian migrants'. *Cancer Causes & Control* **2.2** (1991), 133–140.
 3. M McCredie. 'Cancer epidemiology in migrant populations'. *Genes and environment in cancer*. Ed. by M Schwab, H Rabes, K Munk, and H P Hofschneider. **154**. Recent Results in Cancer Research. Springer, 1998, 298–305.
 4. J H Lassetter and L C Callister. 'The impact of migration on the health of voluntary migrants in Western societies: a review of the literature'. *Journal of Transcultural Nursing* **20.1** (2009), 93–104.
 5. M G Marmot, S L Syme, A Kagan, H Kato, J B Cohen, and J Belsky. 'Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: prevalence of coronary and hypertensive heart disease and associated risk factors'. *American Journal of Epidemiology* **102.6** (1975), 514–525.
 6. M Arnold, O Razum, and J-W Coebergh. 'Cancer risk diversity in non-Western migrants to Europe: an overview of the literature'. *European Journal of Cancer* **46.14** (2010), 2647–2659.
 7. A P Bird. 'DNA methylation patterns and epigenetic memory'. *Genes & Development* **16.1** (2002), 6–21.
 8. W Reik. 'Stability and flexibility of epigenetic gene regulation in mammalian development'. *Nature* **447** (2007), 425–432.
 9. S E Johnstone and S B Baylin. 'Stress and the epigenetic landscape: a link to the pathobiology of human diseases?' *Nature Reviews Genetics* **11** (2010), 806–812.

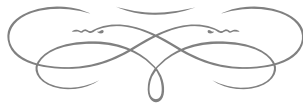
10. R Feil and M F Fraga. 'Epigenetics and the environment: emerging patterns and implications'. *Nature Reviews Genetics* **13** (2012), 97–109.
11. A P Bird and A P Wolffe. 'Methylation-induced repression – belts, braces, and chromatin'. *Cell* **99.5** (1999), 451–454.
12. R J Klose and A P Bird. 'Genomic DNA methylation: the mark and its mediators'. *Trends in Biochemical Sciences* **31.2** (2006), 89–97.
13. A Meissner, T S Mikkelsen, H Gu, M Wernig, J Hanna, A Sivachenko, X Zhang, B E Bernstein, C Nusbaum, D B Jaffe, A Gnirke, R Jaenisch, and E S Lander. 'Genome-scale DNA methylation maps of pluripotent and differentiated cells'. *Nature* **454** (2008), 766–770.
14. S Zeilinger, B Kühnel, N Klopp, H Baurecht, A Kleinschmidt, C Gieger, S Weidinger, E Lattka, J Adamski, A Peters, K Strauch, M Waldenberger, and T Illig. 'Tobacco smoking leads to extensive genome-wide changes in DNA methylation'. *PLOS ONE* **8.5** (2013), e63812.
15. G Egger, G Liang, A Aparicio, and P A Jones. 'Epigenetics in human disease and prospects for epigenetic therapy'. *Nature* **429** (2004), 457–463.
16. W Haenszel and M Kurihara. 'Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States'. *Journal of the National Cancer Institute* **40.1** (1968), 43–68.
17. R G Ziegler, R N Hoover, M C Pike, A Hildesheim, A M Y Nomura, D W West, A H Wu-Williams, L N Kolonel, P L Horn-Ross, J F Rosenthal, and M B Hyer. 'Migration patterns and breast cancer risk in Asian-American women'. *Journal of the National Cancer Institute* **85.22** (1993), 1819–1827.
18. C Bonifazi and F Heins. 'Long-term trends of internal migration in Italy'. *International Journal of Population Geography* **6.2** (2000), 111–131.

19. D Rasulo, T Spadea, R Onorati, and G Costa. 'The impact of migration in all-cause mortality: the Turin Longitudinal Study, 1971–2005'. *Social Science & Medicine* **74.6** (2012), 897–906.
20. P Francalacci, L Morelli, A Angius, R Berutti, F Reinier, R Atzeni, R Pili, F Busonero, A Maschio, I Zara, D Sanna, A Useli, M F Urru, M Marcelli, R Cusano, M Oppo, M Zoledziwska, M Pitzalis, F Deidda, E Porcu, et al. 'Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny'. *Science* **341.6145** (2013), 565–569.
21. P Vineis, R Tumino, S Sieri, S Salvini, C Saieva, C Sacerdote, S Panico, D Palli, V Pala, G Masala, V Krogh, M C Giurdanella, R Galasso, G Frasca, L Fiorini, A Decarli, M Ceroti, E Celentano, and F Berrino. 'A molecular epidemiology project on diet and cancer: the EPIC-Italy prospective study. Design and baseline characteristics of participants'. *Tumori* **89.6** (2003), 586–593.
22. P Pisani, F Faggiano, V Krogh, D Palli, P Vineis, and F Berrino. 'Relative validity and reproducibility of a food frequency dietary questionnaire for use in the Italian EPIC centres'. *International Journal of Epidemiology* **26**.Suppl. 1 (1997), S152–S160.
23. G Fiorito, S Guarrera, C Valle, F Ricceri, A Russo, S Grioni, A Mattiello, C Di Gaetano, F Rosa, F Modica, L Iacoviello, G Frasca, R Tumino, V Krogh, S Panico, P Vineis, C Sacerdote, and G Matullo. 'B vitamins intake, DNA methylation of one-carbon metabolism and homocysteine pathway genes and myocardial infarction risk: the EPICOR study'. *Nutrition, Metabolism and Cardiovascular Diseases* **24.5** (2014), 483–488.
24. S L P Ferrari and F Cribari-Neto. 'Beta regression for modelling rates and proportions'. *Journal of Applied Statistics* **31.7** (2004), 799–815.

25. P Vineis, F Faggiano, E Riboli, F Berrino, P Pisani, and P Crosignani. 'Dietary habits, internal migration and social class in a sample of a northern Italian population'. *Tumori* **78.4** (1992), 235–238.
26. B Langmead and S L Salzberg. 'Fast gapped-read alignment with Bowtie 2'. *Nature Methods* **9** (2012), 357–359.
27. M E Price, A Cotton, L Lam, P Farre, E Emberly, C Brown, W Robinson, and M Kobor. 'Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array'. *Epigenetics & Chromatin* **6.1** (2013), 4.
28. The 1000 Genomes Project Consortium. 'An integrated map of genetic variation from 1,092 human genomes'. *Nature Methods* **491** (2012), 56–65.
29. L E Hall, S E Mitchell, and R J O'Neill. 'Pericentric and centromeric transcription: a perfect balance required'. *Chromosome Research* **20.5** (2012), 535–546.
30. R Valgardsdottir, I Chiodi, M Giordano, A Rossi, S Bazzini, C Ghigna, S Riva, and G Biamonti. 'Transcription of Satellite III non-coding RNAs is a general stress response in human cells'. *Nucleic Acids Research* **36.2** (2008), 423–434.
31. R A Waterland and K B Michels. 'Epigenetic epidemiology of the developmental origins hypothesis'. *Annual Review of Nutrition* **27** (2007), 363–388.
32. C W Kuzawa and E A Quinn. 'Developmental origins of adult function and health: evolutionary hypotheses'. *Annual Review of Anthropology* **38** (2009), 131–147.



This page intentionally left blank.



Chapter 4

Epigenome-wide study of adiposity and future risk of obesity-related diseases

An abridged version of this chapter has been submitted as a paper to the *Journal of the American Medical Association*.

§ 4.1 Introduction

Obesity is a growing global public health challenge with an estimated 1.9 billion adults classified as overweight or obese worldwide [1]. Obesity is a risk factor for a variety of chronic diseases including type 2 diabetes, cardiovascular disease (CVD), and a number of malignancies such as post-menopausal breast, colorectal, liver, kidney, and endometrial cancer [2, 3]. Obesity-associated metabolic abnormalities such hyperinsulinaemia, hyperglycaemia, and dyslipidaemia have also been associated with CVD and a number of malignancies [4]. However, the contribution of obesity and its physiological consequences to the development of these diseases remains poorly understood.

The aetiology of obesity is multifactorial, with both genetic and environmental determinants. Numerous genetic variants linked to obesity have been discovered by large meta-analyses of genome-wide association studies [5–7]; however, taken together, these variants explain only a modest fraction of the phenotypic variation, and do not appear to directly modulate phenotypes that are thought to be involved in obesity-related complications, making them unlikely to directly affect risk of these conditions [8].

We hypothesised that potentially reversible epigenetic changes, in particular altered DNA methylation patterns, may lie at the intersection of obesity and obesity-related diseases, and could offer clues as to underlying biological mechanisms. To test this hypothesis, we obtained genome-wide DNA methylation levels in peripheral blood leukocytes (PBLs) from over 1,900 individuals from four prospective cohorts in Italy, the Netherlands, Norway, and Sweden. First, we conducted a series of association studies with adiposity measures including body mass index (BMI), waist circumference, waist-to-hip ratio (WHR), and waist-to-height ratio (WHtR), and defined a set of informative CpG loci. These

were functionally characterised by relating DNA methylation levels to transcriptional activity of neighbouring genes in PBLs. Furthermore, concurrent and independent associations with all investigated adiposity measures were replicated using bisulphite pyrosequencing in an independent Italian sample. Second, we examined associations between DNA methylation levels at these loci and fasting blood concentrations of several biomarkers of lipid and glucose metabolism, including total and HDL cholesterol, triglycerides, glucose, and insulin. Finally, we investigated whether DNA methylation levels at these loci could predict (alone or in combination with adiposity measures) future development of obesity-related diseases including post-menopausal breast cancer, colorectal cancer, and myocardial infarction.

§ 4.2 Methods

4.2.1 Study population and sample selection

Discovery set

The study included participants from four large population-based cohort studies: the Italian and Dutch components of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy [9] and EPIC-Netherlands [10]), the Norwegian Women and Cancer Study (NOWAC) [11], and the Northern Sweden Health and Disease Study (NSHDS) [12]. All participants gave written informed consent, and study protocols were approved by the ethical review boards of the relevant institutions and local participating centres.

A total of 2,560 genome-wide DNA methylation profiles from PBLs were available as part of:

- (i) three separate prospective case-control studies nested within EPIC-Italy on breast cancer (BC, $N = 332$), colorectal cancer (CRC, $N = 338$), and myocardial infarction (EPICOR [13], $N = 552$);
- (ii) two separate prospective case-control studies nested within EPIC-Italy and NSHDS on breast cancer and non-Hodgkin lymphoma (EPIC-Italy and NSHDS components of the EnviroGenoMarkers project, $N = 241$ and 400 , respectively);
- (iii) a longitudinal study in healthy women nested within EPIC-Netherlands ($N = 148$ pairs of baseline and follow-up samples, of which $N = 66$ at five years and $N = 82$ at ten years from baseline);
- (iv) a prospective case-control study nested within NOWAC on breast cancer ($N = 384$).

A single profile was retained (on the basis of technical quality) for the few EPIC participants included in more than one study (separately at each time point for EPIC-Netherlands), or whose samples had been analysed multiple times for quality control purposes. Participants from prospective nested case-control studies who developed the relevant condition less than one year after blood draw (to guard against reverse causation), or who developed any kind of haematological malignancy at any time after enrolment, were excluded, as were those with no anthropometric data. Characteristics of the 1,941 participants (588 men and 1,353 women) included in the study are reported in table 4.1.

Replication set

The replication sample ($N = 384$) was drawn from EPIC-Italy with the following exclusion criteria: (i) prevalent dyslipidaemia or diabetes (self-reported);

	EPIC-Italy			EnviroGenoMarkers			
	BC	CRC	EPICOR	EPIC-Italy	NSHDS	EPIC Netherlands	NOWAC
Samples analysed	332	338	552	241	400	313	384
Samples excluded	31	35	67	87	176	25	54
Failed QC	1	2	5	3	4	6	4
Duplicates	3	12	29	2	—	16	—
Diagnosis ≤ 1 year	21	10	15	0	0	—	43
Haematological malignancy	0	1	3	74	158	—	0
Incomplete information	6	10	15	8	14	3	7
Participants	301	303	485	154	224	144	330
Men	0 (0%)	161 (53%)	326 (67%)	23 (15%)	78 (35%)	0 (0%)	0 (0%)
Women	301 (100%)	142 (47%)	159 (33%)	131 (85%)	146 (65%)	144 (100%)	330 (100%)
Controls	149 (49%)	132 (44%)	195 (40%)	101 (66%)	182 (81%)	—	181 (55%)
Age (years)	54.12 (10.35)	56.28 (9.36)	51.95 (11.77)	52.32 (12.37)	50.13 (10.13)	57.73 (7.13)	55.60 (6.40)
BMI (kg/m ²)	25.10 (4.57)	25.90 (5.06)	26.42 (4.34)	25.35 (5.07)	25.34 (4.74)	25.73 (4.97)	25.10 (4.73)
< 18.5	3 (1%)	1 (0%)	2 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)
[18.5, 25)	141 (47%)	124 (41%)	149 (31%)	72 (47%)	106 (47%)	60 (42%)	158 (48%)
[25, 30)	111 (37%)	131 (43%)	255 (53%)	68 (44%)	93 (42%)	59 (41%)	126 (38%)
[30, 35)	36 (12%)	38 (13%)	61 (13%)	12 (8%)	19 (8%)	20 (14%)	37 (11%)
[35, 40)	6 (2%)	7 (2%)	14 (3%)	2 (1%)	5 (2%)	3 (2%)	7 (2%)
≥ 40	4 (1%)	2 (1%)	4 (1%)	0 (0%)	1 (0%)	2 (1%)	0 (0%)
Waist circumference (cm)	80.00 (13.00)	88.70 (19.10)	92.00 (14.50)	80.35 (14.53)	—	87.50 (11.69)	—
WHR	0.79 (0.08)	0.88 (0.15)	0.90 (0.11)	0.81 (0.10)	—	0.85 (0.07)	—
WHtR	0.50 (0.09)	0.54 (0.09)	0.55 (0.08)	0.50 (0.08)	—	0.54 (0.08)	—

Table 4.1: Sample selection and characteristics of participants in the discovery set (at baseline for EPIC-Netherlands). Counts (percentages) and medians (IQR) are reported for categorical and continuous variables, respectively.

(ii) diagnosis of any non-haematological malignancy less than five years after blood draw; (iii) diagnosis of any haematological malignancy at any time after enrolment. Adiposity was assessed using a combination of BMI and waist circumference [14]; 288 (75%) participants were randomly selected from the obese group, with the remaining 96 (25%) participants sampled from the reference, normal-weight group. Samples that did not pass quality control procedures, as well as participants with incomplete anthropometric information, were excluded from further analyses, leaving a total of 358 participants (127 men and 231 women) whose characteristics are reported in table 4.2.

Adiposity measures

Four adiposity measures were considered: BMI (kg/m^2), waist circumference (cm), WHR, and WHtR. Anthropometric measurements concurrent with blood sampling were obtained by a trained nurse at enrolment for all EPIC and NSHDS participants; they were self-reported for NOWAC. Waist circumference and related measures were only available for EPIC participants.

4.2.2 Laboratory analyses and data pre-processing

DNA methylation

Genome-wide DNA methylation analyses were carried out using the Illumina Infinium[®] HumanMethylation450 (HM450) BeadChip. Genomic DNA extracted from PBLs was bisulphite-converted using the Zymo Research EZ-96 DNA Methylation-Gold[™] Kit, and hybridised to HM450 BeadChips according to the manufacturer's protocol. Micro-arrays were subsequently scanned using the Illumina HiScanSQ system, and raw intensity data were exported from Illumina

EPIC-Italy	
Samples analysed	384
Samples excluded	26
Failed QC	25
Incomplete information	1
Participants	358
Men	127 (35%)
Women	231 (65%)
Age (years)	49.25 (12.10)
BMI (kg/m ²)	34.08 (14.47)
< 18.5	0 (0%)
[18.5, 25)	95 (27%)
[25, 30)	34 (9%)
[30, 35)	60 (17%)
[35, 40)	105 (29%)
≥ 40	64 (18%)
Waist circumference (cm)	101.75 (24.00)
WHR	0.86 (0.11)
WHtR	0.62 (0.16)

Table 4.2: Sample selection and characteristics of participants in the replication set. Counts (percentages) and medians (IQR) are reported for categorical and continuous variables, respectively.

GenomeStudio (version 2011.1). Control probes included in the micro-array were used to assess bisulphite conversion efficiency and exclude lower-quality samples from further analyses. Data pre-processing was carried out using in-house software written for the R statistical computing environment. In particular, for each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction and dye bias correction (for probes using the Infinium[®] II design) were also performed. The subset of 470,870 probes targeting autosomal CpG loci was selected for further analyses. DNA methylation levels at each locus were assumed proportional to the ratio of intensities arising from methylated cytosines over total intensities.

DNA methylation levels in the replication sample were assessed using bisulphite pyrosequencing. Primers were designed using QIAGEN PyroMark Assay Design (version 2.0). Preliminary PCR reactions were performed using the following cycling protocol: 95°C for 10 minutes, followed by 45 cycles of denaturation at 95°C for 30 seconds, annealing at 58°C for one minute, and extension at 72°C for one minute; a final extension was performed at 72°C for 10 minutes. All samples were analysed two or three times (following bisulphite conversion) to assess and account for technical variability.

Gene expression

Samples from EnviroGenoMarkers and NOWAC additionally underwent genome-wide gene expression profiling using either the Agilent Whole Human Genome 4×44K Microarray (EnviroGenoMarkers), or the Illumina HumanWG-6 and HumanHT-12 BeadChips (NOWAC), as previously reported [15, 16]. Samples that failed quality control procedures were excluded from further analyses, leav-

ing a total of 672 genome-wide gene expression profiles ($N = 353$ in EnviroGenoMarkers, and $N = 319$ in NOWAC) with associated DNA methylation data. For both studies, pre-processing steps included within- and between-array normalisation, followed by removal of unannotated and non-specific probes, leaving a total of 14,698 transcripts for EnviroGenoMarkers, and 14,980 transcripts for NOWAC.

Blood biomarkers

Fasting blood levels of circulating total and HDL cholesterol, triglycerides, glucose, and insulin were available for a subset of non-diabetic EPIC-Italy participants ($N = 563$) for whom DNA methylation profiles were also available. Differential blood counts were available for a subset of EPIC-Netherlands participants at baseline ($N = 56$) for whom DNA methylation profiles were also available.

4.2.3 Statistical analyses

Discovery set

Separate analyses were conducted for each study in the discovery set and each adiposity measure (if available). DNA methylation levels at each assayed CpG locus were modelled as dependent variable in a generalised linear model with beta-distributed response using the parametrisation of Ferrari et al. [17]. DNA methylation levels of paired samples from EPIC-Netherlands were analysed using linear mixed models with a participant random effect after logit transformation. All models were adjusted for micro-array and position on the micro-array, sex (if applicable), age at blood draw, and case-control status (if applicable); analyses of waist circumference and related measures were additionally conducted

with adjustment for BMI. For each adiposity measure, estimated effect sizes (on the logit scale) and associated standard errors were obtained, and meta-analyses were performed across studies using fixed effects models with inverse-variance weighting. For each CpG locus and each adiposity measure, effect sizes, standard errors, and corresponding p -values for the meta-analysis were obtained in addition to the I^2 heterogeneity statistic measuring the percentage of variance due to differences between studies rather than sampling error [18]. Associations were declared statistically significant based on the p -value for the effect size combined across studies, taking into account multiple comparisons by considering a 5% Bonferroni-corrected significance threshold $\alpha = 0.05/470870 \approx 1.1 \times 10^{-7}$.

Subgroup analyses included stratification by sex and case-control status; multiple comparisons were taken into account by adopting a 5% significance threshold after Bonferroni correction for the number of informative CpG loci identified in the meta-analysis for each adiposity measure.

Replication set

DNA methylation measurements in the replication sample were analysed using generalised linear mixed models with a participant random effect after logit transformation, using the same parametrisation and confounding variables as described above.

Gene expression

Informative CpG loci were further characterised by relating DNA methylation levels to transcriptional activity of neighbouring (*cis*) genes using a $\log_2$ -linear model adjusted for technical confounders, cohort and sex (EnviroGenoMarkers data only), age, case-control status, and BMI; linear mixed models were used

for the EnviroGenoMarkers data, as previously reported [19]. Associations were declared statistically significant at a 5% significance threshold. Associations of transcription levels of the identified genes with BMI were also investigated using similarly parametrised models.

Blood biomarkers

Fasting levels of circulating total and HDL cholesterol, triglycerides, glucose, and insulin, as well as two indicators of cardiometabolic risk (total-to-HDL cholesterol ratio and triglycerides-to-HDL cholesterol ratio) [20], and two indicators of β -cell function and insulin resistance (HOMA-%B and HOMAIR [21]), were analysed in relation to DNA methylation levels at informative CpG loci using a log-linear model adjusted for sex, age, case-control status (separately for cancers and myocardial infarction), BMI, and WHR. Given the interdependency among these phenotypes, multiple comparisons were taken into account by adopting a 5% significance threshold after Bonferroni correction for the total number of informative CpG loci identified in the meta-analysis across adiposity measures.

Differential blood counts were analysed in relation to DNA methylation levels at informative CpG loci using linear models adjusted for age. Multiple comparisons were taken into account by adopting a 5% significance threshold after Bonferroni correction for the total number of informative CpG loci identified in the meta-analysis across adiposity measures.

Future risk of obesity-related diseases

Case-control analyses for post-menopausal breast cancer included only women who had reached menopause at enrolment, or who were aged 60 years or older at diagnosis.

For each disease separately, DNA methylation levels at each informative CpG locus identified in the meta-analysis were entered in logistic regression models for (incident) post-menopausal breast cancer, colorectal cancer, and myocardial infarction, adjusted for age, sex (except for post-menopausal breast cancer), BMI, WHR, and established disease-specific risk factors, as detailed below:

Post-menopausal breast cancer

Alcohol intake (continuous, g/day); contraceptive use (binary); hormone replacement therapy (HRT) use (binary); smoking status (categorical: never, former, current smoker); physical activity (categorical: five levels).

Colorectal cancer

Alcohol intake (continuous, g/day); processed meat intake (continuous, g/day); fibre intake (continuous, g/day); smoking status (categorical: never, former, current smoker); physical activity (categorical: five levels).

Myocardial infarction

Smoking status (categorical: never, former, current smoker); physical activity (categorical: five levels); total-to-HDL cholesterol ratio (continuous, with interaction term for fasting status); triglycerides-to-HDL cholesterol ratio (continuous, with interaction term for fasting status).

To maximise power despite limited sample sizes, DNA methylation levels at all informative CpG loci identified in the meta-analysis across adiposity measures were subjected to principal component analysis (PCA), and multiple comparisons were taken into account by adopting a 5% significance threshold after Bonferroni correction for the number of PCs explaining more than 95% of the variance. The discrimination performance of the resulting models was characterised by the area under the receiver operating characteristic curve (AUC of the ROC).

§ 4.3 Results

4.3.1 Associations with adiposity measures

The meta-analysis of studies in the discovery set identified 40 informative CpG loci associated with one or more adiposity measures (tables 4.3 to 4.6). In particular, 26 statistically significant associations were found with BMI (p -values ranging between 9.89×10^{-8} and 2.76×10^{-18}), of which 18 were characterised by low to moderate heterogeneity ($I^2 < 50\%$). Two associations were found on the same gene (*ABCG1*, cg27243685 and cg06500161). Analyses for waist circumference and derived measures identified fewer associations ($N = 12$, 9, and 12 for waist circumference, WHR, and WHtR, respectively); 10 associations with waist circumference, five associations with WHR, and all 12 associations with WHtR exhibited low to moderate heterogeneity ($I^2 < 50\%$). Two associations with WHR were found on the same gene (*PHGDH*, cg16246545 and cg14476101). After adjustment for BMI, only four associations (all with WHR) retained statistical significance: the corresponding CpG loci were cg16246545 (*PHGDH*, $\beta = -0.48$, p -value 7.84×10^{-9} , $I^2 = 0.00\%$), cg14476101 (*PHGDH*, $\beta = -0.78$, p -value 1.10×10^{-11} , $I^2 = 0.00\%$), cg19472611 (*PRKAR2A*, $\beta = -0.79$, p -value 1.79×10^{-8} , $I^2 = 51.97\%$), and cg06730756 (*DENND3*, $\beta = -1.72$, p -value 7.57×10^{-9} , $I^2 = 80.69\%$).

DNA methylation levels at nine CpG loci were consistently associated with BMI and at least one among waist circumference and derived measures. Two associations were found with BMI and waist circumference: cg01538605 (intergenic, inversely associated) and cg19750657 (*UFM1*, directly); one with BMI and WHtR: cg06164260 (*BCL6*, inversely); six with BMI, waist circumference, and WHtR: cg09831562 (*SOX2OT*, directly), cg03619256 (*ATXN7L1*, inversely), cg11024682 (*SREBF1*, directly), cg24174557 (*TMEM49*, inversely), and cg26950531 (*DPF1*, in-

versely). Finally, DNA methylation levels at a single CpG locus (cg06500161, *ABCG1*) were directly associated with all adiposity measures (p -values ranging between 3.07×10^{-11} and 2.76×10^{-18}).

Subgroup analyses

Upon stratification by sex, all identified associations were replicated among women. Among men, 15 associations with BMI did not reach statistical significance; however, seven were nominally statistically significant, with effect size estimates similar to those obtained in the main analysis. DNA methylation levels at the eight remaining informative CpG loci were not consistently associated with BMI in men. Six associations with waist circumference, seven with WHR, and six with WHtR did not reach statistical significance in the subgroup analysis; nonetheless, only one association with waist circumference, three with WHR, and two with WHtR were strongly attenuated and did not reach nominal statistical significance (tables 4.7 to 4.10).

One association with BMI, two with WHR, and one with WHtR lost statistical significance after stratification by case-control status. In particular, the association of DNA methylation levels at cg16797901 (*DOCK4*) and BMI was strongly attenuated (p -value 6.10×10^{-3}), as were those involving cg06730756 (*DENND3*) and WHR (p -value 0.66), cg25799109 (*ARHGEF3*) and WHR (p -value 0.02), and cg07037944 (*DAPK2*) and WHtR (p -value 0.01); nonetheless, the direction of the associations was preserved in all cases.

Replication and dose-response assessment

The associations with DNA methylation levels at cg06500161 (*ABCG1*) were confirmed in the replication experiment (p -values 5.61×10^{-10} , 8.17×10^{-11} , 7.68×10^{-9} ,

CpG locus	Chr	Location	Gene(s)	N	β	$SE(\beta)$	p -value	I^2
cg23172671	1	203,482,523		2,072	9.46×10^{-3}	1.76×10^{-3}	7.80×10^{-8}	19.67%
cg16258657	3	29,338,542	<i>RBMS3</i>	2,078	-8.31×10^{-3}	1.54×10^{-3}	6.59×10^{-8}	0.00%
cg15059608	3	62,305,003	<i>C3orf14</i>	2,080	8.67×10^{-3}	1.58×10^{-3}	3.94×10^{-8}	30.98%
cg14492061	3	63,637,129	<i>SNTN</i>	2,071	-9.42×10^{-3}	1.66×10^{-3}	1.36×10^{-8}	57.11%
cg18098839	3	167,742,700	<i>GOLIM4</i>	2,085	-6.53×10^{-3}	1.16×10^{-3}	1.72×10^{-8}	0.00%
cg09831562	3	181,327,125	<i>SOX2OT</i>	2,053	1.33×10^{-2}	2.13×10^{-3}	3.94×10^{-10}	0.00%
cg06164260	3	187,454,439	<i>BCL6</i>	2,085	-4.03×10^{-3}	7.09×10^{-4}	1.27×10^{-8}	14.59%
cg11661512	5	170,864,692	<i>FGF18</i>	2,083	-4.75×10^{-3}	8.12×10^{-4}	5.03×10^{-9}	60.48%
cg01538605	6	169,285,957		2,085	-5.85×10^{-3}	9.84×10^{-4}	2.79×10^{-9}	0.00%
cg06087826	6	170,582,184		2,083	-5.78×10^{-3}	1.05×10^{-3}	3.78×10^{-8}	11.59%
cg13286614	7	96,655,501	<i>DLX5</i>	2,084	7.28×10^{-3}	1.28×10^{-3}	1.20×10^{-8}	54.63%
cg03619256	7	105,515,219	<i>ATXN7L1</i>	2,085	-5.49×10^{-3}	8.24×10^{-4}	2.79×10^{-11}	0.00%
cg16797901	7	111,513,558	<i>DOCK4</i>	2,061	-9.32×10^{-3}	1.73×10^{-3}	6.90×10^{-8}	29.37%
cg19249811	10	29,747,315	<i>SVIL</i>	1,982	-3.23×10^{-2}	3.77×10^{-3}	1.00×10^{-17}	94.82%
cg17061862	11	9,590,431		2,084	-7.23×10^{-3}	1.18×10^{-3}	9.16×10^{-10}	0.00%
cg00574958	11	68,607,622	<i>CPT1A</i>	2,083	-7.62×10^{-3}	1.41×10^{-3}	6.15×10^{-8}	0.00%
cg27117792	12	102,330,180		2,083	-9.90×10^{-3}	1.81×10^{-3}	4.29×10^{-8}	53.99%
cg19750657	13	38,935,967	<i>UFM1</i>	2,083	8.25×10^{-3}	1.23×10^{-3}	2.15×10^{-11}	0.00%
cg02119938	15	78,505,051	<i>ACSBG1</i>	2,071	-1.33×10^{-2}	2.48×10^{-3}	9.07×10^{-8}	54.49%
cg01115923	16	57,793,728	<i>KIFC3</i>	2,085	-6.07×10^{-3}	1.00×10^{-3}	1.42×10^{-9}	20.33%
cg11024682	17	17,730,094	<i>SREBF1</i>	2,085	5.36×10^{-3}	7.64×10^{-4}	2.29×10^{-12}	21.98%
cg24174557	17	57,903,544	<i>TMEM49</i>	2,077	-8.36×10^{-3}	1.49×10^{-3}	2.14×10^{-8}	0.00%
cg04927537	17	76,976,091	<i>LGALS3BP</i>	2,085	6.71×10^{-3}	1.26×10^{-3}	9.89×10^{-8}	54.38%
cg26950531	19	38,704,515	<i>DPF1</i>	2,071	-1.44×10^{-2}	2.06×10^{-3}	3.40×10^{-12}	0.00%
cg27243685	21	43,642,366	<i>ABCG1</i>	2,085	6.88×10^{-3}	1.18×10^{-3}	5.04×10^{-9}	48.25%
cg06500161	21	43,656,587	<i>ABCG1</i>	2,085	6.88×10^{-3}	7.89×10^{-4}	2.76×10^{-18}	53.62%

Table 4.3: Significant associations with BMI in the discovery set. The table reports the total number of samples N included in the meta-analysis, the beta regression model coefficient β and the associated standard error (SE), the p -value for $\mathcal{H}_0 : \beta = 0$, and the I^2 heterogeneity statistic.

CpG locus	Chr	Location	Gene(s)	N	β	$SE(\beta)$	p -value	I^2
cg09831562	3	181,327,125	<i>SOX2OT</i>	1,501	6.19×10^{-3}	1.02×10^{-3}	1.13×10^{-9}	0.00%
cg01538605	6	169,285,957		1,531	-2.47×10^{-3}	4.63×10^{-4}	1.04×10^{-7}	0.00%
cg03619256	7	105,515,219	<i>ATXN7L1</i>	1,531	-2.24×10^{-3}	3.91×10^{-4}	9.54×10^{-9}	0.00%
cg11376147	11	57,261,198	<i>SLC43A1</i>	1,530	-2.20×10^{-3}	3.90×10^{-4}	1.53×10^{-8}	0.00%
cg19750657	13	38,935,967	<i>UFM1</i>	1,530	3.38×10^{-3}	5.92×10^{-4}	1.16×10^{-8}	59.96%
cg07037944	15	64,290,807	<i>DAPK2</i>	1,524	-2.73×10^{-3}	4.63×10^{-4}	3.65×10^{-9}	10.86%
cg04583842	16	88,103,117	<i>BANP</i>	1,515	3.16×10^{-3}	5.84×10^{-4}	5.90×10^{-8}	0.00%
cg11024682	17	17,730,094	<i>SREBF1</i>	1,531	1.99×10^{-3}	3.59×10^{-4}	3.01×10^{-8}	0.00%
cg24174557	17	57,903,544	<i>TMEM49</i>	1,523	-3.94×10^{-3}	7.13×10^{-4}	3.26×10^{-8}	0.00%
cg26950531	19	38,704,515	<i>DPF1</i>	1,520	-5.60×10^{-3}	1.00×10^{-3}	2.44×10^{-8}	50.19%
cg06500161	21	43,656,587	<i>ABCG1</i>	1,531	3.17×10^{-3}	3.68×10^{-4}	8.36×10^{-18}	49.58%
cg21307484	22	37,546,220	<i>IL2RB</i>	1,531	2.35×10^{-3}	4.19×10^{-4}	2.07×10^{-8}	0.00%

Table 4.4: Significant associations with waist circumference in the discovery set. The table reports the total number of samples N included in the meta-analysis, the beta regression model coefficient β and the associated standard error (SE), the p -value for $\mathcal{H}_0 : \beta = 0$, and the I^2 heterogeneity statistic.

CpG locus	Chr	Location	Gene(s)	N	β	$SE(\beta)$	p -value	I^2
cg17901584	1	55,353,706	<i>DHCR24</i>	1,531	-5.79×10^{-1}	1.07×10^{-1}	5.53×10^{-8}	0.00%
cg16246545	1	120,255,941	<i>PHGDH</i>	1,531	-4.39×10^{-1}	7.49×10^{-2}	4.44×10^{-9}	0.00%
cg14476101	1	120,255,992	<i>PHGDH</i>	1,527	-6.83×10^{-1}	1.02×10^{-1}	2.38×10^{-11}	0.00%
cg19472611	3	48,885,488	<i>PRKAR2A</i>	1,513	-6.84×10^{-1}	1.27×10^{-1}	6.70×10^{-8}	12.56%
cg25799109	3	57,102,900	<i>ARHGEF3, SPATA12</i>	1,519	-6.08×10^{-1}	1.11×10^{-1}	4.14×10^{-8}	79.04%
cg20898587	7	51,969,244		1,530	-3.08×10^{-1}	5.64×10^{-2}	4.60×10^{-8}	63.03%
cg06730756	8	142,198,993	<i>DENND3</i>	1,478	-1.41	2.62×10^{-1}	8.25×10^{-8}	84.90%
cg11376147	11	57,261,198	<i>SLC43A1</i>	1,530	-3.32×10^{-1}	5.91×10^{-2}	1.87×10^{-8}	41.25%
cg06500161	21	43,656,587	<i>ABCG1</i>	1,531	3.65×10^{-1}	5.49×10^{-2}	3.07×10^{-11}	62.98%

Table 4.5: Significant associations with waist-to-hip ratio in the discovery set. The table reports the total number of samples N included in the meta-analysis, the beta regression model coefficient β and the associated standard error (SE), the p -value for $\mathcal{H}_0 : \beta = 0$, and the I^2 heterogeneity statistic.

CpG locus	Chr	Location	Gene(s)	N	β	$SE(\beta)$	p -value	I^2
cg12593793	1	156,074,135		1,530	-4.37×10^{-1}	8.21×10^{-2}	1.01×10^{-7}	0.00%
cg09831562	3	181,327,125	<i>SOX2OT</i>	1,501	9.34×10^{-1}	1.61×10^{-1}	6.69×10^{-9}	0.00%
cg06164260	3	187,454,439	<i>BCL6</i>	1,531	-2.75×10^{-1}	5.03×10^{-2}	4.59×10^{-8}	0.00%
cg03619256	7	105,515,219	<i>ATXN7L1</i>	1,531	-3.79×10^{-1}	6.16×10^{-2}	7.55×10^{-10}	28.48%
cg11376147	11	57,261,198	<i>SLC43A1</i>	1,530	-3.53×10^{-1}	6.16×10^{-2}	9.43×10^{-9}	0.00%
cg07037944	15	64,290,807	<i>DAPK2</i>	1,524	-4.09×10^{-1}	7.35×10^{-2}	2.58×10^{-8}	4.90%
cg11024682	17	17,730,094	<i>SREBF1</i>	1,531	3.10×10^{-1}	5.68×10^{-2}	4.77×10^{-8}	11.53%
cg24174557	17	57,903,544	<i>TMEM49</i>	1,523	-6.29×10^{-1}	1.12×10^{-1}	2.29×10^{-8}	37.14%
cg05091997	17	60,897,721		1,513	-9.63×10^{-1}	1.75×10^{-1}	3.97×10^{-8}	0.00%
cg26950531	19	38,704,515	<i>DPF1</i>	1,520	-1.04	1.59×10^{-1}	7.51×10^{-11}	17.36%
cg26470501	19	45,252,955	<i>BCL3</i>	1,531	-3.21×10^{-1}	5.49×10^{-2}	5.10×10^{-9}	0.00%
cg06500161	21	43,656,587	<i>ABCG1</i>	1,531	5.07×10^{-1}	5.83×10^{-2}	3.32×10^{-18}	19.08%

Table 4.6: Significant associations with waist-to-height ratio in the discovery set. The table reports the total number of samples N included in the meta-analysis, the beta regression model coefficient β and the associated standard error (SE), the p -value for $\mathcal{H}_0 : \beta = 0$, and the I^2 heterogeneity statistic.

CpG locus	Men				Women			
	N	β	$SE(\beta)$	p -value	N	β	$SE(\beta)$	p -value
cg23172671	565	7.42×10^{-3}	3.75×10^{-3}	4.75×10^{-2}	1,484	1.12×10^{-2}	2.06×10^{-3}	5.68×10^{-8}
cg16258657	564	-1.66×10^{-2}	3.30×10^{-3}	4.83×10^{-7}	1,491	-7.88×10^{-3}	1.74×10^{-3}	5.78×10^{-6}
cg15059608	565	4.34×10^{-3}	3.29×10^{-3}	1.87×10^{-1}	1,492	8.23×10^{-3}	1.86×10^{-3}	9.94×10^{-6}
cg14492061	561	-1.78×10^{-2}	3.75×10^{-3}	2.07×10^{-6}	1,487	-9.69×10^{-3}	1.88×10^{-3}	2.38×10^{-7}
cg18098839	565	-1.02×10^{-2}	2.54×10^{-3}	6.10×10^{-5}	1,497	-7.09×10^{-3}	1.33×10^{-3}	9.72×10^{-8}
cg09831562	564	1.41×10^{-2}	5.37×10^{-3}	8.67×10^{-3}	1,466	1.40×10^{-2}	2.36×10^{-3}	2.85×10^{-9}
cg06164260	565	-3.76×10^{-3}	1.60×10^{-3}	1.87×10^{-2}	1,497	-4.29×10^{-3}	7.98×10^{-4}	7.68×10^{-8}
cg11661512	564	-1.89×10^{-3}	1.69×10^{-3}	2.63×10^{-1}	1,496	-5.77×10^{-3}	9.04×10^{-4}	1.77×10^{-10}
cg01538605	565	-6.14×10^{-3}	2.11×10^{-3}	3.61×10^{-3}	1,497	-5.93×10^{-3}	1.13×10^{-3}	1.61×10^{-7}
cg06087826	565	-8.18×10^{-3}	2.23×10^{-3}	2.38×10^{-4}	1,495	-6.21×10^{-3}	1.20×10^{-3}	2.07×10^{-7}
cg13286614	565	-2.82×10^{-3}	2.74×10^{-3}	3.04×10^{-1}	1,496	7.87×10^{-3}	1.46×10^{-3}	6.26×10^{-8}
cg03619256	565	-8.59×10^{-3}	1.81×10^{-3}	2.15×10^{-6}	1,497	-5.17×10^{-3}	9.50×10^{-4}	5.42×10^{-8}
cg16797901	562	-3.54×10^{-3}	3.84×10^{-3}	3.56×10^{-1}	1,476	-1.12×10^{-2}	2.02×10^{-3}	3.26×10^{-8}
cg19249811	545	3.84×10^{-4}	8.43×10^{-3}	9.64×10^{-1}	1,416	-2.76×10^{-2}	4.00×10^{-3}	5.92×10^{-12}
cg17061862	565	-6.82×10^{-3}	2.58×10^{-3}	8.26×10^{-3}	1,496	-7.86×10^{-3}	1.32×10^{-3}	2.69×10^{-9}
cg00574958	565	-1.16×10^{-2}	3.23×10^{-3}	3.18×10^{-4}	1,495	-6.93×10^{-3}	1.62×10^{-3}	1.97×10^{-5}
cg27117792	565	-6.74×10^{-3}	3.83×10^{-3}	7.82×10^{-2}	1,495	-1.17×10^{-2}	2.13×10^{-3}	4.04×10^{-8}
cg19750657	565	9.92×10^{-3}	2.88×10^{-3}	5.74×10^{-4}	1,495	7.57×10^{-3}	1.42×10^{-3}	8.80×10^{-8}
cg02119938	564	-1.32×10^{-2}	5.90×10^{-3}	2.49×10^{-2}	1,484	-1.43×10^{-2}	2.82×10^{-3}	4.06×10^{-7}
cg01115923	565	-2.36×10^{-3}	2.46×10^{-3}	3.36×10^{-1}	1,497	-5.92×10^{-3}	1.10×10^{-3}	7.10×10^{-8}
cg11024682	565	5.95×10^{-3}	1.53×10^{-3}	1.05×10^{-4}	1,338	5.70×10^{-3}	9.79×10^{-4}	5.99×10^{-9}
cg24174557	562	-1.05×10^{-2}	3.49×10^{-3}	2.65×10^{-3}	1,492	-8.74×10^{-3}	1.67×10^{-3}	1.67×10^{-7}
cg04927537	565	3.72×10^{-3}	2.69×10^{-3}	1.66×10^{-1}	1,497	7.21×10^{-3}	1.49×10^{-3}	1.32×10^{-6}
cg26950531	561	-1.97×10^{-2}	4.75×10^{-3}	3.41×10^{-5}	1,487	-1.40×10^{-2}	2.33×10^{-3}	1.69×10^{-9}
cg27243685	565	8.71×10^{-3}	2.30×10^{-3}	1.56×10^{-4}	1,497	7.03×10^{-3}	1.42×10^{-3}	7.71×10^{-7}
cg06500161	565	1.01×10^{-2}	1.66×10^{-3}	1.26×10^{-9}	1,497	6.94×10^{-3}	9.18×10^{-4}	4.15×10^{-14}

Table 4.7: Sex-specific analyses of associations between DNA methylation levels and BMI identified in the discovery set. For men and women separately, the table reports the number of samples N included in the subgroup analysis, the beta regression model coefficient β (on the logit scale), the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

CpG locus	Men				Women			
	N	β	$SE(\beta)$	p -value	N	β	$SE(\beta)$	p -value
cg09831562	486	7.30×10^{-3}	2.18×10^{-3}	7.91×10^{-4}	992	6.84×10^{-3}	1.19×10^{-3}	8.76×10^{-9}
cg01538605	487	-2.61×10^{-3}	9.08×10^{-4}	4.02×10^{-3}	1,021	-2.78×10^{-3}	5.62×10^{-4}	7.58×10^{-7}
cg03619256	487	-2.31×10^{-3}	7.35×10^{-4}	1.66×10^{-3}	1,021	-2.42×10^{-3}	4.82×10^{-4}	5.23×10^{-7}
cg11376147	486	-2.66×10^{-3}	7.04×10^{-4}	1.56×10^{-4}	1,021	-2.06×10^{-3}	4.86×10^{-4}	2.30×10^{-5}
cg19750657	487	3.16×10^{-3}	1.15×10^{-3}	6.05×10^{-3}	1,020	3.60×10^{-3}	7.30×10^{-4}	8.40×10^{-7}
cg07037944	482	-2.38×10^{-3}	8.34×10^{-4}	4.38×10^{-3}	1,019	-3.00×10^{-3}	5.79×10^{-4}	2.25×10^{-7}
cg04583842	487	1.65×10^{-3}	1.08×10^{-3}	1.29×10^{-1}	1,006	3.59×10^{-3}	7.34×10^{-4}	1.00×10^{-6}
cg11024682	487	1.89×10^{-3}	6.31×10^{-4}	2.82×10^{-3}	1,021	2.05×10^{-3}	4.60×10^{-4}	8.70×10^{-6}
cg24174557	484	-3.72×10^{-3}	1.43×10^{-3}	9.39×10^{-3}	1,016	-4.45×10^{-3}	8.44×10^{-4}	1.34×10^{-7}
cg26950531	483	-5.49×10^{-3}	1.96×10^{-3}	5.03×10^{-3}	1,014	-7.24×10^{-3}	1.22×10^{-3}	3.05×10^{-9}
cg06500161	487	2.63×10^{-3}	7.10×10^{-4}	2.07×10^{-4}	1,021	3.89×10^{-3}	4.54×10^{-4}	1.05×10^{-17}
cg21307484	487	2.05×10^{-3}	7.69×10^{-4}	7.80×10^{-3}	1,021	2.12×10^{-3}	5.16×10^{-4}	4.05×10^{-5}

Table 4.8: Sex-specific analyses of associations between DNA methylation levels and waist circumference identified in the discovery set. For men and women separately, the table reports the number of samples N included in the subgroup analysis, the beta regression model coefficient β (on the logit scale), the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

CpG locus	Men				Women			
	N	β	$SE(\beta)$	p -value	N	β	$SE(\beta)$	p -value
cg17901584	487	-2.96×10^{-1}	2.15×10^{-1}	1.67×10^{-1}	1,021	-7.01×10^{-1}	1.31×10^{-1}	8.77×10^{-8}
cg16246545	487	-3.43×10^{-1}	1.48×10^{-1}	2.06×10^{-2}	1,021	-4.68×10^{-1}	9.01×10^{-2}	2.08×10^{-7}
cg14476101	485	-4.60×10^{-1}	2.07×10^{-1}	2.62×10^{-2}	1,019	-7.32×10^{-1}	1.22×10^{-1}	1.92×10^{-9}
cg19472611	487	-9.39×10^{-1}	2.72×10^{-1}	5.56×10^{-4}	1,003	-5.78×10^{-1}	1.47×10^{-1}	8.58×10^{-5}
cg25799109	486	-4.58×10^{-1}	2.33×10^{-1}	4.98×10^{-2}	1,010	-5.73×10^{-1}	1.34×10^{-1}	1.89×10^{-5}
cg20898587	487	-1.51×10^{-3}	1.30×10^{-1}	9.91×10^{-1}	1,020	-3.64×10^{-1}	6.88×10^{-2}	1.22×10^{-7}
cg06730756	486	-7.99×10^{-1}	5.67×10^{-1}	1.59×10^{-1}	969	-2.51	3.12×10^{-1}	9.96×10^{-16}
cg11376147	486	-2.99×10^{-1}	1.13×10^{-1}	8.01×10^{-3}	1,021	-3.29×10^{-1}	7.18×10^{-2}	4.67×10^{-6}
cg06500161	487	3.58×10^{-1}	1.15×10^{-1}	1.79×10^{-3}	1,021	4.41×10^{-1}	6.59×10^{-2}	2.25×10^{-11}

Table 4.9: Sex-specific analyses of associations between DNA methylation levels and waist-to-hip ratio identified in the discovery set. For men and women separately, the table reports the number of samples N included in the subgroup analysis, the beta regression model coefficient β (on the logit scale), the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

CpG locus	Men				Women			
	N	β	$SE(\beta)$	p -value	N	β	$SE(\beta)$	p -value
cg12593793	487	-4.63×10^{-1}	1.57×10^{-1}	3.24×10^{-3}	1,020	-3.83×10^{-1}	1.00×10^{-1}	1.28×10^{-4}
cg09831562	486	1.21	3.72×10^{-1}	1.13×10^{-3}	992	9.88×10^{-1}	1.81×10^{-1}	4.67×10^{-8}
cg06164260	487	-1.19×10^{-1}	1.10×10^{-1}	2.78×10^{-1}	1,021	-3.11×10^{-1}	5.76×10^{-2}	6.33×10^{-8}
cg03619256	487	-5.03×10^{-1}	1.24×10^{-1}	5.11×10^{-5}	1,021	-3.58×10^{-1}	7.31×10^{-2}	9.49×10^{-7}
cg11376147	486	-4.46×10^{-1}	1.20×10^{-1}	2.04×10^{-4}	1,021	-3.59×10^{-1}	7.35×10^{-2}	1.03×10^{-6}
cg07037944	482	-3.44×10^{-1}	1.43×10^{-1}	1.61×10^{-2}	1,019	-4.43×10^{-1}	8.79×10^{-2}	4.65×10^{-7}
cg11024682	487	2.49×10^{-1}	1.08×10^{-1}	2.13×10^{-2}	1,021	3.21×10^{-1}	7.00×10^{-2}	4.46×10^{-6}
cg24174557	484	-6.21×10^{-1}	2.43×10^{-1}	1.07×10^{-2}	1,016	-6.52×10^{-1}	1.28×10^{-1}	3.62×10^{-7}
cg05091997	486	-8.85×10^{-1}	3.56×10^{-1}	1.29×10^{-2}	1,004	-9.50×10^{-1}	2.04×10^{-1}	3.07×10^{-6}
cg26950531	483	-1.14	3.34×10^{-1}	6.39×10^{-4}	1,014	-1.16	1.85×10^{-1}	4.02×10^{-10}
cg26470501	487	-1.49×10^{-1}	1.10×10^{-1}	1.74×10^{-1}	1,021	-3.60×10^{-1}	6.49×10^{-2}	2.94×10^{-8}
cg06500161	487	5.19×10^{-1}	1.21×10^{-1}	1.68×10^{-5}	1,021	5.72×10^{-1}	6.90×10^{-2}	1.10×10^{-16}

Table 4.10: Sex-specific analyses of associations between DNA methylation levels and waist-to-height ratio identified in the discovery set. For men and women separately, the table reports the number of samples N included in the subgroup analysis, the beta regression model coefficient β (on the logit scale), the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

and 4.71×10^{-11} for BMI, waist circumference, WHR, and WHtR, respectively).

DNA methylation measurements obtained using bisulphite pyrosequencing were additionally used to examine the shape of the dose-response curve with each adiposity measure; this showed generally linear direct relationships (fig. 4.1).

Genotype data for a single nucleotide polymorphism (SNP) located 7 bp upstream of cg06500161 (rs9982016) were also obtained as part of the experiment. Of 358 participants, 327 were homozygous for the major allele (AA), and 31 were heterozygous (AT); the observed minor allele frequency (4.3%) was in good agreement with that reported by the 1000 Genomes project [22] for the European population (4.6%). No statistically significant associations were found between heterozygosity at rs9982016 and adiposity indices (all p -values > 0.05). DNA methylation levels at cg06500161 were higher in the presence of the minor allele ($\beta = 0.14$ for AT compared with AA, p -value 1.37×10^{-5}); nonetheless, associations between DNA methylation and adiposity measures remained strongly statistically significant after adjustment for heterozygosity (p -values 7.12×10^{-10} , 6.61×10^{-11} , 6.68×10^{-9} , and 7.61×10^{-11} for BMI, waist circumference, WHR, and WHtR, respectively), and upon stratification (p -values 1.16×10^{-9} , 1.82×10^{-10} , 3.69×10^{-8} , and 3.49×10^{-10} for BMI, waist circumference, WHR, and WHtR, respectively, in participants homozygous for the major allele). In addition, we note that no SNPs with minor allele frequencies above 1% in the European population (as reported by the 1000 Genomes project [22]) located within 500 kb windows centred at each informative CpG locus were associated at a genome-wide significance level $\alpha = 10^{-7}$ with BMI or WHR in large meta-analyses by the GIANT consortium [5, 6] (minimum reported p -values 6.36×10^{-5} and 5.42×10^{-6} for BMI and WHR, respectively).

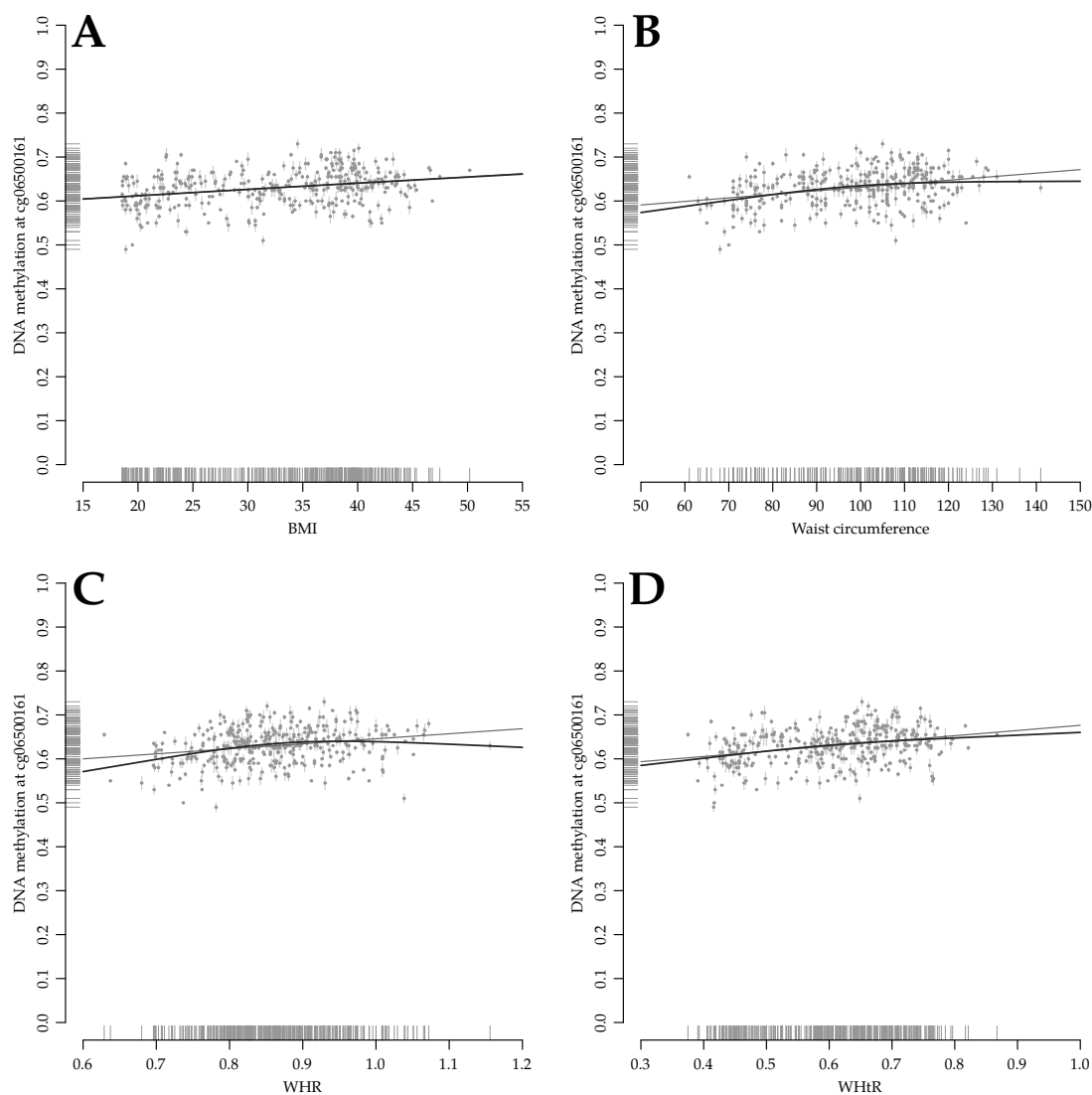


Figure 4.1: Dose-response relationships between DNA methylation levels at cg06500161 (*ABCG1*) in the replication set and adiposity measures.. (A) BMI; (B) waist circumference; (C) WHR; (D) WHtR. Points and bars represent the mean and range of DNA methylation levels across technical replicates, respectively. Dark grey lines are from linear regression models, and black lines from thin plate spline regression models.

4.3.2 Associations with transcriptional activity

DNA methylation levels at six informative CpG loci were statistically significantly associated with transcriptional activity in PBLs of the nearest gene in both EnviroGenoMarkers and NOWAC data (tables 4.11 and 4.12). With the exception of DNA methylation levels at cg00574958, which were inversely associated with transcription levels of *CPT1A* in EnviroGenoMarkers, but directly associated in NOWAC, all associations showed consistent directions in both studies. In particular, DNA methylation levels at cg11024682 were inversely associated with transcriptional activity of *SREBF1*, and DNA methylation levels at both pairs of informative CpG loci found in close proximity (cg27243685 and cg06500161 on *ABCG1*, and cg16246545 and cg14476101 on *PHGDH*) were consistently inversely associated with transcriptional activity of the corresponding genes. In addition, transcription levels of *ABCG1* were independently inversely associated with BMI ($\beta = -0.05$ per kg/m², p -value 7.38×10^{-7}) in the NOWAC data; an association in the same direction was also found in the EnviroGenoMarkers data (p -values 1.18×10^{-2}).

4.3.3 Associations with biomarkers of lipid and glucose metabolism

Statistically significant associations between DNA methylation levels at the 40 informative CpG loci and biomarkers of lipid and glucose metabolism are reported in table 4.13. Fasting blood levels of HDL cholesterol were associated with DNA methylation levels at cg06500161 (*ABCG1*) and cg17901584 (*DHCR24*); in contrast, no associations were declared statistically significant with total cholesterol levels. Triglyceride levels were associated with DNA methylation levels at

mRNA	Gene	CpG locus	Association with DNA methylation			Association with BMI		
			β	SE(β)	p -value	β	SE(β)	p -value
NM_001031847	<i>CPT1A</i>	cg00574958	-3.72	1.46	1.15×10^{-2}	2.75×10^{-2}	7.53×10^{-3}	3.00×10^{-4}
NM_016617	<i>UFMI</i>	cg19750657	2.68	1.13	1.89×10^{-2}	8.70×10^{-3}	1.16×10^{-2}	4.52×10^{-1}
NM_001005291	<i>SREBF1</i>	cg11024682	-1.90	0.82	2.10×10^{-2}	-9.25×10^{-3}	7.45×10^{-3}	2.15×10^{-1}
NM_004915	<i>ABCG1</i>	cg27243685	-3.55	1.12	1.73×10^{-3}	-2.02×10^{-2}	7.96×10^{-3}	1.18×10^{-2}
		cg06500161	-2.81	0.94	2.93×10^{-3}			
NM_003627	<i>SLC43A1</i>	cg11376147	-3.61	1.65	2.89×10^{-2}	7.04×10^{-3}	9.53×10^{-3}	4.61×10^{-1}
NM_006623	<i>PHGDH</i>	cg16246545	-2.60	0.77	9.92×10^{-4}	-3.03×10^{-3}	8.76×10^{-3}	7.29×10^{-1}
		cg14476101	-2.03	0.60	8.49×10^{-4}			
NM_001128615	<i>ARHGEF3</i>	cg25799109	1.05	0.51	4.22×10^{-2}	2.48×10^{-3}	8.47×10^{-3}	7.70×10^{-1}

Table 4.11: Associations of DNA methylation levels at informative CpG loci with *cis* transcriptional activity in EnviroGenoMarkers data. Associations between transcription levels of the identified transcripts and BMI are also presented. The tables report the $\log_2$ -linear regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

mRNA	Gene	CpG locus	Association with DNA methylation		Association with BMI		
			β	SE(β)	p -value	β	SE(β) p -value
NM_020685	<i>C3orf14</i>	cg15059608	2.29	0.53	1.74×10^{-5}	-2.31×10^{-3}	5.23×10^{-3} 6.59 $\times 10^{-1}$
NM_001876	<i>CPT1A</i>	cg00574958	3.39	1.27	7.90×10^{-3}	3.12×10^{-3}	7.70×10^{-3} 6.85 $\times 10^{-1}$
NM_001031847	<i>CPT1A</i>	cg00574958	2.37	0.93	1.15×10^{-2}	2.26×10^{-4}	5.65×10^{-3} 9.68 $\times 10^{-1}$
NM_001005291	<i>SREBF1</i>	cg11024682	-2.99	1.41	3.47×10^{-2}	-3.63×10^{-2}	1.10×10^{-2} 1.08 $\times 10^{-3}$
NM_004915	<i>ABCG1</i>	cg27243685	-6.21	1.54	6.79×10^{-5}	-5.25×10^{-2}	1.04×10^{-2} 7.38 $\times 10^{-7}$
		cg06500161	-5.37	1.27	2.89×10^{-5}		
NM_014326	<i>DAPK2</i>	cg07037944	-3.06	1.14	7.68×10^{-3}	-3.27×10^{-2}	9.03×10^{-3} 3.40 $\times 10^{-4}$
NM_006623	<i>PHGDH</i>	cg16246545	-3.74	0.84	1.06×10^{-5}	-1.59×10^{-2}	9.21×10^{-3} 8.45 $\times 10^{-2}$
		cg14476101	-2.80	0.61	6.18×10^{-6}		

Table 4.12: Associations of DNA methylation levels at informative CpG loci with *cis* transcriptional activity in NOWAC data. Associations between transcription levels of the identified transcripts and BMI are also presented. The tables report the $\log_2$ -linear regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

two informative CpG loci, cg00574958 (*CPT1A*) and cg06500161 (*ABCG1*). A significant association was found between total-to-HDL cholesterol ratio and DNA methylation levels at cg06500161 (*ABCG1*), and between triglycerides-to-HDL cholesterol ratio and three informative CpG loci: cg27243685 and cg06500161 (both on *ABCG1*), and cg17901584 (*DHCR24*). Finally, DNA methylation levels at cg06500161 (*ABCG1*) were associated with fasting blood glucose levels; no significant associations were found with fasting insulin levels or HOMA indices.

4.3.4 Associations with differential blood counts

No significant associations were found between differential blood counts and DNA methylation levels at informative CpG loci, with the exception of cg27117792 (intergenic). DNA methylation levels at this locus were directly associated with the proportion of eosinophil granulocytes (p -value 1.19×10^{-3}), and inversely associated with the proportion of neutrophil granulocytes (p -value 1.70×10^{-5}); these associations lost statistical significance when considering totals instead of proportions.

4.3.5 Associations with obesity-related diseases

After adjustment for known risk factors, BMI, and WHR, DNA methylation levels at six, nine and six of the 40 informative CpG loci were nominally associated with post-menopausal breast cancer, colorectal cancer, and myocardial infarction, respectively (tables 4.14 to 4.16). Of these, none of the post-menopausal breast cancer loci survived Bonferroni correction assuming 30 independent tests (as suggested by PCA of DNA methylation levels at the 40 informative CpG loci). In contrast, two CpG loci associated with colorectal cancer (cg21307484, *IL2RB* and cg11661512, *FGF18*), and one associated with myocardial infarction (cg12593793,

CpG locus	Gene	Biomarker	β	SE(β)	p -value
cg00574958	<i>CPT1A</i>	Triglycerides	-2.92	0.89	1.06×10^{-3}
cg27243685	<i>ABCG1</i>	Triglycerides/HDL cholesterol	3.44	1.01	6.76×10^{-4}
cg06500161	<i>ABCG1</i>	HDL cholesterol	-1.37	0.31	1.01×10^{-5}
		Triglycerides	3.27	0.61	1.13×10^{-7}
		Total/HDL cholesterol	1.29	0.33	1.10×10^{-4}
		Triglycerides/HDL cholesterol	4.65	0.78	4.77×10^{-9}
		Glucose	0.76	0.22	7.83×10^{-4}
cg17901584	<i>DHCR24</i>	HDL cholesterol	0.61	0.15	9.71×10^{-5}
		Triglycerides/HDL cholesterol	-1.33	0.40	8.74×10^{-4}

Table 4.13: Associations of DNA methylation levels at informative CpG loci with biomarkers of lipid and glucose metabolism. Investigated associations are with fasting blood levels of total and HDL cholesterol, triglycerides, glucose (all in mmol/L), and insulin (μ U/mL) after adjustment for sex, age, case-control status (separately for cancers and myocardial infarction), and BMI. The table reports the log-linear regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$.

CpG locus	N	β	$SE(\beta)$	p -value
cg01115923	182	-5.06×10^{-2}	1.91×10^{-2}	8.03×10^{-3}
cg03619256	182	-4.18×10^{-2}	1.78×10^{-2}	1.86×10^{-2}
cg26470501	182	-3.73×10^{-2}	1.62×10^{-2}	2.13×10^{-2}
cg13286614	182	-6.80×10^{-2}	2.96×10^{-2}	2.17×10^{-2}
cg25799109	182	-8.41×10^{-2}	3.87×10^{-2}	2.99×10^{-2}
cg21307484	182	-4.05×10^{-2}	1.91×10^{-2}	3.45×10^{-2}

Table 4.14: Nominally significant associations with post-menopausal breast cancer. The table reports the regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$. No associations are declared significant after Bonferroni correction assuming 30 independent tests.

intergenic), survived Bonferroni correction. We compared the ROC curves obtained from logistic regression models including known risk factors for colorectal cancer and myocardial infarction with those from models additionally including the identified DNA methylation markers (fig. 4.2). Inclusion of these CpG loci moderately but systematically improved discriminatory performances of the models (around 2.2% increase in AUC for colorectal cancer, and 1.4% for myocardial infarction).

§ 4.4 Discussion

We conducted a series of genome-wide association studies of DNA methylation levels for four adiposity measures in over 1,900 individuals from large population-based cohorts across four European countries (Italy, the Netherlands, Norway, and Sweden), and synthesised the results in a meta-analytical framework.

CpG locus	N	β	$SE(\beta)$	p -value	
cg21307484	276	-7.90×10^{-2}	2.26×10^{-2}	4.76×10^{-4}	*
cg11661512	275	-7.48×10^{-2}	2.37×10^{-2}	1.61×10^{-3}	*
cg26470501	276	-5.06×10^{-2}	1.94×10^{-2}	9.14×10^{-3}	
cg26950531	276	-1.29×10^{-1}	4.97×10^{-2}	9.22×10^{-3}	
cg14476101	275	-9.73×10^{-2}	4.14×10^{-2}	1.88×10^{-2}	
cg16258657	276	-8.25×10^{-2}	3.68×10^{-2}	2.50×10^{-2}	
cg06164260	276	-3.86×10^{-2}	1.74×10^{-2}	2.63×10^{-2}	
cg19249811	273	1.98×10^{-1}	9.44×10^{-2}	3.56×10^{-2}	
cg17901584	276	-8.04×10^{-2}	3.84×10^{-2}	3.63×10^{-2}	

Table 4.15: Nominally significant associations with colorectal cancer. The table reports the regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$. Associations marked with * are declared significant after Bonferroni correction assuming 30 independent tests.

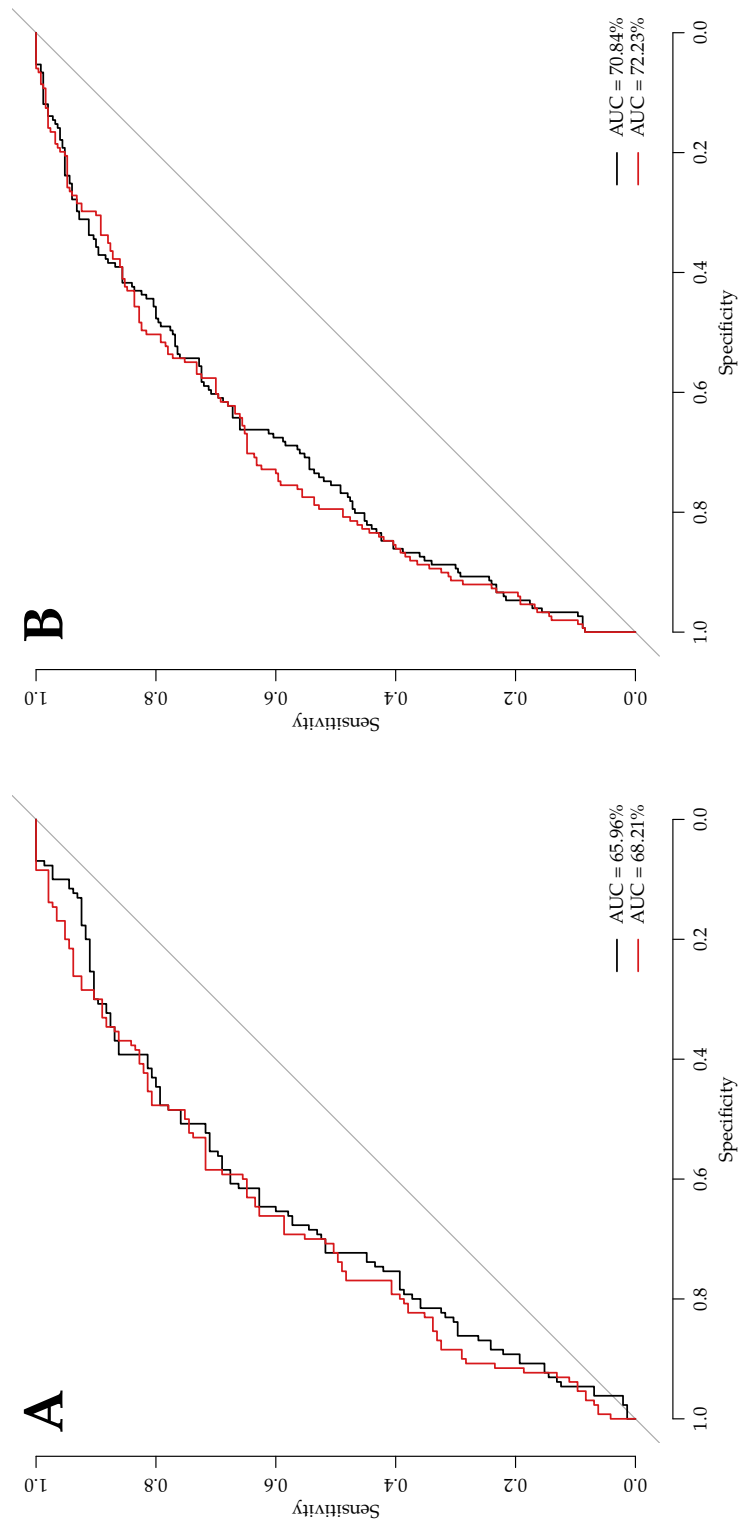


Figure 4.2: Receiver operating characteristic (ROC) curves for case-control logistic regression models. (A) Colorectal cancer; **(B)** myocardial infarction. Black lines are from logistic regression models including adipsity measures and established disease-specific risk factors; red lines are from models additionally including the DNA methylation markers highlighted in tables 4.15 and 4.16.

CpG locus	N	β	$SE(\beta)$	p -value	
cg12593793	401	-7.38×10^{-2}	2.29×10^{-2}	1.25×10^{-3}	*
cg13286614	402	-8.38×10^{-2}	2.95×10^{-2}	4.50×10^{-3}	
cg17901584	402	-8.78×10^{-2}	3.65×10^{-2}	1.61×10^{-2}	
cg26470501	402	-3.89×10^{-2}	1.70×10^{-2}	2.24×10^{-2}	
cg27117792	402	-8.69×10^{-2}	4.11×10^{-2}	3.45×10^{-2}	
cg06164260	402	-3.50×10^{-2}	1.76×10^{-2}	4.63×10^{-2}	

Table 4.16: Nominally significant associations with myocardial infarction. The table reports the regression model coefficient β , the associated standard error (SE), and the p -value for $\mathcal{H}_0 : \beta = 0$. Associations marked with $\star$ are declared significant after Bonferroni correction assuming 30 independent tests.

We identified associations between one or more adiposity measures (among BMI, waist circumference, WHR, and WHtR) and DNA methylation levels at 40 informative CpG loci, nine of which were consistently associated with BMI and at least one among waist circumference and derived measures. Among these, we found DNA methylation levels at a single intronic CpG locus (cg06500161, *ABCG1*) to be directly and independently associated with all adiposity measures, and replicated these results in an independent sample using bisulphite pyrosequencing.

Upon stratification by sex, all associations were replicated in women; in men, some lost statistical significance but preserved effect size estimates, whereas strong attenuations were observed for a small number of associations. The skewed sex ratio in our study may explain the limited replication attained in men-only analyses, though sex-specific effect modification may also plausibly underlie this occurrence.

To understand the functional relevance of informative CpG loci, we related DNA methylation levels to transcriptional activity in PBLs of the nearest gene. Using data from two studies, we identified and replicated associations between increased DNA methylation levels at five CpG loci and down-regulation of *ABCG1*, *PHGDH*, and *SREBF1*. In addition, we found transcription levels of *ABCG1* to be independently inversely associated with BMI. DNA methylation levels at two CpG loci on *ABCG1* were also associated with multiple lipid and glucose metabolism phenotypes, consistently with the function of this gene (as described below), and with a recent report in the literature [23]. Associations with transcriptional activity of *PHGDH* and *SREBF1* are of particular interest, for different reasons. *PHGDH* catalyses the oxidation of the glycolytic intermediate 3-phospho-D-glycerate, the first step of the serine biosynthetic pathway; it has also been reported to divert the glycolytic flux and contribute to oncogenesis [24], possibly through production of the oncometabolite D-2-hydroxyglutarate [25]. *SREBF1* encodes a membrane-bound transcription factor that regulates genes involved in cholesterol metabolism and homeostasis [26, 27], and thus plays a pivotal role in the sterol biosynthetic and regulatory networks.

As reported above, DNA methylation levels at cg06500161 (*ABCG1*) were associated with fasting triglyceride levels, HDL cholesterol levels, total-to-HDL cholesterol ratio, and triglycerides-to-HDL cholesterol ratio; the latter was also associated with DNA methylation levels at cg27243685, the other informative CpG locus found on *ABCG1*. These results are consistent with the role of *ABCG1* in cellular cholesterol efflux and reverse cholesterol transport [28]. In particular, *ABCG1* facilitates removal of excess cholesterol by promoting maturation of nascent HDL particles (generated by ABCA1) to larger, cholesterol-rich lipoproteins [29]. *ABCG1* is highly expressed in cholesterol-loaded macrophages, where it guards against lipid overloading and consequent foam cell formation [30], and

targeted disruption of *Abcg1* in mice results in substantial lipid accumulation in macrophages in multiple tissues [31]. *ABCG1* also plays a role in lipoprotein lipase (LPL) regulation; knockdown of *ABCG1* expression in cultured human macrophages leads to decreased LPL secretion and bioavailability [32], and may thus result in less efficient uptake of circulating triglycerides. These findings reinforce the biological plausibility of our results, which are also in agreement with the ‘conjoint trait’ hypothesis of a combined low HDL cholesterol/high triglycerides phenotype that is regulated pleiotropically [33]. In addition, DNA methylation levels at cg06500161 (*ABCG1*) were also associated with fasting glucose levels, which is in agreement with a recent study that found DNA methylation levels at this CpG locus to be associated with parameters of glucose metabolism in non-diabetic participants [34].

We also found an association between DNA methylation levels at one CpG locus on *CPT1A* and fasting triglyceride levels, consistent with recent reports in the literature [35–37] and with previous results in animal models [38], and two associations of DNA methylation levels at one CpG locus on *DHCR24*, and HDL cholesterol and triglycerides-to-HDL cholesterol ratio. Expression of this gene is known to be under epigenetic control [39], and its products are thought to mediate the anti-inflammatory effect of HDL cholesterol in endothelial cells [40, 41]; they also appear to regulate cellular response to oncogenic and oxidative stress [42].

Among the 40 informative CpG sites, two loci in *IL2RB* and *FGF18*, and one CpG locus in an intergenic region of chromosome 1 were associated with colorectal cancer and myocardial infarction development, respectively, after controlling for anthropometric parameters and established risk factors. None of the obesity-related CpG loci were associated with post-menopausal breast cancer following Bonferroni correction, though we note that the number of post-menopausal

breast cancer cases included in the study was relatively small.

Lower DNA methylation levels at *IL2RB* and *FGF18* were associated with increased risk of colorectal cancer, which suggests that differential activity of these genes may be contributing to development of colorectal cancer. The *IL2RB* gene encodes the interleukin(IL)-2 receptor- β which binds IL-2, a cytokine which plays a major role in T-cell differentiation and proliferation. Increased expression of lymphocyte *IL2RB* has been linked to risk of colorectal cancer in small clinical case-control studies, and was shown to accurately predict risk when combined with six additional genes [43, 44]. Given that obesity is accompanied by alterations in T-cell populations, inflammation and immune deregulation [45], our results provide a potential link between obesity, T-cell proliferative pathways and colorectal cancer development that warrants further study. The *FGF18* gene, which encodes the fibroblast growth factor 18, has also been previously linked to colorectal tumorigenesis [46]. FGF is a downstream target of the Wnt signalling and β -catenin pathways which are central to cell growth and proliferation, and frequently mutated in colorectal cancer [47]. Alterations in the Wnt/ β -catenin pathway have been associated with obesity-related disorders such as insulin resistance and type 2 diabetes [48, 49], and *FGF18* itself interacts with PI3K/mTOR and growth factor signalling pathways. Our findings may provide new evidence that obesity is associated with specific changes in the Wnt/ β -catenin signalling pathways, that could, in turn, drive colorectal cancer development.

DNA methylation levels at cg12593793 were found to be inversely associated with risk of myocardial infarction, independent of obesity and established risk factors. This locus lies in an intergenic region of chromosome 1, and has not been linked to any biological function to date; therefore, we are unable to speculate on any potential role in myocardial infarction risk. Nevertheless, our study offers evidence that DNA methylation markers may offer improved predictive

ability for future risk of this condition, extending previous results for prevalent (hospitalised) myocardial infarction [23]. DNA methylation levels at cg12593793 were also found to be inversely associated with smoking, even after adjustment for BMI ($\beta = -0.11$, p -value 2.93×10^{-6} for current-versus-never smokers), and remained significantly associated with myocardial infarction following control for smoking status. We may speculate that both smoking and adiposity influence myocardial infarction risk through a mechanism that alters DNA methylation status at this locus.

To summarise, for both colorectal cancer and myocardial infarction, the inclusion of disease-related CpG loci improved predictive performances of the risk model. These improvements, though modest, were independent of obesity and known risk factors.

There are also limitations to our approach. DNA methylation levels in peripheral blood leukocytes may be affected by high turnover rates and inter-individual tissue heterogeneity [50, 51]; however, we only found evidence for this type of confounding for one informative CpG locus. The observed associations may also be confounded by genetic variants, which have been shown to contribute to inter-individual variation in DNA methylation levels [52]. Nonetheless, we did not find convincing evidence for a strong confounding effect.

In conclusion, our findings substantiate the involvement of DNA methylation in the complex interplay between adiposity, metabolic health, and risk of downstream sequelae. We have identified a number of CpG sites that are associated with adiposity and metabolic traits, and which appear to have functional effects on gene expression. Among these, two CpG sites in *IL2RB* and *FGF18*, and one intergenic CpG locus on chromosome 1 were associated, independently of obesity and known risk factors, with subsequent risk of colorectal cancer and myocardial infarction, respectively. This suggests a potential role for DNA methylation

as intermediate biomarker. While causality can only be formally assessed in an experimental setting, our observational study provides novel and plausible evidence for a hitherto unknown mechanistic link involving DNA methylation alterations, which may partly explain the association of adiposity with obesity-related chronic diseases.

References

1. G A Stevens, G M Singh, Y Lu, G Danaei, J K Lin, M M Finucane, A N Bahalim, R K McIntire, H R Gutierrez, M Cowan, C J Paciorek, F Farzadfar, L Riley, and M Ezzati. 'National, regional, and global trends in adult overweight and obesity prevalences'. *Population Health Metrics* **10** (2012), 22.
2. K Bhaskaran, I Douglas, H Forbes, I dos-Santos-Silva, D A Leon, and L Smeeth. 'Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults'. *The Lancet* **384**.9945 (2014), 755–765.
3. M Arnold, N Pandeya, G Byrnes, A G Renehan, G A Stevens, M Ezzati, J Ferlay, J J Miranda, I Romieu, R Dikshit, D Forman, and I Soerjomataram. 'Global burden of cancer attributable to high body-mass index in 2012: a population-based study'. *The Lancet Oncology* **16**.1 (2015), 36–46.
4. V T Samuel and G I Shulman. 'Mechanisms for insulin resistance: common threads and missing links'. *Cell* **148**.5 (2012), 852–871.
5. E K Speliotes, C J Willer, S I Berndt, K L Monda, G Thorleifsson, et al. 'Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index'. *Nature Genetics* **42** (2010), 937–948.

6. I M Heid, A U Jackson, J C Randall, T W Winkler, L Qi, et al. 'Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution'. *Nature Genetics* **42** (2010), 949–960.
7. J Yang, R J F Loos, J E Powell, S E Medland, E K Speliotes, et al. 'FTO genotype is associated with phenotypic variability of body mass index'. *Nature* **490** (2012), 267–272.
8. R A Scott, V Lagou, R P Welch, E Wheeler, M E Montasser, et al. 'Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways'. *Nature Genetics* **44** (2012), 991–1005.
9. P Vineis, R Tumino, S Sieri, S Salvini, C Saieva, C Sacerdote, S Panico, D Palli, V Pala, G Masala, V Krogh, M C Giurdanella, R Galasso, G Frasca, L Fiorini, A Decarli, M Ceroti, E Celentano, and F Berrino. 'A molecular epidemiology project on diet and cancer: the EPIC-Italy prospective study. Design and baseline characteristics of participants'. *Tumori* **89.6** (2003), 586–593.
10. J W J Beulens, E M Monninkhof, W M M Verschuren, Y T van der Schouw, J Smit, M C Ocke, E H J M Jansen, S van Dieren, D E Grobbee, P H M Peeters, and H B Bueno-de-Mesquita. 'Cohort profile: the EPIC-NL study'. *International Journal of Epidemiology* **39.5** (2010), 1170–1178.
11. E Lund, M Kumle, T Braaten, A Hjartåker, K Bakken, E Eggen, and I T Gram. 'External validity in a population-based national prospective study: the Norwegian Women and Cancer Study (NOWAC)'. *Cancer Causes & Control* **14.10** (2003), 1001–1008.

12. G Hallmans, Å Ågren, G Johansson, A Johansson, B Stegmayr, J-H Jansson, B Lindahl, O Rolandsson, S Söderberg, M Nilsson, I Johansson, and L Weinehall. 'Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort – evaluation of risk factors and their interactions'. *Scandinavian Journal of Public Health* **31.61** suppl (2003), 18–24.
13. G Fiorito, S Guarrera, C Valle, F Ricceri, A Russo, S Grioni, A Mattiello, C Di Gaetano, F Rosa, F Modica, L Iacoviello, G Frasca, R Tumino, V Krogh, S Panico, P Vineis, C Sacerdote, and G Matullo. 'B vitamins intake, DNA methylation of one-carbon metabolism and homocysteine pathway genes and myocardial infarction risk: the EPICOR study'. *Nutrition, Metabolism and Cardiovascular Diseases* **24.5** (2014), 483–488.
14. Expert panel on the identification, evaluation, and treatment of overweight in adults. 'Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: executive summary'. *The American Journal of Clinical Nutrition* **68.4** (1998), 899–917.
15. V Dumeaux, A-L Borresen-Dale, J-O Frantzen, M Kumle, V Kristensen, and E Lund. 'Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study'. *Breast Cancer Research* **10** (2008), R13.
16. D G A J HeBELS, P Georgiadis, H C Keun, T J Athersuch, P Vineis, R Vermeulen, L Portengen, I A Bergdahl, G Hallmans, D Palli, B Bendinelli, V Krogh, R Tumino, C Sacerdote, S Panico, J C S Kleinjans, T M C M de Kok, M T Smith, and S A Kyrtopoulos. 'Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research'. *Environmental Health Perspectives* **121.4** (2013), 480–487.

17. S L P Ferrari and F Cribari-Neto. 'Beta regression for modelling rates and proportions'. *Journal of Applied Statistics* **31.7** (2004), 799–815.
18. J P T Higgins and S G Thompson. 'Quantifying heterogeneity in a meta-analysis'. *Statistics in Medicine* **21.11** (2002), 1539–1558.
19. M Chadeau-Hyam, R C H Vermeulen, D G A J Hebels, R Castagné, G Campanella, L Portengen, R S Kelly, I A Bergdahl, B Melin, G Hallmans, D Palli, V Krogh, R Tumino, C Sacerdote, S Panico, T M C M de Kok, M T Smith, J C S Kleinjans, P Vineis, and S A Kyrtopoulos. 'Prediagnostic transcriptomic markers of chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis'. *Annals of Oncology* **25.5** (2014), 1065–1072.
20. W B Kannel, R S Vasan, M J Keyes, L M Sullivan, and S J Robins. 'Usefulness of the triglyceride-high-density lipoprotein versus the cholesterol-high-density lipoprotein ratio for predicting insulin resistance and cardiometabolic risk (from the Framingham Offspring cohort)'. *The American Journal of Cardiology* **101.4** (2008), 497–501.
21. D R Matthews, J P Hosker, A S Rudenski, B A Naylor, D F Treacher, and R C Turner. 'Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man'. *Diabetologia* **28.7** (1985), 412–419.
22. The 1000 Genomes Project Consortium. 'An integrated map of genetic variation from 1,092 human genomes'. *Nature Methods* **491** (2012), 56–65.
23. L Pfeifferm, S Wahl, L C Pilling, E Reischl, J K Sandling, S Kunze, L M Holdt, A Kretschmer, K Schramm, J Adamski, N Klopp, T Illig, Å K Hedman, M Roden, D G Hernandez, A B Singleton, W E Thasler, H Grallert, C Gieger, C Herder, et al. 'DNA methylation of lipid-related genes affects blood lipid levels'. *Circulation: Cardiovascular Genetics* (2015).

24. J W Locasale, A R Grassian, T Melman, C A Lyssiotis, K R Mattaini, A J Bass, G Heffron, C M Metallo, T Muranen, H Sharfi, A T Sasaki, D Anastasiou, E Mullarky, N I Vokes, M Sasaki, R Beroukhim, G Stephanopoulos, A H Li-gon, M Meyerson, A L Richardson, et al. 'Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis'. *Nature Genetics* **43** (2011), 869–874.
25. J Fan, X Teng, L Liu, K R Mattaini, R E Looper, M G Vander Heiden, and J D Rabinowitz. 'Human phosphoglycerate dehydrogenase produces the on-cometabolite D-2-hydroxyglutarate'. *ACS Chemical Biology* **10.2** (2015), 510–516.
26. M S Brown and J L Goldstein. 'The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor'. *Cell* **89.3** (1997), 331–340.
27. J D Horton, J L Goldstein, and M S Brown. 'SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver'. *The Journal of Clinical Investigation* **109.9** (2002), 1125–1131.
28. G Li, H-M Gu, and D-W Zhang. 'ATP-binding cassette transporters and cholesterol translocation'. *IUBMB Life* (2013).
29. A M Vaughan and J F Oram. 'ABCA1 and ABCG1 or ABCG4 act sequentially to remove cellular cholesterol and generate cholesterol-rich HDL'. *Journal of Lipid Research* **47.11** (2006), 2433–2443.
30. J Klucken, C Büchler, E Orsó, W E Kaminski, M Porsch-Özcürümez, G Liebisch, M Kapinsky, W Diederich, W Drobnik, M Dean, R Allikmets, and Gerd Schmitz. 'ABCG1 (ABC8), the human homolog of the *Drosophila white* gene, is a regulator of macrophage cholesterol and phospholipid transport'. *Proceedings of the National Academy of Sciences* **97.2** (2000), 817–822.

31. M A Kennedy, G C Barrera, K Nakamura, Á Baldán, P Tarr, M C Fishbein, J Frank, O L Francone, and P A Edwards. 'ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation'. *Cell Metabolism* **1.2** (2005), 121–131.
32. M Olivier, M W Tanck, R Out, E F Villard, B Lammers, L Bouchareychas, E Frisdal, A Superville, T Van Berkel, J J Kastelein, M Van Eck, J W Jukema, M J Chapman, G M Dallinga-Thie, M Guerin, and W Le Goff. 'Human ATP-binding cassette G1 controls macrophage lipoprotein lipase bioavailability and promotes foam cell formation'. *Arteriosclerosis, Thrombosis, and Vascular Biology* **32.9** (2012), 2223–2231.
33. M C Mahaney, J Blangero, A G Comuzzie, J L VandeBerg, M P Stern, and J W MacCluer. 'Plasma HDL cholesterol, triglycerides, and adiposity: a quantitative genetic test of the conjoint trait hypothesis in the San Antonio Family Heart Study'. *Circulation* **92.11** (1995), 3240–3248.
34. B Hidalgo, M R Irvin, J Sha, D Zhi, S Aslibekyan, D Absher, H K Tiwari, E K Kabagambe, J M Ordovas, and D K Arnett. 'Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network Study'. *Diabetes* **63.2** (2014), 801–807.
35. F Gagnon, D Aïssi, A Carrié, P-E Morange, and D-A Trégouët. 'Robust validation of methylation levels association at *CPT1A* locus with lipid plasma levels'. *Journal of Lipid Research* **55.7** (2014), 1189–1191.
36. A C Frazier-Wood, S Aslibekyan, D M Absher, P N Hopkins, J Sha, M Y Tsai, H K Tiwari, L L Waite, D Zhi, and D K Arnett. 'Methylation at *CPT1A* locus is associated with lipoprotein subfraction profiles'. *Journal of Lipid Research* **55.7** (2014), 1324–1330.

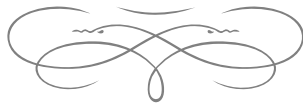
37. E W Tobi, J J Goeman, R Monajemi, H Gu, H Putter, Y Zhang, R C Slieker, A P Stok, P E Thijssen, F Müller, E W van Zwet, C Bock, A Meissner, L H Lumey, P E Slagboom, and B T Heijmans. 'DNA methylation signatures link prenatal famine exposure to growth and metabolism'. *Nature Communications* **5** (2014), 5592.
38. M Stefanovic-Racic, G Perdomo, B S Mantell, I J Sipula, N F Brown, and R M O'Doherty. 'A moderate increase in carnitine palmitoyltransferase 1A activity is sufficient to substantially reduce hepatic triglyceride levels'. *American Journal of Physiology – Endocrinology and Metabolism* **294.5** (2008), E969–E977.
39. J Drzewinska, A Walczak-Drzewiecka, and M Ratajewski. 'Identification and analysis of the promoter region of the human *DHCR24* gene: involvement of DNA methylation and histone acetylation'. *Molecular Biology Reports* **38.2** (2011), 1091–1101.
40. K C Y McGrath, X H Li, R Puranik, E C Liong, J T M Tan, V M Dy, B A DiBartolo, P J Barter, K A Rye, and A K Heather. 'Role of 3 β -hydroxysteroid- Δ 24 reductase in mediating antiinflammatory effects of high-density lipoproteins in endothelial cells'. *Arteriosclerosis, Thrombosis, and Vascular Biology* **29.6** (2009), 877–882.
41. B J Wu, K Chen, S Shrestha, K L Ong, P J Barter, and K-A Rye. 'High-density lipoproteins inhibit vascular endothelial inflammation by increasing 3 β -hydroxysteroid- Δ 24 reductase expression and inducing heme oxygenase-1'. *Circulation Research* **112.2** (2013), 278–288.
42. C Wu, I Miloslavskaya, S Demontis, R Maestro, and K Galaktionov. 'Regulation of cellular response to oncogenic and oxidative stress by seladin-1'. *Nature* **432** (2004), 640–645.

43. K W Marshall, S Mohr, F El Khettabi, N Nossova, S Chao, W Bao, J Ma, X-J Li, and C-C Liew. 'A blood-based biomarker panel for stratifying current risk for colorectal cancer'. *International Journal of Cancer* **126.5** (2010), 1177–1186.
44. Y-T Chang, C-S Huang, C-T Yao, S-L Su, H-J Terng, H-L Chou, Y-C Chou, K-H Chen, Y-W Shih, C-Y Lu, C-H Lai, C-E Jian, C-H Lin, C-T Chen, Y-S Wu, K-S Lin, T Wetter, C-W Chang, and C-M Chu. 'Gene expression profile of peripheral blood in colorectal cancer'. *World Journal of Gastroenterology* **20.39** (2014), 14463–14471.
45. S Nishimura, I Manabe, M Nagasaki, K Eto, H Yamashita, M Ohsugi, M Otsu, K Hara, K Ueki, S Sugiura, K Yoshimura, T Kadowaki, and R Nagai. 'CD8⁺ effector T cells contribute to macrophage recruitment and adipose tissue inflammation in obesity'. *Nature Medicine* **15** (2009), 914–920.
46. T Shimokawa, Y Furukawa, M Sakai, M Li, N Miwa, Y-M Lin, and Y Nakamura. 'Involvement of the *FGF18* gene in colorectal carcinogenesis, as a novel downstream target of the β -catenin/T-cell factor complex'. *Cancer Research* **63.19** (2003), 6116–6120.
47. M Bienz and H Clevers. 'Linking colorectal cancer to Wnt signaling'. *Cell* **103.2** (2000), 311–320.
48. S F A Grant, G Thorleifsson, I Reynisdottir, R Benediktsson, A Manolescu, J Sainz, A Helgason, H Stefansson, V Emilsson, A Helgadottir, U Styrkarsdottir, K P Magnusson, G Bragi Walters, E Palsdottir, T Jonsdottir, T Gudmundsdottir, A Gylfason, J Saemundsdottir, R L Wilensky, M P Reilly, et al. 'Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes'. *Nature Genetics* **38** (2006), 320–323.
49. H J Walters and R N Kulkarni. 'Wnt signaling: relevance to β -cell biology and diabetes'. *Trends in Endocrinology & Metabolism* **19.10** (2008), 349–355.

50. E A Houseman, W P Accomando, D C Koestler, B C Christensen, C J Marsit, H H Nelson, J K Wiencke, and K T Kelsey. 'DNA methylation arrays as surrogate measures of cell mixture distribution'. *BMC Bioinformatics* **13** (2012), 86.
51. K B Michels, A M Binder, S Dedeurwaerder, C B Epstein, J M Grealley, I Gut, E A Houseman, B Izzi, K T Kelsey, A Meissner, A Milosavljevic, K D Siegmund, C Bock, and R A Irizarry. 'Recommendations for the design and analysis of epigenome-wide association studies'. *Nature Methods* **10** (2013), 949–955.
52. J T Bell, A A Pai, J K Pickrell, D J Gaffney, R Pique-Regi, J F Degner, Y Gilad, and J K Pritchard. 'DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines'. *Genome Biology* **12** (2011), R10.



This page intentionally left blank.



Chapter 5

Effects of diet-induced metabolic acidosis on blood pressure

§ 5.1 Introduction

The influence of nutrition on acid-base balance has long been recognized, and significant pathophysiological effects of diet-induced metabolic acidosis (DIMA) on bone resorption, nephrolithiasis, muscle function, insulin resistance, and other endocrine systems are now acknowledged [1–3], making it a potential target for preventive dietary interventions [4–6]. While these effects can be observed in the

short term, they become clinically relevant pre-eminently in the long term, possibly because of the physiological age-related decline in kidney function [7] that is thought to aggravate a chronic (lifelong) state of mild metabolic acidosis such as that induced by a diet rich in acid-forming foods. The kidneys play a pivotal role in acid-base homeostasis, and DIMA is known to affect progression to end-stage renal disease in patients affected by diabetic nephropathy or chronic kidney disease [8, 9]. Since the kidneys are also involved in blood pressure (BP) regulation, DIMA may be similarly associated with systemic hypertension and cardiovascular risk [10]. Nevertheless, recent cross-sectional and prospective studies have failed to provide consistent evidence supporting these associations [11–15].

We assessed the hypothesis that two validated DIMA indices were associated with systolic and diastolic BP in the context of the cross-sectional International Study of Macro- and Micro-nutrients and Blood Pressure (INTERMAP) [16]. Compared to previous studies, INTERMAP affords greater precision in nutrient intake estimates owing to repeated measurements and the definition of standardised procedures to collect and process dietary recall data, and a large variability in daily dietary intake due to its multi-country design.

First, we explored the role of dietary diversity by computing partial correlation coefficients of these indices with broadly defined food groups. Second, we investigated associations with BP, and conducted additional analyses to understand the role of body mass index (BMI) in these associations. We also related the indices to urinary concentrations of 22 proteinogenic amino acids and related compounds. Finally, we conducted two metabolome-wide association studies to identify potential metabolic signatures of DIMA as quantified by these indices, and to characterise the biological pathways it may act on.

§ 5.2 Methods

5.2.1 Study population

INTERMAP is a cross-sectional epidemiological study including 4,680 participants aged 40 to 59 years from 17 populations in four countries (China, Japan, UK, and USA). Its general aim is to clarify nutritional influences on BP, and to explore associations with urinary metabolites quantified by means of ^1H nuclear magnetic resonance (NMR) spectroscopy [17, 18]. Data collected by centrally trained staff during the course of two visits includes anthropometric measurements, extensive lifestyle and medical history information, two standardised 48-hour dietary recalls (including dietary supplement use), two standardised seven-day histories of alcohol intake, and two timed 24-hour urine samples corresponding to each of the two dietary recalls. Urine aliquots (5 mL) were preserved in boric acid and frozen at -20°C within 24 hours of collection.

5.2.2 Sample selection

For the purpose of the present study, participants who declared use of BP-lowering medications ($N = 685$) were excluded from subsequent analyses. In addition, participants were categorised into nine groups according to their ethnicity and country of residence; 108 participants belonging to groups smaller than 50 individuals were also excluded from further analyses. Information on the 3,887 participants included in the study is reported in table 5.1.

	China	Japan	UK	USA
Study centres	3	4	2	8
Participants	776	1074	409	1628
Men	394	540	217	822
Women	382	534	192	806
Ethnicity				
African-American	0	0	0	247
Chinese	776	0	0	0
Hispanic	0	0	0	214
Japanese	0	1074	0	195
White	0	0	409	972
Age (years)	49.38 (10.08)	49.49 (8.44)	49.43 (9.71)	49.08 (8.81)
BMI (kg/m ²)	22.46 (4.29)	23.15 (3.71)	26.44 (5.13)	27.20 (6.43)
Blood pressure (mmHg)				
Systolic	116.75 (17.31)	114.50 (17.19)	116.50 (15.75)	115.25 (16.75)
Diastolic	71.75 (11.75)	72.25 (13.44)	75.75 (11.75)	72.00 (12.50)

Table 5.1: Characteristics of participants included in the study. Counts and medians (IQR) are reported for categorical and continuous variables, respectively.

5.2.3 DIMA indices

The potential of each participant's diet to induce metabolic acidosis was assessed using two previously validated indices, the potential renal acid load (PRAL, mEq/day) and the net endogenous acid production (NEAP, mEq/1.73m²/day) [19], computed from daily nutrient intakes estimated from dietary recalls [17].

PRAL summarises the acid-forming potential of the diet only by taking into account the three main sources of diet-related acid production, i.e. the metabolism of: (i) sulphur-containing amino acids; (ii) organic acids; (iii) phosphate esters and dietary phosphoproteins [20]. It is computed as follows (substituting actual amounts of sulphur-containing amino acids in place of total protein, as in the work of Sebastian et al. [21]),

$$\begin{aligned} \text{PRAL (mEq)} = & 12.48 \times \text{cystine (g)} + \\ & 10.05 \times \text{methionine (g)} + \\ & 0.0366 \times \text{phosphorus (mg)} + \\ & -0.0205 \times \text{potassium (mg)} + \\ & -0.0263 \times \text{magnesium (mg)} + \\ & -0.0125 \times \text{calcium (mg)}. \end{aligned}$$

NEAP extends PRAL to account for endogenous acid production, which is assumed to be proportional to the body surface area (BSA), and thus gives a better indication of the balance between dietary and endogenous acid sources. It is computed as follows,

$$\text{NEAP (mEq/1.73m}^2\text{)} = \text{PRAL (mEq)} + \text{BSA (m}^2\text{)} \times 41/1.73, \quad (5.1)$$

where BSA is approximated using the formula by Mosteller [22],

$$\text{BSA (m}^2\text{)} \approx \sqrt{\text{weight (kg)} \times \text{height (m)}/6}. \quad (5.2)$$

5.2.4 Laboratory analyses

Urinary electrolytes (calcium, magnesium, potassium, and sodium), creatinine, urea, and albumin were measured at the Central Laboratory (Leuven, Belgium), as were 18 proteinogenic amino acids (alanine, arginine, asparagine, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tryptophan, tyrosine, and valine), and four related compounds (1- and 3-methylhistidine, ethanolamine, and taurine); analytical methods have been described elsewhere [16]. Concentrations of amino acids and related compounds were measured in 3,875 urine samples collected at the first visit, and in 1,486 urine samples collected at the second visit; data from both visits were available for 1,478 participants. Glutamic acid was mostly undetectable, and was thus excluded from further analyses.

^1H NMR spectra were acquired at Imperial College London (London, United Kingdom) using a standard one-dimensional pulse sequence with water suppression on a Bruker Avance 600 spectrometer operating at 600.29 MHz in flow-injection mode, as previously reported [18]. One specimen was lost, and the ^1H NMR spectra were unusable for either the first or second specimen in 49 participants.

5.2.5 Statistical analyses

The role of dietary diversity was investigated across and within groups of participants by computing partial correlation coefficients for PRAL and its dietary constituents (sulphur-containing amino acids, phosphorus, potassium, magnesium, and calcium) with seven broadly defined food groups (meat, fish, dairy products, eggs, fruits, vegetables, and grains); for each food group, correlation coefficients were adjusted on study centre (if applicable), sex, age, and all remaining food groups.

Associations of DIMA indices with systolic and diastolic BP were investigated using linear mixed models with a random intercept per study participant nested within study centre and group, and adjusted on sex, age (years), family history of hypertension (categorical), body mass index (BMI, kg/m²), amount of moderate and heavy physical activity (h/day), smoking status (categorical), alcohol intake (g/week), 24-hour urinary sodium and potassium excretion (mmol/day), length of education (years) as proxy for socioeconomic status, and use of dietary supplements (categorical) as proxy for health-conscious behaviour; all analyses were additionally repeated without adjustment on BMI, and stratifying by sex. Associations were declared statistically significant at a 5% significance level.

Associations of DIMA indices with BMI were investigated using similarly parametrised linear mixed models adjusted on sex, age, amount of moderate and heavy physical activity, smoking status, energy intake (MJ/day), length of education, and use of dietary supplements; subgroup analyses included stratification by sex. Associations were declared statistically significant at a 5% significance level.

Associations of DIMA indices with urinary amino acid concentrations were investigated using similarly parametrised linear mixed models adjusted on sex, age, BMI, and dilution factor with three additional random intercepts corresponding to the laboratory technician who prepared the sample ($N = 2$), the laboratory technician who performed integration of chromatographic peaks ($N = 2$), and the analytical instrument used ($N = 3$). Associations were declared statistically significant at a 5% significance level after Bonferroni correction for 22 analytes ($\alpha = 0.05/22 \approx 0.0023$).

Separate metabolome-wide association studies were conducted for DIMA indices at each time point using linear models adjusted on group, sex, and age. Exploratory analyses were also conducted using similarly parametrised models

for systolic and diastolic BP, BMI, and dietary constituents and food groups as defined above. Associations were declared statistically significant at a 5% significance level after Bonferroni correction for 7100 spectral variables ($\alpha = 0.05/7,100 \approx 7.04 \times 10^{-6}$). Spectral variables consistently associated at both time points were designated for structural metabolite identification using available spectral databases and statistical total correlation spectroscopy [23].

§ 5.3 Results

5.3.1 Dietary diversity

The partial correlation coefficients described above reflect the relative contribution of each food group (or of an associated dietary pattern) to PRAL and its dietary constituents; results are depicted in fig. 5.1.

Positive correlations between PRAL and intake of meat, fish, eggs, and grains were observed across groups, with evidence for group-specific preferences: in particular, the contribution of meat intake was more pronounced for participants from the UK and USA (independently of ethnicity), that of fish for Japanese participants from Japan, that of eggs for northern Chinese participants, and that of grains for southern Chinese participants. Consumption of fruit and vegetables was negatively correlated with PRAL consistently across groups, whereas dairy products did not appear to contribute significantly to PRAL.

Dietary intake of sulphur-containing amino acids followed similar group-specific patterns as for PRAL, with the exception of dairy products that represented a substantial source of cysteine and methionine for all participants but the northern Chinese, and of fruits that did not appear to contribute significantly.

Dietary intake of phosphorus was also strongly correlated with consumption

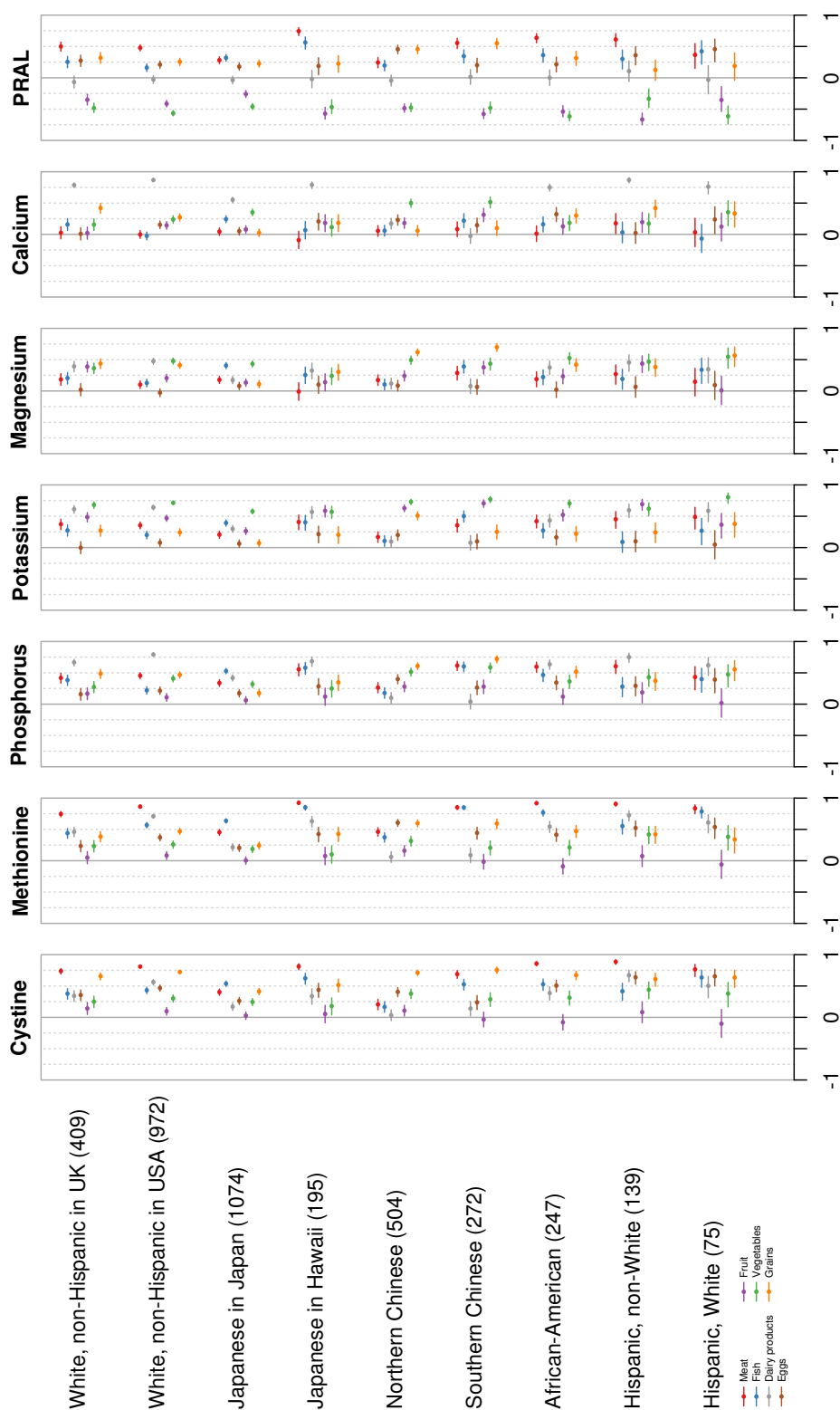


Figure 5.1: Partial correlation coefficients and 95% CIs for PRAL and its constituents with seven broadly defined food groups. For each food group, correlation coefficients are adjusted on study centre (if applicable), sex, age, and all remaining food groups.

of dairy products across groups (with the exception of northern and southern Chinese), and the remaining correlations mostly duplicated those observed for sulphur-containing amino acids.

Dietary intake of potassium was consistently found correlated with consumption of fruits and vegetables, with dairy products as secondary sources but in Chinese participants.

With the exception of vegetables, dietary sources of magnesium showed greater heterogeneity across groups: grains played an important role but for Japanese participants from Japan, for whom a strong association was found with fish, whereas dairy products were identified as sources characteristic of participants from Western countries, irrespective of ethnicity; meat and eggs contributed only marginally, while fruits were found to be considerable sources of magnesium for participants from the UK and southern China.

Dietary intake of calcium was strongly associated with consumption of dairy products in participants from Western countries and Japan, whereas in China the main source appeared to be vegetables; the contribution of grains was notable in Western countries (irrespective of ethnicity), and that of fish among Japanese participants from Japan and the southern Chinese.

5.3.2 Associations with blood pressure and BMI

Regression coefficient estimates and 95% CIs for the associations of PRAL and NEAP with BP are reported in tables 5.2 and 5.3, respectively, for systolic and diastolic BP. Analyses unadjusted for BMI identified statistically significant positive associations between NEAP and systolic BP in women, and between NEAP and diastolic BP in men. These associations were not replicated in the opposite sex, but were still significant when both sexes were analysed together; however, they lost significance after adjustment for BMI.

	Pooled		Men		Women	
	β	95% CI	β	95% CI	β	95% CI
PRAL	0.173	−0.044 to 0.389	0.095	−0.169 to 0.359	0.288	−0.080 to 0.657
+ BMI	0.042	−0.173 to 0.258	−0.014	−0.277 to 0.249	0.127	−0.239 to 0.494
NEAP	0.402	0.173 to 0.630	0.244	−0.035 to 0.523	0.642	0.254 to 1.030
+ BMI	0.064	−0.165 to 0.294	−0.011	−0.292 to 0.269	0.170	−0.220 to 0.560

Table 5.2: Associations of PRAL and NEAP with systolic BP. Regression coefficients and 95% CIs are presented before and after adjustment for BMI, and upon stratification by sex.

Regression coefficient estimates and 95% CIs for the associations of PRAL and NEAP with BMI are reported in table 5.4. All associations were found statistically significant, and effect size estimates indicated a slightly stronger effect in women.

5.3.3 Associations with urinary amino acid concentrations

Regression coefficient estimates and *p*-values for the associations of PRAL and NEAP with urinary concentrations of amino acids and related compounds are reported in table 5.5. Statistically significant direct associations were found consistently between both DIMA indices and urinary concentrations of 12 proteinogenic amino acids (alanine, asparagine, glutamine, glycine, histidine, lysine, methionine, serine, threonine, tryptophan, tyrosine, and valine) and three related compounds (1- and 3-methylhistidine, and taurine).

	Pooled		Men		Women	
	β	95% CI	β	95% CI	β	95% CI
PRAL	0.079	−0.080 to 0.239	0.123	−0.078 to 0.325	−0.022	−0.285 to 0.240
+ BMI	−0.019	−0.178 to 0.141	0.034	−0.168 to 0.235	−0.134	−0.395 to 0.128
NEAP	0.260	0.092 to 0.428	0.259	0.046 to 0.471	0.227	−0.048 to 0.503
+ BMI	0.013	−0.156 to 0.182	0.054	−0.160 to 0.268	−0.098	−0.376 to 0.180

Table 5.3: Associations of PRAL and NEAP with diastolic BP. Regression coefficients and 95% CIs are presented before and after adjustment for BMI, and upon stratification by sex.

	Pooled		Men		Women	
	β	95% CI	β	95% CI	β	95% CI
PRAL	0.023	0.010 to 0.036	0.021	0.006 to 0.037	0.026	0.005 to 0.048
NEAP	0.033	0.019 to 0.047	0.030	0.013 to 0.046	0.039	0.016 to 0.062

Table 5.4: Associations of PRAL and NEAP with BMI. Regression coefficients and 95% CIs are additionally presented upon stratification by sex.

	<i>N</i>	PRAL		NEAP	
		β	<i>p</i> -value	β	<i>p</i> -value
Alanine	4538	0.92	7.64×10^{-14}	0.87	9.24×10^{-13}
Arginine	2843	0.07	0.00512	0.06	0.0118
Asparagine	4535	0.32	0.00049	0.30	0.0011
Cysteine	4204	0.12	0.00278	0.12	0.00278
Glutamine	4536	1.04	1.1×10^{-10}	0.99	5.53×10^{-10}
Glycine	4537	2.63	1.81×10^{-5}	2.62	1.77×10^{-5}
Histidine	4538	2.21	4.23×10^{-13}	2.13	2.19×10^{-12}
Isoleucine	945	0.01	0.11	0.01	0.321
Leucine	1800	0.09	0.0109	0.09	0.0103
Lysine	4537	1.20	3.52×10^{-7}	1.17	6.54×10^{-7}
Methionine	3760	0.06	0.000258	0.05	0.000964
Phenylalanine	3929	0.09	0.00283	0.08	0.0059
Serine	4535	0.84	8.87×10^{-9}	0.80	3.59×10^{-8}
Threonine	4536	0.75	7.11×10^{-16}	0.73	4.07×10^{-15}
Tryptophan	3217	0.10	0.000352	0.09	0.00136
Tyrosine	4298	0.26	2.57×10^{-11}	0.25	3.87×10^{-10}
Valine	2837	0.09	1.08×10^{-7}	0.08	5.61×10^{-7}
1-methylhistidine	4462	4.01	5.48×10^{-15}	3.92	1.3×10^{-14}
3-methylhistidine	4536	0.51	6.98×10^{-25}	0.49	5.45×10^{-24}
Ethanolamine	4538	0.23	0.00298	0.21	0.00593
Taurine	4533	4.22	1.28×10^{-14}	4.18	1.63×10^{-14}

Table 5.5: Associations of PRAL and NEAP with urinary amino acid concentrations. The total number of samples *N* with concentration above the analytical detection limit is also presented.

5.3.4 Metabolome-wide association studies

As depicted in figs. 5.2 and 5.3, the metabolome-wide association studies for PRAL and NEAP highlighted a number of (mostly overlapping) metabolic signals, some of which were additionally associated with BP, BMI, and/or dietary constituents and food groups in exploratory analyses.

The metabolome-wide association studies confirmed direct associations with urinary concentrations of alanine, glutamine, glycine, tyrosine, and valine. In addition, they unveiled direct associations with urinary concentrations of betaine, creatine and creatinine, dimethylamine, trimethylamine and trimethylamine *N*-oxide, and inverse associations with urinary concentrations of citrate, hippurate, proline betaine, and trigonelline. The direct association with creatinine was also confirmed using additional measurements obtained via the Jaffe reaction (PRAL: $\beta = 0.059$, 95% CI: 0.004 to 0.114; NEAP: $\beta = 0.158$, 95% CI: 0.099 to 0.216).

§ 5.4 Discussion

The analysis of correlations of PRAL with its dietary constituents and food groups suggested that some variability may be attributed to dietary differences between groups, with geographical location also playing a role. For example, we confirmed dietary dissimilarities between northern and southern China that had been previously characterized in the context of INTERMAP [24]. Similarly, we observed a divergence in dietary habits of Japanese participants in Japan compared to those in the USA, and of these two groups with respect to NHWs in the USA.

Most dietary constituents of PRAL followed comparable correlation patterns, with some exceptions: for instance, sources of magnesium secondary to veget-

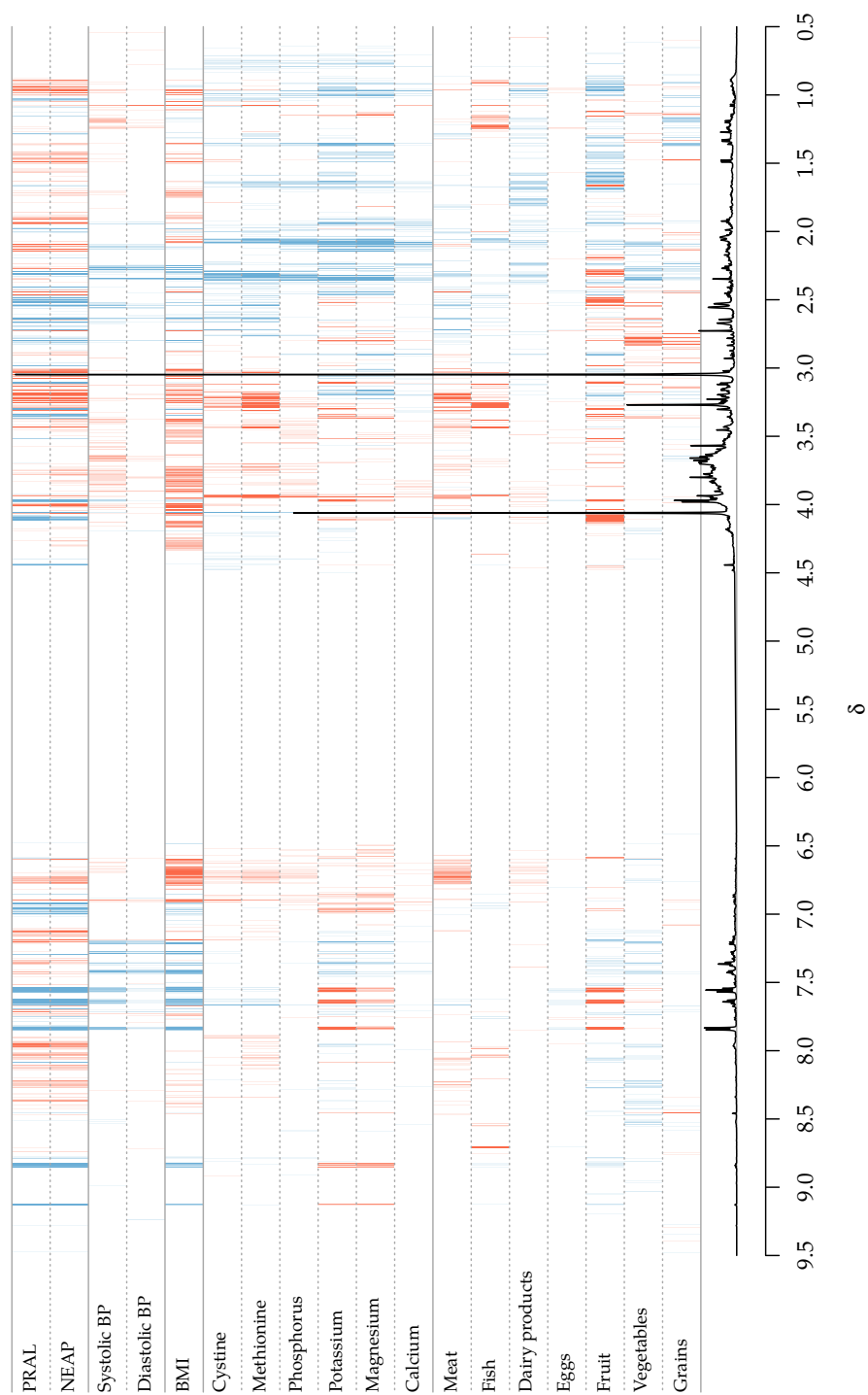


Figure 5.2: Metabolome-wide association study based on samples from the first visit. Direct and inverse associations are depicted in red and blue, respectively; the colour scale reflects the p -value (in $-\log_{10}$ scale) of each association. The median spectrum across samples is also plotted.

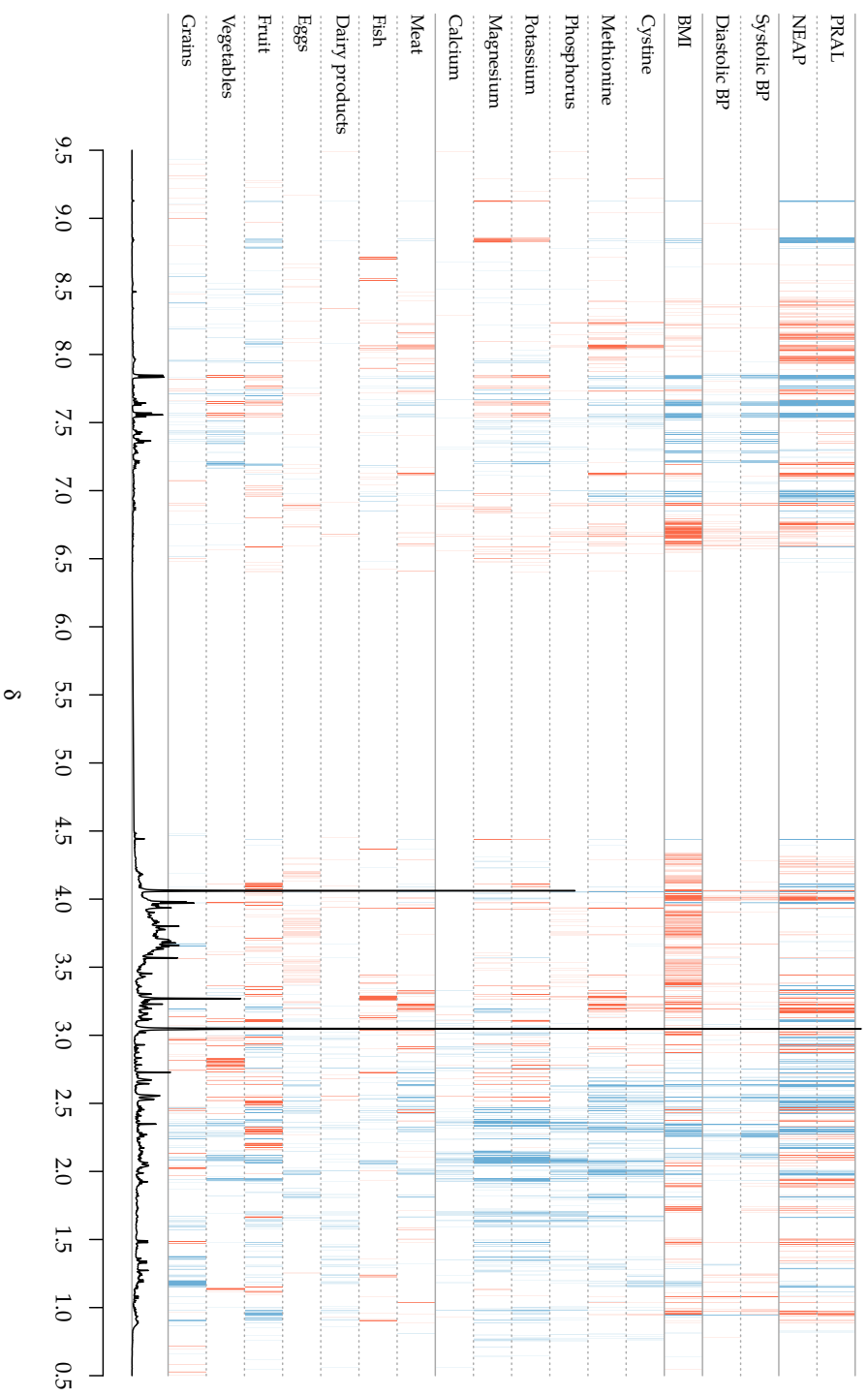


Figure 5.3: Metabolome-wide association study based on samples from the second visit. Direct and inverse associations are depicted in red and blue, respectively; the colour scale reflects the p -value (in $-\log_{10}$ scale) of each association. The median spectrum across samples is also plotted.

ables exhibited greater heterogeneity among groups, and calcium intake was strongly correlated with consumption of dairy products but in Chinese participants. This suggests that differences observed across groups are related to ethnicity- and country-specific dietary patterns. It should be mentioned that some of the described correlations may reflect dietary patterns (several food groups being frequently consumed together) rather than actual nutrient sources.

Overall, the major upward drivers of PRAL were animal source foods, with country-specific preferences, and its major downward drivers were fruit and vegetables, independently of group [25]. These associations suggest a prominent role of animal source foods — mainly meat in Westernized countries, fish in Japan, and eggs in China — towards induction of metabolic acidosis, and a similar important role of fruits and vegetables in its prevention.

Our results thus confirm the role of the balance between animal and vegetable source food consumption in determining DIMA. As noted, since potassium intake is strongly correlated with fruit and vegetable intake, it may serve as proxy for consumption of these foods. This is in line with the observation that the dietary protein-to-potassium ratio can serve as reliable DIMA index when dietary data are limited to fewer nutrients [26], substituting animal protein in place of total protein if available [27].

Investigation of the effects of DIMA on BP revealed consistent significant associations with NEAP, but not with PRAL. Effect size estimates were generally greater in women than in men, which could be due to sex-specific effects of DIMA, or to sex differences in the regulation of BP [28]. To our knowledge, the present study is the first to investigate the effect of DIMA on BP in both sexes, and generalizes previous results found in women only [11, 12].

Associations with BP lost significance after adjustment for BMI. One possible explanation relates to the definition of NEAP itself, which includes an approx-

imation of BSA depending on height and weight. NEAP and BMI are therefore correlated, and inclusion of the latter in models for BP may constitute over-adjustment. Furthermore, in light of the strong associations between NEAP and BMI found consistently across groups and sexes, BMI may have a mediating (rather than confounding) role in the associations of NEAP with BP. Nevertheless, due to the cross-sectional nature of INTERMAP, our data cannot reveal whether the identified relationship is reflective of specific dietary patterns (animal source foods versus fruits and vegetables), and/or of an effect of BMI on BP.

While our data do not seem to consistently support associations between PRAL and BP as previously reported [11], we found associations with NEAP, therefore confirming the role of DIMA in BP regulation. The lack of associations with PRAL may be explained by its reliance on dietary information only, which is overcome by NEAP through the inclusion of an additional term accounting for endogenous acid production.

Analyses of urinary concentrations of amino acids and other small molecules revealed a complex metabolic phenotype associated with DIMA.

Remarkably, levels of commonly excreted amino acids such as alanine, glutamine, glycine, histidine, serine, threonine, and the related compounds 1- and 3-methylhistidine as well as taurine [29] were all directly associated with both DIMA indices.

Concentrations of some of these compounds are strongly affected by diet: for example, both 1- and 3-methylhistidine have been proposed as biomarkers of meat intake [30, 31].

Associations between DIMA and urinary levels of other molecules may also reflect dietary patterns: for instance, proline betaine has been proposed as biomarker of citrus consumption [32], and trigonelline has been previously identified as biomarker of coffee intake [33–35].

The inverse association with urinary citrate levels is of particular interest: citrate excretion is remarkably sensitive to acid-base balance [36], and it is clinically relevant because of its role in preventing nephrolithiasis by way of calcium complexation [37].

There is natural uncertainty in nutrient measurements (regression dilution bias), which is, however, limited owing to repeated measurements available in INTERMAP and the choice of statistical models. Additionally, there is potential for systematic bias, which is again controlled for in INTERMAP through the definition of standardised procedures, as well as a large variability in daily dietary intake. Despite these limitations that have the potential to underestimate effect sizes, our results suggest a promising potential of DIMA indices to determine BP.

Clinical relevance of these indices may only be established if these results are replicated in other cross-sectional, and possibly prospective, studies, which are by nature more suited to investigate long-term dietary effects of DIMA on BP. Such an extension of this work will constitute a way forward in the understanding, and potentially in the prevention, of hypertension and related risks.

References

1. M Wiederkehr and R Krapf. 'Metabolic and endocrine effects of metabolic acidosis in humans'. *Swiss Medical Weekly* **131** (2001), 127–132.
2. J Pizzorno, L A Frassetto, and J Katzinger. 'Diet-induced acidosis: is it real and clinically relevant?' *British Journal of Nutrition* **103.8** (2010), 1185–1194.
3. G Souto, C Donapetry, J Calviño, and M M Adeva. 'Metabolic acidosis-induced insulin resistance and cardiovascular risk'. *Metabolic Syndrome and Related Disorders* **9.4** (2011), 247–253.

4. A Kanbara, M Hakoda, and I Seyama. 'Urine alkalization facilitates uric acid excretion'. *Nutrition Journal* **9** (2010), 45.
5. A Kanbara, Y Miura, H Hyogo, K Chayama, and I Seyama. 'Effect of urine pH changed by dietary intervention on uric acid clearance mechanism of pH-dependent excretion of urinary uric acid'. *Nutrition Journal* **11** (2012), 39.
6. J J Scialla and C A M Anderson. 'Dietary acid load: a novel nutritional target in chronic kidney disease?' *Advances in Chronic Kidney Disease* **20.2** (2013), 141–149.
7. B R Hemmelgarn, J Zhang, B J Manns, M Tonelli, E Larsen, W A Ghali, D A Southern, K McLaughlin, G Mortis, and B F Culleton. 'Progression of kidney dysfunction in the community-dwelling elderly'. *Kidney International* **69** (2006), 2155–2161.
8. E van den Berg, F A P Hospers, G Navis, M F Engberink, E J Brink, J M Geleijnse, M A van Baak, R O B Gans, and S J L Bakker. 'Dietary acid load and rapid progression to end-stage renal disease of diabetic nephropathy in Westernized South Asian people'. *Journal of Nephrology* **24.1** (2011), 11–17.
9. N Goraya and D E Wesson. 'Dietary management of chronic kidney disease: protein restriction and beyond'. *Current Opinion in Nephrology and Hypertension* **21.5** (2012), 552–556.
10. M M Adeva and G Souto. 'Diet-induced metabolic acidosis'. *Clinical Nutrition* **30.4** (2011), 416–421.
11. K Murakami, S Sasaki, Y Takahashi, and K Uenishi. 'Association between dietary acid-base load and cardiometabolic risk factors in young Japanese women'. *British Journal of Nutrition* **100.3** (2008), 642–651.

12. L Zhang, G C Curhan, and J P Forman. 'Diet-dependent net acid load and risk of incident hypertension in United States women'. *Hypertension* **54.4** (2009), 751–755.
13. M F Engberink, S J L Bakker, E J Brink, M A van Baak, F J A van Rooij, A Hofman, J C M Witteman, and J M Geleijnse. 'Dietary acid load and risk of hypertension: the Rotterdam Study'. *The American Journal of Clinical Nutrition* **95.6** (2012), 1438–1444.
14. D Krupp, A Ströhle, and T Remer. 'Dietary acid load and risk of hypertension'. *The American Journal of Clinical Nutrition* **96.4** (2012), 942–943.
15. D Krupp, L Shi, C Maser-Gluth, M Pietzarka, and T Remer. '11 β -hydroxysteroid dehydrogenase type 2 and dietary acid load are independently associated with blood pressure in healthy children and adolescents'. *The American Journal of Clinical Nutrition* **97.3** (2013), 612–620.
16. J Stamler, P Elliott, B Dennis, A R Dyer, H Kesteloot, K Liu, H Ueshima, and B F Zhou. 'INTERMAP: background, aims, design, methods, and descriptive statistics (nondietary)'. *Journal of Human Hypertension* **17** (2003), 591–608.
17. B Dennis, J Stamler, M Buzzard, R Conway, P Elliott, A Moag-Stahlberg, A Okayama, N Okuda, C Robertson, F Robinson, S Schakel, M Stevens, N Van Heel, L Zhao, and B F Zhou. 'INTERMAP: the dietary data – process and quality control'. *Journal of Human Hypertension* **17** (2003), 609–622.
18. E Holmes, R-L Loo, J Stamler, M Bictash, I K S Yap, Q Chan, T Ebbels, M De Iorio, I J Brown, K A Veselkov, M L Daviglus, H Kesteloot, H Ueshima, L Zhao, J K Nicholson, and P Elliott. 'Human metabolic phenotype diversity and its association with diet and blood pressure'. *Nature* **453** (2008), 396–400.

19. T Remer and F Manz. 'Estimation of the renal net acid excretion by adults consuming diets containing variable amounts of protein'. *The American Journal of Clinical Nutrition* **59.6** (1994), 1356–1361.
20. N Poupin, J Calvez, C Lassale, C Chesneau, and D Tomé. 'Impact of the diet on net endogenous acid production and acid-base balance'. *Clinical Nutrition* **31.3** (2012), 313–321.
21. A Sebastian, L A Frassetto, D E Sellmeyer, R L Merriam, and R C Morris. 'Estimation of the net acid load of the diet of ancestral preagricultural *Homo sapiens* and their hominid ancestors'. *The American Journal of Clinical Nutrition* **76.6** (2002), 1308–1316.
22. R D Mosteller. 'Simplified calculation of body-surface area'. *New England Journal of Medicine* **317.17** (1987), 1098.
23. O Cloarec, M-E Dumas, A Craig, R H Barton, J Trygg, J Hudson, C Blancher, D Gauguier, J C Lindon, E Holmes, and J K Nicholson. 'Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets'. *Analytical Chemistry* **77.5** (2005), 1282–1289.
24. I K S Yap, I J Brown, Q Chan, A Wijeyesekera, I Garcia-Perez, M Bictash, R L Loo, M Chadeau-Hyam, T Ebbels, M De Iorio, E Maibaum, L Zhao, H Kesteloot, M Daviglus, J Stamler, J K Nicholson, P Elliott, and E Holmes. 'Metabolome-wide association study identifies multiple biomarkers that discriminate north and south chinese populations at differing risks of cardiovascular disease: INTERMAP study'. *Journal of Proteome Research* **9.12** (2010), 6647–6654.
25. A Ströhle, A Waldmann, J Koschizke, C Leitzmann, and A Hahn. 'Diet-dependent net endogenous acid load of vegan diets in relation to food groups

- and bone health-related nutrients: results from the German Vegan Study'. *Annals of Nutrition and Metabolism* **59** (2011), 117–126.
26. L A Frassetto, K M Todd, R C Morris, and A Sebastian. 'Estimation of net endogenous noncarbonic acid production in humans from diet potassium and protein contents'. *The American Journal of Clinical Nutrition* **68.3** (1998), 576–583.
 27. S R Zwart, A R Hargens, and S M Smith. 'The ratio of animal protein intake to potassium intake is a predictor of bone resorption in space flight analogues and in ambulatory subjects'. *The American Journal of Clinical Nutrition* **80.4** (2004), 1058–1065.
 28. J F Reckelhoff. 'Gender differences in the regulation of blood pressure'. *Hypertension* **37.5** (2001), 1199–1208.
 29. P Soupart. 'Urinary excretion of free amino acids in normal adult men and women'. *Clinica Chimica Acta* **4.2** (1959), 265–271.
 30. A J Cross, J M Major, and R Sinha. 'Urinary biomarkers of meat consumption'. *Cancer Epidemiology, Biomarkers & Prevention* **20.6** (2011), 1107–1111.
 31. W Altorf-van der Kuil, E J Brink, M Boetje, E Siebelink, S Bijlsma, M F Engberink, P van 't Veer, D Tomé, S J L Bakker, M A van Baak, and J M Geleijnse. 'Identification of biomarkers for intake of protein from meat, dairy products and grains: a controlled dietary intervention study'. *British Journal of Nutrition* **110.5** (2013), 810–822.
 32. S S Heinzmann, I J Brown, Q Chan, M Bictash, M-E Dumas, S Kochhar, J Stamler, E Holmes, P Elliott, and J K Nicholson. 'Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption'. *The American Journal of Clinical Nutrition* **92.2** (2010), 436–443.

33. R Lang, A Wahl, T Stark, and T Hofmann. 'Urinary N-methylpyridinium and trigonelline as candidate dietary biomarkers of coffee consumption'. *Molecular Nutrition & Food Research* **55.11** (2011), 1613–1623.
34. K A Guertin, S C Moore, J N Sampson, W-Y Huang, Q Xiao, R Z Stolzenberg-Solomon, R Sinha, and A J Cross. 'Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations'. *The American Journal of Clinical Nutrition* **100.1** (2014), 208–217.
35. J A Rothwell, Y Fillâtre, J-F Martin, B Lyan, E Pujos-Guillot, L Fezeu, S Herberg, B Comte, P Galan, M Touvier, and C Manach. 'New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects'. *PLOS ONE* **9.4** (2014), e93474.
36. L L Hamm and E E Simon. 'Roles and mechanisms of urinary buffer excretion'. *American Journal of Physiology – Renal Physiology* **253.4** (1987), F595–F605.
37. M J Nicar, C Skurla, K Sakhaee, and C Y Pak. 'Low urinary citrate excretion in nephrolithiasis'. *Urology* **21.1** (1983), 8–14.

Chapter 6

Conclusions

In the introductory chapter of this thesis, the author has provided a detailed description of the multifaceted field of ‘computational epidemiology’ from the perspectives of molecular biology, measurement theory, and statistical modelling.

The focus was on ‘-omes’ that can be characterised by high-throughput laboratory techniques at present. It may be expected that novel laboratory techniques with increasingly higher throughputs will emerge in the near future. Their introduction will require careful evaluation of pre-processing methods to reduce sensitivity to experimental conditions in a principled way. Similarly, the larger and larger datasets thus generated will require computationally efficient statistical methods to maximise the potential for valid and reliable findings. These

aims could be objectively quantified within the flexible evaluation frameworks presented in chapter 2.

Genomic research has mostly harvested the ‘low-hanging fruits’ by conducting increasingly larger genome-wide association studies, and other ‘-omics’ will soon follow suit. While genomics has bequeathed a certain analytical mindset (and a corresponding set of statistical tools) to the field of computational epidemiology, this legacy must not be merely conserved, but also built upon.

The robustness and redundancy bestowed upon biological regulatory networks by their complexity and phenomena such as epistasis and pleiotropy means that, for example, ill health may result from a variety of disruptions of cellular homeostasis. Current statistical methods are unable to capture the diversity of biological mechanisms that can lead to indistinguishable phenotypes including (early) disease states.

In the context of ‘-omics’ other than genomics, as discussed in section 1.1.6, much remains to be understood about temporal variability and plasticity in response to endogenous and exogenous exposures. In addition to ageing, it is plausible that other factors including nutrition may act as pivotal modifying variables in the pathways from genome to phenotype.

For example, the methyl groups needed to establish and maintain DNA methylation patterns are derived from a network of interdependent biochemical reactions collectively referred to as one-carbon metabolism (OM). OM can be influenced not only by suboptimal availability of one-carbon nutrients, but also by a number of relatively common genetic polymorphisms. A number of small studies have implicated OM disruption in DNA hypomethylation at the level of the entire genome. Given that DNA methyltransferases preferentially target certain sites, it is plausible that loss of DNA methylation following OM disruption should also exhibit some degree of selectivity. Nonetheless, adequately-powered

studies at specific loci across the genome are lacking, and the combined effects of nutrition and genetic susceptibility remain largely unexplored.

The increasing availability of ‘-omics’ measurements at different complexity levels will soon raise questions about biological mediation of exposures and their effects e.g. on health. While considering multiple ‘-omics’ profiles independently is a reasonable first approximation, characterisation of the joint (rather than marginal) variation will be indispensable to understand how complex mixtures of exposures are embedded in complex combinations of ‘-omics’ signatures, and how these interact to affect the cellular micro- and macro-environments in health and disease.